Improving scalability of the AMS-IX network

analyzing load adaptive TE concepts and solutions

Stéfan Deelen & Reinier Schoof
System & Network Engineering
Universiteit van Amsterdam
stefan.deelen@os3.nl, reinier.schoof@os3.nl

February 4, 2008

Contents

Abstract

This document describes the results of our research on concepts and solutions for applying scalability to the AMS-IX platform. The objective of this research is to achieve load distribution by traffic engineering within the current hub and spoke topology of the AMS-IX LAN.

This research continues on preceding research on the automatical creation of Cut-through Paths (CTPs) within the AMS-IX's hub and spoke topology, driven by a load adaptive feedback control architecture. This research is mainly scoped to layer 1 and layer 2 methods and relevant properties of TE protocols within the context of path determination for load adaptive CTPs as the method for creating an efficient usage of the limited 10GE ports within the AMS-IX platform.

Within Ethernet switches, flow based forwarding via a redundant CTP next to destination based forwarding, is without solid TE solutions a non-preferred approach for the AMS-IX. This approach is not expected to perform very well. However, having customers tag their own CTP traffic, a CTP direction within regular Ethernet switches could be feasible. Since this is the case with *Private Interconnect* customers, this traffic type of this group of customers could be transparently isolated from the core within a regular Ethernet switching approach.

We think a move to more solid TE approaches is inevitable in the mid-term for creating scalability for the AMS-IX. Competitive candidates are the Provider Backbone Bridging (PBB) approach of TE within Ethernet and the versatile Multiprotocol Label Switching (MPLS) suite which is proven technology.

Extending MPLS by a GMPLS control plane offers the capability of managing light-paths by controlling Optical Cross Connects (OXCs). How this approach could be integrated with the concept of an sFlow driven control architecture is future work. Whether MPLS or PBB is capable of performing under the extreme AMS-IX circumstances should be evaluated by prove of concepts.

Acknowledgements

We would like to thank the following people for their time and help during our research:

- **Romeo Zwart, AMS-IX** - for his guidance through the research and his criticism and moral support.
- **Jan-Philip Velders, UvA** - for several brainstorm sessions and some clever insight in this matter.
- **Paola Grosso, UvA** - for sharing her visions on optical network with us.
- **Paul Bottorf, Nortel Networks** - for sending us the current 802.1ah draft and providing additional information about Provider Backbone Bridging.
- **Douglas Stewart, Matisse Networks** - for providing in-depth information about Etherburst.
- **Ralf Korschner, Foundry Networks** - for supplying Foundry proprietary information and testing some of our concepts in practice.

1 Introduction

1.1 AMS-IX

Started in the early 1990s as a LAN-based Internet Exchange, the Amsterdam Internet Exchange (AMS-IX), has become one of the world's largest IXes. The AMS-IX provides a layer 2 network to its members over which they exchange traffic. AMS-IXs peak traffic rates are over 400Gbit/s, the average traffic load is approximately 270Gbit/s (as of January 2008). Amongst the members are the largest ISP's in Europe and North America which continually contribute to a rapidly increasing traffic load on the network.

The AMS-IX network offers its members the opportunity to exchange huge amounts of Internet traffic with other members against fixed costs. This enables customers to offload traffic from their transits for destinations reachable to them through the AMS-IX platform.

The ultimate level of scalability for the AMS-IX would be to facilitate unlimited traffic exchange for an unlimited amount of members, with the only limits on throughput being either the capacity of the sending or the receiving party. AMS-IX is dedicated to offering non-blocking peering services over Ethernet infrastructure.

1.2 AMS-IX's infrastructure

In the current AMS-IX topology there is a central "hub", the core switch, to which all "spokes", the edge switches, are connected. Currently, all traffic from one site to another goes through the core.

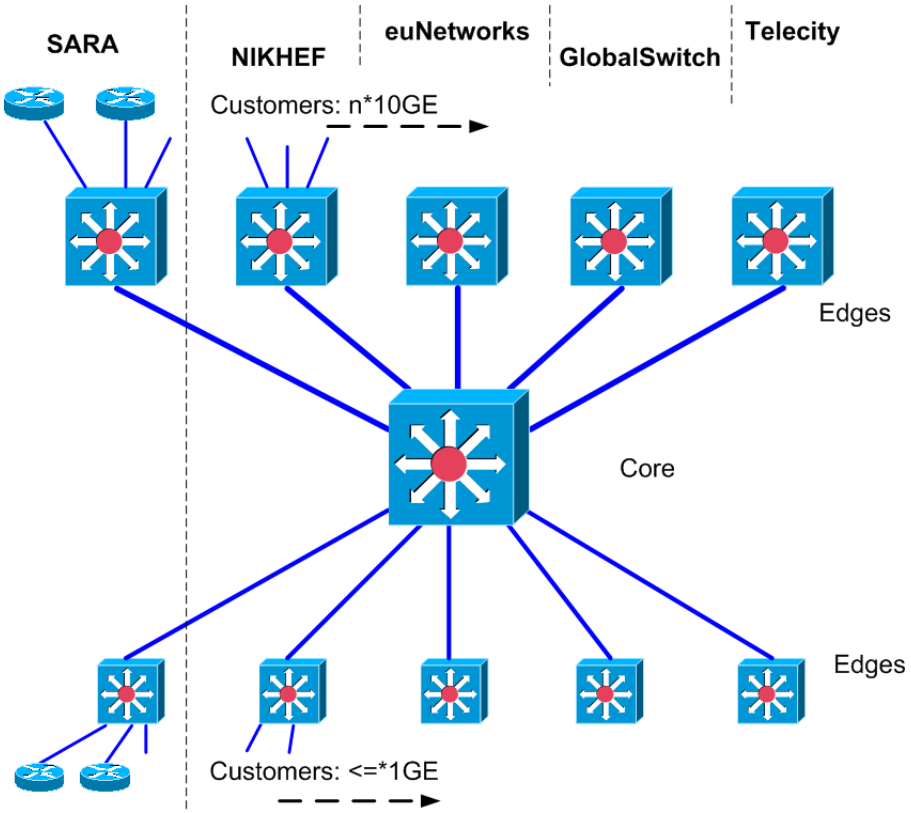


Figure 1: Overview AMS-IX hub/spoke active topology, January 2008.

Edge switches, located at five different sites across Amsterdam[1], which provide the customer access, are connected to the core through multiple aggregated 10 Gigabit Ethernet (10GE) links, so called trunks. These trunks act as a single Ethernet link, providing their accumulative bandwidth to that site.

All traffic besides site-locally switched traffic goes through the core, as can be seen in figure 1. Customer traffic growth leads to an ongoing demand for more bandwidth and new members still connect to the AMS-IX, so additional 10GE links in the backbone trunks are necessary to support this need. Since the number of available 10GE ports on the central core switch is not unlimited, available bandwidth to the core node could become a bottleneck.

The AMS-IX infrastructure has a completely redundant setup with automatic fail-over criteria for activating the standby infrastructure. Swapping the complete infrastructure from backup to active is possible within a short period of about 300ms outage[3], as seen from the customer perspective. This is possible due to the Foundry Networks[2] proprietary Virtual Switch Redundancy Protocol or VSRP which is used, to trigger the reconfiguration of Optical Cross Connects or OXCs. These OXCs can be seen as automated optical patch panels through which customers are connected to a Foundry switch. For a precise overview of the topology we refer to appendix A.

1.2.1 AMS-IX's fail-over methods

In the current AMS-IX network optical cross connect or OXCs are applied to intelligently patch fiber connections between customer routers and edge switches. Customer routers are in fact bound to a local edge switch port via an OXC port, which in fact is - from the Ethernet perspective - a transparent access to an edge switch. The AMS-IX manages the OXC connections and makes sure the customer ports are connected, via the OXC, to the active topology. The AMS-IX topology is resiliently architectured with a permanently available standby topology. The applied OXCs switches are triggered via a self-developed control method. Platform fail-overs may occur, for instance, in case of link or hardware failures, software or hardware upgrades or other maintenance.

When an active connection or hardware element fails, the OXC is automatically instructed, after a VSRP state-change has triggered an SNMP trap to be sent, to swap over to the backup infrastructure. VSRP in fact is AMS-IX's central "topology manager". The generated SNMP trap is used for indirectly triggering OXCs which are controlled via a TL1 instruction shell interface. So the logical connection between VSRP and OXC is designed by the AMS-IX.

VSRP manages state information between its controlled active and standby switches, in such a way that CAM tables keep synchronized. This shortens outage time caused by a fail-over, because no or a minimal flooding and learning is needed. During a fail-over the OXC applies a mechanical operation, it changes the light path by internally rotating the mirror which directs the light to the fail-over output port. This part of the fail-over action is done within a 20 to 50ms timeframe[9].

1.3 Document Layout

In this document several concepts and solutions are described which can be useful for further scaling the AMS-IX. The current situation and its limits is described. These limits

[1] Current points of presence are: SARA, NIKHEF, Global Switch, Telecity & euNetworks
[2] AMS-IX's current switch vendor

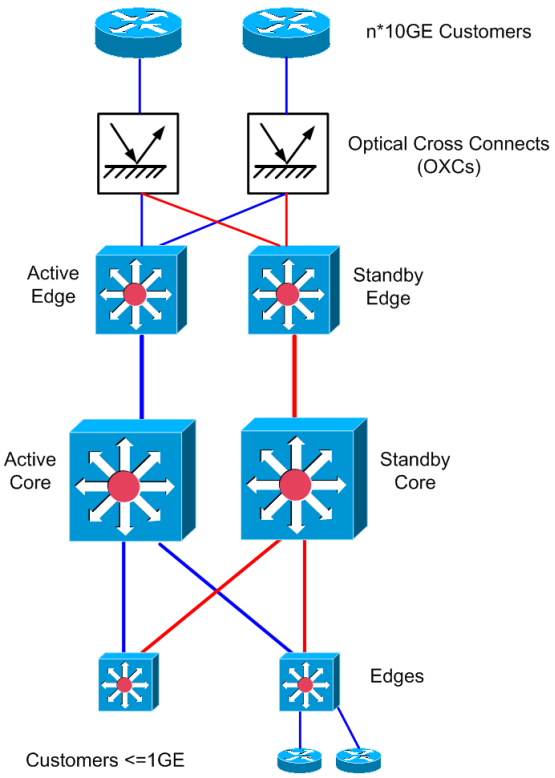


Figure 2: Brief overview of AMS-IX active-standby fail-over topology.

can be bypassed by making the network more *load adaptive*, for which in chapter 2 a control architecture is shown as a starting point for controlling traffic engineering.

