Metropolitan Ethernet Network:

A Move from LAN to MAN

Minh Huynh, Prasant Mohapatra

Computer Science Department, University of California at Davis
Davis, California, USA
Emails: {mahuynh, pmohapatra}@ucdavis.edu

Abstract

Ethernet has been the indisputable technology of choice for the local area networks (LANs) for more than 30 years. Its popularity is due to its versatility, plug-n-play feature, and low cost. It has transformed from a CSMA/CD technology providing low throughput to a full duplex link increasing the throughput 1000 folds. Despite these improvements, Ethernet is still restricted to local area networks, and is not ready to become a carrier grade technology for wider areas. However, there are efforts to assist the transformation of Ethernet from the mainstream LAN technology to the possible adoption for metropolitan area networks (MENs). This paper will introduce the movement from basic Ethernet to the carrier grade Ethernet for MENs. The paper describes the underlying technology, offered services, the state-of-the-art, and the comparison between various technologies. In the context of the move from LAN to MAN, various problems and their corresponding solutions are discussed, along with the future of Metro Ethernet Network.

1 INTRODUCTION

Traditionally, Ethernet is a plug-n-play technology at the link layer intended for Local Area Network (LAN). Its success is in parts due to its standardization that enables the interoperation among different equipment vendors. Thus, mass production drives down the cost of Ethernet and also advances the popularity of basic Ethernet further. From the earlier Ethernet that uses CSMA/CD and runs at 10Mbps on the coaxial cable, Ethernet now can run full-duplex 10Gbps links with backward compatibility. It does not need any special device to convert between equipments running at different speeds displaying a true plug-n-play system.

Initially, Ethernet was designed to operate on a bus topology using the coaxial cable at 10Mbps. It was a broadcast environment where there is the possibility of frame collision. Using the CSMA/CD, Ethernet successfully sent frame to other hosts while entering the exponential backup phase if there is a collision. Several versions of Ethernet technologies existed ranging from 10Mbps to 10Gbps running on coaxial cable, twisted pair copper cable, and fiber optic line. However, all of the different Ethernet versions kept the same frame structure for backward compatibility.

Ethernet evolves from a LAN service interconnecting an enterprise workgroup to running the enterprise backbone. Now extending to the Metro Area Network (MAN), Ethernet provides Ethernet services across MAN. MAN makes up of a metro core network and several access networks. The access networks border with the subscribers networks. Subscribers include business enterprise networks and residential network such as DSL and cable services. The metro core is the backbone of MAN where it interconnects the access networks hauling large trunk of traffic. In addition the metro core provides the subscribers with access to the Internet.

Ethernet in MAN is an alternative to the traditional Time Division Multiplexing (TDM) technology. Although TDM is used to delivered voice and leased-line services, it is inefficient for delivering the emerging data oriented applications. Ethernet services can offer the point-to-point line service or multipoint-to-multipoint LAN services. LAN services connect multiple sites belonging to the same enterprise across different physical locations into a virtual LAN as if all sites exist in a local building.

An example of emerging data oriented applications running over Metro Ethernet Network (MEN) is LAN to network resources [51]. LAN to network resources can offer services such as the backing up data of enterprises at a remote and secured site for disaster recovery. Customers can backup and recover their data constantly across the metro. For residential areas, LAN to network resources can distribute multimedia services. For example, video servers can be deployed at a Points of Presence (POP) where the residents can access for broadband video on demand over an Ethernet connection. Other services that MEN can offer include [51] Internet connection, Extranet, Storage Area Networks (SANs), Metro Transport, and VoIP. Around the world, different applications are the main driving force for MEN. For example, in Korea, the growing game parlor business is the bandwidth hog. Japan focuses on the inter-office connection between large multi-sites enterprises that span across remote physical location. China and India are building a common platform for the residential triple play: voice, video, and data.

In addition, the advantages of Ethernet such as cost effectiveness, flexibility, rapid provision on demand, and ease of interoperability drive the adaptation of Ethernet into MAN. The mass production of Ethernet equipments and the simplicity of Ethernet technicality keep the cost of having Ethernet relatively low compared to others competitive

protocols. Another factor in cost saving is the ease of interoperability without third party converter or sometime without purchasing new equipments. The same Ethernet interface can support a variety of bandwidth unlike legacy technologies. One feature that makes Ethernet stands out than the rest is the flexibility in bandwidth upgrade. With bandwidth increment as fine granularity as 1Mbps, Ethernet offers better bandwidth efficiency than TDM. Therefore, it is able to have rapid provision on demand.

Ethernet in combination with VPLS is the convergence technology that brings mass traffic together from diverse platform. Besides the high speed wired networks, MEN is the cost effective backhaul for the mobile carriers. The Carrier Ethernet will become the common "transport layer" to deliver multiple services over a single connection.

In the remaining parts of this report, we will introduce the basic Ethernet to the carrier grade Ethernet for MEN. We will explore the underlying technologies, offered services, and architectures from both the industry and the academia literatures. Challenges and corresponding solutions are also discussed.

2 DEMAND FOR METRO ETHERNET NETWORK

We are on the verge of witnessing the transformation of Ethernets from the traditional local area networks within buildings to wider metropolitan areas. This gradual expansion of the scope is guided by the growing needs as well as the versatility of the protocol. In this section, we overview the motivations and the characteristics of Ethernet that make it a suitable candidate for this broadening scope of usage.

2.1 Motivation for the MEN transformation

Incumbent technologies such as Private Line (PL), Frame Relay (FR), and Asynchronous Transfer Mode (ATM) cannot respond as fast as Ethernet to the high volume demand for new connections because of the long waiting period to establish a dedicated physical connection. For example, an incumbent carrier takes three to six months to deploy a T1 circuit [51]. In addition, upgrading the current connection exposes the inefficiencies in Time Division Multiplexing (TDM), such as coarse granularity of bandwidth increments resulting in oversubscribing, requirement of new equipments, and changing to new service platforms and protocols. For instance, a T1 connection running at 1.5 Mbps will be upgraded to a DS3 connection at 45Mbps. An alternative solution is to provide multiple T1 connections [51]. Both result in purchasing of new equipments. In contrast, Ethernet can provide bandwidth increment with 1Mbps granularity. The same Ethernet protocol can be used from 10Mbps to 10Gbps. It is 10 times lower cost than high speed SONET interfaces [29]. Therefore, the agility to respond to customers' need and the cost efficiency drive the Ethernet expansion to the carrier grade for Metro Ethernet Network. Figure 1 shows that the worldwide revenue forecast study from the Metro Ethernet Forum (MEF) indicating that currently FR, ATM, and PL combined together take a larger part of the market than Metro Ethernet. However, with the current growth rate of Ethernet, Metro Ethernet will eventually take over the lead as projected. In this growth rate, 14% of Ethernet services result from new services deployment while the remaining 86% result from the replacement of legacy services from a study by the Vertical System Group [57]. In addition, Figure 2 shows that Ethernet can save more than 50% over a 3 year period in a business case study from the MEF [44]. This operational cost includes Internet access, Private Data, and Monthly Recurring Cost.

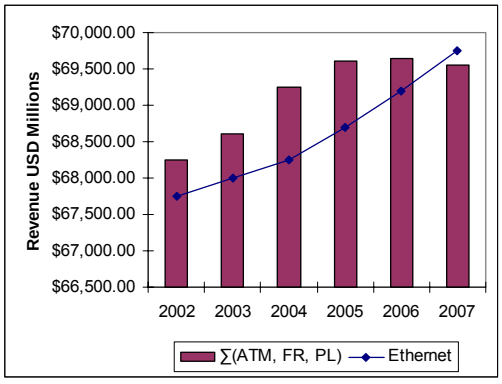

Figure 1: Worldwide revenue forecast Metro Ethernet vs. FR, ATM, and Private Line [43]

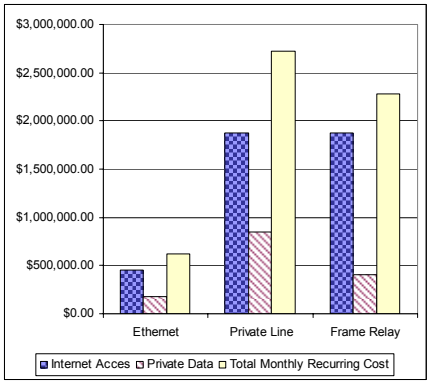

Figure 2: Recurring Cost of Operation in a 3 year period study [44]

2.2 Advantages of MEN

Over the last decade, bandwidth has increased significantly in the backbone network making the metro as the bottleneck. Other legacy services such as T1, T3, or ATM do not provide the flexibility in bandwidth increment that MEN need. Furthermore, Ethernet is a popular protocol in all enterprise LANs. The choice of using Ethernet to interconnect remote sites of an enterprise is appealing for the following reasons: cost effectiveness, flexibility, rapid provision on demand, and ease of interoperability [27].

Since Ethernet equipments are very common on the market, its material and development costs have been kept competitively low. Ethernet's level of technical complexity is relatively lower than the others. Therefore, the capital expenditure and operational expenditure are also low. This cost effectiveness drives the expenses on Ethernet services in MEN down compared to other technologies. A business case study by the MEF in July 2003 showed that compared to legacy SONET/SDH, Ethernet based services save 49% on operational expense and 39% on capital expenses [50]. Using Carrier Ethernet as the common platform to backhaul mobile traffic, the saving on operational expense ranges from 15% by the Yankee Group study to 40% by the MEF study [57]. TABLE 1 shows the list prices for a monthly leasing bandwidth by two leading carriers. From 10Mbps and up, Ethernet is 3-5 times cheaper than E1/T1 line.

TABLE I. LEASING PRICE QUOTES FOR MOBILE CARRIERS

Bandwidth	***Verizon USA***	***BT UK***
E1/T1 (~2.048Mbps/1.544Mbps)	780	500-1090
Ethernet 4Mbps	N/A	1000
Ethernet 10Mbps	1430	1120
Ethernet 50Mbps	2130	1450

Some of the problems with the legacy services are the long wait for the service to be installed and activated and the coarse bandwidth granularity. What the enterprises need is rapid provision on demand of services. For example, an enterprise might need high bandwidth provisioning during the day and low bandwidth provisioning in the evening and on the weekend. However, it is not possible to do so with the legacy technologies. The enterprise ends up paying for the peak bandwidth for all time because the service installation is not on demand. In addition, enterprises have to buy the bandwidth in large chunks. It is not possible for them to have fine granularity of bandwidth increment. In contrast, Ethernet service offers bandwidth increments in term of 1Mbps. The same

Ethernet interface, e.g. 1Gpbs, can support a variety of bandwidths. Therefore, bandwidth-on-demand can be easily provisioned.

The plug-n-play feature of Ethernet enables a simple migration from low speed to high speed without any third party converter. Also, Ethernet services reduce the complexity of protocol translation between different platforms and systems.

2.3 The move from LAN to MAN

After recognizing the expectation for MEN, the next step is the transformation of Ethernet from the LAN environment to the MAN environment. Traditional Ethernet is used to be deployed on small segments of workstations. These segments are then connected to create an intranet that traditionally exists within the same geographical site. This setup is referred to as Local Area Network (LAN) where the geographical area is relatively small, and all of the traffic belongs to the same enterprise. In contrast, Metro Ethernet Network (MEN) spans across a metropolitan area. MEN is comprised of a core network and several access networks as shown in Figure 3. All the access networks connect to the core at one or two aggregation Ethernet switches. The customers' networks are connected to the access network; and the core helps in interconnecting the access networks. Packets hop through multiple switches in both access and core networks. Redundant links are used both in the core as well as the access networks. Since traffics coming from different enterprises traverse the same network, a traffic isolation mechanism is needed inside MEN. On the other hand, there is a need to merge traffic belonging to the same enterprise but coming from multiple geographic locations.

To deploy MAN, there are three options. The first option is to extend the core technologies such as IP/MPLS into the access network. This creates complexity in operation of one large network that is very difficult to configure. There are also incompatibility issues between different software environments. Also, core technologies have high equipment cost. The second option is to deploy Ethernet in the access and the core. One major drawback is that Ethernet does not support such large and flat network, that is, no hierarchical structure is defined. In addition, Ethernet lacks traffic management functions, Service Level Agreement (SLA) mechanism, and security protection. The last option is the hybrid of deploying Ethernet in the access network and using MPLS, Resilient Packet Ring (RPR), or other core technologies for the core. This has the benefit of providing Ethernet at career class with the simplicity and low cost of the traditional Ethernet.

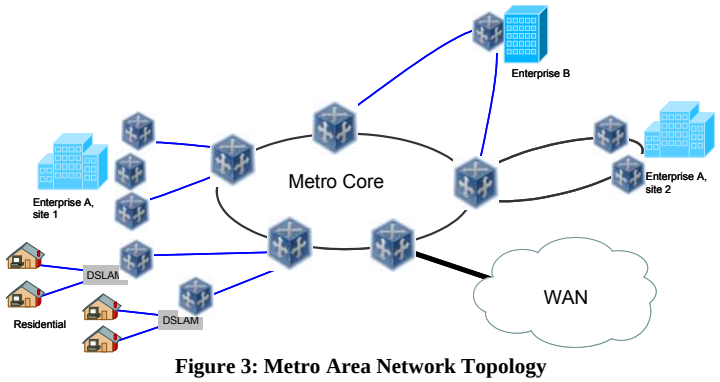

Figure 3: Metro Area Network Topology

3 MEN TECHNOLOGY

In this section, we will discuss the standardized technologies that are used to support Ethernet in MEN. Since

Ethernet by itself does not have all of the features to support services required by the customers, it relies on those that can complement it. The difficulty in these inter-networking is the mapping from one platform to another. As always, interoperability requires the use of tunneling. In MEN, there are two main tunneling approaches: Virtual Private Wire Service (VPWS) for point-to-point and Virtual Private LAN Service (VPLS) for any-to-any or multipoint. In general, point-to-point behaves like a single connection while multipoint behaves like a LAN.

