Network High Availability for Ethernet Services Using IP/MPLS Networks

Matthew Bocci, Ian Cowburn, and Jim Guillet, Alcatel-Lucent

ABSTRACT

Enterprises are increasingly using Ethernet as the foundation for transforming their networks to Internet Protocol. Simultaneously, service providers are deploying Ethernet to exploit the demand for wide-area Ethernet services and as the infrastructure for new residential services such as IPTV. This is due to Ethernet's low cost per bit and ubiquity in local area networks. Recent years have seen the widespread deployment of IP/MPLS networks to address this opportunity. IP/MPLS enables enhanced flexibility over the same converged network for IP and legacy services, avoiding the need to build separate per-service networks. It also adds carrier-grade capabilities such as quality of service, traffic engineering, and resiliency, thereby enabling new multipoint services such as virtual private LAN service. However, using Ethernet for "always-on" and other mission-critical services has resulted in new resiliency requirements, in both the access and the network core. Two novel developments address these high expectations by enabling significant improvements in service availability. These are pseudowire redundancy and Ethernet multi-chassis link aggregation. This article reviews the current redundancy mechanisms typically deployed in Ethernet and MPLS networks. We show how additional enhancements are required in both the network core and the access to the Ethernet service. We describe new pseudowire redundancy and MC-LAG mechanisms, showing how they work together to enable end-to-end protection for Ethernet virtual private wire services and VPLS.

INTRODUCTION

The accelerating pace of IP network transformation reflects our industry's unlimited capacity to innovate both enterprise and residential applications. With its attractive combination of low cost and high bandwidth, Ethernet is rapidly emerging as the layer 2 technology of choice to support this transformation. However, with corporate governance and business process automation foremost in their minds, enterprise chief information officer (CIO) expectations for always-on network availability to support business critical applications have become table stakes. On the residential side, similar expectations now also exist thanks to triple play offerings of video entertainment and telephony, and the increasing dependence on the Internet in our daily lives. And as operators embrace these new requirements via new network and service architectures, the industry's overall benchmark for network performance is raised accordingly, driving yet another cycle of application innovation.

The foundation of IP technologies continues to expand: multiprotocol label switching (MPLS) is widely deployed by operators to broaden IP's multiservice capabilities. More recently, MPLS has put essential "carrier" attributes into Ethernet, enabling operators to leverage the desirable benefits of Ethernet throughout their networks, without those that have propagated its perception as an enterprise-only technology. Indeed, the MPLS control plane reduces operational costs by adding stability and control to Ethernet bandwidth, thus simplifying large-scale carrier deployments. Now, with public operators undertaking massive transformation projects throughout the world (e.g., BT [1]), technology advances have become focused on further improvements and cost optimization of end-to-end networking requirements, and the elimination of any residual failure conditions. Specifically, two challenges need to be addressed in order for the MPLS network to fully meet these expectations for Ethernet services:

- How to provide resiliency against catastrophic node failures in the core of the MPLS network
- How to provide resilient access to Ethernet services delivered by the MPLS network

This article presents two recent breakthroughs that address these issues: end-to-end pseudowire (PW) redundancy and multi-chassis link aggregation (MC-LAG), respectively.

We first review the mechanisms that enable MPLS to deliver wide-area Ethernet services for both point-to-point and multipoint-to-multipoint applications: virtual private wire service (VPWS) and virtual private LAN service (VPLS), both of which make use of pseudowires. Established

0163-6804/08/$25.00 © 2008 IEEE

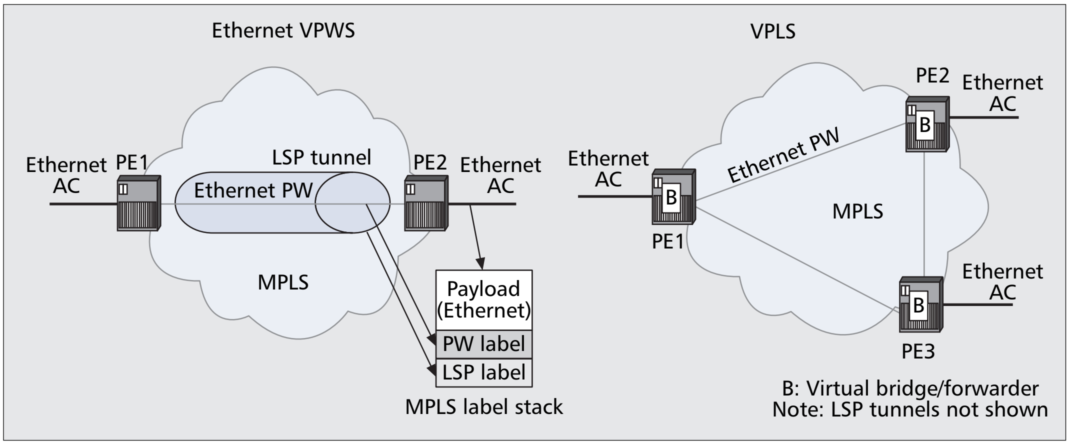


Figure 1. *Architectures for Ethernet VPWS and VPLS.*

The objective of node level redundancy is to prevent failures of particular components of a node from impacting the externally observable protocol behaviour. This is typically achieved through hot standby operation of components implementing critical routing protocols.