In the following chapter 3, some alternative concepts are evaluated for addressing AMS-IX's scalability problem, a more dynamic use of Optical Cross Connects is also evaluated here. Chapter 4 continues by describing how traffic engineering with minimal impact on the current network can be realized. At layer 2, the regular Ethernet methods will turn out insufficient for accomplishing this. Also desired link adaptivity methods for controlling layer 1 resources will be considered. Moving on to more solid layer 2 solutions, we will take a look at the possibilities of MPLS in chapter 5 and finally we will describe how PBB could be a solution in chapter 6. We will conclude our report by a round up of all concepts and a conclusion which concept or solution deserves preference of solving AMS-IX's issues.

2 Problem description

2.1 Traffic characteristics

Currently, there are 291 members, or customers, of the AMS-IX. Hypothetically, when all customers peer with all other customers, this could create a total amount of 42.195 peerings[3]. However, typically, every AMS-IX member peers only with a subset of the other parties present. The number of peering relations will vary as a consequence of, for instance, customer network topology changes, changes in company policies and the connection of new customers. Members of the exchange connect their router(s) to the AMS-IX platform. We may assume that data traffic is only exchanged between members with a BGP peering relation between their routers (i.e. between BGP peers). Each customer router has its own unique address. In this report we define a data flow between two AMS-IX members as a data stream specified by a source MAC address and a specific destination MAC address.

Rates of data flows between peers are variable during the day, these rates are demand driven. However, over the cause of several days or weeks, it is generally possible to identify certain patterns of 'stable' flows. These data flows are monitored by the AMS-IX with the use of sFlow[23].

2.2 Research question and problem space

According to statistics, about 75 to 80% of inter-customer traffic is exchanged through the core, which leaves 20 to 25% site-locally switched traffic. Most traffic is transported via 10GE ports. The number of these ports within a single switch is limited and the expectation is that the number 10GE ports on the core may become a limiting factor to the exchange as a whole, before hardware with greater 10GE port density or an upgrade to 100GE ports is available on the market. Optimizing network usage could postpone the moment AMS-IX reaches this bottleneck. The general research question of our research on the AMS-IX is:

How can the scalability of the AMS-IX network be improved?

Which is divided into multiple sub-questions which reflect our approach:

1. *What other relevant researches have been conducted previously preceded and what is their relevance to the current research project?*
2. *Which potential solutions can be found to address AMS-IX's problem in scalability and what are their respective cons and pros?*
3. *Is there a solution which deserves preference?*
4. *How could this solution be deployed on the AMS-IX network?*

One of the objectives is to improve scalability, independent from specific vendor limitations. The direction of this research is to find load distribution means and methods which are feasible within the AMS-IX. The intention is to find solutions that can be applied transparently to customer connection(s). The currently used hardware systems are amongst the most powerful systems nowadays available. The solution AMS-IX is looking

[3]The maximum amount of peerings is: $\{p(p-1)\}/2$, where p is the number of peers.

for, is scoped to the field of networking technologies by which load distribution across redundant active Ethernet links and devices becomes possible for the AMS-IX.

Comparable research has been conducted two years ago for the AMS-IX[19], but the focus then was on a load adaptivity control mechanism. The general outcome of this previous project did not include a description of the applicability of the proposed method within the AMS-IX.

2.3 Research approach

Our approach existed of studying AMS-IXs current network and services. Also, we have studied relevant papers through Internet research and contacted vendors and people with an extensive protocol knowledge for the required inside information.

Data flows between peers are fluctuating between peers in size and frequency. Based on historic information from AMS-IX traffic monitoring, we assume that the characteristics of these flows are, at least partially, predictable. Creating scalability by distribution calls for means and methods for load adaptive traffic engineering. To be more specific, the network topology at the physical and logical level should be adapted to the demanding peers to enable off loading the core backbone connections. Within this context we will specically look for realistic methods for creating Cut-through Paths (CTP). A CTP is a shortcut between AMS-IX sites which should be temporarily created, driven by high demands for throughput between two particular peers on the peering LAN. The 'lifespan' of such CTPs is to be thought of in hours or days, rather than minutes or even seconds.

As mentioned above, we have found preceding research[19] on the automatical creation of CTPs within the AMS-IX hub and spoke topology. Our approach is partly based on this preceding research. Based on this previous work, the question whether demand driven CTP creation would help to off-load the core backbone is considered to be already answered. Also, a high-level model for a control architecture for controlling and calculating CTPs is already available. This control server architecture is shown in figure 3.

2.4 Project scope

This research is mainly intended to find solutions how the needed topology changes and traffic engineering for creating CTPs within the AMS-IX can be applied. This means this research is scoped to layer 1 and layer 2 mechanisms and relevant properties of TE protocols within te context of path determination. This research is not totally restricted to CTP solutions, so we will also consider other concepts which potentially adds scalability and judge them on their usability. At the moment within the field of physical data networking, a 10Gbit/s connection is the highest available speed for Ethernet connections. Our research is all about these connections, about the efficient usage of these 10GE ports since their available number or density is limited on hardware.

Because of the limited impact on the AMS-IX platform of traffic to and from lower speed customer connections, this project will focus on 10GE customer connections only. In the discussed concepts, solutions and pictures, these customers and their corresponding AMS-IX switches are omitted in this document.

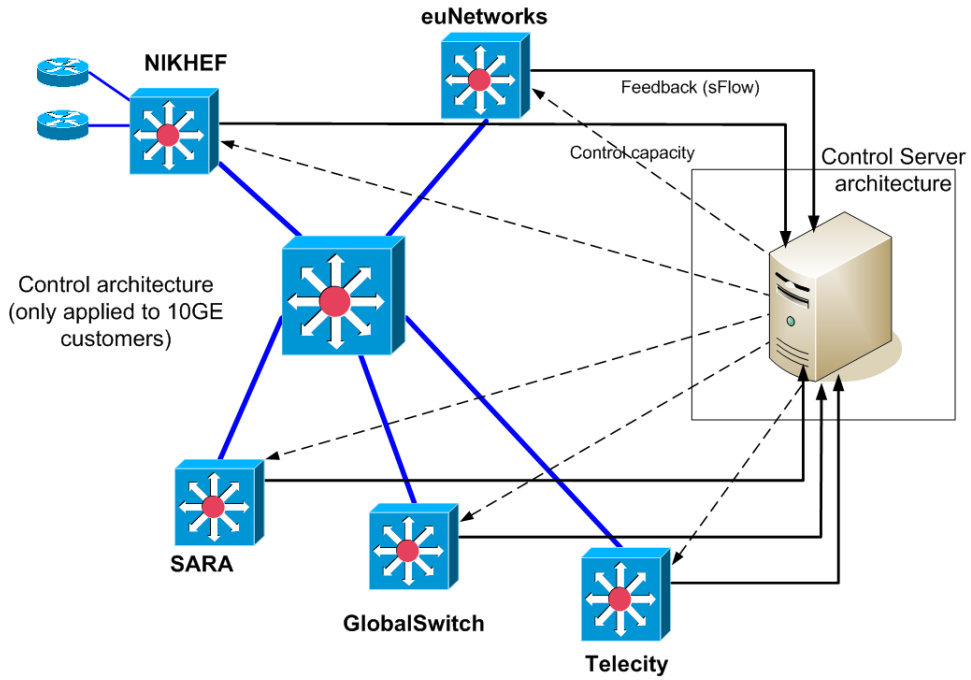


Figure 3: Control Server architecture, acting as feedback control with sFlow input.

3 Alternatives and non approaches

This chapter describes some approaches that have been evaluated after some initial research, and were decided to be less effective or unusable solutions for the problem of the AMS-IX. Furthermore we describe some approaches for which we were unable to collect sufficient information to conclude whether they are usable to the AMS-IX or not. In these cases further investigation is needed to completely decide upon their usability.

3.1 Up-scaling the current hardware

All non-locally switched traffic from one site to another has to pass the core. As mentioned before, the core switch creates a potential bottleneck. A quick fix to address this issue would seem to be installing hardware with greater capacities. Since no such hardware is available at the moment this is not really an option. Another approach could be installing additional hardware to balance the load, for instance by duplicating the core. By grouping customers that exchange high volumes of traffic on one of the two switches, more locally switchable traffic is created. This approach however requires an interlink between two cores, to transport traffic from customers located on different cores. To avoid this interlink from becoming a new bottleneck, requires a lot of links which will cost even more ports on each core. Up-scaling the hardware is therefore not an option.

3.2 Applying redundant links between destinations

The only way to offload the core is to let traffic not to pass through it which requires installing direct links between edge switches. This however creates a potential loop in the network and a protocol like STP should be used to avoid traffic from looping and congesting the network. This can have two outcomes: either the interlink path is blocked and nothing is gained, or the original path is blocked. This makes all traffic from one edge go over the interlink to the other edge switch, even when traffic destinations are not directly connected behind that edge. This link becomes a bottleneck if it is not able to handle the demanded bandwidth sufficiently. Even when the link is capable of handling this amount of traffic, nothing is gained either: on both switches additional ports are used but not necessarily filled with traffic, also one of the edge switches has in fact become another core since it is forwarding transiting traffic.

3.3 Applying VLANs on the links between stubs

To utilize both links to the core and links between edges, different VLANs[4] could be applied to the inter-edge-link which will prevent loops in the network. Traffic that meets requirements to bypass the core and use the interlink should be tagged with the VLAN of a particular interlink. Since the AMS-IX already uses some VLANs, frames need to be "untagged" and "re-tagged" when they meet the given requirements, instead of the normal procedure in which the traffic is tagged based on its entry port. Usually frames are tagged once they enter the VLAN-network and remain untouched from there on. Some vendors might support features in which frames are tagged based on certain parameters in the frame headers. In section 4.3 this approach is further described.

[4]Virtual LAN, a virtual layer 2 Ethernet network segment or a so called "broadcast domain"

3.4 Applying Provider Bridging

In 2005 the IEEE standardized Provider Bridging, 802.1ad or *Q-in-Q*. This enables service providers to create a service VLAN for each customer, in which customers can use the full 12-bit in available VLANs without conflicting with other customers or the service provider. When a frame enters the service provider's network, based on the customer entry port, the header is extended with an additional service VLAN field. Since this standard only provides an extra layer of VLANs and provides no further flexibility how to apply these VLANs, this is of no further use to the AMS-IX.

3.5 Optimizing the exchange for private peerings

In the case of a heavy traffic flow between two customers, an approach that could work is creating a layer 1 end-to-end connection between both of them. This way, the exchanged traffic is directly transported to the destination without causing any load on the network. This results in a private peering network, with no connection to any other customer of the AMS-IX. The AMS-IX is currently offering the use of *Private Interconnects*, in which two customers share a VLAN. This also results in a private peering network, only on layer 2. Private peering traffic is forwarded along the same path as other traffic. The routers of customers of this private peering service, have to tag the private peering traffic. Since customers do the VLAN tagging, the edge switch might be capable to apply a forwarding exception by forwarding this specific traffic over an inter-edge-switch link. The former section "Applying VLANs on the links between stubs" now becomes quite interesting. In 4.5, the explained method of CTP forwarding can be applied as a solution for this kind of solution.

