Delivering Ethernet services in the Metro

MICHAEL S. BERGER
Research Center COM
Technical University of Denmark
DTU Building 343, 2800 Kgs. Lyngby
DENMARK

Abstract: - Today, Ethernet is widely deployed in enterprise Local Area Networks, and is now emerging as the access interface of choice for delivery of data services in the Metro. The European Union (EU) project Ethernet Switching at Ten gigabit and Above (ESTA) has among its objectives a study of Ethernet services in the Metropolitan Area Network (MAN). This paper discusses Metro Ethernet services such as point-to-point and multipoint-to-multipoint services. Furthermore, different technologies for deployment of Ethernet services including pure Ethernet, SONET/SDH, RPR and MPLS, are presented and compared.

Key-Words: - Metro Ethernet, MAN, RPR, SONET, SDH, MPLS.

1 Introduction

Previously, telecom operators mainly supplied leased line services based on SDH, Frame Relay and ATM e.g. 2 Mbit/s VC12, 45 Mbit/s VC 3, 155 Mbit/s VC4, 622 Mbit/s VC4-4c and 2.5 Gbit/s VC4-16c. In this case, the customers would have to invest in the necessary knowledge and equipment to build their desired application.

Today, telecom operators must support a broad range of services ranging from layer 0 to layer 3. e.g.: Dark fibre, Grey fibre (Wavelength), Leased Line, Layer 2 VPN and Virtual Private LAN Service (VPLS) and Layer 3 IP VPN.

Layer 1 leased line services offer point-to-point connectivity. e.g. connectivity between branch offices and headquarters, or to give an enterprise access to its Internet Service Provider (ISP). Physical connectivity is provided in increments of the E0, E1, E3 and sometimes even STM-1 rates, either in a dedicated channel or in association with virtual connectivity over ATM or Frame Relay paths. LAN extension is achieved by bridging Ethernet interfaces over these leased lines to link the different sub-LANs or LAN sections.

Layer 2 Ethernet VPNs offer enterprise customers an efficient way to extend their Virtual LAN (VLAN)-based enterprise networks over the MAN/WAN. Accommodating enterprise broadcast and multicast traffic over the shared public network allows the operator to emulate an Ethernet switch, with its ports extending over regional and national boundaries. Layer 3 IP VPNs offer a similar service by emulating a distributed IP router. In this case, it is possible to offer IP-based value-added services, such as packet filtering, firewall services and Internet services. Leaving VPN responsibility up to the IP layer can simplify operation, scalability and multi-service provider scenarios (e.g., service provision to an international corporation).

Today, the majority of data traffic originates and terminates in LAN Ethernets, and therefore it is reasonable to assume that Ethernet services delivered by the MAN (Layer 2 services) could become important. Ethernet VPNs offer a simple way of interconnecting LANs belonging to a large enterprise customer, and it is therefore reasonable to believe that Ethernet transport will play an important role in future MAN networks. The MAN should be designed in a way that takes the role of Ethernet into account. There is a lot of focus in this area, and Metro Ethernet Forum (MEF) [1] has been formed to define Ethernet Services and Transport technology.

Ethernet services defined by the MEF are transport agnostic. The Metro Ethernet services can be communicated using two umbrella service types "Ethernet Line Services" (E-Line Services - point-to-point services) and Ethernet LAN Services” (E-LAN Services - multipoint-to-multipoint services). See Fig. 1. As part of the E-Line Service and E-LAN Service definitions, the MEF has defined the service attributes and parameters recommended for successful implementation of these services. This means that for each service based on these Service definitions, there are associated service attributes that have parameters, which must be set accordingly, so as to meet the customer service expectations. Typical service attributes include performance, bandwidth, speed, mode and class of service.

The goal is to provide carrier-class Ethernet service with: Protection; QoS (Quality of Service); OAM (Operation and Management) and SLA/SLS
(Service Level Agreements/ Specification). MEF specifies requirements to an Ethernet service, but does not specify the actual physical transport technology.

Fig. 1: E-Line (Point-to-point) and E-LAN (multipoint-to-multipoint) service types.