mechanisms that provide resiliency and protection for MPLS-based Ethernet services are then surveyed. We then describe how pseudowire redundancy and MC-LAG can be combined to offer enhanced resiliency for both VPWS- and VPLS-based Ethernet services.

MPLS Support for Ethernet Services

MPLS has evolved from a suite of protocols intended to enhance the forwarding performance of IP routers to encompass applications including traffic engineering and IP virtual private networks (VPNs). More recently, MPLS has been widely deployed to enable Ethernet and other layer 2 services to be delivered from a converged IP network. It achieves this using applications known as layer 2 VPNs (L2VPNs). Two types of L2VPN are defined by the Internet Engineering Task Force (IETF) [2]. The VPWS is used for point-to-point services, such as leased lines, while the VPLS is essentially a bridged Ethernet service that enables a service provider's MPLS network to emulate a large number of customer LANs [3].

L2VPNs are based on pseudowires [4], which form the basis of connectivity between provider edge (PE) nodes. Pseudowires are well documented elsewhere [5], so only a brief introduction is provided here. Each pseudowire (PW) provides discrete point-to-point layer 2 connectivity, and many PWs are multiplexed into an MPLS label switched path (LSP). In an MPLS network label stacking is used; an inner PW label is pushed on the encapsulated layer 2 payload and identifies the PW, and then a further outer label is pushed, which identifies the MPLS label switched path (LSP) which carries the PW across the MPLS network (Fig. 1). For an Ethernet VPWS, each Ethernet PW [6] is associated with an Ethernet attachment circuit (AC) on the PE. This may be a virtual LAN (VLAN) or an Ethernet port. For a VPLS, PWs interconnect virtual bridging and forwarding instances on the PEs.

Pseudowires are typically established using an extension of the MPLS label distribution protocol (LDP) that operates in a targeted mode between the PEs [7]. Targeted LDP (TLDP) enables PW labels to be exchanged, as well as PW status and other maintenance information to be signaled.

Traditionally, each PW has only spanned a single LSP. However, the architecture has recently been extended to allow PWs to be switched from one LSP to another LSP at a PE. This multisegment PW architecture reduces the number of LSPs needed in large networks, and is particularly useful for interprovider L2VPNs. In this architecture the PE that switches the PW is known as a switching PE (S-PE), while the PEs that forward packets between the PW and the AC or virtual bridge are known as terminating PEs (T-PEs).

Redundancy Options for MPLS-Based Ethernet Services

Figure 2 illustrates the techniques available to provide resilient Ethernet services from an MPLS network. These can be broadly categorised into node level redundancy and network level redundancy.

The objective of node level redundancy is to prevent failures of particular components of a node from impacting the externally observable protocol behavior. This is typically achieved through hot standby operation of components implementing critical routing protocols such as LDP, Resource Reservation Protocol (RSVP), Open Shortest Path First (OSPF), Border Gateway Protocol (BGP), Intermediate System to Intermediate System (IS-IS), or Protocol Independent Multicast (PIM). It can also be applied to components implementing other higher-level aspects of services such as VPWS and VPLS. Graceful restart also falls in this category by minimizing the disruption caused to network operation by the failure and restart of a node [8].

However, nodal redundancy does not protect against failures of network links or catastrophic failures of network nodes, such as power failures or widespread disasters. For this, network level redundancy is also required. Network level redundancy has typically been applied at either the Ethernet layer or the MPLS layer.

A key aspect of the scheme described here is that the dual homing mechanism for the CE (MC-LAG) is coupled to the forwarding state of the PWs or the VPLS. This enables end-to-end protection to be provided, while avoiding the need for the CEs on both ends of the service to switch to a backup AC when a single failure occurs.