It also might be an interesting option for the AMS-IX to supply direct inter-customer links at layer 1 on request. Customers could be delivered a layer 1 circuit between each other. Since no layer 2 switch port reservation is needed for these customers, the AMS-IX LAN will get offloaded. This approach, of course, offers a solution which is totally not transparent to customers.

3.6 Rbridges

The IETF workgroup led by Radia Perlman and Joe Touch is developing the concept of RBridges or Routing Bridges, which should provide a way to efficiently utilize an Ethernet network with redundant links. By encapsulating Ethernet frames in so called TRILL headers [18] a multi-hop route could be determined for data in layer 2 network. After the new header is added, the frame looks like normal Ethernet frames to conventional Ethernet switches. The destination address of the new header contains the address of the next RBridge hop, instead of the original destination. Using a shortest path algorithm like IS-IS, RBridges are able to prevent traffic from looping. To prevent frames from looping whilst the RBridges are calculating the shortest path to every other RBridge, a TTL field is added to the TRILL-header.

With the focus of the control plane on using the shortest path through an Ethernet network, this concept does not actually offer a load adaptive aspect. The concept, still being a draft, was not yet tested or implemented by any major vendor. IEEE's version of a protocol concerning the routing of Ethernet frames is called Provider Backbone Bridging, PBB or 802.1ah. While very similar to RBridges, this protocol only offers a data plane, with the ability of using a custom control plane or another existing protocol. This protocol is already implemented by a few vendors and we will discuss this later.

3.7 Optical Burst Switching

Optical Burst Switching uses Optical Cross Connects or OXCs to create lightpaths between endpoints on the fly. Data that arrives at an Optical Burst Switch is buffered until resources to transmit the data like paths and wavelengths are available. A lightpath is set up rapidly and the buffer bursts its data onto the network. This technique combines the end-to-end possibilities of circuit-switched networks with the flexibility of packet-based networks. Matisse Network [13] announced a product line called Etherburst in 2006, using such technology.

Etherburst is based on a fiber ring with on every site one or more so called service nodes. These service nodes operate on different wavelengths, or colors, of light. Each service node provides 20Gbps of bandwidth to the network. For one node to communicate to another, data must be sent on the wavelength corresponding to the receiver. The sender has to wait for the channel to become idle, which in this example means no one is sending on that particular wavelength, before data can be burst onto the ring. The receiver uses a *Wavelength Selective Switch* or WSS, a switch that can forward packets based on wavelength. The product line for this technique supports a range of 32 different wavelengths, providing a total of 640Gbps to the network. Etherburst provides a transparent fully meshed layer 1 network.

The total bandwidth this technique is offering is more than the current amount of bandwidth the AMS-IX core is handling, so it would seem like a decent alternative. However, using this technique requires the replacement of all network hardware and given the trend in growth of the traffic on the AMS-IX, the benefits of Etherburst would only suffice shortly. According to correspondence with Matisse's head of sales, the total bandwidth capacity of Etherburst will be doubled by using lasers of higher precision, for which a product line will come out in 2009.

Another issue is the trunking of links. Currently, the 10GE customer edge switches have trunks of up to 16 aggregated links, allowing traffic throughputs of up to 160Gbps to the core. To make such trunks possible, both the edge and core switch must be configured properly to enable the balancing of frames over the individual links part of a trunk. Based on a hashing algorithm, frames are transported over a certain link. This particular requires a certain beginning and end of the trunk. Etherburst technology is based on a fiber ring, which per site allows the use of up to 8 service nodes, supplying enough links to accommodate trunks of 16 fibers and simultaneously consuming 8 different wavelengths or colors, so layer 1 is capable of trunking to such levels. The beginning and end of the trunk is however not that evident. Since each service node operates on a different wavelength, it is unclear how the trunking configuration of the edge switch, connected to several service nodes, behaves in this setup. Connecting the edge switch to multiple service nodes without trunking the links, will cause $n - 1$ loops on that site, where n is the number of fibers.

3.8 Increase locally switched traffic

The only traffic that does not pass the core is locally switched traffic: traffic with a destination which is connected to the same switch. Increasing the amount of locally switched takes load of the core. One way to achieve that is by grouping customers that exchange traffic onto the same switch. Since the relations and traffic flows between customers constantly change, this grouping must be made dynamical. With the use of OXCs, based on information from the control architecture, customers can be patched to the same switch on the fly. The OXC hardware AMS-IX is using can create a connection in less than 50 ms[10], which has proven to be short enough to prevent the loss of BGP sessions due to

timeout for the majority of the customers.

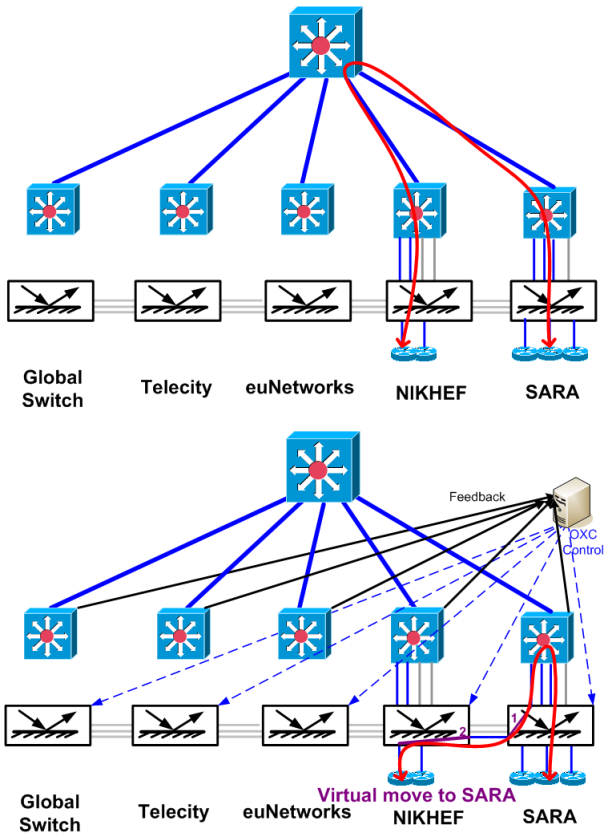


Figure 4: Moving customer routers on the fly with OXCs.

In figure 4 this concept is illustrated. Depending upon customer router hardware and configuration, the re-patching of a customer's connection may or may not interrupt the BGP sessions over this link. Particularly, customer routers must be configured to ignore the (short) layer 2 link flaps introduced by this topology change. Also the edge switch has to learn new locations of addresses when a re-patch has taken place. This leads to suboptimal traffic throughput and should happen as infrequently as possible, so re-patch intervals should be chosen wisely. What could make this approach interesting, is in combination with optical burst switching explained in 3.7. If the OXCs could buffer data during a router movement, data loss can be prevented. This buffering prevent data loss but can't prevent delays during a switch. Delays shall be noticed by specific users, also time critical applications can be affected. Another drawback is the redundancy which is compromised. Customers with redundant routers choose to install these at different AMS-IX sites for lowering risks in case a whole AMS-IX site becomes out of service. Though this approach is not transparant to customers, it might be interesting to investigate whether it is viable or applicable to a subcategorie of customers with non-redundant connections without time critical traffic demands.

3.9 Dedicated Lightpaths

Project StarPlane[21] also uses WSSs to create dedicated lightpaths between endpoints. When two endpoints need to exchange data they demand a lightpath, which involves

both their WSSs to only forward the data transmitted on a certain wavelength. The difference with Etherburst is that this path is only available to these two endpoints; other endpoints will not forward data from or send data on this wavelength. For each endpoint to communicate to all others, multiple wavelengths are required at the same time. This requires the WSS to be able to focus on different wavelengths fast enough and to support a wide enough range of wavelengths to create the necessary lightpaths. To apply such a network to the AMS-IX, 10 different wavelengths[5] for half-duplex connections and 20 different wavelengths for full-duplex connections are required. To meet the total bandwidth used on the AMS-IX, the ring must either be trunked or multiple rings including their hardware must be stacked. This should be further investigated.

3.10 Optical Label Switching

Optical Label Switching or OLS is a technique that brings the efficiency of label switching to layer 1[25]. Optical labels are sent along with their data payload over the network where Optical Label Switching Routers or OLSRs are able to detect labels of different kinds as they enter at their transport interfaces. While the data payload is delayed in so called fiber delays, the OLSR looks the label up in its forwarding table, before making a routing decision. The delay that is created also gives the OLSR the time to optionally adapt the packet. When the path or wavelength it should be forwarded to is congested, the OLSR changes the path or the wavelength so it can proceed in the network. Since the label and the data payload is handled separately, the OLSR is able to be less critical about specific space, time or wavelength difference between label and payload. Data entering the OSLR from a client interface is treated the same way, after it has been labeled and formatted. Labels from data leaving the OSLR through client interfaces are removed and reformatted to a usable format for the client. Figure 5 shows a schematic image of an OSLR.

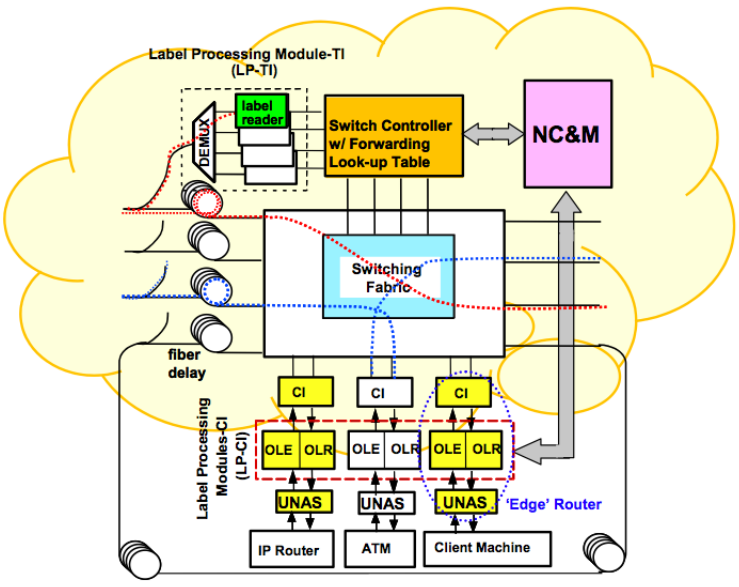


Figure 5: Schematic image of Optical Label Switching Router[25].