Based on the requirements to Metro Ethernet services, the interesting question naturally becomes: What technology should deliver these services? This is not a simple question, and the answer depends on who is asked. People with a telecommunication background would probably claim, that SONET/SDH based transport could deliver the required carrier class functionality, and that Ethernet should be encapsulated and carried over SONET/SDH. Data communication people could argue that the cheapest and easiest way of carrying Ethernet traffic in the Metro would be by building the Metro as a pure Ethernet switched network, thereby avoiding expensive SONET equipment. In this case, however, a number of modifications and upgrades have to be carried out in order to deliver carrier class quality and circuit emulation services.

Another candidate is MPLS based networks. Ethernet services delivered by MPLS is under standardisation, and the so called Martini tunnels can deliver E-Line service and the drafts on Virtual Private LAN Service (VPLS) specifies E-LAN service over MPLS. Furthermore, MPLS offers support for traffic engineering, service differentiation, fast reroute and also support for Layer3 VPNs. Also RPR (Resilient Packet Ring) is a candidate for delivery of Metro Ethernet services. RPR supports fast SONET-like ring protection schemes, and the utilisation of ring resources is potentially more efficient due to the possibility of spatial reuse. The following sections will provide a more detailed description of the technologies mentioned here, i.e pure Ethernet, SONET, RPR and MPLS.

2 Pure Ethernet

In order to provide Ethernet based services, e.g. E-Line and E-LAN as defined by MEF, a straightforward approach would be to build a Metro network based on Layer 2 Ethernet switching. An Ethernet switch learns the Medium Access Control (MAC) address of hosts and associates them with ports and in this way, Ethernet is a simple network to manage. The Spanning Tree Protocol ensures that there are no loops in the topology, and this ensures that broadcast traffic is not looped around several times.

The IEEE 802.1q standard enables implementation of VLANs (Virtual LAN) over a shared network such as Ethernet. The standard was developed to split larger networks into smaller parts such that broadcast and multicast traffic would not take more bandwidth than necessary. Basic E-Line/E-LAN service can be directly supported. However, a number of problems arise when VLAN is scaled to larger geographic areas covered by Metropolitan Area Networks. This is mainly due to the limited size Virtual LAN ID (VID) of 12 bits giving a maximum of 4096 different VLANs. and also due to the possible MAC address table explosion in the Metro switches. The Metro Ethernet switches potentially have to learn all MAC addresses of hosts connected to all customer VLANs.

2.1 VID scalability

A number of proposals have been made to overcome the limitations imposed by the small VID size of 12 bits. One solution is to use VLAN stacking defined in IEEE 802.1ad Provider Bridge standard. A new VLAN tag is appended to the packet as shown in Fig. 2.

![Fig 2: VLAN stacking. The shaded fields are added to the frame.](image-url)
able to support more than 4096 VLANs. This approach is not backward compatible.

Another approach discussed in the 802.1 workgroup is to define a new provider VLAN tag of 24-bit size, thereby avoiding combining the inner and outer tag.

2.2 MAC address scalability

In order to avoid that the Metro Ethernet switches have to learn all the MAC addresses of the clients on customer VLANs, the hierarchical MAC-in-MAC scheme has been proposed. It is basically based on stacking of MAC addresses as shown in Fig. 3.

<table>
<thead>
<tr>
<th>Field name</th>
<th>Number of bytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Destination Address</td>
<td>6</td>
</tr>
<tr>
<td>Source Address</td>
<td>6</td>
</tr>
<tr>
<td>Length/type : 802.1q</td>
<td>2</td>
</tr>
<tr>
<td>VID info</td>
<td>2</td>
</tr>
<tr>
<td>Length/type</td>
<td>2</td>
</tr>
<tr>
<td>Destination Address</td>
<td>2</td>
</tr>
<tr>
<td>Source Address</td>
<td>2</td>
</tr>
<tr>
<td>Length/type : 802.1q</td>
<td>2</td>
</tr>
<tr>
<td>VID info</td>
<td>2</td>
</tr>
<tr>
<td>Length/type</td>
<td>4</td>
</tr>
<tr>
<td>MAC data</td>
<td>46-1500</td>
</tr>
<tr>
<td>Padding</td>
<td></td>
</tr>
<tr>
<td>Frame check sequence</td>
<td>4</td>
</tr>
<tr>
<td>Frame check sequence</td>
<td>4</td>
</tr>
</tbody>
</table>