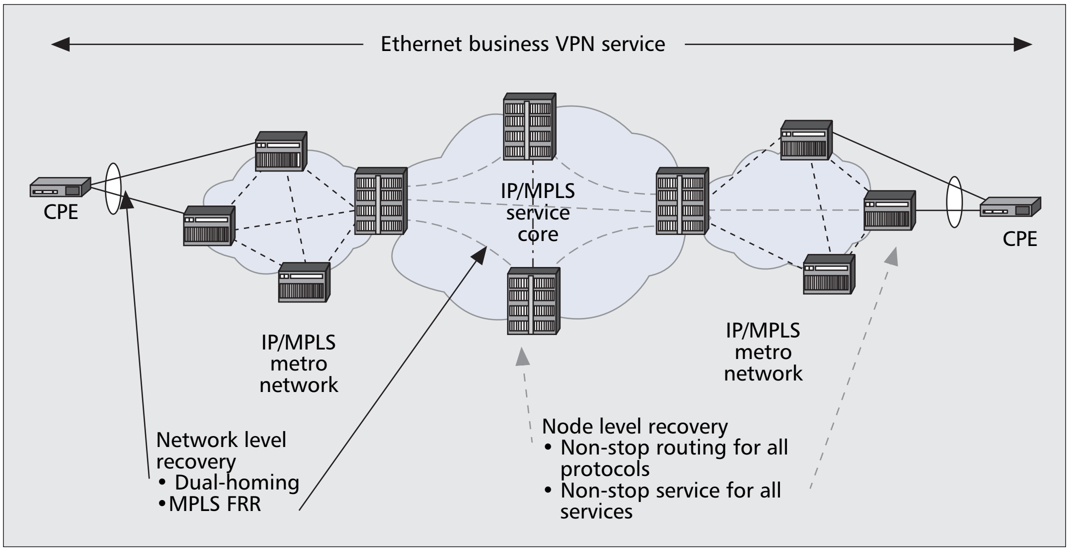


Figure 2. *Network and node level redundancy.*

At the Ethernet layer, IEEE 802.3ad (now incorporated into IEEE 802.3-2005 [9]), otherwise known as link aggregation (LAG), was initially introduced to provide both redundancy and extra capacity for point-to-point connections between two systems. Combining multiple Ethernet links into a group and representing the group as a single bundle, a LAG, on the connected systems accomplishes this. LAGs provide extra capacity and redundancy in that a LAG remains active with a reduced capacity even if some of its composite links fail. LAGs can be used between multiple systems, and combined with both VPLS point-to-multipoint and Ethernet VPWS point-to-point services to allow providers to deliver highly redundant services to their customers.

Spanning Tree Protocol (STP) and Rapid STP (RSTP)[10] could also be run between the customer premises equipment (CPE) and the provider network where multihoming at the Ethernet layer is used. These protocols enable a single active link to be chosen, avoiding loops and removing failed links from the Ethernet domain, enabling protection against failures of the PEs as well as the attachment circuits. However, performance concerns have meant that service providers are reluctant to use STP or RSTP for VPWS and VPLS services. For example, even RSTP can take several seconds to converge, particularly with large networks. Furthermore, it may not be desirable for a service provider's PE to participate in a customer STP because oscillations in the customer STP could impact the stability, performance, and scalability of the service provider's network.

At the MPLS layer, network level redundancy has focused on the MPLS LSP tunnel. Here, mechanisms such as MPLS fast reroute (FRR) [11] or LSP backup can be used to provide sub-50-ms protection to all of the Ethernet PWs carried by an LSP. However, this is insufficient to protect against failures of the PEs or attachment circuits, in the case of either T-PEs (where dual homing is required) or S-PEs.

Network and Access PE Protection for Ethernet Services

To meet the increased network demands to transport business-critical applications and residential triple play services, operators rely on a variety of protection mechanisms. In single-segment PW (SS-PW)-based services where there is no access redundancy, such as VPWS and VPLS, protection for the PW is provided by the MPLS layer as described above. However, there are a number of scenarios where this level of protection is insufficient to cover all possible failure modes. For example, it cannot protect against the failure of the PE or the attachment circuits since these represent single points of access to the emulated service. No alternative path to the service exists.

A detailed set of failure scenarios that require additional protection is described in [3]. In the remainder of this article we consider a subset of these scenarios to illustrate how end-to-end resiliency can be provided. These are:

- Dual homing of a CE via two separate ACs into redundant PEs, with SS-PWs or multi-segment PWs used for a VPWS service
- Dual homing of a CE via two separate ACs into redundant PEs for a VPLS, using SS-PWs or multisegment PWs

These scenarios rely on two mechanisms to provide end-to-end protection for the Ethernet service:

- PW redundancy
- Access and PE redundancy using MC-LAG

A key aspect of the scheme described in this article is that the dual homing mechanism for the CE (MC-LAG) is coupled to the forwarding state of the PWs or VPLS. This enables end-to-end protection to be provided, while avoiding the need for the CEs on both ends of the service to switch to a backup AC when a single failure occurs. The protection mechanisms of the MPLS network are utilized to localize the impact of a failure, which is an important consideration

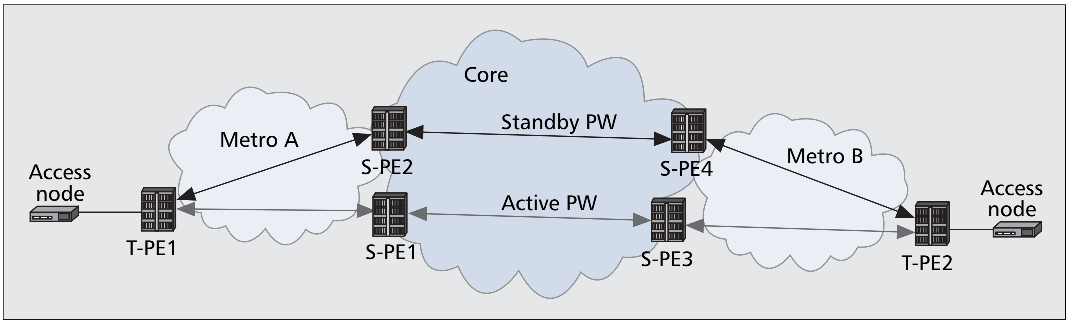


Figure 3. *Pseudowire redundancy.*

In steady state, one LAG sub-group connected to one PE is set to active, and one is set to standby. This choice is by configuration, or based on administrative parameters such as weight or the sub-group containing the most links that are currently up.

when designing large networks. This behavior is to be distinguished from traditional end-to-end protection for layer 2 services that use dual homing, which can require both CEs to switch to a redundant path if the active path fails.

Pseudowire Redundancy

Pseudowire redundancy enables one or more redundant PWs to be configured to protect the traffic on an active PW. Each redundant set of PWs is associated by configuration with a single Ethernet service at each end. PW redundancy relies on extensions to the PW control protocol [13] that use LDP status messages to indicate the active or standby state of a PW. When a PE signals to a remote PE that a given PW is active, and other PWs in the redundant set are signaled for standby, the remote PE should use the active PW to forward packets from the AC to which it is bound.

Figure 3 shows an example of the use of PW redundancy. T-PE1 and T-PE2 are configured with a pair of PWs per service, and one is configured to be the primary PW to be used for forwarding packets when both PWs are in the operational UP state. PW status messages are exchanged end-to-end to notify the PEs of the operational state of both the PWs and the ACs (PW status messages generated by T-PEs and S-PEs are passed transparently by the S-PEs). A T-PE switches to the standby PW if an unrecoverable failure is detected within the network. It learns about this by either locally detecting the failure or receiving a PW status message indicating a remote failure. A PW status message of “active” is sent to a remote PE to request switching to the standby PW.

Multi-Chassis LAG

Historically, the concept of a LAG has been a single connection, comprising more than one physical link, running between two systems. These links are grouped together to form the LAG, and traffic is distributed across them using a hashing algorithm that ensures that each traffic flow maintains frame sequence integrity. A failure of one or more links in the LAG results in its traffic being redistributed to other links, hence ensuring that connectivity remains, albeit with reduced total bandwidth.

Clearly a complete system failure on one end will bring down the LAG. Today’s redundancy requirements in provider networks have created the need for a LAG to maintain connectivity even on complete failure of a single system. In order to achieve this, the concept used in the LAG subgroups is extended such that one end of the LAG is split between two systems instead of, for example, two router blades, thereby creating a multi-chassis LAG.

Multi-chassis LAG thus provides redundant Ethernet access connectivity that extends beyond link level protection by allowing two systems to share a common LAG endpoint. Figure 4 shows the MC-LAG function.

The Ethernet edge device is connected by multiple links toward a redundant pair of PE nodes such that both link and node level redundancy are provided. The LAG between the Ethernet edge device and the PEs is controlled using the Link Aggregation Control Protocol (LACP) [9]. LACP is used to manage the available LAG links into active and standby states such that only links from one PE node are active at a time to and from the Ethernet edge device. A further MC-LAG control protocol runs only between the redundant pair of PEs. This is an IP-based protocol that synchonizes the LAG state between the MC-LAG peers. It ensures a synchronized forwarding plane to and from the Ethernet edge device and is used to synchronize the link state information between the two PE nodes such that proper LACP messaging is provided to the Ethernet edge device. It also includes a keepalive function that enables a PE to detect whether or not its peer is functioning.