The forwarding table of OSLRs is populated by the Network Control & Management

[5]The number of links can be found in $\Sigma_{n=1}^{p-1} n$, where p is the number of sites.

module or NC&M. Inside the OSLR, every label switching decision is communicated over a dedicated channel to the NC&M and provides traffic statistics that way. The NC&M constantly monitors traffic conditions and alters the routing table to it and when connectivity problems occur in the network, tries to find an alternative path, avoiding the problem. The information from the forwarding table is periodically exchanged between OSLRs, providing a dynamical data and control plane which provides flexible and controlled traffic forwarding.

Although this technique offers a efficient control and data plane, like Optical Burst Switching this technique requires replacement of the current network hardware. It is also not clear what kind of bandwidth such platforms can handle, but future research could clear this out.

4 Cut-through Path engineering

According to the previous research[19] the need for cut-through paths (CTPs) at the AMS-IX is evident. These paths need to be arranged at the physical level for enabling a lightpath as well as within the Ethernet. To avoid the core switch efficiently, the AMS-IX needs these paths to be both functional and robust, to support the enormous and dynamic throughput that applies to the AMS-IX LAN. A few interesting approaches for redundant active Ethernet paths are imaginable and described in this chapter. First we describe how to create and tear down lower level CTPs automatically. Second, the feasibility of each of the potential solutions is discussed.

4.1 Expanding OXC automation

The OXCs on the AMS-IX are currently primarily used to switch from the active to the standby network, in case of hardware or link failure. These OXCs are controlled by a process called "Photonic Switch Control Daemon" or PSCD. This process can, after being instructed by, for instance an SNMP trap, reconfigure the patching of OXC ports and in this way dynamically switch traffic from one network to another. This mechanism could also provide a way to dynamically add and remove CTPs. The control architecture could trigger this process and instruct it to prepare a new or remove an obsolete CTP. After new CTPs are created, their function should be checked to ensure it is ready for service and traffic engineering at the Ethernet level can continue directing specific traffic over the new CTPs. The layer 1 infrastructure has to meet certain requirements to function in this method: each AMS-IX 10GE edge switch port should be connected via an OXC port and enough fiber capacity should be available between every different site. This way, the dynamic managing control architecture should instruct an OXC at each site for patching the layer 1 CTP endpoints to the switch ports.

In the current approach the instructions for altering the OXC configuration are sent via SNMP to the PSCD. These SNMP traps should be sent by the control architecture, although it is not clear whether the PSCD supports SNMP driven configuration messages, rather than a signal to switch from preset A to preset B. It might be required to extend the set of SNMP traps which are interpreted by the PSCD.

4.2 Switch operation manipulation limits

The hardware design of a switch is based on fulfilling its intended functions and logical operations as efficient as possible. The supported functions are mostly based on international standards, for instance the 802.1Q standard[2] described by the IEEE. Stress points within logical operations of the line cards in a multi-port 10GE switch can be expected, as described in this paper[1].

Although the potential stress points are vendor dependent, the paper gives a good impression of the performance impact of the operations that have to be performed handling traffic. The more intelligent the traffic must be handled, the more stress points can be expected at high throughput rates. As part of packet processing operations, so called "classification processes" map the extracted packet header information to information stored in the local switch's CAM table which is controlled by the control-plane. Within the CTP context, we are looking for bridging features that support making forwarding decisions based on a source and destination address for each table lookup, or "flow based forwarding". This is quite unusual within standard Ethernet implementations. The manipulation of processes along the route a packet travels as it goes from an ingress port

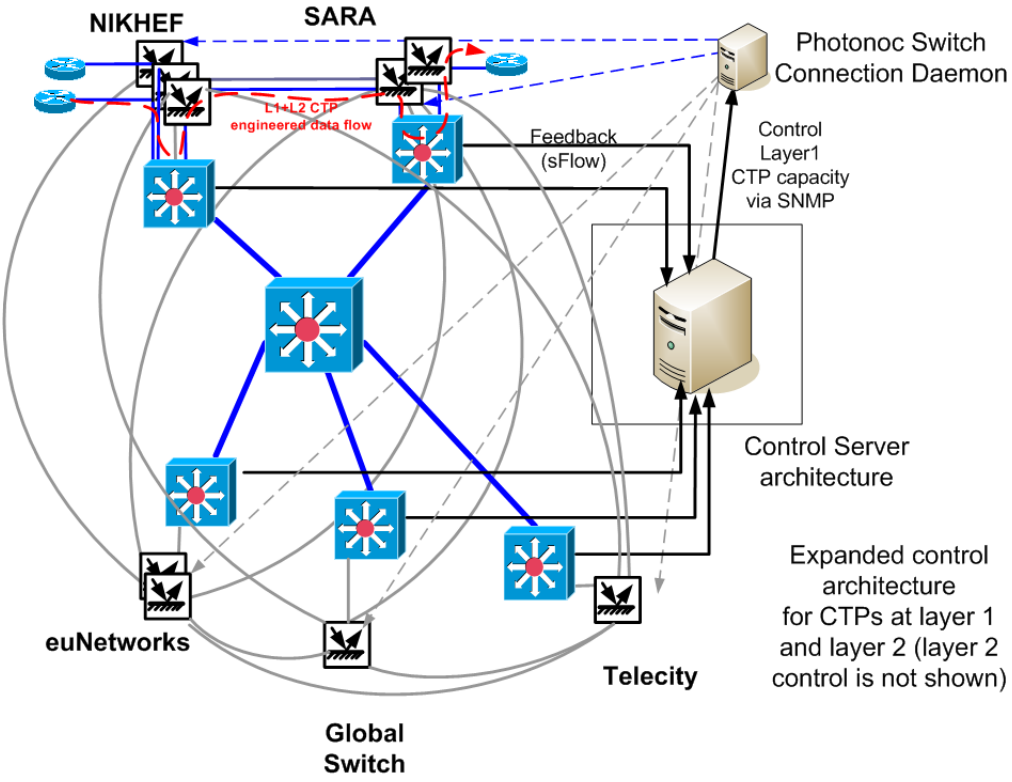


Figure 6: Automated control of OXC for CTP management.

through the switch fabric card to the egress port, can cause an impact on performance. The challenge is to apply only those manipulations that deviate in minimal way from the tasks the hardware is designed for. Actually, during this research little information was available about the internals of AMS-IX's vendor hardware and software. A representative proof of concept of flow based forwarding should demonstrate its viability for the AMS-IX.

4.3 Flow based forwarding within switches

Assuming the possibility to define a VLAN membership on an ingress port based on two criteria, the source and destination MAC address, this would be perfect for a CTP decision. By creating VLANs per CTP, traffic could be forwarded, based on flow conditions. A requirement to make this work is that it should be possible to write two CAM entries with the same destination MAC addresses, differentiated by their VLAN ID's.

On-the-fly traffic engineering by VLANs needs the following sequence of configuration changes that should be applied in the correct order by the control architecture:

1. Externally controlled OXCs should prepare a CTP lightpath between edge switches
2. The selected CTP ports on both switches should be configured as member of a unique CTP VLAN
3. Flooding, broadcast and multicast traffic at CTP ports should be disabled by correct configuration for loop prevention
4. At the CTP egress edge switch the static destination based CAM entry should be applied. This is required because the local destination is noticed within another

VLAN (at this step no traffic is affected yet, data flows are not sent over the CTP VLAN yet)

5. At the CTP ingress edge switch the static destination based CAM entry which only applies within the CTP should be configured. (No traffic is affected yet.)

6. At the CTP ingress edge switch, the CTP VLAN membership based on local source address and remote destination should be configured. This will affect the matched incoming traffic which now will be forwarded over the CTP link.

This way of traffic engineering is quite complex and, since it contains assumptions with uncertainties, might only work in theory. In practice several vendor specific obstacles arise. It is unclear whether it is possible to write duplicate MAC addresses in the same CAM table, only differing from VLAN ID. Some switches might not accept these double entries: It would mean a link identifier which is available and active in two different networks. However, when this approach would be possible within vendor hardware, there remains a risk that these "features" would not obstruct an efficient order of operations within the switch architecture. We were able to check the configurability of this approach on a Cisco[6] switch with the use of VLAN maps[4]. This is described in appendix B. Since Cisco states that this approach is designed for security filtering, it is not to be expected that high packet rates can be supported using this feature. Of course this should be verified during tests. We will look further for solutions suited for high performance.

4.4 Provider VLAN Transport

Provider VLAN Transport or PVT[15], proposed to the IETF by Huawei Technologies and Siemens, enables you to "swap" VLAN IDs within a switch. The destination MAC address is ignored by PVT. Forwarding is based on the ingress port and VLAN ID alone. The consequence of this idea which is proposed for use within carrier networks, is that if PVT is applied, the VLAN ID looses its value within the Ethernet. For the AMS-IX this problem is not relevant, the mechanism could in theory be used as a simple and effective CTP forwarding method. PVT could be interesting if it would be possible to swap the VLAN ID based on a combined source and destination MAC address. Information on whether this specific requirement is supported by PVT was unknown during this research. In the proposed (but expired) draft[16] we read that a GMPLS control-plane should apply the automatic management of swapping the VLAN IDs. VLAN ID swapping for forwardings purposes by switches is considered a violation within the Ethernet architecture. For that reason it is unlikely PVT will be standardized by the IETF, so we don't think this PVT approach will be of any means in the sort term.

4.5 Applying static CAM entries

This CTP method is as simple as effective but involves a risk for underutilizing the CTP or a risk for traffic congestion at the CTP. Assume that edge switches are equipped with CTP links between each other. With this method destinations are selected and forwarded via CTP by applying static CAM entries in edge switches. Of course first it should be checked whether the remaining capacity of a CTP link is larger than the intended traffic flowing to a destination behind the link. As a result all the traffic destined for the static CAM entry and originating from all routers connected to this switch will be forwarded over the CTP from the moment the dynamically learned CAM entry is overruled.

This step can be repeatedly applied to other qualified destination entries for a more effective use of the CTP capacity. The applied control architecture should be capable

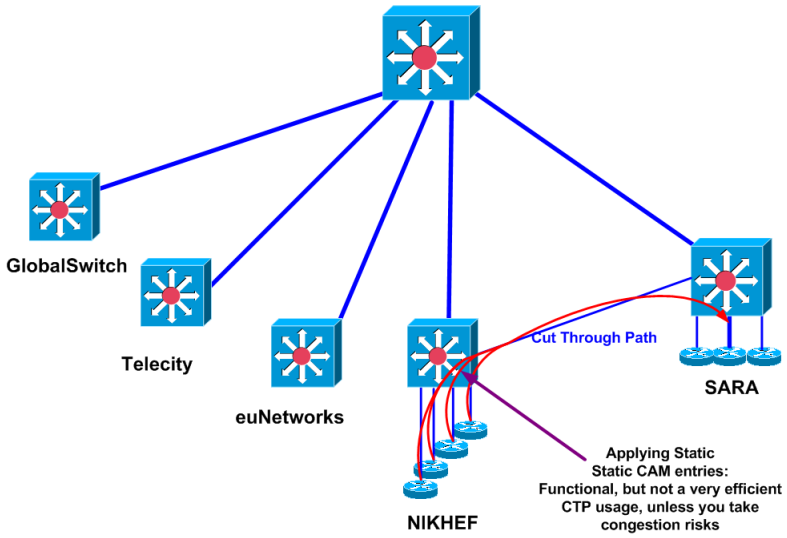


Figure 7: Layer 2 CTP engineering with static CAM entries

of determining the accumulative traffic load sourced from one site and destined for one destination. It should monitor and manage traffic at the CTP continually conform its layer 2 control feedback model.