Fig. 3: MAC-in-MAC scheme

The outer address fields are used by the provider Metro network, and the inner address fields are used by switches in the customer VLANs. Edge switches in the Metro network must append the other address field, and it must select a MAC address that brings the packet to the right Metro Edge switch. Now, the Metro edge switches have only to store MAC addresses for the VLANs connected to that edge, and Metro core switches must only store MAC addresses for Metro edge routers.

The MAC-In-MAC scheme solves the address scalability problem, but the scheme is currently not defined in any standard or implemented in any equipment.

2.3 Protection and restoration

Protection against link and node failures is a key functionality to provide reliable Metro operation. Traditionally, Ethernet restoration was based on the Spanning Tree Protocol (STP) defined in IEEE 802.1d. STP blocks certain links in the network such that loops is avoided and a spanning tree is formed. In case of a failure, a new spanning tree is calculated. The restoration time for STP is in the order of 30s, which is definitely not acceptable for carrier-class service.

The 802.1w Rapid Spanning Tree Protocol (RSTP) is an evolution of the 802.1d standard that proposes modifications in order to reduce the restoration time. With RSTP, a restoration time of a few seconds can be obtained. This is better than STP but still much slower than SONET protection switching. It is therefore expected that faster protection mechanisms must be introduced in the Ethernet Metro, e.g. link and path protection as discussed by MEF.

3 SONET/SDH

Many carriers have already spent huge amount of money building SONET or SDH infrastructures, and this has paved the way for next generation SONET, which is more optimized for data transport. The Generic Framing Procedure (GFP) [2] together with Virtual Concatenation (VCAT) [3] and Link Capacity Adjustment Scheme (LCAS) [4] has introduced more efficient ways of transporting data in SONET based transport networks. The following sections will discuss these next generation components and show possible Ethernet over SONET (EoS) network scenarios.

3.1 Generic Framing Procedure

GFP provides a generic mechanism to adapt traffic from higher-layer client signals over a transport network. Client signals may be PDU-oriented (such as IP/PPP or Ethernet MAC), block-code oriented constant bit rate stream. GFP uses a variation of the HEC-based frame delineation mechanism defined for Asynchronous Transfer Mode (ATM). Currently, two modes of client signal adaptation are defined for GFP: Frame based (F) and Transparent (T).

In the Frame-Mapped adaptation mode, the Client/GFP adaptation function may operate at the data link layer (or higher layer) of the client signal. Client PDU visibility is required. This visibility is obtained when the client PDUs are received from either the data layer network (e.g., IP router fabric or Ethernet switch fabric), or e.g., a bridge, switch or router function in a transport network element (TNE). In the latter case, the client PDUs are received via, e.g., an Ethernet interface.

For the Transparent adaptation mode, the Client/GFP adaptation function may operate at the data link layer (or higher layer) of the client signal. Client PDU visibility is required. This visibility is obtained when the client PDUs are received from either the data layer network (e.g., IP router fabric or Ethernet switch fabric), or e.g., a bridge, switch or router function in a transport network element (TNE). In the latter case, the client PDUs are received via, e.g., an Ethernet interface.
Units (BPDU), but it is desirable with a frame based approach that carries all frames including control frames. Steps are taken to create the necessary improvements [5].

3.2 Virtual Concatenation
Two methods for concatenation are defined: contiguous and virtual concatenation. Both methods provide concatenated bandwidth of X times container-N at the path termination. The difference is the transport between the path terminations. Contiguous concatenation maintains the contiguous bandwidth throughout the whole transport, while virtual concatenation breaks the contiguous bandwidth into individual VCs, transports the individual VCs and recombines these VCs to a contiguous bandwidth at the end point of the transmission. Virtual concatenation requires concatenation functionality only at the path termination equipment, while contiguous concatenation requires concatenation functionality at each network element.