In steady state, one LAG subgroup connected to one PE is set to active, and one is set to standby. This choice is by configuration or based on administrative parameters such as weight or the subgroup containing the most links that are currently up. MC-LAG uses a LAG mode where all Ethernet traffic uses the active subgroup, while no traffic is forwarded on a standby subgroup. A failure of the active subgroup is detected using, for example, keepalive messages in LACP, and causes the MC-LAG protocol to switch to the standby PE and the standby subgroup. This state change is reflected in LACP, which forces the Ethernet device to switch the active subgroup.

End-to-end protection for MPLS-based Ethernet services relies on the effective combination of PW redundancy with MC-LAG. The following sections describe how these two mechanisms are applied to ensure resilient VPWS and VPLS services.

VPWS and VPLS Protection

Figure 5 illustrates how MC-LAG and PW redundancy work together to protect an Ethernet VPWS.

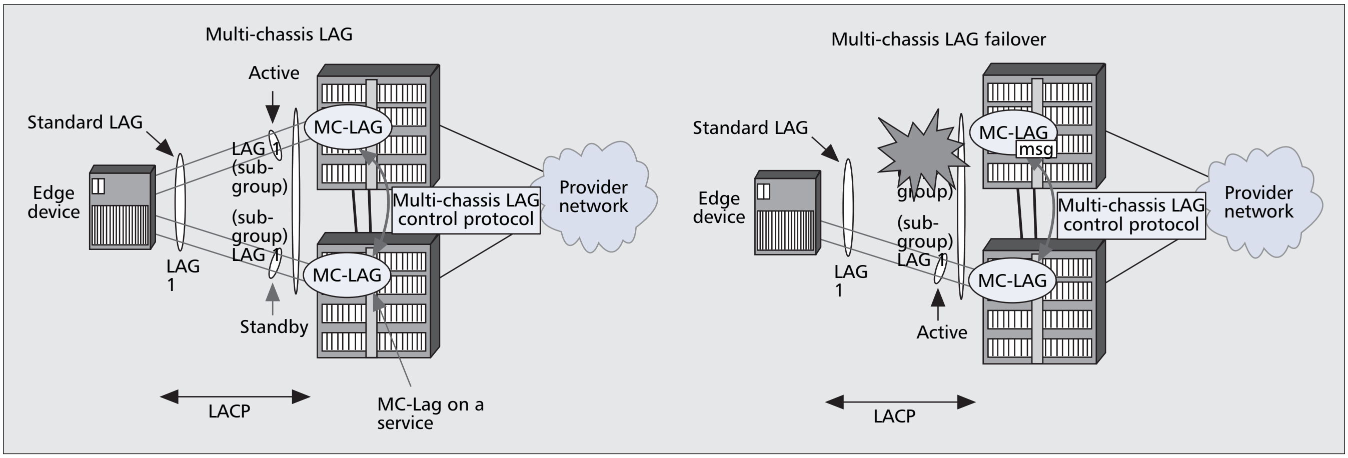


Figure 4. *MC-LAG operation.*

CE1 and CE2 are dual homed to PE1/PE2 and PE3/PE4 by Ethernet ACs, and the PEs are interconnected using Ethernet PWs. LACP operates between each CE and its connected PEs such that one of the LAG subgroups is active and one in standby at any given time. The LAG subgroup status is reflected in the PW status each PE signals to its far-end peer PE. Both the received far-end PW status and the local MC-LAG state for the LAG subgroup of that PW determine which PW to use for data path forwarding. Thus, the MC-LAG state at either end of the service drives the forwarding state of the PWs; the end-to-end path is where the MC-LAG status of both ACs is active and both PW endpoints are active (CE1-PE1-PE4-CE2).

Having considered how an end-to-end forwarding path is constructed, we now describe the sequence of events that enables recovery from a couple of different failure scenarios. Consider first the failure of an active LAG link (e.g., CE1–PE1 in Fig. 5). This failure may be detected either through LACP or using underlying link level failure detection mechanisms, and triggers PE1 to initiate MC-LAG link level convergence. PE1 informs PE2 to transition the MC-LAG link status from standby to active using the MC-LAG control protocol; thus, PE2 assumes active forwarding status. PE1 then changes the PWs to PE3 and PE4 to standby, informing PE4 through a PW status message.

Because PE2 is now the active PE in the pair [PE1/PE2], it changes the LACP link status from standby to active, enabling CE1 to forward using PE2's MC-LAG link. PE2 then connects the local MC-LAG link to the PW to PE2–PE4, advertising this active status. It also changes the status of its PW to PE3 to active (reflecting the MC-LAG state) and updates the PE through PW status (note that PE3 remains in standby state due to the local MC-LAG standby state). On receipt of the PW status message from PE2, PE4 changes its local PW crossconnect to PW PE4–PE2 because both its local status and the remote status received from PE2 are now active. A new active path from CE1-PE2-PE4-CE2 is thus created that avoids the failure.