But what about burst traffic destined for the modified CAM destination(s) behind the CTP? If several customers from one site suddenly increase their data flows to the statically applied destination(s), the limited CTP capacity will cause congestion problems. There are several options to minimize this congestion risk. It would be risk-free to only select those destinations of which the accumulative customer port capacity is equal to the CTP capacity.

This would be inefficient because traffic for one destination originates from all AMS-IX sites, while a CTP only transports traffic that originates from one site. So more static CAM entries are needed which can cause an exponential risk for congestion. Minimizing risks with maximizing efficiency can be achieved by selecting highly predictable destinations together with oversubscribing the CTP. In case of an exceptional increase of traffic load, the static CAM entry or entries should be removed quickly for example one by one by the external control architecture.

Advantages of this approach are the simplicity of the needed adjustments and the correctness of the use of equipment, since switch hardware is used for destination based forwarding without unknown difficult filter classifications. Therefore we expect no risks for performance side effects within hardware.

When a static CAM entry is applied, multi-, broadcast en flooding over the CTP is unnecessary and must be blocked at the CTP ports for preventing loops. If a customer applies ARP flooding, the requested CTP destination will be reached and found via the core since only unicast destinations are affected.

When a static entry is removed, learning the original path again goes unnoticed from a customer perspective. So on the fly engineering of multipoint to point data flows is smoothly possible with this method. The control architecture should apply traffic engineering by continually selecting and adding or removing suitable destinations.

The disadvantage are the potential underutilization of CTP links and corresponding 10GE ports. The desire is to fill these links and ports. Since it is impossible to select

specific data flows, this approach can become quite inefficient when the traffic demand for CTP destinations from sites lowers. However, if CTP paths are automatically prepared on demand by OXCs and dynamically applied or removed as described above, combined with a strict management of static CAM entries this approach migth be interested for further research based on the required traffic characteristics.

4.6 Dedicated CTP access switches

A less desirable approach is to create an extra access level below the edge switches. That way CTP management is isolated from the edge switches at which a large group of routers are connected. Qualified customers, probably large customers with several 10GE trunks, should be selected and moved away to this access switch. The advantage of this method is that specific data flows can be selected for forwarding over a CTP instead of all connected customers at a site which forms an incalculable group. Management of CAM entries can be applied as described in earlier paragraphs. See figure 8 for an overview.

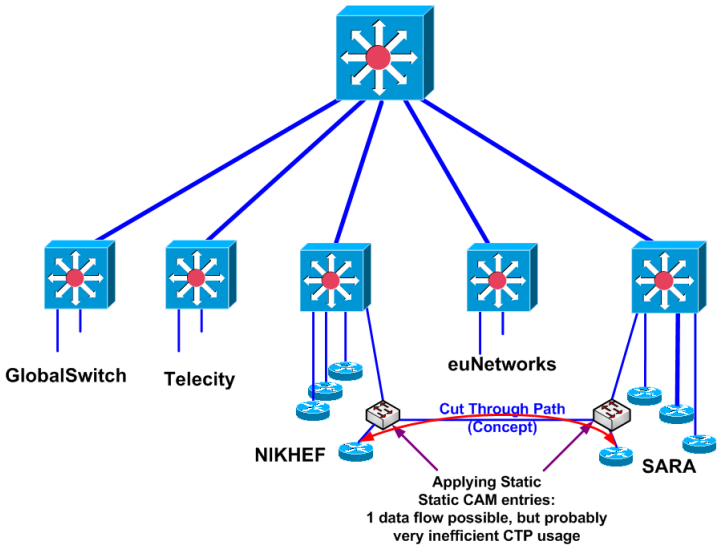


Figure 8: Dedicated CTP access switch.

Paragraph 3.8 describes the concept of virtually moving customer routers on the fly through OXC switching to other sites. Applying this concept between edge switches and CTP access switches within a site, enables data flow selectivity for CTP forwarding. However, there are better alternatives. Disadvantages of this approach are the difficulty of transparently moving customers on the fly, the extra costs of access switches, and the introduction of single point of failures. These however can be overcome by applying the current VSRP triggered OXC fail-over method to these access switches.

4.7 TE with MPLS and PBB

There seems to be no perfect method for creating a CTP with the current Ethernet technology and applied switches within the AMS-IX LAN: or concessions to traffic selection have to be made, or a performance risk has to be taken. (If it would be possible to apply flow based forwarding at all).

Now let's look at another approach. Within carrier environments specific protocols for traffic engineering are used in high performance core environments. With such protocols

pre-provisioned paths and backup paths are possible. At the moment a lot of papers can be found about two competitive protocol suites, MPLS and PBB[14] (or the related PBT). These protocols create a provider domain within and can be used to separate the end-to-end automatic Ethernet addressing characteristics within the CTPs within the AMS-IX LAN. Forwarding over the CTP is applied by PBB through encapsulation so a new address scheme is created, or by label switching with MPLS. MPLS routers are extended with an extra forwarding table for this purpose. An additional reason to look at these protocols in more detail is that the current hardware supports these protocols (or might do so in the near to mid-term future). PBB is a relatively new technology, MPLS is proven technology of over ten years. In the following chapters the potential of these two competitors will be discussed.

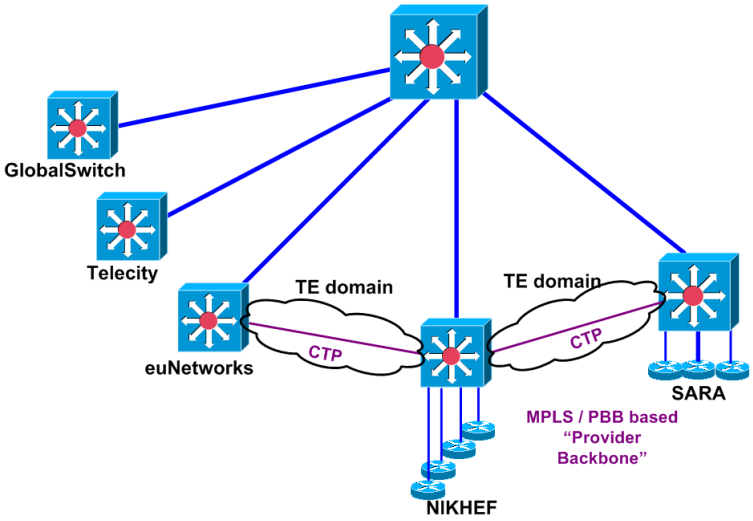


Figure 9: Provider backbone TE protocol based approach for CTPs.

5 TE with MPLS

Multi Protocol Label Switching[24], or MPLS, is a provider WAN technology of which its TE characteristics and its flexibility in combining are capable of establishing the desired Cut-through Path functionality.

5.1 MPLS concepts

MPLS is capable of transporting IP packets, Ethernet frames or ATM cells. It operates by encapsulating customer data and adding a label to it. The label is inserted between the layer 2 header and the layer 3 header. This label is generated based on fields in the headers of the customer data. Once the label is added, MPLS can efficiently forward the labeled packet by a simple lookup in a forward table. The label structure contains four sections:

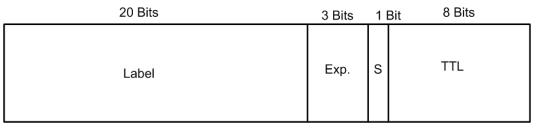


Figure 10: MPLS label structure.

20 bits address space, 3 bits experimental, 'bottom-of-stack' (in case of multiple lables), TTL for loop prevention. Conventional Ethernet switches perform their lookups based on the destination MAC address. MPLS on the other hand performs lookups based on the label a packet is given, or "pushed", at the ingress router (Label Edge Router or LER) of the MPLS network. Selected ingress data is assigned into a Forwarding Equivalence Class or FEC. Packets which belongs to a FEC, will be threatened the same. MPLS routers assign labels based on the FEC to which the packet belongs.

The lookup in MPLS results in a next-hop address, to which the packet is routed. Arrived at this next MPLS node, the label is examined, replaced ("swapped") with the label for the next hop, until the packet reaches an egress router (LER) on the MPLS network, where its label is removed or "popped". Because MPLS switches can perform the label lookup within their switching fabric instead of using its CPU for it like routers would, the lookup can be performed very efficiently.

MPLS routers in a certain MPLS domain exchange label and reachability information using a protocol called Label Distribution Protocol or LDP, so the destinations for each label are consistent throughout the domain. General routing information is distributed using existing protocols like OSPF[6] or IS-IS[7].

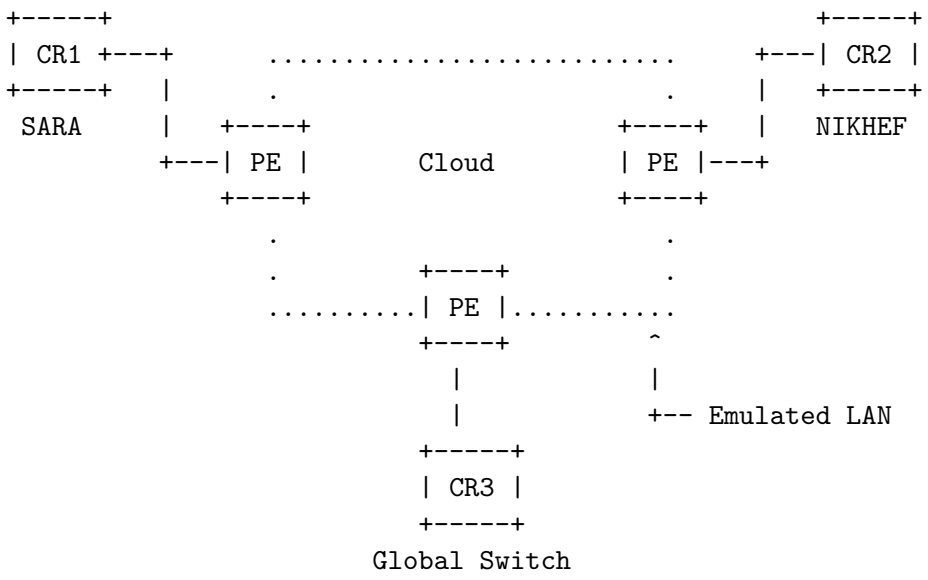
Although MPLS is versatile, it is found complex; MPLS capable devices are also relatively expensive compared with conventional Ethernet switches.