As an example, consider contiguous concatenation of VC-4s. The possible combinations are denoted VC-4-Xc, with X = 4,16,64,256. Thus, the bandwidth granularity is quite low, and this problem is basically solved with the introduction of virtual concatenation in next generation SONET. In case of virtual concatenation of e.g. VC-4s, VC-4-Xv, X= 1,2,3…256, and this allows for higher granularity in bandwidth assignments.

3.3 Link Capacity Adjustment Scheme
The LCAS Recommendation specifies a link capacity adjustment scheme that should be used to increase or decrease the capacity of a container that is transported in an SDH/SONET network using Virtual Concatenation. LCAS in the virtual concatenation source and sink adaptation functions provides a control mechanism to hitless increase or decrease the capacity of a VCG link to meet the bandwidth needs of the application. It also provides the capability of temporarily removing member links that have experienced a failure. This is useful if containers in the virtually concatenated group are routed along diverse paths.

3.4 EoS Location and options
The operator might choose from several options for EOS scenarios with different locations of the EOS interface. The interface between Ethernet and SONET can be located in either the Ethernet switching equipment or in the transport equipment. The two options are illustrated in Fig. 4.

4 RPR
Resilient Packet Ring (RPR) is designed to optimize bandwidth management of data services over a ring network. RPR utilizes SONET/SDH like protection mechanisms to provide a high degree of resiliency, but at the same time, a high utilization can be obtained. Packets can be transmitted on both directions in the ring, thereby ensuring two disjoint ways between source and destination. The Resilient Packet Ring Alliance [6] was established to promote standards based Resilient Packet Ring technology i.e. IEEE 802.17. RPR could be deployed over dark fiber, WDM or SONET but the SONET solution is probably too expensive.

The advantage of RPR compared to normal Ethernet in a ring configuration is that transit traffic is passed through each node without switching. Packets received from the ring are either dropped, if the node is the destination node. Otherwise, it is stored in a transit queue. Typically, two transmit queues are used: A primary transmit queue PTQ for high priority traffic and a Secondary Transmit Queue (STQ) for low priority traffic. With three service
classes A, B and C, class A is queued in PTQ while B and C is queued in STQ. Transit traffic stored in PTQ has the highest priority which means that the size of PTQ can be kept at a minimum. It only have to store up to one transmit packet received during transmission of add traffic or transmission of STQ traffic. The second and third priority is assigned to Class A and Class B add traffic, respectively, while the lowest priority is shared between Class C add traffic and STQ transmit traffic.

Fig. 5: RPR node

In case of congestion, the STQ will start to grow because of the assigned priority. The RPR fairness algorithm, as described in the following section, utilizes the size of STQ.

4.1 RPR Fairness scheme

In case of congestion, a mechanism is required, that can ensure fair access to the ring. Otherwise, a downstream ring node could potentially cause starvation of upstream nodes, and the STQ will eventually overflow. Fairness for the highest priority (Class A) is ensured by a combination of reservation and policing. The excess bandwidth for lower priority traffic should be distributed in a fair manner between ring nodes.

The fairness scheme is activated if a node is congested, that is, if the transmit rate from the node exceeds a certain threshold of if the size of the STQ buffer exceeds a certain threshold. In this case, fairness messages are transmitted in the upstream direction from the congested node. The fairness message contains the estimated fair share of each of the upstream nodes, that causes the congestion, and the upstream nodes will adjust their transmit rate to that announced in the message. A weight is assigned to each ring node, so the bandwidth does not necessarily need to be divided equally.

4.2 Resiliency

As the name implies, resiliency is an important part of the RPR protocol. A combination of steering and wrapping is used to obtain restoration times in the order of 50 ms. In case of a failure, topology messages are transmitted indicating the failure location. Steering means, that a station will steer the packets in the other direction, if the failure is located on the path to the destination. A station adjacent to the failure may optionally wrap the traffic by sending the received frames back in the other direction. This is similar to Bi-directional Line switched Ring protection used in SONET.