3.1 Point-to-Point

Point-to-Point service is used to connect only two User Network Interfaces (UNI) together. Virtual Private Wire Service (VPWS) is an emulation of L2 Virtual Private Network (VPN) for Ethernet tunneling a point-to-point connection between two ends. Figure 4 shows the L2 VPN architecture. VPWS is categorized into Ethernet Relay Service (ERS) or Ethernet Wire Service (EWS) [42]. ERS uses the VLAN number and offers services similar to frame-relay. The services are shared and multiplexed at the UNI. In contrast, EWS is a port-based service where traffic transporting over a port is treated as a private line. VLAN numbering is neglected in EWS. The encapsulation of VPWS uses the draft-Martini approach [26]. Figure 5 shows the protocols needed to support VPWS. A list of VPN-related IETF drafts is listed in [42].

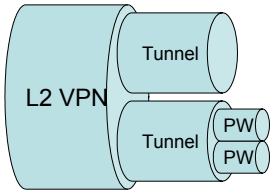

Figure 4: L2 VPN architecture. The pseudowire (PW) is composed of virtual circuit (VC) label and the remote PE. Tunnel serves as header for routing within the provider network.

Draft-Martini [26] is a tunneling protocol for a point-to-point connection. It was intended as a carrier backhaul or high-speed connection between major sites. In MEN, it is a connection between two User Network Interfaces (UNI). One of IETF's working groups defines pseudowire emulation edge to edge (PW3) based on the draft-Martini. The pseudowire is being used to offer layer-2 transport across the MPLS core. It specifies a virtual circuit label and the remote Provider Edge (PE).

When an Ethernet frame enters the provider network, as shown in Figure 6, the ingress router stacks two labels on it: virtual circuit (VC) label and tunnel label. The VC label stays the same as the frame traverses across the network. It is used for multiplexing purposes when frames arrive at the destination PE. Each pair of PE has a unique VC label. The tunnel label is locally significant at each hop for routing purposes within the MPLS domain. It can also provide multipoint-to-multipoint service but it will suffer from the n-squared problem whereas n2 connections are required to connect n locations. Its main purpose is to support the E-line service.

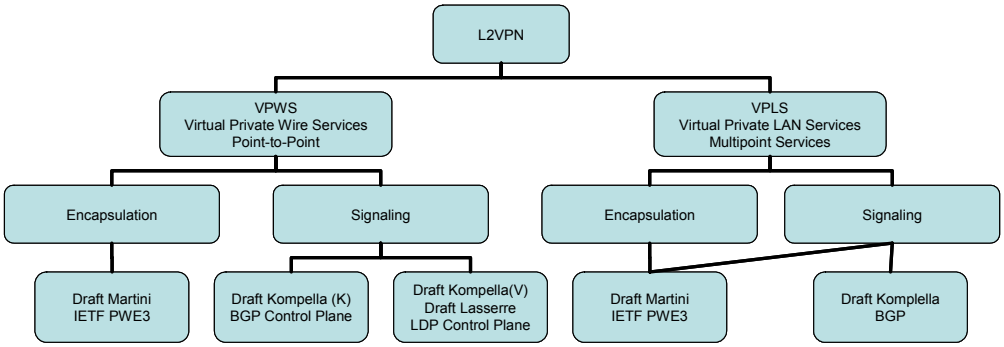

Figure 5: L2 VPN protocol classification

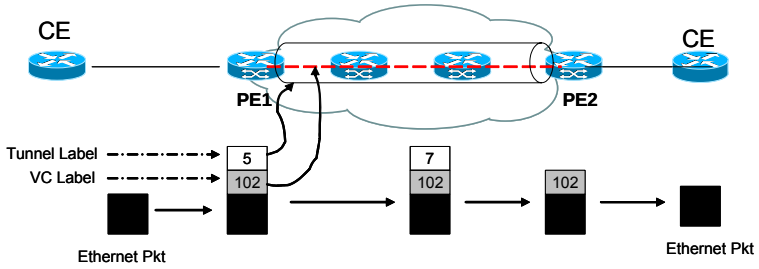

Figure 6: Virtual Private Wire Service (VPWS) for point-to-point [51]

Auto-provisioning of pseudowires is based on the colored pool concept [42]. Different VPNs are assigned with different colors. For example, PE1 and PE2 each have a green pool of attachment circuits (AC). When PE1 discovers PE2, it removes an AC from its green pool and binds the AC to the pseudowire connecting PE2. Similarly, PE2 binds one its ACs from the green pool to the pseudowire connecting PE1. After the auto-discovery, a full mesh of pseudowires is created between all the ACs belonging to the same VPN.

3.2 Multipoint

In multipoint service, more than two UNIs are connected together to form a LAN. A frame sending from one UNI will arrive at all the remaining UNIs within the same virtual LAN. Similar to point-to-point service, multipoint service are categorized into two services: Ethernet Multipoint Service (EMS) and Ethernet Relay Multipoint Service (ERMS) [49]. EMS behaves similar to Ethernet LAN like address learning and unknown address broadcasting. However, each UNI can only receive one service. On the other hand, ERMS can multiplex different services at each UNI.

Carriers have offered legacy services such as ATM, Frame Relay and private line for quite sometime. Therefore, they have a large share in the market. Reports on Virtual Private LAN Service (VPLS) [37], [39], [40] answer the need to integrate new Ethernet services while carrying on legacy services with large market share. VPLS, also known as Layer 2 MPLS, is built on pseudowire as a multipoint tunneling scheme. It offers multi-point connectivity by virtualizing enterprises with remote LAN sites onto the same LAN supporting the E-LAN services, as shown in Figure 7. Similar to MPLS at layer 3, VPLS offers the same services at layer 2. The difference is in the interface between the Customer Edge equipment (CE) and the Provider Edge equipment (PE). In MPLS layer3, the CEs are IP routers as opposed to Ethernet bridge/switch/hub or router in VPLS, allowing both non-IP and IP traffic to be routed. In addition, VPLS can emulate the behavior of Ethernet LAN such as broadcasting of unknown MAC addresses and MAC address learning.

VPLS is defined in two IETF drafts: VPLS-LDP [24] and VPLS-BGP [25]. Figure 5 shows a mapping of the

different protocols for VPLS. VPLS-LDP uses LDP protocol as the signaling protocol to establish a full mesh of LSP between PE nodes. It is backed by majority of the vendors such as Atrica, Cisco, Extreme, Force10, Foundry, Nortel, Riverstone, Cosine Communication Inc, Laurel, Overture, Timetra, Vivace. One advantage that VPLS-BGP has over VPLS-LDP is the discovery of the neighbor PEs since BGP has that capability built in. VPLS-LDP could incorporate BGP into it or using a directory-based approach such as Radius [45] that is being discussed within the IETF. Juniper is the only vendor that supports VPLS-BGP because it is one of the first vendors to develop VPLS. Since Juniper has invested a lot into the BGP approach, it is difficult to switch to a new approach [6].

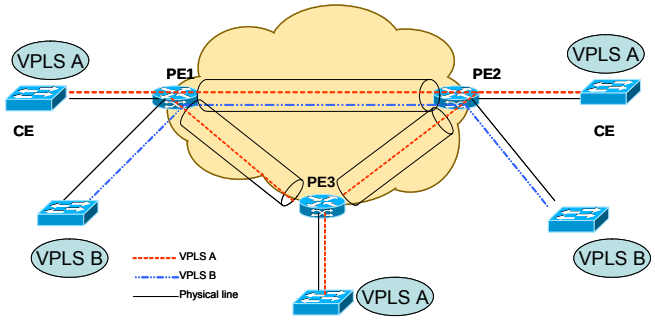

Figure 7: A sample VPLS configuration [51]

For each VPLS, VPLS-LDP creates a full mesh of tunnels by first using UDP to determine neighbors, then establishing a TCP session to request for label mapping. Next, it needs to define a VPLS ID and establish virtual circuit (VC) labels for LSP. The outer tunnel label and the VC label are tagged onto the front of each Ethernet packet header for switching. Each virtual circuit Label Switching Path (LSP) is a bidirectional pseudo-wire inside the outer tunnel. Now the PEs act like bridges and perform the following functions: Learning and aging MAC addresses on a per LSP basis, flooding of unknown frames, and replication for unknown, multicast, and broadcast frames.

To create a loop free routing environment, VPLS uses the split-horizontal technique instead of the Spanning Tree. In split-horizontal, a PE would not forward packets that it had received from one PE to another PE. The packet is still guaranteed to reach the destination because the networked has a full mesh topology. Essentially, the source PE broadcasts the packets to all of its adjacent neighboring PEs.

To better scale VPLS, a hierarchical VPLS (HVPLS) topology is laid out to a hub-and-spoke topology where the PEs act as hubs and the simple switches terminate each spoke, as shown in Figure 8. This approach minimizes the topology of the full mesh, reducing the number of LDP peers. Only the core network is needed to have a fully mesh topology. Another problem is the explosion of MAC addresses since the MAC addresses have a flat structure. One approach is to use routers for the customer/provider edge (CPE) devices. Therefore, each site is reduced to one address that the switches have to learn. The other approach is to limit the number of addresses that can be learned per access circuit.

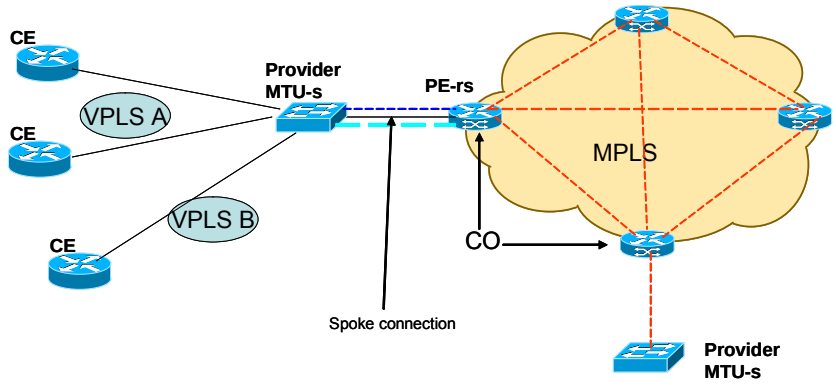

Figure 8: A hierarchy VPLS topology. MTU-s is Multi-Tenant Unit running on switches. PE-rs is Provider Edge equipment running either router or switch. CO stands for Central Office. CE is the Customer Edge [51].

4 METRO ETHERNET SERVICES

There are two services in MEN: E-Line for a point-to-point service and E-LAN for a multipoint service. These services arise from the tunneling approach of MEN, specifically VPWS and VPLS. Both of these services are similar in the parameters for the quality of service. Their difference lies in the connectivity between end-points.

4.1 E-Line Service

Ethernet Line Service (E-Line) is a point-to-point Ethernet Virtual Circuit between two user network interfaces (UNI) [28]. E-Line can provide a simple best effort service on the bi-directional line or with some performance assurances. E-Line performance assurance includes Committed Information Rate (CIR) and the associated Committed Burst Size (CBS), Peak Information Rate (PIR) and the associated Peak Burst Size (PBS), delay, jitter, and loss performance assurances. Each UNI can multiplex more than once Ethernet Virtual Circuit (EVC) if there are more than one EVC connected to it. In Figure 9, CE1 is multiplexing EVC from CE2 and CE3.

4.2 E-LAN Service

Ethernet LAN Service (E-LAN) is a multipoint-to-multipoint service connecting at least two UNIs [28]. Each UNI is connected to the same multipoint EVC so that data sent from one UNI can be received at multiple ends, as shown in Figure 10. In addition, if a new UNI is added, only the new UNI is needed to add to the multipoint EVC. In contrast, if a new UNI is added in E-Line, a new EVC must be added to every existing UNI in the same service. Similar to E-Line, E-LAN can provide best-effort service or quality-assured service with parameters such as CIR, CBS, PIR, PBS, jitter, delay, and loss performance.

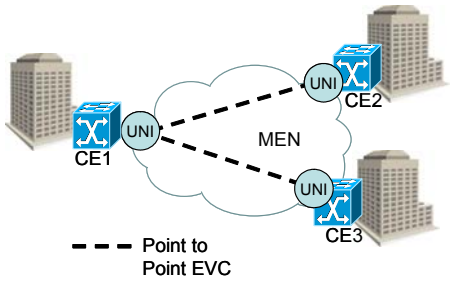

Figure 9: E-Line Service

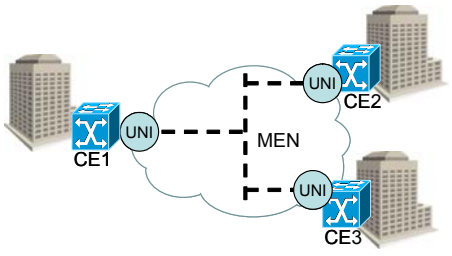

Figure 10: E-LAN Services

5 CHALLENGES

The MEF defined the five key attributes of a Carrier Ethernet service: Standardized Services, Scalability,

Resilience, Quality of Service, and Service Management. Meeting these requirements would truly enable the full potential and benefits of MEN to support a large number of next generation applications and services including: Business Services, Residential Triple-Play, and Mobile Backhaul.