5.2 MPLS Virtual Private LAN Service

The desired functionality for CTPs within the AMS-IX LAN, can be found within the combination of MPLS and VPLS, or "Virtual Private LAN Service". VPLS is a layer 2 VPN; it offers the functionality of coupling two separated Ethernet segments via a transparent MPLS transport tunnel which acts as a pseudowire. If we apply this concept to three AMS-IX sites, the current edge switches can be seen as Provider Edge (PE)

[6]Open Shortest Path First, an Interior Gateway Routing Protocol (IGRP).
[7]Intermediate System to Intermediate System, also an IGRP.

switchs, to which Customer Routers (CR) are connected. The following picture from RFC4762[20] (but adapted to the AMS-IX situation) symbolizes this as follows:



VPLS emulates a LAN segment between remote sites, by which after coupling, the remotely separated LAN segments become integrated. Traditional learning, filtering and forwarding of MAC addresses between the coupled sites is arranged by VPLS. In an MPLS VPLS setup, a customer Ethernet frame is encapsulated as defined in RFC4448 [22]. The label switching technique of MPLS arranges the transport over the intermediary backbone in which customer MAC addresses will be associated with labels. AMS-IX staff should configure these associatons manually, or arrange this by the programmed control architecture. The MPLS labels are placed between the inner MAC header and the outer MAC header of which the addresses are visible within the CTP's Ethernet segment. Since this is a different forwarding technique between AMS-IX edge switches, paths can be Ethernet independently provisioned with the flexibility of MPLS's label switching approach.

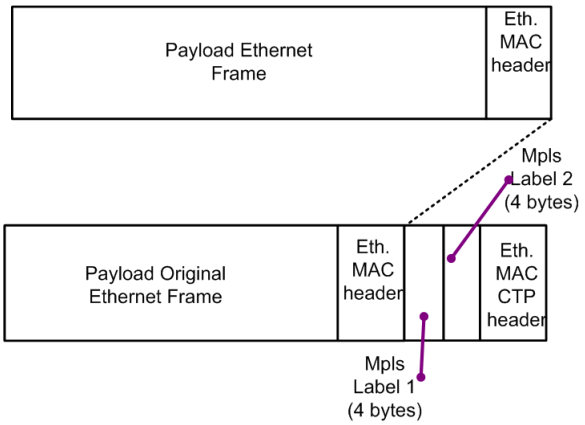


Figure 11: MPLS VPLS label structure.

Because of the encapsulation, an extra MAC domain at the CTP link is created; customer MAC addresses are not visible within the CTP domain. VPLS at the ingress ports of the AMS-IX edge switch, will be capable of selecting the desired data flows. Due

to the fact that the CTP link is a transparent Ethernet link, the accompanied flooding and learning properties are unwanted because of the redundant path via the core (not drawn in the above VPLS figure).

A disadvantage can be the added protocol headers to the original Ethernet frame. In figure 11 we can see that two MPLS labels are added for VPLS between the added outer Ethernet header and the original Ethernet frame. One MPLS VPLS label is used for identifying the tunnel endpoints, and the other for identifying the specific Ethernet environments which are coupled via this layer 2 tunnel. Small packets will cause more inefficiency within MPLS VPLS, compared to large packets [11].

The added protocol overhead is equal to the added overhead in the PBB approach. The total added size is:

Source MAC (6bytes) + Destination MAC (6bytes) + Ethertype (2bytes) + MPLS-Label1 (4bytes) + MPLSLabel2 (4bytes) = 22 bytes.

In appendix C measurements of packet sizes at a particular moment on a particual interface within the AMS-IX are shown. Over 50% of the packets are equal to or less than 128 bytes. This perspective shows that the size of the 22 bytes added header is quite relevant, over 17% [8] is added to the caused througput by these category small-sized packets.

Also broadcast and multicast traffic should be prevented over the pseudo-wire for loop prevention within the AMS-IX infrastructure. There are some uncertainties with MPLS VPLS: for the AMS-IX it's important that VPLS can be transparently applied to an ingress customer interface doing its work parallel next to other ingress traffic on the same customer port, which should be forwarded to the core switch. This specific requirement is n't defined in RFCs and is thereby left to the vendor. In this vendor paper[8] we've found Foundry's presentation regarding VPLS, in which several VPLS applications about applying load distribution to distributed core environments are demonstrated, specifics regarding the VPLS label functions are also explained. For further details about MPLS or VPLS specifics we also refer to the mentioned RFC's in the reference section.

MPLS VPLS is presented in this document as a method for the required TE for CTPs. Therefore, encapsulation only applies to specific data flows. We think that encapsulating all AMS-IX's traffic (which should be forwarded via the core) in MPLS VPLS, should not be necessary. Strictly taken, only the encapsulation of CTP traffic is required. Whether this is possible or not depends on the flexibility of the MPLSVPLS implementation on the edge switches. Further investigation is therefore required.

5.3 GMPLS as multi-layer control plane

In section 3.10 we have described optical label switching briefly. Beside that, in this document we have also looked at an approach in which OXC layer 1 circuits are provisioned on demand via the PSCD, after which layer 2 traffic engineering could continue. This approach would require two separated control planes: one for managing the physical topology and one for directing traffic on the fly along the new paths.

If it would be possible to extend MPLS with the possibility to not only apply traffic engineering but also apply control to OXCs, then it would be possible to manage both layers with one control plane. On the website of AMS-IX's current OXC vendor we have found a case[12] in which OXCs are experimentally controlled with GMPLS.

[8] (100/128)x22=17.19%, if we apply this to 128 bytes packets

MPLambdaS was designed to adopt MPLS traffic engineering as the control plane for OXCs. Its purpose was to provision and reconfigure lightpaths in OXCs.

Today MPLambdaS is referred to as GMPLS, which stands for Generalized Multiprotocol Label Switching. GMPLS in fact is a whole suite of protocol extensions for managing network technologies. It is a multi-platform control plane technology which supports not only devices that perform packet-switching, but also devices that perform switching in time, space, and wavelength domains. The latter is what makes GMPLS interesting for the AMS-IX.

We've described in section 5 that Label Switched Routers (LSRs) are placed along the path of a data flow, of which the path of the flow is determined by the FEC to which the data packets are assigned by the MPLS edge router. Within a path, Label Switching Path (LSP) control messages are exchanged for signaling. These LSP messages are send along the same path within the same layer as in which the data flow is transported. So with MPLS, the data-, and control plane are separated logically, but bounded physically.

GMPLS allows signalling across IP, MPLS and MPLambdaS domains. LSPs are the means by which GMPLS is able to apply this multi-layer control. LSPs now travel separately for the same purpose through both layers, controlled by one GMPLS control plane. So with GMPLS also a physical separation is possible. LSPs which are sent along a path from end-to-end, can now be translated into a different physical environment which is done by GMPLS. A GMPLS controlled path might consist of a layer 1 path combined with a layer 2 path, both with the same purpose: Controlling physically and logical paths for applying traffic engineering.

This means that GMPLS should be pre-programmed with the knowledge of the lightpaths which can be provisioned. GMPLS could setup potential lightpaths on demand, send LSPs across it, and proceed with the upper layer LSPs for traffic engineering at the next level. (lightpaths which can possibly be created by OXCs on the command of GMPLS), GMPLS becomes capable of using this path when there is a demand for it.

Since our research time is short, further work is needed in the field of how the interfacing to-, and controlling by a feedback control model should be done.

5.4 New fail-over scenarios

As is described in section 1.2.1, the current platform is redundantly applied in which VSRP manages the active and standby network. VSRP manages state information of CAM tables between active and standby switches. Since there are no redundant connections, this approach is straightforward and effective.

When CTPs would be applied by TE provider backbone protocols, as is drawn in figure 9, redundancy within the active topology becomes feasible. When a CTP is activated, it is unnecessary during a CTP link failure to apply a VSRP swap and activate the standby infrastructure. The MPLS or PBB protocols are capable in delivering this desired local link fail-over option. PBB properties are described in depth in section 6.

However, in case a VSRP swap is needed when provider backbone protocols are applied, more factors regarding state information have to be taken into account. We think there are two possible directions for implementing fail-over in the AMS-IX primary topology with provider backbone protocols. Direction A: When MPLS or PBB is applied in the active topology, the standby topology stays unchanged as hub and spoke topology without TE intelligence. This is straightforward but can cause potential congestion problems during a fail-over. Another difficulty is the handling of VSRP state information since both topologies aren't symmetrical anymore.

Or direction B: The redundant topology stays symmetrical, PBB and TE will be identically applied. This means that if a TE decision is made by a control plane or control architecture in the primary environment, it should synchronous and symmetrically be applied in the backup infrastructure. This could become quite complex qua design.

6 TE with PBB

Since 2005 a work group for the IEEE is developing a new technology that should enable carrier network providers to benefit from the flexibility and simplicity of Ethernet, which originally is not designed for large networks as those of carrier providers. It is called Provider Backbone Bridging or PBB and focuses on bringing these benefits to the carrier networks while delivering the same service levels and guarantees as common circuit switched networks can.

6.1 Technical details

PBB, or as assigned 802.1ah by the IEEE, is based on encapsulating customer Ethernet frames in provider Ethernet frames, completely separating both MAC-domains. Customer traffic is encapsulated when it enters the Provider Backbone on ingress switches and decapsulated when it leaves through egress switches. The PBB header contains the following additional fields:

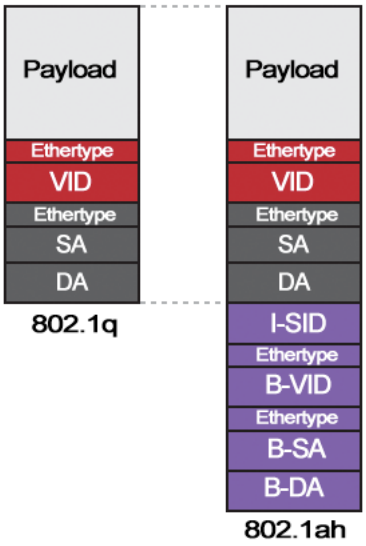


Figure 12: PBB frame compared to Ethernet frame

Some new fields define:

- **B-SA** or Backbone source address, The MAC address of the port of the ingress switch
- **B-DA** or Backbone destination address, The MAC address of the port of the egress switch
- **B-VID** or Backbone VLAN Identifier, The VLAN on the Provider Backbone this frame is in. This VLAN is in no way connected to the VLANs the customer defines on its network.
- **I-SID** or Service Instance identifier, either a 32 or a 128 bit field, defining the type of service of the frame.