Note that two key objectves are achieved through the use of MC-LAG and PW redundancy in this manner:

- The Ethernet service stays operationally up, despite the failure of an attachment circuit. This is to be distinguished from traditional VPWS services where there is no redundancy of the ACs.

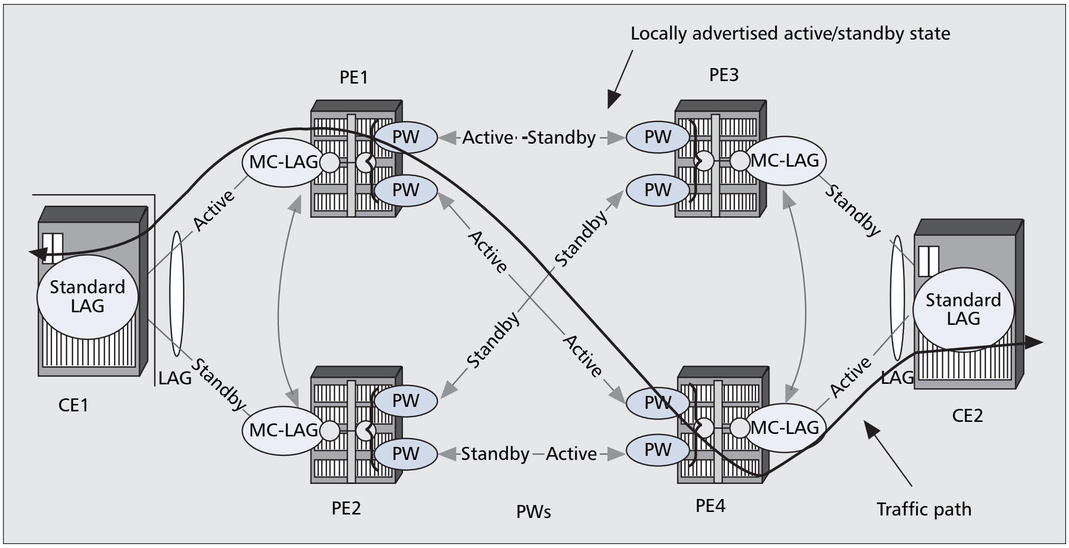
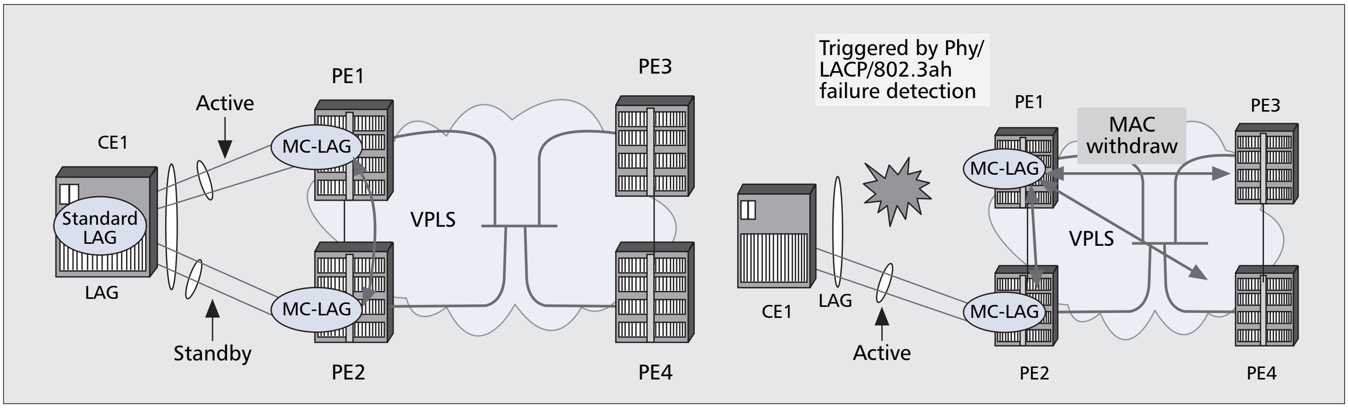


Figure 5. *MC-LAG and PW redundancy for VPWS.*



■ **Figure 6.** *VPLS access protection using MC-LAG.*

- The failover operation is transparent to the far end CE. That is, only the PEs and the CE where the failure occurred are aware of the switchover. This is important for large-scale deployments where it is desirable to localize any failover operations in order to minimize the load on the network and minimize the failover time.