5.1 Standardization

Currently, there are solutions that aim at resolving issues in MEN remaining proprietary such as VLAN stacking [22], MAC-in-MAC [36], Extreme Standby Router Protocol (ESRP) [17], or Virtual Switch Redundancy Protocol [19]. However, in order for E-Line and E-LAN services to be provided transparently, it requires a ubiquitous service where little or no changes to customer equipments on existing networks. For this global compatibility to exist across different platforms, equipment vendors must be able to supply standardized equipments. Ideally, equipments from different vendors should be able to work together to converge services such as voice, video, and data and to converge legacy technologies and emerging technologies. When customers need to change their subscriptions or switch to a different carrier, they would not need to completely replace their existing equipments. Therefore, there should be a set of standardized requirements that require any Metro Ethernet certified equipment to pass before it can be rolling off the assembly line. As of this writing, there are 17 MEF specifications [56] (MEF1, MEF5, and MEF5 are superseded by MEF10.1) that define the requirements for MEN ranging from service definition to test suites. However, these specifications are recommendations and are not enforced as standardizations.

5.2 Scalability

With the fast growing Ethernet Services in MAN, a single MAN is expected to support thousands of equipments that in turn providing services for million of users. The second key attribute of a Carrier Ethernet require a robust and dynamic ability to support growing number of users converging on a network where voice, video and data applications aggregate from variety of business enterprises and residential areas. Furthermore, Metro Ethernet services expand from one Metro access network to other Metro access network globally to support a wide variety of VLAN that exist in multiple remote geographical regions. In addition, emerging applications increase the bandwidth usage so that any MAN technologies must be able to scale from 1Mbps to 10Gbps. Unfortunately, very few Ethernet solutions are scalable. The traditional Ethernet management protocol, the Spanning Tree Protocol (STP), would not be able to scale in MAN. At most, STP can span to 7 hops. Beyond that, STP's behavior is unpredictable. Most technologies that were designed for Ethernet were intended to run in a LAN environment. Therefore, the move from LAN to MAN requires new considerations in MAN technology.

5.3 Resilience

The next key attribute of Carrier Ethernet is resilience that is defined by the detection and recovery ability of an Ethernet technology. The ideal network would perform smoothly and transparently to the user in the face of failures meeting the demanded quality and availability as agreement in the subscription. After the fault detection, the network should autonomously recover from it. Recovery means that an alternate path is provided through a network reconfiguration or a backup path. Optical networks set the bar at sub-50ms for recovery time and it has become the industry standard. There are debates going on whether this sub-50ms is needed. For some applications, a recovery of longer than 50ms is acceptable such as file transfer or email.

Despite its popularity and simplicity, the traditional Ethernet does not meet this requirement. The recovery time of the Spanning Tree Protocol (STP) that manages the Ethernet is in the range from 1 second to 60 seconds depending

on the version of the STP, which is considered drastically slow. Such performance hits can interrupt or slow down applications, in turn cause great financial loss to enterprises. Other Ethernet solutions have mixed success in this area. Some are very resilient but they require specific network configurations or have high complexity that is costly. Others have lower cost in exchange for performance.

5.4 Quality of Service

One of the most important attributes that a Carrier Ethernet must have is the support of end-to-end Quality of Service (QoS). It includes, but not limited to, bandwidth, delay, jitter, and packet loss guarantees. These guarantees must be made from an end-to-end point of view. Service Level Agreements (SLAs) are the agreements on these metrics that are negotiable between the client and the carrier to assure the end-to-end performance of voice, video, and data applications. MEF defines QoS specification via the following metrics: CIR, frame loss, delay, and jitter.

In supporting QoS, there exist several traffic engineering mechanisms such as traffic policing and traffic shaping. Traffic policing is the act of dropping customers' packets when they exceed the service level agreement. It can be softening through the marking of packets that reach a certain threshold such as the packet coloring scheme. The marked packets are more likely to be dropped than the unmarked ones. Traffic shaping involves the isolation of traffic in each queue in order to protect from the burstiness of another queue by placing an upper bound on the maximum bandwidth available to a traffic class. Often included with QoS is network load balancing that is the redirection of traffic flows to prevent network load imbalance that leads to traffic congestion. Load balancing can be done at the micro level where the individual links are controlled in a distributed manner. Alternatively, it can also be done at the macro level where the load is controlled per traffic class or domain.

In Ethernet, the widely adopted STP lacks the support for assured QoS capability and load balancing capability. At the most, it can provide marking for class of service via the priority bits as defined in 802.1p. Therefore, without further enhancement, Ethernet is not fit to be a carrier class technology. Many equipment vendors have implemented their proprietary schemes in light of STP's drawbacks but none can provide a complete off-the-shelf solution.

5.5 Service Management

The last requirement for Carrier Ethernet as defined by MEF is Operation, Administration, and Maintenance (OAM). OAM is the ability to monitor, diagnose, and manage the network autonomously or through a central standard interface implementation that is vendor independence. Since the original intention of Ethernet aims at LAN environment, it does not include the OAM capability in the standard. However, the large scale MAN environment requires the crucial OAM feature that exists in optical network. Little works have focused on this area except for the on going standardizing process by the standard bodies ITU, IETF, and MEF.

6 ARCHITECTURE and PERFORMANCE

In this section, we will look at solutions developed for Ethernet in response to the existing challenges to push forward the transformation of Carrier Ethernet. These challenges include resilience, load balancing, quality of service (QoS), and scalability. Efforts are ongoing in both academia and industry to enhance each of these areas and to supplement the standardized protocols. In the MEN context, the presented solutions are categorized based on where they would better be deployed: the access network or the metro core. There is one class of solutions that would work in both the access network or the core network but they are not stand alone architecture. These are

categorized as supplement solutions as they can be used concurrently with other protocols.

6.1 Metro Access Solutions

As earlier described and depicted in Figure 3, a typical topology of a metro access network is the mesh topology. In the dense mesh topology, there are many redundant links to route traffic. Therefore, a suitable protocol to manage the access network must have a high utilization to take advantages of the redundant links. It must also be highly scalable to support millions of subscribers and to aggregate large volume of traffic from diverse platforms onto a common platform. A summary of the solutions in the metro access with respect to the 5 key attributes as defined by the MEF is shown in TABLE II.

TABLE II. SUMMARY OF THE PROTOCOLS FOR METRO ACCESS

Solutions	***Standard***	***Resilience***	***Scalability***	***QoS***	***OAM***	***Ready as stand-alone***
PESO	Academia Publication	high	scalable	assured bandwidth	none	no
AREA	Industry Proprietary	high	low: limited by VLAN tag	MPLS supported	none	no
Ethereal	Academia Publication	low	low	assured end-to-end QoS metrics	none	no
SmartBridge	Academia Publication	low	low	none	none	no
STAR	Academia Publication	low	low	none	none	no

6.1.1 PESO

A proposed scheme aims to protect Ethernet traffic over SONET with a low overhead is called PESO, proposed by Acharya et al. [1]. In traditional SONET, voice traffic is supported by a primary and backup path providing 100% protection. This approach provides fast recovery upon failure but it imposes high operation cost. However, in data traffic, whenever there is a failure, it is not necessary to have 100% protection because it can tolerate the failure by running at a reduced rate. Depending on the protection requirements, PESO will compute an optimum routing path and using virtual concatenation (VC), as shown in Figure 11, and Link Capacity Adjustment Scheme (LCAS) to make the necessary recovery. For the scenario where a single failure should not affect more than x% of the bandwidth, PESO transforms the link capacity in the topology to the equivalent STS-y line. Each chosen line cannot carry more than x% protected bandwidth. PESO determines the number of members in the VC. Using path augmentation maximum flow algorithm such as Ford & Fulkerson [47] or Edmonds & Karp [48], PESO determines the routes that the virtual concatenation group (VCG) will take. Upon failure, LCAS removes the failed member resulting in a continuous connection with the destination but the throughput has been reduced not less than x% protected bandwidth. A variance of this protection is that the provider wishes to minimize the performance degradation. Then first PESO must find the value for the protected bandwidth capacity which is the bandwidth capacity remained after a failure between two extreme cases: if all VCG on disjoint paths and if they are on the same path. Then PESO proceeds as earlier. The last case is when PESO must calculate the routing so that over provision is supported to reconstruct the data if there is a fault so that the connection is still at full throttle in the face of failure.

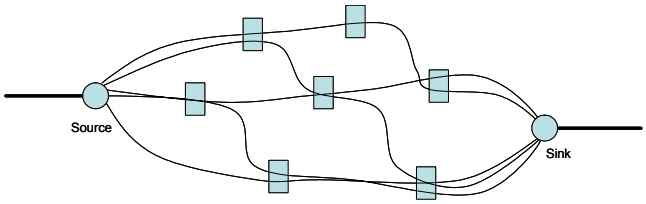

Figure 11: Virtual Concatenation (VC)

Even though PESO is not ready to be used as-is, it has satisfied many of the 5 key attributes of Carrier Ethernet. VC and LCAS provide high resilience for the network and guaranteed bandwidth for each flow. Since assured bandwidth can be calculated, PESO can be extended to guarantee other metrics as well. In addition, PESO is better suited for the metro access network due to its ability to utilize the redundant links when determining the route for the VCG. What PESO missing is the OAM feature before it can be used as a stand-alone architecture.

6.1.2 Atrica Resilient Ethernet Access (AREA)

Atrica Resilient Ethernet Access (AREA) [4] is the innovation from Atrica Network. To attain high resilience, AREA uses two recovery mechanisms: tunneling and hardware-based. The hardware-based approach uses a Hello messages to monitor for failure. The tunneling protects aggregated link and nodes by redirecting traffic at wirespeed upon failure detection to a pre-configured tunnel. Therefore, it can achieve the recovery time of sub-50ms. It supports both MPLS and VLAN tagging. For end-to-end, it uses MPLS; and for next hop, it uses Ethernet VLAN. Because of the VLAN tag, AREA has a low scalability limited by the VLAN space. The tunnel is identified by an MPLS or VLAN protection label generated by the ingress devices. Using MPLS, AREA can support the sophisticated QoS scheme that MPLS can provide. However, it is costly to use AREA in the metro access network in term of operational expenses and capital expenses due to the supporting of MPLS. It is claimed to be compatible with Ethernet Spanning Tree Protocol family. Furthermore, AREA does not support OAM in the original design. Therefore, it can be used as a stand-alone architecture for Carrier Ethernet

6.1.3 Ethereal

Ethereal [15] retains the distributed feature of Spanning Tree Protocol with some new improvements. Ethereal, a real time connection oriented architecture supporting best effort and assured service traffic at the link layer, proposes to use the propagation order spanning tree for fast re-converge of the ST once a failure has been detected. Ethereal switches periodically send out hello message to the immediate neighbor switches. Absence of any previous received hello message indicates a node/link failure. A hello message from a new switch indicates an addition or substitution of a switch. In either case, the switches that detect the fault immediately discard all best effort traffic, tears down the established QoS-assured connections that traverse through the offending link, re-converge the spanning tree, and reestablish any torn down connections. All of the best effort traffics are discarded because all destinations are unknown after a fault, which will require a flooding to deliver the traffic; and flooding without a spanning tree structure causes loop in the network. The established QoS-assured connections that are not on the offending link can continue to be forwarded on the established path without causing any problem. The spanning tree re-converges using the propagation order spanning tree. The initiated switch sends out an invitation to its neighbors to join its spanning tree. If a node accepts, it propagates the invitation to its neighbors. If a node receives all rejection from its invitation, it assumes that it is the leaf. Then it will send a start up phase complete (SPC) to its parent. After a parent receives

SPC from all of its children, it sends an SPC to its parent. This SPC propagates until it reaches the root. The root will then send topology discovery phase complete (TDC) to all to indicate that it is now safe to forward on the established ST. If multiple nodes compete for the root role and send out multiple invitations, the other nodes pick the one with the lowest bridge id and propagate the invitation. This is an improvement over the standard STP because the root role competition is among groups of disjoint ST and not between every single switch.

Ethereal [15] demonstrate an example of running a flow reservation mechanism. The Ethereal switch architecture is designed to meet the QoS requirements for real time multimedia applications via hop-by-hop reservation. When an application makes a request for connection, it sends QoS parameters, the destination IP address, and the destination IP port number. A Real Time Communication Daemon (RTCD), developed by Ethereal, contacts the neighbor Ethereal switch and give it a generated connection id. The connection id is unique on a per hop basic, similar to MPLS label. If the Ethereal switch can make the QoS commitment, then it contacts the next switch on the path with a unique connection id for this hop. The reservation propagates until it reaches the destination and the periphery switch at the destination returns a reply. If the reservation is successful, all the switches on the path bind the routing entries with the connection id and the QoS parameters. The RTCD at the source binds the proxy Ethernet address with a proxy IP address into the Address Resolution Protocol (ARP) cache. Then it returns the proxy IP address to the request application. The proxy IP address has the format 1.1.XX.YY where XX.YY is the connection id. Similarly, the proxy Ethernet address has the format FF-FF-FF-FF-XX-YY where XX-YY is the connection id. The application then opens a UDP connection to the destination with the proxy IP address. The ARP cache will translate the proxy IP address into the proxy Ethernet address so that each switch on the path can extract the connection id and remaps it to their locally unique connection id until the frames reach the destination. The last switch will need to make a translation from the proxy address to the real address before it can deliver the frame to the end host. Ethereal requires some cooperation with the IP layer and the application. Furthermore, there is no mechanism for traffic priority to guard the QoS commitments.