5 Ethernet over MPLS.

The Internet Engineering Task Force (IETF) is currently defining E-Line and E-LAN services running over MPLS. The E-Line service, denoted “draft Martini” [7], named after the author, uses MPLS connections as Pseudo Wires (PW) to transport Ethernet frames. The PW concept is defined by IETF. An Ethernet PW allows Ethernet protocol data units to be carried over e.g. MPLS networks. The operation of MPLS Martini tunnels (PW) is shown in Fig. 6.

Fig. 6: MPLS Pseudo Wire

The Customer Edge (CE1) transmits an Ethernet packet to CE2. The ingress Provider Edge (PE1) pushes a label (52) that indicates this specific PW. Then, the packets are tunnelled from PE1 to PE2 across the provider (P) routers P1 and P2. The tunnel is a label switched path with labels 5 and 13. The outer label is popped in the penultimate hop. The IETF defined Virtual Private LAN Service (VPLS) is an E-LAN service based on MPLS. It is an extension to the Martini draft, which is only point-to-point. With VPLS support, the MPLS network basically constitutes an emulated Ethernet bridge. Bridging functionality, such as MAC address learning, is performed in the PE routers. The core P routers do not contain any VPLS functionality. The following section describes in more detail, the mechanisms required to establish a VPLS service.

5.1 VPLS signalling

IETF have two different drafts on VPLS: VPLS-LDP [8] and VPLS-BGP [9]. The main difference is on how circuits are established between PE routers.
MPLS-LDP is basically an extension of Martini-tunnels that also uses LDP for signalling. Tunnels must be specified manually between all the PE routers that participate in the VPLS instance. On the other hand, VPLS-BGP uses the Border Gateway Protocol (BGP) for topology discovery and signalling; All PE routers are BGP peers in this case, and when e.g. a new site joins a VLAN, then BGP announces this information to other PE routers. Furthermore, BGP announces the label used for that VLAN, so it is both used for discovery and signalling at the same time. Tunnels between PE routers are typically established based on ordinary IP/MPLS procedures, e.g. OSPF and LDP.

There is a potential scalability problem with the full mesh of BGP connections. This problem can, however, be solved by introducing what is called a BGP Routing Reflector (RR) as shown in Fig. 7.

![Fig. 7: BGP Route Reflector (RR)](image)

Actually, there are many similarities between VPLS and L3 MPLS/BGP VPNs, and one of the main benefits from using IP/MPLS is the broad range of services that can be offered. Furthermore, MPLS has support for traffic engineering and there exists schemes for fast recovery from failures.

These features can partly be provided with existing technology; SONET/SDH has well-established procedures for protection, and it can carry data traffic efficiently with next gen. features as LCAS and Virtual Concatenation. Given the large deployment of SONET/SDH, this approach will surely receive attention.

Greenfield operators will probably optimise their network for data services, and RPR could be a viable solution. It combines the fast SONET like protection scheme with efficient packet transport with spatial reuse support.

Of cause, it is of great importance to build a MAN infrastructure that can efficiently support Ethernet services, but at the same time support other types of services e.g. Leased Line, L3VPN c.q. MPLS is a promising solution due to the separation of routing and forwarding, which makes it possible to implement L2VPNs in a scalable and automated way. But at the same time, other services can easily be provided. But it is important to notice, that this solution requires operation of complex protocols like BGP.

In reality, we might expect to see a mix of these technologies, with e.g. MPLS in the Metro core and pure Ethernet or RPR in the Metro access part. Or SONET/SDH rings in the access part if they are already deployed.

7 Acknowledgements
This work was carried out within the partly EU funded IST project Ethernet Switching at Ten Gigabit and Above (ESTA). For further information see [10].

References:
[6] [http://www.rpralliance.com](http://www.rpralliance.com)
[7] draft-ietf-pwe3-hdlc-ppp-encap-mpls-03.txt
[8] draft-ietf-l2vpn-vpls-ldp-03.txt
[10] [http://www.ist-esta.org](http://www.ist-esta.org)