As an optimization, additional protection can be provided using interchassis backup (ICB) PWs between each of the redundant PE pairs (for simplicity these were ignored in the above discussion and Fig. 5). These PWs enable a PE to forward "in-flight" packets received from the MPLS network over a PW destined for a locally failed MC-LAG subgroup to the local redundant PE during the transient period when the MC-LAG has switched over before the remote PW status. They also enable an unrecoverable failure in the core of the MPLS network to be avoided by allowing a PE to send packets toward the MPLS network by using an alternative path via a local redundant PE.

PW redundancy and MC-LAG can also protect the VPWS service where one of the PEs (e.g., PE1) experiences a catastrophic failure. Such a failure can be detected by the other PEs in a number of ways, such as a failure of T-LDP hello messages, an operations, administration, and maintenance (OAM) protocol such as LSP Ping on the LSP tunnel between the PEs, or the MC-LAG control protocol between PE2 and PE1 (the failed PE will no longer respond to MC-LAG control keepalive messages). This triggers PE3 and PE4 to place their PWs to PE1 in an operationally down state, and PE2 to assume the active forwarding status. PE2 thus advertises a PW status of active for its PWs to PE3 and PE4. PE3 will remain in standby status (because its local MC-LAG state is standby), but PE4 will now forward packets on the PW PE4-PE2 as both ends now show active status. In order to ensure the correct flow of frames to and from CE1, PE2 changes the LACP link status from standby to active. As in the case of the failure of an AC, the failover operation only impacts locally connected PEs and the local CE. There is no switchover forced on the remote CE.

Now consider how MC-LAG can be used to enhance the resiliency of VPLS services. Figure 6 illustrates how MC-LAG protects the access to a VPLS.

CE1 is connected by a MC-LAG to two PEs in a VPLS. Standard LACP is used to select which LAG subgroup is active and which is on standby. Consider the failover operation when one of the LAG subgroups fails. Initially, the subgroup from CE1 to PE1 is active, and the subgroup from CE1 to PE2 is on standby. A failure of the link between CE1 and PE1 is detected by PE1 through physical layer, LACP, or Ethernet OAM mechanisms. This triggers the MC-LAG control protocol to make PE2 the active PE. Because PE2 is now the active PE in the redundant pair [PE1/PE2], it changes the LACP link status from standby to active to CE1 (CE1 may now forward using PE2's MC-LAG link).

VPLS PEs contain virtual bridges with MAC tables that provide forwarding information for all of the Ethernet MAC addresses known to the PE. Therefore, a failure of the active LAG sub-group on a PE will render the MAC forwarding information for that PE invalid. In order to prevent Ethernet frames for CE1 being misdelivered to PE1, PE1 sends a MAC withdraw message to its connected PEs. In VPLS this message is carried in the LDP signaling used for the constituent PWs, and causes the PEs to remove those MAC addresses from their forwarding tables. PEs participating in the VPLS will then learn the identity of the new PE to which frames should be forwarded for CE1 by flooding any packets destined for unknown MAC addresses to all active PEs and installing the source MAC address for packets received from the new PE in their forwarding tables.

As well as providing redundancy at the service provider network edge, MC-LAG can also be used to protect the interconnection between service providers' Ethernet networks. For example, Fig. 7 shows an application where MC-LAG allows redundant PEs and Ethernet links to interconnect two metro Ethernet networks that use VPLS. One pair of redundant PEs assumes a slave role with respect to the other. LACP is then used between the redundant PE pairs to signal the active or standby state of the subgroups in the LAG, in a similar manner to the access redundancy case shown above.

CONCLUSIONS

Multi-chassis LAG and pseudowire redundancy provide a reliable and simple end-to-end protection scenario for point-to-point and point-to-

Multi-chassis LAG and pseudowire redundancy provide a reliable and simple end-to-end protection scenario for point-to-point and point-to-multipoint data services re-using existing LACP mechanisms in Ethernet access nodes.