The nice feature about Ethereal is its support for flow reservation to guarantee the QoS metrics which is one of the required key attribute for Carrier Ethernet. However, Ethereal has poor scalability because of the number of connections it can make. The total number of connections it can commit to in the worse case is 2^16 = 65536 connections. This number is far too small for a metro access network where the number of subscribers can reach millions. In addition, Ethereal dependency on Spanning Tree gives it a low resilience status. Together with the lack of OAM support, Ethereal is not a complete package for Carrier Ethereal. Ethereal can be deployed in the metro core as well as the metro access. Although the scalability issue will mitigated in the metro core, it still not ready for MAN due to its low resilience.

6.1.4 SmartBridge

Because the Spanning Tree Protocol tends to forward frames toward the root, as the network size grows, the amount of inter-LAN traffic increases causing a bottleneck at some bridges. Realizing the congestion problem of inter-LAN forwarding, SmartBridge [10] was developed as a new architecture to scale traffic in LANs. Retaining the good properties of STP and combining with some good features of IP routing, SmartBridge proposes to forward frames along the shortest paths. It requires a full knowledge of the topology so that forwarding can be done between

hosts of known location along the calculated shortest path. The inventory construction and topology acquisition processes maintain the complete description of the topology. SmartBridge keeps track and update any change in network topology such as addition or removal of bridges. A host's location information is kept in a table inside the SmartBridge. A host location revision mechanism keeps track of all the hosts and updates the table if necessary. For the purpose of consistency, frames with unknown source address are dropped automatically and a topology acquisition process will be initiated. Frames with unknown destination address are flooded like the standard STP but with slight modification to update the host location table. Frames with known source and destination addresses are guaranteed to be forwarded on the shortest path that is calculated based on an assignment of weights so that any least-weight path from source to destination is a shortest path in the topology and the least-weight path from source to destination is unique.

SmartBridge introduces more complexity into the link layer in order to enhance the performance of a large network. This complexity will have a direct affect on the processing power of a bridge/switch and increase the controller traffic. The usage of IP routing increases SmartBridge's topology utilization by employing the redundant links. Therefore, it can be deployed in the metro access network. However, the need to have the global topology with the storage for all destination addresses yields a low scalability making SmartBridge less desirable for the metro access. SmartBridge also lack the support for QoS and OAM feature. Therefore, it is not ready to be deployed as the main architecture for MEN.

6.1.5 STAR

Exploiting the fact that frames traveling on the standard ST is not necessary the shortest path, Spanning Tree Alternate Routing [7] proposes a new forwarding scheme to enhance the forwarding performance. The idea behind STAR is that the performance of a flow is affected by the length of the forwarding path. Therefore, STAR finds an alternate route that is shorter than the corresponding path on the spanning tree. The metric that can be used to determine the path might be delay, bandwidth, or any other required metric. Each STAR-aware bridge has two routing tables: bridge forwarding table (BF table), and host location table (HL table). The BF table indicates the forwarding port to other STAR-awared bridge along the shortest path, and the HL table maps an end host to a STAR-awared bridge. The BF table is found by using a modified version of the distance vector algorithm. STAR uses BF table to find the shortest path to another STAR bridge that is closest to the destination; and then that bridge will deliver frame to the destination. Since STAR use distance vector algorithm, for a given topology, it produces static paths for any computation. STAR is designed to coexist with legacy STP bridges so that STAR deployment can be incremental. However, STAR has no distinction between different classes of traffic. It forwards frames along the shortest route as if they belong to the same class. It also does not provide any guaranteed quality of service or establishes any service level agreements.

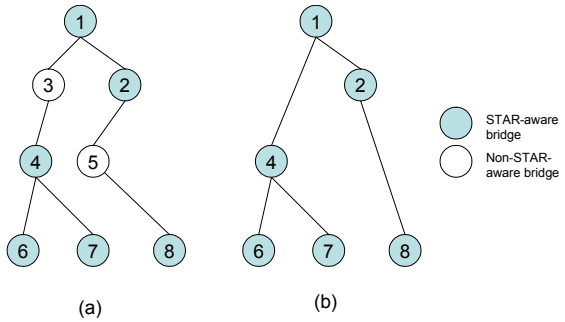

Figure 12: (a) Physical LAN topology (b) STAR overlay topology

Similar to SmartBridge, STAR finds alternate paths to destination that is faster than going along the Spanning Tree path. However, this overlay approach fails to distinct itself as the next generation of Ethernet protocol for MEN. In the metro access, STAR runs the traditional Spanning Tree Protocol that has the reconvergence time of 30 to 60 seconds. This delay is not acceptable for the voice and video applications. Lacking the support for QoS and OAM, STAR cannot be the stand-alone solution to run MEN.

6.2 Metro Core Solutions

The metro core topology in most cases is a ring [5] which is simpler than the mesh access network. Unlike a mesh topology, a ring topology is generally simpler; therefore, the management is lighter and certain behaviors are more predictable such as the direction of traffic flows yielding a faster reconvergence time. The drawback with the ring topology is scalability. The latency of traversing the ring is proportional to the number of equipment on the ring. One solution is to break the large ring into smaller ones. However, the difficulty lies on the management plane of interconnecting multi-rings. The advantages of using the Spanning Tree Protocol to manage a network are low cost and simple to manage. Therefore, it is more suitable for the core where it only hauls large trunks of traffic from one end to the other. The metro core topology is simple enough for used with a Spanning Tree Protocol while link state or MPLS are more complex than necessary to set up. A summary of the solutions in the metro core with respect to the 5 key attributes as defined by the MEF is shown in TABLE III.

TABLE III. SUMMARY OF THE PROTOCOLS FOR METRO CORE

Solutions	***Standard***	***Resilience***	***Scalability***	***QoS***	***OAM***	***Ready as stand-alone***
STP	IEEE 802.1d	poor	low	none	none	no
RSTP	IEEE 802.1w	low	low	none,	none	no
MSTP	IEEE 802.1s	low	low	manual load balance	none	no
Viking	Academia Publication	high	low	some guaranteed metrics	central server	no
RRSTP	Industry, Proprietary	med	low	none	none	no
EAPS	IETF RFC 3619	med to high	Low (VLAN space dependent)	none	none	no
MRP	Industry, Proprietary	med	N/A	none	none	no
H-VPLS	IETF draft	high	scalable	N/A	none	no

6.2.1 Spanning Tree Protocol Family

Traditionally, Ethernet-based networks use the standard spanning tree protocol (IEEE 802.1d) for routing packets in the network. Spanning Tree Protocol [20] is standardized in IEEE 802.1d. It is a layer2 protocol that can be

implemented in switches and bridges. The spanning tree protocol (STP) essentially uses a shortest-path to the central root approach in forming a tree that is overlaid on top of the mesh-oriented Ethernet networks, as shown in Figure 13. Spanning tree is used primarily to avoid formation of cycles or loops in the network. Unlike IP packets, Ethernet packets do not have a time-to-live (TTL) field. STP prevents loop in the network by blocking redundant links. Therefore, the load is concentrated on a single link which leaves it at risk of failures and with no load balancing mechanism. The root of the tree is chosen based on the bridge priority, and the path cost to the root is propagated throughout so that each switch can determine the state of its ports. Only the ports that are in the forwarding state can forward incoming frames. This ensures a single path between a source and a destination. Whenever there is a change in the topology, switches rerun the protocol that can take up 30 to 60 seconds. At any one time only one spanning tree dictates the network.

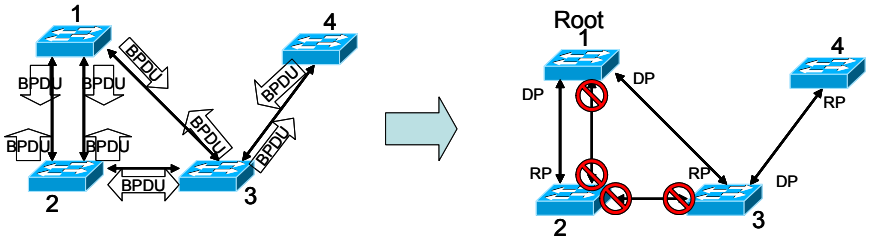

Figure 13: The distributed STP protocol broadcast its control packet, BPDU, to elect a root switch and select the shortest path from each switch to the root. Redundant links are block to prevent loop.

An improvement of STP is the Rapid Spanning Tree Protocol RSTP [21] specified in IEEE 802.1w. RSTP reduces the number of port states to three: discarding, learning, and forwarding. Through faster aging time and rapid transition to forwarding state, RSTP is able to reduce the convergence time to between 1 and 3 seconds. In addition, the topology change notification is propagated throughout the network simultaneously, unlike STP, in which a switch first notifies the root, and then the root broadcast the changes, as shown in Figure 14. The left of Figure 14 shows that STP topology change process takes 2 phases so that the delay lies in the propagation of the TCN message to the root. However, the improvement in RSTP as shown on the right of Figure 14, uses the source of the TCN message as the root and broadcasts the TCN message from it. Similar to STP, there is only one spanning tree over the whole network. RSTP still blocks redundant links to ensure loop free paths leaving the network underutilized, vulnerable to failures, and supports no load balancing.

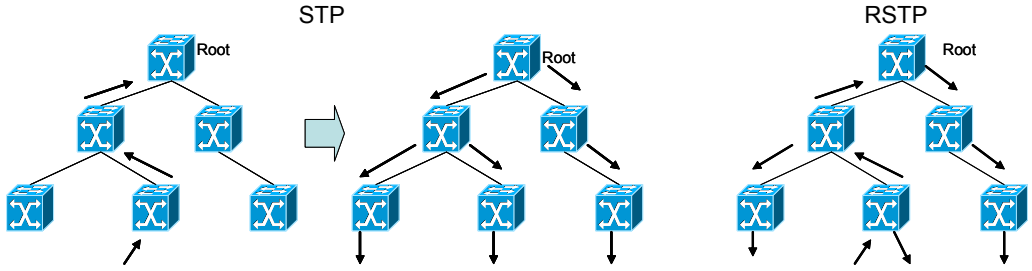

Figure 14: Topology Change notification in STP (left) and RSTP (right)

The latest upgrade to the Spanning Tree Protocol is Multiple Spanning Tree Protocol (MSTP) [23], defined in IEEE 802.1s. MSTP uses a common spanning tree that connects all of the regions in the topology called the Internal Spanning Tree (IST). The regions in MSTP are multiple instances of the spanning tree. Each instance (MSTI) is an

instance of the RSTP. An instance of RSTP governs a region, where each region has its own regional root. The regional roots are in turn connected to the common root that belongs to the common spanning tree, as shown in Figure 15. One or more VLAN can be assigned into a MSTI. By assigning different traffic flows into different VLAN, the network load is balanced under the assumption that different VLAN will take different path.

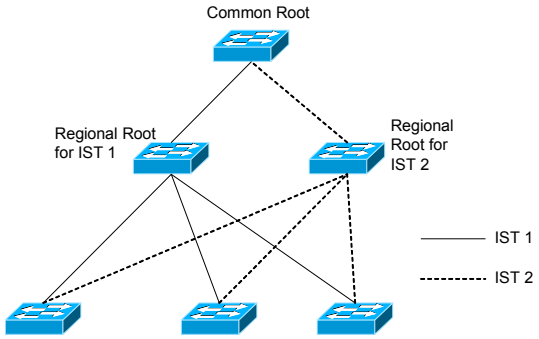

Figure 15: A sample MSTP configuration

Although STP family has been used for most Ethernet networks and it could be very effective to be used in the metro core, it has several shortcomings in the context of its use for MEN. These shortcomings that are the reason why the Spanning Tree Protocol family is not ready to be used as-is for MEN are enumerated as follows:

1. Spanning trees restrict the number of ports being used. In high-capacity Ethernets, this restriction translates to a very low utilization of the network.
2. Poor resiliency: a very high convergence time (STP: 30s to 60s, RSTP: 1s to 3s).
3. No mechanisms to balance load across the network.
4. Lack of support QoS and OAM.

6.2.2 Viking

Recognizing the poor resilience in the standard Spanning Tree, STP and RSTP, Viking, proposed by Sharma et al. [12], enhances the resiliency with a Multiple Spanning Tree architecture. Viking precomputes multiple spanning trees so that it can change to a backup spanning tree in the event of a failure. To precompute the spanning tree, Viking finds k-shortest primary path. For each of the primary path, Viking finds k backup path. For each pair of primary and backup paths, it rejects the path if the bandwidth requirement is not satisfied. The spanning trees are created when these paths are merged together. The Viking server receives monitored information for the network condition. It uses this information to compute the spanning tree periodically so that it can reply to any client query for spanning tree information.

Viking performs load balance by precomputing the path to avoid the heavily used links. This can be done by a cost formula that assigns the cost to the network as more paths are added. The cost formula relies on the expected cost of each link which is the fraction of all the possible paths between all source and destination pairs that passes through that link. The acceptance or rejection of new path requests depends on the output of the cost formula that will determine if network will be overloaded.

Viking delivers some quality of services by satisfying bandwidth or delay requests. Although it does provide an end-to-end bandwidth or delay guarantee mechanism, it does not provide traffic policing, traffic shaping, traffic priority, or drop precedence in case of network congestion.

With respect to the 5 key attributes defined by MEF, Viking comes close to being the standalone technology for Carrier Ethernet to be deployed in MEN. Viking is versatile in that it can be deployed in the metro core as well as the metro access where it can take advantages of the redundant links. Having high resilience, Viking has no problem meeting the sub 50ms criteria. However, more works are needed to improve Viking's support for QoS. It still lacks required features in QoS like differentiation of service and assured end-to-end quality.