Since customer frames are encapsulated in Ethernet frames, PBB frames are completely transparent to non-PBB Ethernet hardware. Instead of the original MAC destination the customer sent its traffic to, the B-DA specifies the MAC address of the correct egress switch to reach the original destination. Therefore, non-PBB enabled switches can operate normally as core switches in this configuration. Since the only addresses these switches learn are the addresses of the edge switches, their lookup tables remain very slim. Based on pre-configured policies, traffic is assigned to a certain Service Instance, which then determines to which B-VLAN the traffic is tagged. This allows providers to provide circuit-switched like connections between two sides of the provider network. These policies for instance can contain certain source/destination pairs. Since the backbone can be divided in multiple VLANs, providers are able to create a network containing loops at the physical layer while at the same time provide a loop-free end-to-end connection to customers. With loops in the network, flooding or broadcasting of frames can become a big congestion risk, since frames can endlessly loop through the network, potentially duplicating themselves. Since flooding plays a key role in the learning process of a switch, it can not just be turned off. It is however required that it is turned of on CTP ports. Broadcasting can also not be turned off since mechanisms depending on it, such as ARP request, will break. Broadcasts over a PBB network however are only provisioned to egress switches associated in the same B-VLAN.

6.2 PBB on the AMS-IX

PBB could provide a way for the AMS-IX to utilize alternative paths than through the core. Based on input of the control architecture certain flows could be assigned to a certain service instance, assuming that service instance is lead through such paths, to reach its destination without burdening the core.

In order to make traffic bypass the core switch, the ingress switch for that traffic needs to be instructed to assign the frames to a specific unique Service Instance, which than can be put in the correct Backbone VLAN: that of the link between this particular ingress switch and the corresponding egress switch. Obviously, one needs to make sure that regular traffic, i.e. all remaining traffic that to flow through the core to be able to reach the correct destination port, is not affected.

Since the backbone only transports frames with backbone MAC addresses, the customer destination MACs have to be pre-configured in order for the ingress switches to be aware of the location of each destination on the network. It depends on the implementation of this protocol by hardware vendors how manageable this turns out, but by using an external control architecture most of this work can be done automated.

6.2.1 Control plane

Certain protocols exist to extend PBB with management features, *Link Layer Discovery Protocol*, LLDP or 802.1AB and *Connection Fault Management Protocol*. These protocols and *Shortest Path Bridging*, 802.1aq, create together with PBB a suite of protocols called *Provider Backbone Transport* or *Provider Backbone Bridging - Traffic Engineering*, which offers a data plane (PBB) and a control plane (LLDP, CFMP & SPB) to have a completely dynamic network.

6.2.2 Link Layer Discovery Protocol, 802.1AB

This protocol describes a way for layer 2 devices on a LAN to discover each other and exchange management configurations. This is done by periodically sending LLDP data units

to adjacent switches. These data units contain information about like management addresses, port VLANs and link aggregations. Since we are rather interested in the discovery of customer MAC addresses, this protocol is not making our external control architecture obsolete.

6.2.3 Connection Fault Management Protocol

To provide a reliable carrier service over Ethernet, traffic management and protection is needed. Ethernet does not facilitate this by itself, so an Ethernet extension called Connection Fault Management Protocol, CFMP or 802.1ag was developed. To ensure customer traffic is transported correctly, this protocol facilitates a way to detect and anticipate on connectivity problems within the provider network.

6.2.4 CFMP Operations

As explained in [7] the protocol operates by using 3 types of messages:

- **Continuity Check Message** or CCM: Each edge switch in a provider network send on a interval CCMs to all other edge switches. When a certain switch does not receive a CCM from another switch within a certain period of time, a connection fault is detected. It then proceeds to the next stage.
- **Loopback Message** or LBM: When a switch detects a fault in a connection, it starts sending LBMs to verify the occurred problem. It sends the LBM to each switch on the path where the fault was detected. Each switch replies with a **Loopback Reply** or LBR. When no reply is received from a certain switch, the connection fault lies between the switch that last replied and the next. Now that the switch has verified there is a connection fault, it is time to propagate this information through the network.
- **Linktrace Message** or LTM: The switch sends a LTM to the first switch on the path to the connection fault. This switch replies with a **Linktrace Reply** or LTR and forwards the LTM to the next switch. As soon as a switch does not respond, the switch before that switch knows about the faulty connection.

When it is possible to generate a signal to the external control architecture, CFMP could be of great use. It could detect failures on CTP links and the control architecture could prevent data loss by not sending traffic over these links but use default paths instead, on the AMS-IX for instance through the core, until a connection is re-established. When no alternative route is possible, a fail-over to the backup network could be triggered.

6.3 Impact of PBB on the network

Before enabling PBB, it is important to estimate the impact it will have on the network of the AMS-IX. When estimating, the following items are involved:

- Since edge switches in a PBB enabled network encapsulate customer Ethernet frames and use only edge switch source and destination address, the core switch learns only those addresses and its CAM table will be very small. This could allow faster CAM table lookups.

- The PBB enabled edge switches however will have to encapsulate non-local destined traffic and decapsulate incoming traffic from the PBB network. Also it has to not only look up addresses in the CAM table and Q-tag frames correctly, but it also needs to determine which Service Instance each frames belongs to. This is expected to limit throughput of the edge switch.

- The encapsulation of customer Ethernet frames add 34 bytes of header to the frame (when the Long I-TAG or ISID would be applied, the Short I-TAG gives a 22 bytes PBB header). Compared to the usual 26 bytes of typical Q-tagged Ethernet headers on the AMS-IX, this means an increase of 138% (or 84% using the Short I-TAG) in framing overhead. Since the impact of this overhead depends on the average frame size distributed on the platform, further investigation is required.

6.4 Creating a hybrid PBB/non-PBB network

Looking at the current situation of the AMS-IX, there are also 5 edge switches located on each site, which connect 10Mbit to 1GE customers (<1GE customers) to the network. These switches are not expected to support 802.1ah on a short term, so the possibilities of creating a hybrid network, with both PBB enabled and non-PBB supporting switches, should be considered. For a situation like this to work, it must be possible to define policies to whether PBB must be enabled to certain traffic or not. Traffic to non-PBB switches inside the backbone should be treated as normal traffic, while other traffic does need to be encapsulated.

One could set a policy to determine whether traffic should be encapsulated or not, depending on whether it is directed to <1GE customers or not. Unicast traffic from a 10GE customer to <1GE customers should not be encapsulated and will be directed to the core, which switches it to the correct customer. Unicast traffic the other way around could only be possible when PBB enabled switches also allow Ethernet frames not directed to their MAC address, but to those of the customers behind them. Broadcasts, for instance ARP requests, coming from <1GE customers should be let through and broadcasts from 10GE customers should be duplicated: one for the PBB domain and one for the non-PBB domain.

Much simpler would be situation in where these customers make part of the PBB network, for instance by connecting them on the ingress side of the PBB enabled 10GE customer edge on that location.

6.5 Vendor support

At the moment, several hardware vendors are announcing their upcoming support for PBB, but only a few actually do ship hardware supporting it[17][5]. Both Nortel and Cisco have employees involved in standardizing IEEE 802.1ah. Foundry Networks, AMS-IX's current hardware vendor, let us know, although they announced the support for 802.1ah in the MLX-series, no implementation is present in current releases of their firmware. Despite this, the industry sees 802.1ah as promising technology.

7 Conclusions

Due to expected scalability issues, the AMS-IX wanted a research to be conducted upon how to improve the scalability of their platform. This report describes our findings throughout this research, trying to answer the multiple sub-questions we stated in chapter 2.2 in order to provide an answer to the AMS-IX. Being the world's largest Internet Exchange, the AMS-IX faces a unique problem. Two years earlier however, two students examined the same problem for the AMS-IX and came up with Cut-Through Paths, designed to offload the core switch by making traffic bypass the core switch on its way to destinations on the network. They described the requirements of the control architecture that should control the utilization of these paths and provided some potential solutions, being the conceptual RBridges and the usage of different VLANs on the CTPs. In chapter 3 we describe why these approaches, among others, do not deserve our preference. For instance, one could consider filtering certain CTP-credible traffic by applying certain policies, which is described in chapter 4.3. This approach is not expected to perform very well, since not all vendors support this and if so sell it as a security feature.

Chapter 4 describes how we think CTPs could be utilized using either layer 1 or layer 2 approaches. Layer 1 can be used for statically assigned CTPs and dynamical on-demand preparation of CTPs using OXCs, while we show that layer 2 based on Ethernet by itself is not capable of utilizing CTPs efficiently. Therefore we think that the solution lies in the combination of layer 1 CTP preparation and layer 2 traffic engineering and that the move to more solid TE approaches is inevitable in the mid-term for creating scalability for the AMS-IX. Customers using the *Private Interconnect* service can be handled over CTPs already. Traffic from this group of customers could be transparently isolated from the core within a regular Ethernet switching approach, since these customers perform their own VLAN-tagging.

Competitive candidates are the Provider Backbone Bridging (PBB) approach of TE within Ethernet and the versatile Multiprotocol Label Switching (MPLS) suite which has proven itself for over ten years. PBB applies MAC-in-MAC encapsulation and creates a provider domain at the CTP level. MPLS applies label switching as forwarding technique via the CTP. Both protocols are capable of selecting specific data flows for CTP forwarding, both can apply link fail-over between redundant paths and both offer a solid control for path determination.

When MPLS VPLS is applied as L2VPN approach for CTPs, MPLS VPLS adds the same amount of protocol overhead to an Ethernet frame as PBB does, but much more can not be said about the comparison in performance between PBB and MPLS. Both concepts should be benchmarked on a test-setup like the current AMS-IX platform. This way usable conclusions about the performance of both concepts can be drawn and a preference can be chosen. Chapter 5 and 6 describe how these techniques should be deployed on the platform of the AMS-IX respectively.

When OXCs supports GMPLS, GMPLS offers a way of managing the layer 1 CTP preparation and layer 2 traffic engineering from one GMPLS control plane, providing a solid all-in-one data and control plane solution.

7.1 Future Work

In order for the AMS-IX to be able to grow along with the increasing traffic demands of their customers, some additional research has to be conducted. This can be divided into two aspects: research on the performance and actual implementation of PBB or MPLS in combination with layer 1 demand-based CTP-preparation is still required. Also the

impact on the network by enabling such protocols, the possibility to enable MPLS VPLS only to CTPs to minimize impact should be investigated. Further work is needed in the field of how the interfacing to-, and controlling by a feedback control model should be done.

On the other hand, promising optical techniques such as Optical Label Switching and Optical Burst Switching can become very interesting once those techniques become available on 40GE or 100GE networks. Also, what can dedicated lightpaths between edge switches offer the AMS-IX platform, does this scale beyond current amounts of throughputs? Besides that, a prototype for the feedback and control architecture must be developed to be able to dynamically engineer traffic at all.