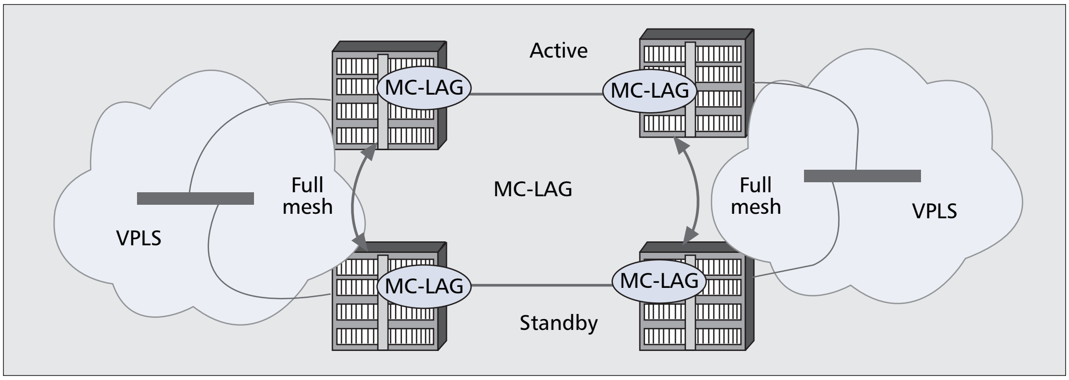


Figure 7. *Inter-metro resilience using MC-LAG.*

multipoint data services reusing existing LACP mechanisms in Ethernet access nodes. Using the techniques described, providers can go beyond using LAG technology as a simple way to increase capacity by using LAGs to provide increased redundancy, both at the network edge and within the service delivery infrastructure. The pseudowire redundancy in conjunction with multi-chassis LAG capability provides a unique way of extending redundant connections to the network access, increasing uptime in triple play services when used in conjunction with Ethernet-based DSLAMs, or in business services when used with Ethernet CPE. Having maximized the edge redundancy, providers can combine multi-chassis LAG with both VPLS and Ethernet VPWS services to achieve end-to-end redundancy across their network.

References

[1] Reeve *et al.*, "Networks and Systems for BT in the 21st Century," *IEE Commun. Eng.*, Oct./Nov. 2005.
[2] Andersson *et al.*, "Framework for Layer 2 Virtual Private Networks (L2VPNs)," IETF RFC 4664, Sept. 2006.
[3] Lasserre *et al.*, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling," IETF RFC 4762, Jan. 2007.
[4] Bryant *et al.*, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture," IETF RFC 3985, Mar. 2005.
[5] Bates *et al.*, *Converged Multimedia Networks*, Wiley, Aug. 2006.
[6] Martini *et al.*, "Encapsulation Methods for Transport of Ethernet over MPLS Networks," IETF RFC 4448, Apr. 2006.
[7] Martini *et al.*, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)," IETF RFC 4447, Apr. 2006.
[8] Leelanivas *et al.*, "Graceful Restart Mechanism for Label Distribution Protocol," IETF RFC 3478, Feb. 2003.
[9] IEEE 802.3, "Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications — Section Three," 2005.
[10] IEEE 802.1D-2004, "IEEE Standard for Local and Metropolitan Area Networks Media Access Control (MAC) Bridges," 2004.
[11] Pan *et al.*, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels," IETF RFC 4090, May 2005.
[12] Muley *et al.*, "Pseudowire (PW) Redundancy," Internet draft, draft-muley-pwe3-redundancy-01.txt, Mar. 2007.
[13] Muley *et al.*, Internet draft, "Preferential Forwarding Status Bit Definition," Internet draft, draft-muley-dutta-pwe3-redundancy-bit-01.txt, July 2007.

Biographies

Matthew Bocci (Matthew.Bocci@alcatel-lucent.co.uk) is director of technology and standards with Alcatel-Lucent's IP Division. He is a regular contributor to the IETF, where he co-chairs the ANCP working group and is secretary of the PWE3 working group, and the IP/MPLS Forum where he chairs the interworking working group. He is co-author of a number of publications, and IETF drafts and RFCs on MPLS-based converged networks. Previously, he provided advanced technical consulting in traffic management, signaling, and network performance. He holds a Ph.D. in ATM network modeling from Queen Mary & Westfield College, London, and a B.Eng. (hons) (1st class) degree in electrical and electronic engineering from University College London. He is a member of Alcatel-Lucent Technical Academy and the Institution of Engineering and Technology.

Ian Cowburn is a consulting engineer with Alcatel-Lucent. He has been in the networking industry for over 25 years and has a Master's degree in computer science from Manchester University, United Kingdom. His previous networking experience includes various posts within Digital Equipment, Cabletron Systems and Riverstone Networks, all following the industry's technology evolutions to today's carrier class MPLS networks.

Jim Guillet is senior director of IP marketing activities to service providers at Alcatel-Lucent. He leads a team focused on the transformation of operators' single-service IP networks to multiservice IP/MPLS networks. His team communicates Alcatel-Lucent's Triple Play Service Delivery Architecture (TPSDA) to operators, media and industry analysts. He has held a number of marketing and product management roles during his 25 years in the telecommunications industry. He holds a degree in engineering from Queen's University.