6.2.3 The Specialized Ring Protocols

The following three protocols are specifically designed to manage ring topologies by taking advantages of the unique characteristic of a ring. Although developed independently, they result in similar concept. They are best deployed in the metro core where it is most likely to be ring structure. With the appropriate configuration, they can achieve good resilience having sub-second reconvergence. However, they lack the support for QoS and OAM. Although suitable for the metro core, all three protocol cannot be deployed as-is in MEN.

Riverstone Network leverages MSTP and RSTP to improve the turnaround time for port state for Ethernet switches using the Rapid Ring Spanning Tree Protocol (RRSTP), adhering strictly to ring topologies [5]. Each spanning tree instance stays on its designated ring. A node will be the root switch for that instance with a primary and alternate port. Initially, traffics enter the root will be sent on the primary port. If a link on the primary path is broken, the alternate port will open for use. Each node on the ring topology will be a root for an instance of a MSTP. Since VLANs are used to distinguish spanning tree, the VLAN space limits its scalability. The recovery time after failure is approximately equal to the BPDU hello time, which can be from 0.5 second to 1 second.

For link protection, Ethernet Automatic Protection Switching (EAPS) is Extreme Network's proprietary solution for resilience in the link layer that also is currently specified in IETF RFC 3619 [11], [16]. It is designed only for a ring topology. A ring makes up of at least two switches. One of the nodes on the ring must be a master. The master switch has a primary and backup port where initially, traffic is sent on the primary port and the backup port is blocked. An EAPS domain is configured to protect a group of VLANs as seen in Figure 16a. Multiple EAPS domains can exist on the same ring protecting different set of VLANs. Each domain reserves one VLAN as the control VLAN. The control VLAN only sends and receives EAPS specific control message. Layer 2 switching and address learning behave normally. Traffic belonging to a VLAN only flows through one direction on the ring preventing loops from occurring. The master sends out periodic poll from the primary port on the control VLAN and to be received on the secondary testing the ring connectivity. A non-master switch for a domain can send a link-down message to the master if it detects a fault, as shown in Figure 16b. If a poll is timed out or a link-down message is received, the master declares a failed state, unblocks its secondary port to allow traffic going through, flushes its forwarding database, and sends a flush DB message forcing the other switches to flush their forwarding database. The master continues to send out polls on its primary port. If the link is restored, the master blocks the secondary port and forces a database flush on all the other switches. If any of the other switches detect the recovery before receiving the notification from the master switch, it puts the traffic on the recovered port in blocked state, sets the state of the temporarily blocked port to pre-forwarding. When it receives the flush-DB message from the master, it flushes the entire forwarding database. If the state is set to pre-forwarding, it begins to forward traffic on that port. By setting the timeout to be sub-second, the fault detection and

recovery can be sub-second. Besides being limited by the VLAN space of 4096 VLANs, a maximum of 64 EAPS domains can be defined on a single switch/ring. Initial tests show that the failover of EAPSv2 is less than 50ms for 10,000 layer2 flows and 100 protected VLANs [46].

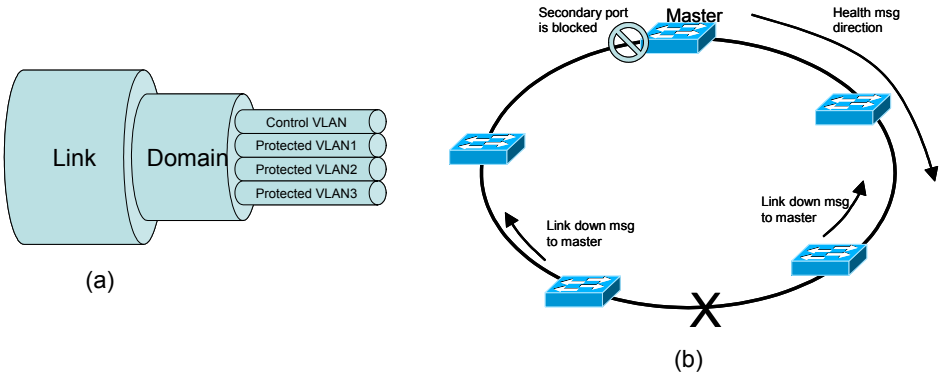

Figure 16: (a) EAPS architecture (b) Fault Detection in EAPS

As an alternative to the standard STP, Foundry delivered Metro Ring Protocol (MRP) [18]. MRP prevents loops and provides fast re-convergence for ring topology only. A master node will be picked for each ring. On a given ring, the master has two interfaces: primary and secondary. Initially, data traffic travels on the primary path unless it fails then the master unblock the secondary port. The primary interface generates the Ring Health Packets (RHP). If the RHP messages reach the secondary port then the primary path is working properly. Multi ring can be merged to create a large topology but an MRP instance only runs on one ring and not the entire topology as seen in Figure 17. The convergence time after the fault detection is said to be sub-second, and there is no load balance mechanism built-in.

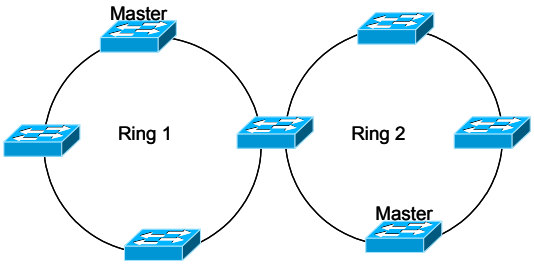

Figure 17: A multi-ring topology

6.2.4 Hierarchical VPLS

As an emerging technology for layer 2, VPLS uses the Martini encapsulation standard to transport Ethernet, ATM, and Frame Relay over the same core network. A VPLS network can support over a million unique labels, which means it is possible to support over a million customers. VPLS also suffers from the MAC address explosion problem [38]. One solution is to use a router at the customer edge device that results in a single MAC address per site. Another solution is to limit the number of MAC addresses that can be learned at the provider edge per access circuit. A unique scalability issue with VPLS is that it uses MPLS tunnels to create a fully meshed network and thus potentially requires a very large number of individual connections. To cope with this problem, a hierarchical VPLS (HVPLS) approach is needed as described earlier.

Although HVPLS is not ready to be a Carrier Ethernet technology, it is a good candidate. Having high resilience as the result of taking after MPLS, HVPLS is also scalable. Even though QoS is not defined yet in VPLS, it should not be too hard to extend it to support QoS similar to MPLS. The one critical attribute that HVPLS misses is the OAM capability. With the appropriate configuration, HVPLS can be deployed to manage the whole metro topology including the access and the core.

6.3 Supplement Solutions

The following solutions that can be used as supplement to other protocols focus on enhancing one of the 5 key attributes of the Carrier Ethernet. They would not be sufficient to serve as a stand alone MEN technology but they can be use concurrent with any of the protocols that manage the metro access and metro core. A summary of the supplement solutions is shown in TABLE IV.

TABLE IV. SUMMARY OF THE SPECILIZED PROTOCOLS

Solutions	***Standard***	***Enhanced Key Attribute***
LACP	IEEE 802.3ad	Load balance and Resilience
TBTP	Academia Publication	QoS: Utilization and load balance
ESRP	Industry, Proprietary	Resilience: standby node
VSRP	Industry Proprietary	Resilience: standby node
SuperSpan	Industry Proprietary	Scalability: Spanning Tree Scope
Q-in-Q	Industry Proprietary	Scalability: VLAN
MAC-in-MAC	Industry Proprietary	Scalability: MAC addresses
3bit priority	IEEE 802.1P	QoS: Class of service
MEF bandwidth profile	MEF specification	QoS: bandwidth profiling

6.3.1 Link Aggregation Control Protocol (LACP)

The Link Aggregation Control Protocol (LACP) or IEEE 802.3ad proposes to group multiple physical ports together on a switch to create a single logical port. The benefit being that more ports can be managed as one connection; service provider can add or remove bandwidth to the current connection in chunk relative to the physical port bandwidth; load sharing and load balancing performs between links within a logical connection; and high resiliency at the cost of reduced bandwidth if some of the physical ports go down.

6.3.2 TBTP

Recognizing the inefficient utilization of the bandwidth that is caused by the overly restrictive Spanning Tree protocol, Pellegrini et al. proposes a novel scheme, Tree-Based Turn-Prohibition (TBTP) [9], to loosen the restriction but still keep the network operational. STP prevents loops in the topology by pruning it down to a tree-structure path imposing severe penalty on performance. Given a topology and a formed spanning tree, TBTP constructs a less restrictive spanning tree by blocking a small number of pairs of links around nodes, called turn, so that all cycles in a network can be broken. TBTP does not prohibit turns that are on the original spanning tree in order to be backward compatible with the STP. In addition, the upper bound on the number of prohibited turns is at most half of the turns in the topology. Therefore, the algorithm guarantees that the total weight of the permitted turns is always greater than the total weight of the prohibited turns in the network. By opening up more turns, TBTP provides more paths to route Ethernet frames and lessen the congestion on the main spanning tree. The benefit of TBTP is proportional to the degree of the nodes. However, TBTP did not improve on the recovery time of the

standard spanning tree protocol. Since TBTP relies on the standard STP to re-converge before it can re-compute its routing, the recover time is in the order of seconds.

6.3.3 ESRP

A straight forward scheme for failure protection is to provide backup node that normally in standby mode. In Ethernet, a switch is placed on standby ready to take over the primary equipment in case the primary fails. Figure 18 shows a typical example of an equipment redundancy protocol.

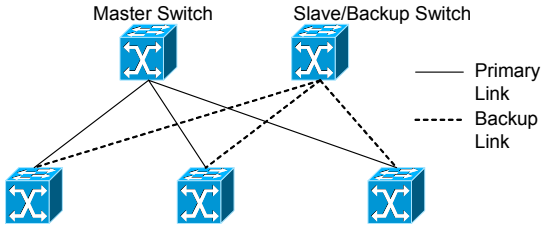

Figure 18: An example of providing switch redundancy

In providing node protection, Extreme has developed the master and slave switch model called Extreme Standby Router Protocol (ESRP) [17] for the mesh topology. In ESRP, there is a slave switch that is in standby mode and all data frames to it are blocked unless the master switch fails. When the master switch fails, the slave switch immediately assumes all functions of the master switch including forwarding frames and MAC address learning without waiting for re-converge of the new topology. The recovery time after failure depends on the communication speed between the master and the slave switch. ESRP can scale up to 64 VLANs per ESRP domain. ESRP is not backward compatible with STP on the master and slave switch because both protocols cannot be configured to run on the same interface. Upon failure, the new master needs to send out the Extreme Discover Protocol (EDP) messages announcing the new master. The downstream switches then discard all address entries associated with former master. If there is no bi-directional connection between the new master and the downstream switches, the recovery time could take up to five minutes. If the downstream switches are not Extreme switches, another mechanism is required to inform the switches of the new master.

6.3.4 VSRP

Similar to Extreme, Foundry also has a proprietary resilience protocol based on the principle of a standby switch. Foundry's Virtual Switch Redundancy Protocol (VSRP) [19] has one master and at least one backup switch. Initially, the master switch forwards all traffic. When it goes down, the downstream switch will accept the backup switch as the new master and the backup switch unblocks its ports for forwarding. The failover time is sub-second if all of the switches are VSRP-aware. Failover can also occur even if the master switch is not completely offline. The master switch's priority is reduced each time a port fails. Therefore, over time if the master's priority is reduced to lower than the backup switch, failover will occur.

6.3.5 SuperSpan™

To enhance the scalability of the standard STP, Foundry Network has developed SuperSpan™ [41] that is based on the concept of divide-and-conquer. Intuitively, the network topology is divided into smaller, easy-to-manage, and fast converging domains, as shown in Figure 19. Each domain will run a separate RSTP instance. A link going down in one domain will not cause the entire network to reconverge but its own domain. Since SuperSpan™ uses STP, it prevents loops in the topology and forwards frames on a single path. Between domains

are boundary interfaces that connect adjacent domains together. They filter out local BPDU so that one domain cannot affect another. For providers that use Per VLAN Spanning Tree (PVST) to isolate customers' traffic for secured connection, Foundry claims that SuperSpan can scale it to more than 100 VLANs where normal system resources limit is.

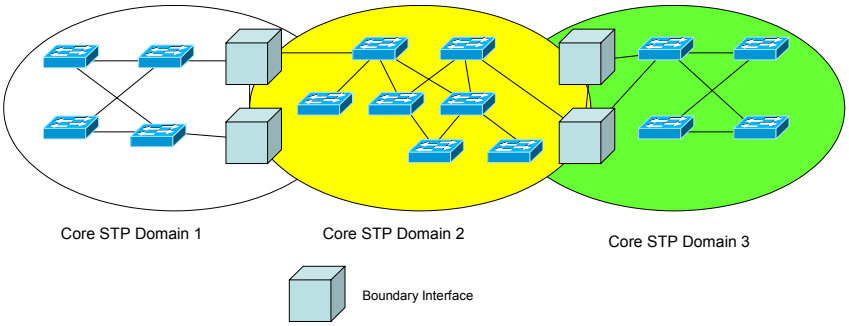

Figure 19: Foundry's SuperSpan divides the provider network

6.3.6 VLAN Tags

In Ethernet switching protocol, the IEEE 802.1Q [22] defines VLAN tag on each Ethernet frame. The VLAN tag includes a VLAN ID to distinguish the frames belonging to different VLANs. Within a VLAN tag, a three-bit Class of Service (CoS) field provides up to eight classes of service for a VLAN. Light traffic policing can be enforced with VLAN ID. Enterprises use the VLAN ID to restrict broadcast storms and to separate different services. Carriers use the VLAN ID to separate different customers' traffic. The mapping is one to one correspondence of one VLAN per customer per service type so that one customer cannot snoop on another customer's traffic. However, the VLAN ID field is limited to 12 bits or 4096 VLANs. This limits a carrier to serve much less than 4096 customers because the carrier uses more than one VLAN on a single customer to identify different services that the customer is subscribed to. Furthermore, the VLAN space is also shared with all the customers since each customer will want to use the VLAN ID to identify their own services within their own network. If the customer edge equipments are layer 2 switches, then all the customers MAC addresses will have to be learned and registered by the provider equipments [38]. This results in MAC table explosion problem, and whenever the Source Address Table periodically refreshes, broadcast storm is possible. Another drawback for VLAN is that the BPDUs are not transparent to the provider network that could cause undesirable results as explained later.