The AMS-IX could consider alternative business cases, such as layer 1 private interconnects in stead of layer 2 VLAN based private interconnects.

8 Glossary

AMS-IX Amsterdam Internet Exchange

BGP Border Gateway Protocol, the core routing protocol on the Internet

Control plane Maintains routing and/or label information exchange between adjacent devices

CTP Cut-through Path

Data plane Forwards traffic based on destination addresses or labels

FEC Forwarding Equivalence Class

LAN Local Area Network

L2TP Layer Two Tunneling Protocol

MPLS Multi-protocol Label Switching

OXC Optical Cross Connect, intelligent optical layer 1 switch

PBB Provider Backbone Bridging, protocol for encapsulating Ethernet frames and tunnel them through a provider network

PE Provider Edge device, bridge or router in provider domain at border position

PEB Provider Edge Bridge, participating bridge in a PBB setup

PXC Photonic Cross Connect, alias for OXC

SNMP Simple Network Management Protocol

STP Spanning Tree Protocol, control protocol for preventing loops

TE Traffic Engineering, applying control for balancing network resources

TL1 Transaction Language 1, management protocol to translate configuration to machine instructions

VLAN Virtual LAN

VPLS Virtual Private LAN Service, a L2TP available for use with MPLS

VSRP Virtual Switch Redundancy Protocol, Foundry proprietary protocol that is used to determine flaws in the network

A AMS-IX's topology

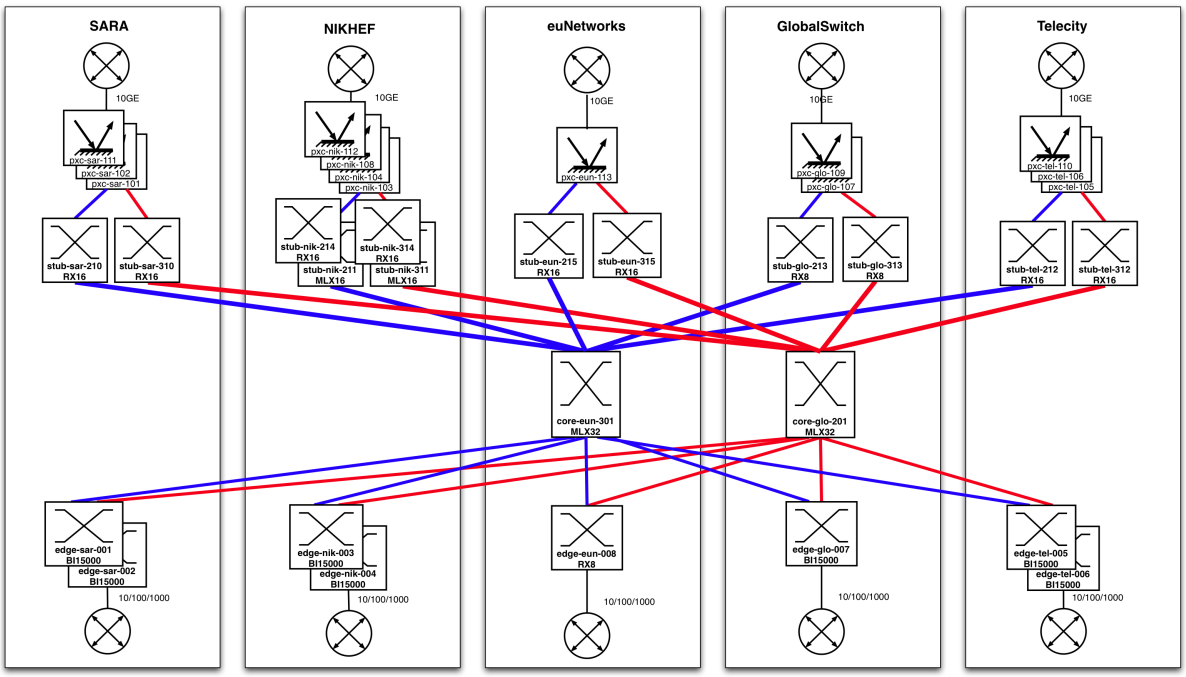


Figure 13: Actual AMS-IX topology

B Cisco CTP layer2 VLAN map configuration

In this example we show the method for security filtering which is used for creating Quarantine VLANs. This method selects a flow apart from other traffic, for example for the duration of the needed removal time of for example a virus. There are many filter criteria from layer 2 to 4, which can be used for input for VLAN membership criteria. This approach offers the desired functionality for forwarding specific flows during a specific time along CTPs.

```
Regular entry:

BEA04TST#sh mac-address-table | inc 532f
 20    0012.79cf.532f    DYNAMIC     Fa0/1

Adding a static entry in CTP vlan:
BEA04TST(config)#mac address-table static 0012.79cf.532f vlan 550 interface fastEthernet
BEA04TST#sh mac-address-table | inc 532f
 20    0012.79cf.532f    DYNAMIC     Fa0/1
 550    0012.79cf.532f    STATIC      Fa0/3

Configuring MAC pair-based VLAN member-ship:
BEA04TST(config)#mac access-list extended pair-list
BEA04TST(config-ext-macl)#permit 1111.2222.3333 255.255.255 0012.79cf.532f 255.255.255
BEA04TST(config-ext-macl)#end
```

```
BEA04TST(config)#vlan access-map mac-pair 10
BEA04TST(config-access-map)#match mac address pair-list
BEA04TST(config-access-map)#action forward

Activate by applying filter to VLAN:
BEA04TST(config)#vlan filter mac-pair vlan-list 550
```

This sequence of configuration statements could be applied by the control architecture. However, we don't think the desired throughputs are feasible by this solution.

Foundry did n't succeed in applying this procedure/feature.

C Packet distribution within AMS-IX

These are samples taken at three AMS-IX switch interfaces at a particular moment, so not quite representative as an average. However, this information is intended as an indication for the impact of exta TE protocolheaders.

```
Interface A:
> Distribution of     8.38396e+12 packets:
> <=    64:     84478055360 (1.0%)
> <=   128:   4580339324273 (54.6%)
> <=   256:    460266925230 (5.5%)
> <=   512:    305746032003 (3.6%)
> <=  1024:    468512662130 (5.6%)
> <=  1500:   2484615058189 (29.6%)
Interface B:
> Distribution of     4.50616e+12 packets:
> <=    64:      4643769338 (0.1%)
> <=   128:   2647158679473 (58.7%)
> <=   256:    430654605179 (9.6%)
> <=   512:    196292851003 (4.4%)
> <=  1024:    260100739727 (5.8%)
> <=  1500:    967312322442 (21.5%)
Interface C:
> Distribution of      5.4064e+12 packets:
> <=    64:       589246240 (0.0%)
> <=   128:   2853635375959 (52.8%)
> <=   256:    590110645741 (10.9%)
> <=   512:    246995960671 (4.6%)
> <=  1024:    293062826056 (5.4%)
> <=  1500:   1422004226546 (26.3%)
```

References

[1] 10-Gigabit Ethernet Switch Performance Testing.
http://www.ixiacom.com/pdfs/library/white_papers/10ge.pdf.

[2] 802.1Q - Virtual LANs, IEEE.
http://www.ieee802.org/1/pages/802.1Q.html.

[3] Amsterdam Internet Exchange.
http://www.ams-ix.net/.

[4] Block ARP Packets with Use of MAC Access Lists, Cisco.
http://www.cisco.com/warp/public/473/mac_acl_block_arp.pdf.

[5] Ciena, The Network Specialist.
http://www.ciena.com/.

[6] Cisco Systems Inc.
http://www.cisco.com.

[7] Ethernet now offers the most comprehensive oam for packet-based solutions.
www.nortel.com/solutions/collateral/nn119660.pdf.

[8] Foundrynet, Offering Scalable L2 Services, VPLS.
http://www.foundrynet.com/pdf/an-offering-scalable-l2-services-vpls-vll.pdf.

[9] Glimmerglass Intelligent Optical Switch.
http://www.glimmerglass.com/.

[10] Glimmerglass Intelligent Optical Switch System 100, Data Sheet.
http://www.glimmerglass.com/documentbank/Glimmerglass_System_100.pdf.

[11] Internet Packet Size Distributions: Some Observations.
ftp://ftp.isi.edu/isi-pubs/tr-643.pdf.

[12] Internet2 Collaborates with Glimmerglass to Provision and Test On Demand Optical Paths.
http://www.glimmerglass.com/PDF/Internet2%20Glimmerglass%20Case%20Study.pdf.

[13] Matisse Networks Optical Burst Switching.
http://matissenetworks.com/.

[14] Provider Backone Transport, Lightwave, Peter Lunk.
http://lw.pennnet.com/display_article/291600/13/ARTCL/none/none/Traffic-engineering-for-Ethernet:-PBT-vs-T-MPLS.

[15] Provider VLAN Transport, Huawei Technologies.
http://www.huawei.com/file/download.do?f=735.

[16] Provider VLAN Transport, IETF draft (expired).
https://datatracker.ietf.org/drafts/draft-sprecher-gels-ethernet-vlan-xc.

[17] SMARTPACK PBT Switch Solution.
`http://www.tpack.com/fileadmin/user_upload/Product_Briefs/SMARTPACK_PBT_v1_web.pdf`.

[18] Transparent Interconnection of Lots of Links (TRILL).
`http://www.ietf.org/html.charters/trill-charter.html`.

[19] René Jorissen & Lourens Bordewijk. Automated Cut-Through Paths. 2006.
`http://staff.science.uva.nl/~delaat/snb-2005-2006/p33/report.pdf`.

[20] M. Lasserre and V. Kompella. Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling. RFC 4762 (Proposed Standard), January 2007.

[21] Jason Maassen, Paola Grosso, and Li XU. Het StarPlane-project. May 2006.
`http://www.cs.vu.nl/~jason/papers/0604-28Maa.pdf`.

[22] L. Martini, Ed., E. Rosen, N. El-Aawar, and G. Heron. Encapsulation Methods for Transport of Ethernet over MPLS Networks. RFC 4448 (Proposed Standard), April 2006.

[23] P. Phaal, S. Panchen, and N. McKee. InMon Corporation's sFlow: A Method for Monitoring Traffic in Switched and Routed Networks. RFC 3176 (Informational), September 2001.

[24] E. Rosen, A. Viswanathan, and R. Callon. Multiprotocol Label Switching Architecture. RFC 3031 (Proposed Standard), January 2001.

[25] S. J. B. Yoo. Optical-Label Switching, MPLS, MPLambdaS, and GMPLS. 2002.
`http://sierra.ece.ucdavis.edu/documents/Yoo_ONM_OLS_2002_submission.pdf`.