In the attempt to fix the VLAN ID shortage problem, one approach, called Stacked VLAN [36] or Q-in-Q [38], is to stack another VLAN tag in front of the original VLAN in the Ethernet header. Since Q-in-Q is not standardized, many vendors offer this solution as proprietary. As an Ethernet frame enters the provider network from the customer network, the provider VLAN tag is added to the frame by the provider ingress switches, as shown in Figure 20. The second VLAN tag is used by the carriers to isolate traffic among different customers. The egress switches of the provider network strip the provider VLAN tag from a frame before it leaves the provider network. Leaving the first VLAN tag untouched, the customers are free to assign its own VLAN ID on its local network. Different enterprises can overlap the VLAN IDs. However, Q-in-Q does not separate providers' and customers' MAC addresses. Therefore, the provider switches must learn all MAC addresses in the network including the customers', creating the MAC table explosion problem. The providers' switches see both its own network and the customers' network as one big network. Since there is no separation between provider's MAC addresses and customers' MAC addresses,

complications arise for Ethernet control protocols (e.g. BPDU). For example, a customer's BPDU must not interact with the provider's network. However, the Spanning Tree protocol is identified by the fixed MAC address 01-80-C2-00-00-00. This means that a spanning tree recalculation request for a customer's network might trigger a ST recalculation in the provider's network as well. From the scalability perspective, the provider is able to support up to 4096 customers and still able to support separation of traffic among different customers. Each incoming C-VLAN ID and C-VLAN CoS is mapped to P-VLAN and P-VLAN CoS while keeping C-VLAN ID and C-VLAN CoS unchanged. Therefore, the customer VLAN ID and VLAN CoS is preserved. Another approach is that the provider uses both VLAN fields. For every incoming customer VLAN, it is translated to the provider 24bit VLAN. Therefore, the provider can support up to 16 millions customers but the difficulty lies in the translation between VLAN IDs. If separation among different customers' traffic is not required, the provider can set the P-VLAN to be of service type and aggregate customer's traffic into P-VLAN ID based on the service needed e.g. E-Line, E-LAN, WAN, VoIP, and so on.

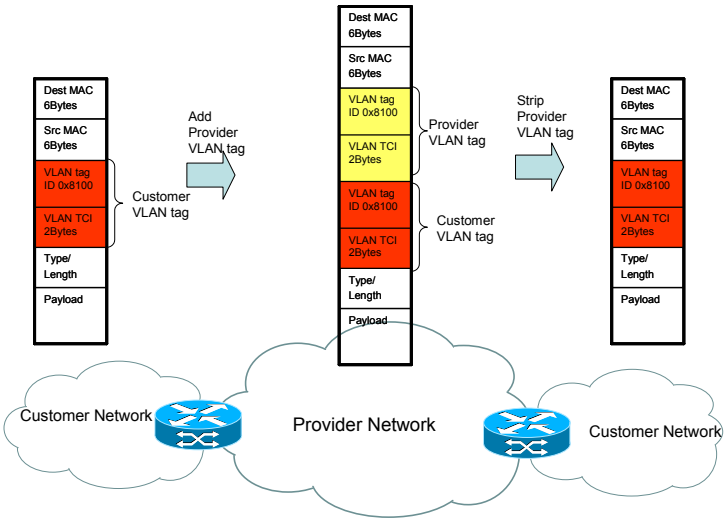

Figure 20: VLAN stacking scheme or Q-in-Q

6.3.7 MAC-in-MAC

Another tunneling scheme is the MAC-in-MAC (M-in-M) [36] that can isolate the provider's MAC addresses from the customers' MAC addresses. Hence, it resolves some drawbacks of 802.1q and Q-in-Q, such as customer control protocol transparency and enhancing the scalability. Similar to the Q-in-Q concept, M-in-M prepends the provider's MAC source and destination addresses to the Ethernet header at the provider ingress switch, as shown in Figure 21. Switches within the provider network use the provider MAC addresses to deliver frames to the egress switches. At the egress switches, the provider MAC addresses are stripped before any frames leave the provider network. M-in-M also has not been standardized. Nortel's M-in-M proprietary solution prepends provider's MAC source address, MAC destination address, P-EtherType, P-VLAN tag, and P-Service Label into the Etherframe frame. Since MAC addresses are allowed to overlap between the provider and customer, the spanning tree protocol's fixed MAC address no longer poses a problem for the provider network. In addition, the provider network needs to learn only MAC addresses from its own switches and not all the customer's MAC addresses as before. Therefore, the MAC table explosion is mitigated. However, the complication is at the ingress and egress switches where the translation/mapping between customers' MAC addresses and provider's MAC addresses. These ingress and egress

switches still need to learn all of the customers' MAC addresses in order to make the translation. It is possible for Q-in-Q and M-in-M to coexist as a hybrid solution. As suggested by Nortel Network [36], the Q-in-Q is used in the Metro Access and M-in-M is used in the Metro Aggregation Network, as shown in Figure 22. When the customer frames pass to the metro access network, the provider VLAN tag is added on top of the original frame. The provider MAC labels are stacked on the frames coming from the metro access. Both additional labels are stripped off as they move out of their respective area.

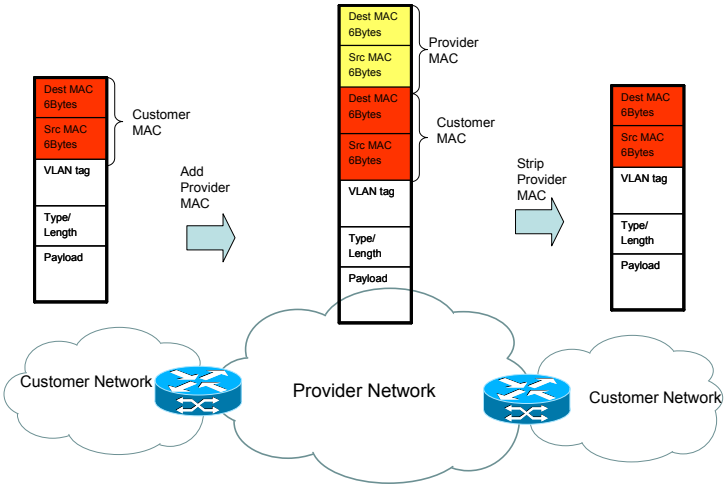

Figure 21: MAC-in-MAC approach

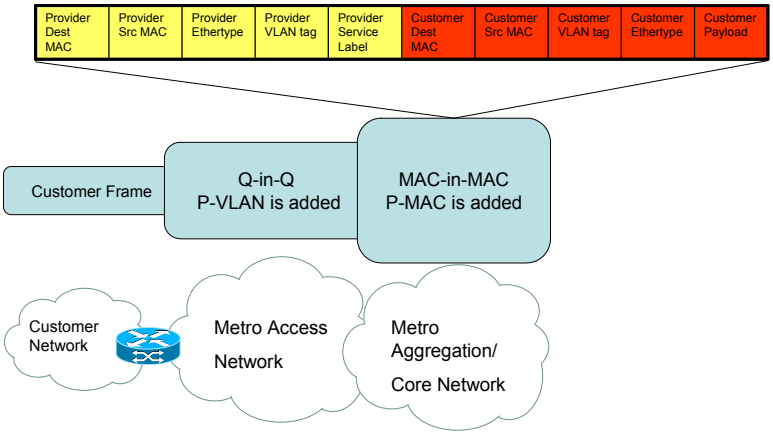

Figure 22: A hybrid approach of M-in-M and Q-in-Q. The provider service label contains the service id for which the customer frame is mapped to

6.3.8 Traffic Class of Service

As a simple protocol, the standard IEEE 802.1D does not describe any sophisticated scheme for QoS, admission control, or traffic policing as in DiffServe, IntServe, and MPLS. Light traffic policing is possible by defining VLAN configurations that map traffic type to priority queues. However, there are no guaranteed services such as bandwidth reservation. The standard defines eight traffic types, in order of priority: background, spare, best effort, excellent effort, controlled load, video, voice, and network control. Network control has the "no loss" requirement to maintain and support the network infrastructure. Voice must be less than 10ms delay and video must be 100ms delay. Control load is important business application traffic that is subjected to some form of "admission control". Depending on the number of queues on a switch, 802.1D divides these traffic types among the priority

queues TABLE V. For example, if there are three queues, then the traffic types are divided as follow: {best effort, excellent effort, background}, {controlled load, video}, and {voice, network control}. The 802.1P defines a 3bit user priority field within the Ethernet header for differentiation of services. Then there are a maximum of eight priorities can be supported on a switch.

TABLE V. RECOMMENED USER PRIORITY TO TRAFFIC CLASS MAPPINGS [22].

User Priority	Number of Available Traffic Classes (Queues)							
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***
0 (default)	0	0	0	1	1	1	1	2
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	1
3	0	0	0	1	1	2	2	3
4	0	1	1	2	2	3	3	4
5	0	1	1	2	3	4	4	5
6	0	1	2	3	4	5	5	6
7	0	1	2	3	4	5	6	7

Note-User priority value mapping to traffic types are as following: 0-best effort, 1-background, 2-spare, 3-excellent effort, 4-controlled load, 5-video, 6-voice, 7-network control.

To cope with the QoS insufficiency in Ethernet, MEF defines traffic management mechanisms that include bandwidth profiling at the edge and inside the MEN [30]. At the MEN ingress, the Customer Equipment VLAN ID (CE-VLAN ID) is mapped to an EVC. Then, the EVC or a combination of EVC and the Customer Equipment VLAN Class of Service (CE-VLAN-CoS) determine the Class of Service (CoS) instance to be used inside the MEN. Bandwidth profiling is a set of parameters that the service providers use to control the incoming traffic into the MEN so that they meet the Service Level Specification (SLS) that is an agreement with the customers. These parameters are input into a bandwidth profile algorithm that verifies the conformance of the traffic and performs the necessary operations including dropping or recoloring the traffic frames to indicate the drop precedence. The frame color and their meaning are: green indicates that the frame is in-profile, yellow indicates that the frame is out-of-profile and drop if the network is congested, and red means that the frame is to be dropped immediately. The six parameters to control the traffic rate are Committed Information Rate (CIR), Committed Burst Size (CBS), Excess Information Rate (EIR), Excess Burst Size (EBS), Coupling Flag (CF), and Color Mode (CM). CM indicates whether or not the frame color is taken into account when determining the conformity of the traffic. The value of CF has the effect of controlling the volume of the yellow frames traffic. The algorithm uses token buckets to determine if the traffic conforms to the SLS. Initially, there are two buckets that are full of tokens. As frames enter the provider's network, tokens from the first bucket (green bucket) is decremented by the size of the frames. If the green bucket is not empty, the frame is CIR-conformant and is allowed into the network. If it is empty, the second bucket (yellow bucket) is decremented. If yellow tokens are available, then the frame is colored yellow and allowed into the network. If there are no more yellow token, then the frame is declared red and discarded. Bandwidth profiling can be applied per ingress UNI, per EVC, or per CoS. The profiling granularity goes in the order of increasing granularity from per ingress UNI to per CoS since each UNI is composed of multiple EVCs, and each EVC is composed of multipled CoS. In addition, the document also defines the frame delay performance, frame jitter performance, and frame loss ration. Frame delay performance is the P-percentile of the delay for all green frames successfully delivered for a UNI pair within a time interval. Frame jitter performance is the P-percentile of the difference of the

one-way delay of green frame pairs that arrive at the ingress UNI within a time interval. The frame loss ratio is the percentage of the number of green frames that are loss over the total number of green frames that arrive at the UNI.

7 Operation, Administration, and Maintenance (OAM)

One of the features that imposes a high degree of obstruction to Ethernet from being a standalone carrier-grade technology is the lack of Operation, Administration, and Maintenance (OAM). Other technologies such as SONET and ATM have OAM capabilities within the data link layer [14]. OAM deals with the network performance monitoring and maintenance. It includes predefined variables about the status of the network such as delay, loss, jitter, and bandwidth availability that can be sent in-band or out-of-band. In-band signaling, like SONET/SDH, sends OAM information attached to the data. This is a closer measurement of the network performance. However, there is a risk that the OAM information might affect and bias the measurement. Out-of-band signaling sends the OAM information on a different path than the data path like MPLS echo request. Therefore, careful implementation must be enforced to ensure that the OAM frames are treated as closely as possible to the data frames in order to get an accurate reading. Reference [2] speculates that out-of-band will likely be the case for Ethernet OAM. To measure frame delay and jitter, [2] suggests one way to measure delay/jitter but it requires clock synchronization at end hosts. One approach is to use primary reference clock (PRC) like in SONET that is derived from GPS. However, this technique is still complex and not widely available for Ethernet. Roundtrip measurement is not accurate because the reverse direction might not take the same path. Frame loss and throughput are also important and can be measured by counting packet arrival at one end in a given time. Late frames must be dropped from the counting for certain applications such as voice and video. Link level failure detection uses hello messages and passes on the information about the failure in both up and down stream. Service failure detection is also needed to in case nodes and links are online but flows get interrupted such as in the case of poison forwarding table. A connectionless approach is hard for OAM because its network resources are spread throughout the network as opposed to the nailed down path in a connection-oriented paradigm [2]. Moreover, Ethernet runs a risk of broadcast storm that is a challenge for traffic management such as enforcing SLA on customers.

7.1 MEF Standards

The MEF current work on defining OAM for Ethernet does not focus on single link OAM mechanisms. That would overlap with some of the IEEE's drafts. However, MEF defines OAM mechanisms on multilink such as edge-to-edge intra-carrier OAM, edge-to-edge inter-carrier OAM, and end-to-end customer OAM. They suggest that the measurements must be per VLAN and be with the data plane. This means that in-band signaling will be used for more accurate measurements and user data is mixed with OAM. Currently, the draft includes connectivity, latency, loss, and jitter for SLA metrics. The defined OAM frame is the same as the data frame but is differentiated by the multicast address for OAM discovery and the Ethertype field. An OAM barrier filters out OAM at the edges of the domain to prevent leaking OAM from one provider to another provider or customer. Domains are defined as intra-provider, inter-provider, and customer-to-customer. For discovery operations, an edge switch sends a multicast ping request. Other edge switches response to the ping. Then the requester constructs a list of all the edge switches. Automatic discovery is useful for plug-n-play and diagnostic. Loss measurement is performed by unicast ping n times, the packet loss is m/n where m – n requests are responded. Latency is measured through roundtrip time. For

delay measurement, the source attaches a “relative” timestamp in the request. The receiver calculates the delay by the inter-transmit time which is recorded in the timestamp or the inter-received time via the actual ping received time.

7.2 ITU Standards

ITU works on the requirements for Ethernet OAM or Y.1730 and Ethernet OAM mechanism Y.17ethoam [3]. Y.1730 defines the motivation and requirements for user-plane OAM including the required OAM functions for point-to-point and multipoint-to-multipoint in dedicated and shared access. Y.17ethoam defines the mechanisms for fault management, performance measurement, and discovery. Ethernet OAM frames format is also included.

7.3 IETF Standards

IETF [13] is working on a draft entitled “Ethernet in the First Mile (EFM) OAM MIB” for single Ethernet link OAM. It is expected to complement SNMP management by defining the basic functions at layer 2 supporting directly connected Ethernet stations. The draft focuses on three areas: link fault indicator, link monitor, and control remote loopback. Link fault indicator enables one Ethernet end host to signal the other end that the path is non-operational. In addition, it allows a mechanism to operate in unidirectional mode so that the link continues to operate in one direction even though the reverse direction has failed. Link monitoring incorporates into SNMP the ability for an end host to signal the occurrences of certain important events via layer two. There are also mechanisms for an Ethernet station to query its adjacent neighbor for the status of its interface. Remote loopback is when an Ethernet host station echoes back every received packet onto the link. The draft defines object controlling the loopback and reading the status of the loopback state.

8 Security

In the past, LANs have been under the control of an organization in a small and contained area that would not span across different networks. Therefore, layer 2 is considered to be trusted and little work have been done in security for layer 2. However, this assumption is invalidated as long as there are inside-attackers. The movement of extending Ethernet to the metro core opens up more opportunities for attackers to exploit the vulnerabilities of the network. Current intrusion detection mechanisms, filtering rules, and firewall only work at layer 3 and above. Therefore, it does not directly protect the vulnerable Metro Ethernet network where layer 2 is the underlying technology. Marro [8] shows that by exploiting the lack of authentication in BPDU messages, an inside-attacker can perform Denial of Service (DOS) attacks by creating loop in the spanning tree or preventing the tree formation. A host can also snoop the network by impersonate as a switch to gain confidential data.

8.1 Vulnerabilities

The most obvious and crucial weakness in Ethernet is the lack of authentication for Bridge Protocol Data Unit (BPDU) messages. BPDU messages are used to administrate the operations of an Ethernet network. Any end host connected to a switch can generate well-formed BPDU message forcing the switch to process. This enables the host to masquerade as a working switch to join in the active topology. As the result, the attackers can perform a DOS attack, disrupting data forwarding, and Man-in-the-Middle attack, snooping traffic going through that originally was not intended for the attacker.

Ethernet runs the standard 802.1d Spanning Tree Protocol family that includes vulnerability in which the root role

is not fully monitored. For example, when a switch discovers a change in topology, it keeps sending a Topology Change Notification (TCN) BPDU up the tree toward the root until it receives an acknowledgement from the immediate neighbor. In STP, after the root receives this BPDU, it is up to root to generate subsequence Configuration Message BPDUs with the Topology Change (TC) flag on. However, the originator does not keep track of the pending operation. Therefore, a compromised root can acknowledge the TCN BPDU and not generate subsequence BPDU with the TC flag on, the rest of the switches will not detect any further changes aiding the success of a DOS attack.

There are two categorized attack-approach resulted from the combination of the two mentioned vulnerabilities: flooding attack and topology engagement attacks [8]. The following subsections describe the variants and attack scenarios.

8.2 Flooding Attacks

A flooding attack is a brute force attack that sends a steady flood of bogus BPDUs disrupting the normal behavior of the Spanning Tree Algorithm. The first variant of flooding attack is a flood of Configuration Message BPDUs with the TC flag on. The second variant of flooding attack sends a steady flow of bogus Topology Change Notification message propagating up the tree. The last variant of flooding sends special messages from the attacker claiming to be new to the topology and having root path cost of zero. Its purpose is to poison the forwarding table of the target switch. These approaches force the spanning tree algorithm to continuously recalculating so that it would not go into forwarding state. Thus, a DOS attack is successful when no data packet is forwarded.

The experiments from [8] suggest that the higher the target switch is in the tree hierarchy, the more effective the flooding attack is to degrade the network performance. One possible reason is that the higher level switches handle more trunk traffic than switches that are lower in the hierarchy. Therefore, they consume more computational resources so that they are more susceptible to resource draining attacks.

While under a flooding attack, it is possible for the topology to form a loop. When a switch under attack is computationally compromised, it absorbs all incoming BPDUs and does not generate any new ones. This is switch1 from Figure 23. Consequently, it is unable to participate in the ST protocol. The ports of the uncompromised switches that face the compromised switch believe that they are ports connecting to a non-switch node. Then, the original tree topology changes to a new one neglecting the compromised switch. However, the logical connectivity still uses the old path. This means that the ports of the uncompromised switches that face the compromised switch do not change roles and stay forwarding. When a station that sends ICMP requests stop receiving consistent ICMP replies traffic through the compromised switch, it issues ARP requests that open up the loop for the topology.

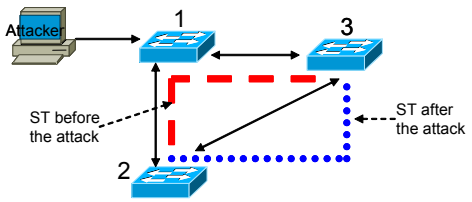

Figure 23: Loop forms when topology is under attack

8.3 Topology Engagement Attacks

In this class of attack, the attacker claims to be one of the switches participating in the spanning tree protocol. It is different from flooding in that it only sends a single BPDU per hello time period. In the BPDU, the original root

bridge ID and other parameters are preserved except for the root path cost which is incremented by a hop. By gaining a role in the topology, it is able to snoop traffic going through. The closer to the root the attacker is, the higher the volume of traffic the attacker can snoop. In addition, the attacker can stop propagating the TCN information up the tree defeating the resilience mechanism of STP. The general case of this attack is called internal node role claiming. A more specific case of the attack is when the attacker claims the root role. A host can claim to be a root by sending a configuration BPDU with the bridge ID lower than the current root's ID. When the fake root receives a TCN BPDU from a regular switch, it acknowledges that switch but it fails to set the TC flag on in its subsequent configuration BPDUs. As a result, the TC information does not propagate down the tree to invoke the ST recalculation. This attack strikes at the resilience mechanism of the Spanning Tree Protocol. It is the direct result of the vulnerability in which the root role is not fully monitored. However, it is not effective against RSTP because RSTP does not rely on the root to propagate the topology change information. While being the root, the attacker can change any parameter in the BPDU to cause further instability in the network. A variant of this attack engages the attacker to more than one host through multi-home. It is similar to the single host attack except, the attacker send the BPDU message per hello time per interface (home) targeting multiple switches. Both of these variants can do snooping of traffic as well. As the result of snooping, the attacker defeats the purpose of VLAN separation. Confidential traffic between different enterprises can be screened by the attacker.

Finally, the attacker can segment the network topology by having two or more hosts claim to have the same bridge ID that is lower the root bridge ID. Each attacking host target a different switch in the network sending a single BPDU message per hello time per interface. All switches in the network receive more than one advertisement on the same new root. Because the shortest path is picked, the topology is segmented as shown in Figure 24.

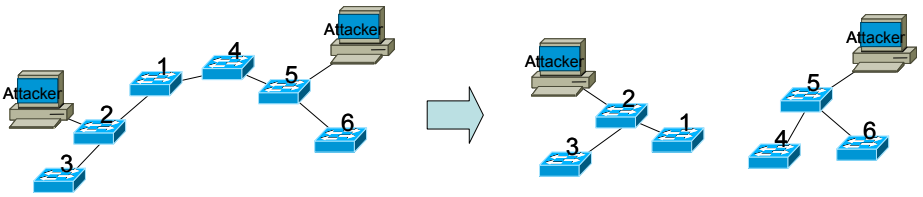

Figure 24: Topology segmentation is created by topology engagement attack

9 CURRENT DEPLOYMENTS

At the present, service providers tend to offer Ethernet over some optical network in MEN so that the services can enjoy the de-facto sub-50ms recovery time. The common trend is to have Ethernet deploys at the access and aggregation part of the metro area network. PRP or MPLS is the technology for the metro core. The customer premise equipment can be an Ethernet switch or a router.

There are several deployments of Metro Ethernet Network around the globe supporting a wide range of applications. AT&T installs a multipoint VPN E-LAN providing high-speed connection among 12 locations of the Clarian Health Center in Indiana, USA [31]. The services include voice and data interconnecting their healthcare systems providing accesses to up-to-date medical research, clinical expertise, and patient values to reduce the variance of care between physicians giving the same patient and physical condition. The city of Roanoke, Virginia, USA, deploys Ethernet over SONET network using RPR [35]. The city has OC-48 SONET backbone between ten sites. They are also running two RPRs at OC-12 to minimize disruption from future expansions and new

applications. Their application is a web portal that allows citizens to pay parking tickets and tax online, obtain registrations and permits, view parks recreation information, and satellite images of properties for real estates and development purposes. A utility company in Idaho, USA, the Idaho Falls Powers (IFP), creates a communication network of their own after realizing the flexibility and scalability of metro Ethernet technology avoiding the constrained of the incumbent carriers for new leased services [32]. Their network includes an installation of Luminous Networks Packetwave™ platforms deploying in RPR rings with MPLS abstraction. It can provide the transportation of multiple services with various CoS. To its customer, IFP offers transport of mission critical data, Ethernet private line, multiple services such as VoIP, and video surveillance. IFP leases wavelength instead of fiber strands with resiliency and automatic protection for the fiber. At the same time, IFP requires little service level staff to maintain the network.

A Korean-based wholesale provider of metro Ethernet, PowerComm, offers Ethernet services to residents in Seoul and integrating with their existing cable network at the same time [33]. PowerComm wants to provide gigabit Ethernet with the reliability as the traditional SONET/SDH but at a lower cost. It requires one metro backbone infrastructure to unify its existing cable network with the new last-mile Ethernet services. Gigabit Ethernet links are deployed in parallel between all of its Points of Presence (POP). The Gigabit Ethernet POPs are connected parallel with the central backbone POPs and to the Hybrid Fiber Coax cable network. The Metro Access connections aggregate into districts and regional POPs and then into the redundant Gigabit Ethernet Network backbone. The result is a ring of rings and each of the Ethernet rings runs RRSTP from Riverstone (see section 4.1.3). PowerComm also implements layer 2 MPLS E-line services and VPLS E-LAN services. The access ports and regional rings use hardware-based rate limiting and shaping to control bandwidth and QoS in a single management system.

In Spain, a company, called Al-Pi, saw the potential of Metro Ethernet in 2001 and decided to deploy Gigabit Ethernet services in Barcelona. At the start, it offers LAN to LAN services using the enterprises class Ethernet switches. These switches lack features such as sub-50ms resilience, SLA control, high scalability, and end-to-end QoS guarantee. Then it makes a move toward optical Metro Ethernet using dense fibre network. Now it can offer carrier class services to its customer.

10 ACRONYM

1. **AC**: Attachment Circuit
2. **AREA**: Atrica Resilient Ethernet Access
3. **ATM**: Asynchronous Transfer Mode
4. **BGP**: Border Gate Protocol
5. **BPDU**: Bridge Protocol Data Unit
6. **CBS**: Committed Burst Size
7. **CE**: Customer Edge/Equipment
8. **CIR**: Committed Information Rate
9. **CoS**: Class of Service
10. **CPE**: Customer/Provider Equipment
11. **CSMA/CD**: Carrier Sense Multiple Access with Collision Detection
12. **CWDM**: Coarse Wave Division Multiplexing
13. **DOS**: Denial of Service
14. **DWDM**: Dense Wave Division Multiplexing
15. **EAPS**: Ethernet Automatic Protection Switching
16. **EDP**: Extreme Discover Protocol
17. **E-LAN**: Ethernet LAN service (a multipoint service)

18. **E-Line**: Ethernet Line service (a point-to-point service)
19. **EoA**: Ethernet over ATM
20. **EoMPLS**: Ethernet over MPLS
21. **EoS**: Ethernet over SONET
22. **EoWDM**: Ethernet over Wave Division Multiplexing
23. **ERS**: Ethernet Relay Service
24. **ESCON**: Enterprise System Connection
25. **ESRP**: Extreme Standby Router Protocol
26. **EVC**: Ethernet Virtual Circuit
27. **EWS**: Ethernet Wire Service
28. **FDDI**: Fiber Distributed Data Interface
29. **FR**: Frame Relay
30. **HVPLS**: Hierarchical VPLS
31. **IEEE 802.17**: Resilient Packet Ring Protocol
32. **IEEE 802.1ad**: Q-in-Q or VLAN Stacking
33. **IEEE 802.1D**: Spanning Tree Protocol
34. **IEEE 802.1P**: LAN Layer 2 CoS Protocol for Traffic Prioritization
35. **IEEE 802.1Q**: VLAN tag
36. **IEEE 802.1s**: Multiple Spanning Tree Protocol
37. **IEEE 802.1w**: Rapid Spanning Tree Protocol
38. **IEEE 802.3**: Ethernet Protocol
39. **IEEE 802.3ad**: Link Aggregation
40. **IETF**: Internet Engineer Task Force
41. **IP**: Internet Protocol
42. **IST**: Internal Spanning Tree
43. **ITU-T**: International Telecommunications Union-Telecommunications Standard Sector
44. **L2**: Layer 2
45. **LAN**: Local Area Network
46. **LCAS**: Link Capacity Adjustment Scheme
47. **LDP**: Label Distribution Protocol
48. **LSP**: Label Switching Path
49. **MAC**: Media Access Control
50. **MAN**: Metropolitan Area Network
51. **MEF**: Metro Ethernet Forum
52. **MEN**: Metropolitan Ethernet Network
53. **M-in-M**: MAC-in-MAC
54. **MPLS**: Multi-Protocol Label Switching
55. **MRP**: Metro Ring Protocol
56. **MSTI**: Multiple Spanning Tree Instance
57. **MSTP**: Multiple Spanning Tree Protocol
58. **MTU**: Multi Tenant Unit
59. **OAM**: Operation, Administration, and Maintenance
60. **PBS**: Peak Burst Size
61. **PE**: Provider Edge/Equipment
62. **PIR**: Peak Information Rate
63. **PL**: Private Line
64. **POP**: Points of Presence
65. **PRC**: Primary Reference Clock
66. **PW**: Pseudowire
67. **Q-in-Q**: VLAN Stack
68. **QoS**: Quality of Service
69. **RHP**: Ring Health Packet
70. **RPR**: Resilient Packet Ring
71. **RRSTP**: Rapid Ring Spanning Tree Protocol
72. **RSTP**: Rapid Spanning Tree Protocol
73. **SDH**: Synchronous Digital Hierarchy

74. **SLA**: Service Level Agreement
75. **SLS**: Service Level Specification
76. **SONET**: Synchronous Optical Network
77. **SPC**: Startup Phase Complete
78. **STP**: Spanning Tree Protocol
79. **TCN**: Topology Change Notification
80. **TDM**: Time Division Multiplexing
81. **TTL**: Time To Live
82. **UNI**: User Network Interface
83. **VC**: Virtual Circuit
84. **VCG**: Virtual Circuit Group
85. **VLAN**: Virtual LAN
86. **VPLS**: Virtual Private LAN Service
87. **VPN**: Virtual Private Network
88. **VPWS**: Virtual Private Wire Service
89. **VSRP**: Virtual Switch Redundancy Protocol
90. **WDM**: Wave Division Multiplexing

11 REFERENCES

[1] S. Acharya, B. Gupta, P. Risbood, A. Srivastava. “PESO: Low Overhead Protection for Ethernet over SONET Transport” Proceedings of IEEE INFOCOM 2004.

[2] D. Cavendish. “Operation, Administration, and Maintenance of Ethernet Services in Wide Area Networks” IEEE Communications Magazine. March 2004

[3] S. Clavenna. “Standardizing Ethernet Services.” Jan, 9th, 2004 http://www.lightreading.com/document.asp?doc_id=45328&print=true

[4] T. Gimpelson. “Atrica makes Ethernet resilient” Network World Fusion http://www.nwfusion.com/edge/news/2002/0122atrica.html Jan, 02, 2002

[5] G. Holland. “Carrier Class Metro Networking: The High Availability Features of Riverstone’s RS Metro Routers” Riverstone Networks Technology whitepaper #135. http://www.riverstonenet.com

[6] M. Jander “VPLS: the Future of VPNs?” Mar 21st 2003 http://www.lightreading.com/document.asp?doc_id=30038

[7] K. Lui, W. C. Lee, K. Nahrstedt. “STAR: A Transparent Spanning Tree Bridge Protocol with Alternate Routing” ACM SIGCOMM Computer Communications Review Volume 32, Number 3: July 2002.

[8] G. M. Marro “Attacks at the Data Link Layer” Master Thesis at UC Davis 2003

[9] F. De Pellegrini, D. Starobinski, M. G. Karpovsky, and L. B. Levitin. ”Scalable Cycle-Breaking Algorithms for Gigabit Ethernet Backbones” Proceedings IEEE INFOCOM 2004

[10] T. L. Rodeheffer, C. A. Thekkath, D. C. Anderson. “SmartBridge: A Scalable Bridge Architecture” Proceedings ACM SIGCOMM 2000

[11] S. Shah, M. Yip “Extreme Networks’ Ethernet Automatic Protection Switching EAPS” RFC 3619

[12] S. Sharma, K. Gopalan, S. Nanda, T. Chiueh “Viking: A Multi-Spanning-Tree Ethernet Architecture for Metropolitan Area and Cluster Networks” Proceedings of IEEE INFOCOM 2004.

[13] M. Squire “Ethernet in the First Mile (EFM) OAM MIB” IETF drafts Oct. 2004 http://www.ietf.org/internet-drafts/draft-ietf-hubmib-efm-mib-02.txt

[14] M. Squire, “Metro Ethernet Forum OAM” Hatteras Networks http://www.metroethernetforum.org/presentations/MEF%20OAM%202003-12-01.pdf

[15] S. Varadarajan, T. Chiueh “Automatic Fault Detection and Recovery in Real Time Switched Ethernet Networks” Proceedings of IEEE INFOCOM 1999.

[16] Extreme Networks “Ethernet Automatic Protection Switching EAPS” Whitepaper. http://www.extremenetworks.com

[17] Extreme Networks “Extreme Standby Router Protocol™ and Virtual Routing Redundancy Protocol” Whitepaper http://www.extremenetworks.com

[18] Foundry Networks “Foundry Switch and Router Installation and Basic Configuration Guide - Chapter 13 - Configuring Metro Features” http://www.foundrynet.com/services/documentation/sribcg/Metro.html

[19] Foundry Networks “Foundry Switch and Router Installation and Basic Configuration Guide - Chapter 13 - Configuring Metro Features” http://www.foundrynet.com/services/documentation/sribcg/Metro.html#61625

[20] IEEE Information technology - telecommunications and information exchange between systems - local and metropolitan area networks - common specifications. Part 3: Media Access Control (MAC) bridges, ISO/IEC 15802-3, ANSI/IEEE Std 802.1D, 1998.
[21] IEEE Standard for Local and metropolitan area networks — Common specifications Part 3: Media Access Control (MAC) Bridges — Amendment 2: Rapid Reconfiguration Amendment to IEEE Std 802.1D, 1998 Edition. IEEE Std 802.1w-2001
[22] IEEE Standard for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks. IEEE Std 802.1Q-1998
[23] IEEE Standards for Local and metropolitan area networks Virtual Bridged Local Area Networks — Amendment 3: Multiple Spanning Trees Amendment to IEEE Std 802.1Q™, 1998 Edition. IEEE Std 802.1s-2002
[24] IETF VPLS (LDP) http://www.ietf.org/internet-drafts/draft-ietf-l2vpn-vpls-ldp-03.txt
[25] IETF VPLS (BGP) http://www.ietf.org/internet-drafts/draft-ietf-l2vpn-vpls-bgp-02.txt
[26] IETF Ethernet encapsulation (PWE3) http://www.ietf.org/internet-drafts/draft-ietf--pwe3-ethernet-encap-05.txt
[27] MEF "Metro Ethernet Networks – A Technical Overview" http://www.metroethernetforum.org
[28] MEF "Metro Ethernet Services" http://www.metroethernetforum.org
[29] MEF. "CapEx/OpEx of Traditional TDM or other L2 Services" http://www.metroethernetforum.org/Portal/S_OpExCapEx.asp
[30] MEF "MEF 5: Traffic Management Specification: Phase 1." http://www.metroethernetforum.org/TechSpec.htm
[31] MEF "Clarian Health Deploys a Metro Area E-LAN Solution to Support Patient Care" http://www.metroethernetforum.org/MEF_SuccessStories.htm
[32] MEF "IDAHO FALLS POWER - Century Old Utility Provider Moves in to the 21st Century" http://www.metroethernetforum.org/MEF_SuccessStories.htm
[33] MEF "Korea-based PowerComm, a wholesale provider of metro Ethernet services" http://www.metroethernetforum.org/MEF_SuccessStories.htm
[34] MEF "Spains Al-Pi: delivering broadband metro Ethernet thru Giganet" http://www.metroethernetforum.org/MEF_SuccessStories.htm
[35] MEF "Optical Ethernet and the City of Roanoke Virginia" http://www.metroethernetforum.org/MEF_SuccessStories.htm
[36] Nortel Networks "Service Delivery Technologies for Metro Ethernet Networks" Nortel Networks Whitepaper Sept. 19 2003 http://www.nortel.com/solutions/optical/collateral/nn-105600-0919-03.pdf
[37] Riverstone Networks. "An Overview of Virtual Private LAN Service: A new approach to LAN to LAN Communication." Riverstone Networks Technology whitepaper #200. http://www.riverstonenet.com
[38] Riverstone Networks "Scalability of Ethernet Services Networks" http://www.riverstonenet.com/solutions/ethernet_scalability.shtml
[39] VPLS.ORG "Virtual Private LAN Services (VPLS) Technical Overview." http://vpls.org/vpls_technical_overview.shtml
[40] VPLS.ORG "VPLS Standards." http://vpls.org
[41] Foundry Networks. "SuperSpan™ A Break-Through for Layer 2 Ethernet Networks" Foundry Networks Whitepaper November 2001. http://www.foundrynet.com
[42] P. Knight, and C. Lewis. "Layer 2 and 3 Virtual Private Networks: Taxonomy, Technology, and Standardization Efforts" IEEE Communications Magazine June 2004.
[43] Metro Ethernet Webinar. "Business Benefits & Key Applications for the Enterprise" November 6th 2003
[44] Metro Ethernet Forum. "Ethernet in the Metro: Status; Services; and the role of Metro Ethernet Forum" Presentation.
[45] C. Rigney, A. Rubens, W. Simpson, S. Willens "Remote Authentication Dial In User Service (RADIUS)" IETF RFC2138 April 1997. http://www.ietf.org/rfc/rfc2138.txt
[46] Extreme Networks Ethernet Automatic Protection Switching Evaluation Report. Technical Report Reference: 80056 Issue 1.1 (30/7/03) http://www.extremenetworks.com/technology/competitive/Default.asp
[47] J. L. R. Ford. "Flows in Network" Princeton University Press. 1962.
[48] J. Edmonds and R. M. Karp. "Theoretical Improvements in Algorithmic Efficiency for Network Flow Problems" Journal of ACM, vol 19, No. 2, 1990.
[49] F. Brockners, N. Finn, S. Phillips. "Metro Ethernet – Deploying the Extended Campus Using Ethernet Technology" Proceeding the 28th IEEE International Conference on Local Computer Networks 2003.
[50] MEF "The Metro Ethernet Network: Comparison to Legacy SONET/SDH MANs for Metro Data Service Providers" Metro Ethernet Forum Whitepaper July 2003.

[51] S. Halabi (Sept 2003) “Metro Ethernet” Indianapolos, IN: Cisco Press.
[52] R. Khoussainov, A. Patel ”LAN security: problems and solutions for Ethernet networks” Proceeding Computer Standard and Interfaces 2000
[53] M. Soriano, J. Forne, F. Recacha, J. L. Melus “A Paritcular Solution to Provide Secure Communications in an Thernet Environment” Proceeding ACM 1st conference Computer and Communication Security 1993
[54] N. Hadjina, P. Thompson “Data Security on Ethernet LAN” Proceeding IEEE MEleCon 2000
[55] N. Hadjina “SECURCOM – The Security Solution for Ethernet LANs” Proceeding IEEE MEleCon 2002
[56] MEF. “The Technical Specifications of the Metro Ethernet Forum” http://metroethernetforum.org/techspec.htm
[57] MEF. “Carrier Ethernet the Technology of Choice for Access Networks” March 2006