Provider Link State Bridging (PLSB)

By Don Fedyk and Paul Bottorff , Nortel Networks
January 2007

Introduction

With the evolution of the provider data plane in Provider Backbone Bridges (PBB) (802.1ah) [PBB] and now a new proposed PAR for PBB Traffic engineering (PBB-TE) we are interested in discussing the evolution of the Provider Bridging Control plane. A number of new and old technologies can come together to provide a new feature rich Ethernet core.

A Provider Backbone Bridge Network (PBBN) works with current Spanning tree protocols, STP, RSTP and MSTP as well with MRP with no extensions. However, the recent developments in link state routing applied to Ethernet are of interest to Providers since they allow for maximum use of network resources without some of the restrictions of the other spanning tree algorithms.

Provider Link State Bridging (PLSB), defined in this document, uses link a state protocol and computation to populate forwarding tables to construct shortest path loop free connectivity for an 802.1ah Provider Backbone Bridge Network (PBBN) for a portioned set of Virtual LANs (VLANs). PLSB can be used as the one and only control plane, or it may run as a "ships in the night" mode with others Spanning tree control planes. Unicast and multicast communication is simultaneously created such that the PBBN offers highly scalable transparent LAN service to the Customer MAC (C-MAC) layer. The combination of PBBN using PLSB has better scaling and better operational characteristics than for PBBN by itself. At the same time, PLSB optimizes multicast topology in a way not achievable prior to this with Ethernet or MPLS systems.

PLSB Topology Requirements

A key item, on a per VLAN ID (VID) basis, is to preserve congruency of forwarding across the network for both unicast and unknown/multicast traffic and to use a common path in both directions:

1) Very low probability of reordering frames in a flow during learning.

2) When network changes or outages occur they have a high probability of being bidirectional.

3) Congruency of forwarding of client IEEE 802.1ag multicast Connectivity Fault Management (CFM) frames and the corresponding unicast path across the service network used for responses.

PLSB must support incremental transition, preserving existing Ethernet attributes at the point of attachment to any non-compliant parts of the network. Non-PLSB portions of the network must be able to peer with PLSB at the customer MAC layer, or be surrounded by PLSB at the backbone MAC (B-MAC) layer. To a PLSB network, the surrounded portion of the network has the same connectivity properties allowed by PBB, it may look like a LAN segment or another type of LAN network.

Another requirement is that learning of the PBBN addresses for the provider Unicast and Multicast is achieved by the link state protocol. In other words learning of Provider topology in a PLSB network is achieved through link state protocol exchange. Learning of customer addresses C-MAC and customer topology is left to current PBB methods on the Provider Edge Bridges.

PLSB Building Blocks

It is worth while reviewing some of the key developments that have been progressing in the Ethernet bridging data plane to fully understand the implications with respect to the control plane.

Common Aspects of the Provider Backbone Data Plane

Encapsulation: For the outer addresses, PBB and PBB-TE can use addresses out of the local admin space as domain wide unique without concern for global uniqueness. This creates an independent provider address space that allows providers additional backbone MAC (B-MAC) address space for provider only use.

For the inner addresses, full customer addresses encapsulation of Customer MACs and tags provides a clean separation of customer and provider addresses saving Provider Switch resources, and also preventing undesirable protocol interactions.

PBB Header: PLSB utilizes the PBB header unmodified. Having a distinct Provider address space allows certain freedoms in the control plane that are unique to the provider backbone.

Provider Service Instance ID (I-SID): The I-SID provides an instance identifier that can be used in the data plane to provide the context of the particular frame. The service instance is a unique identifier for a point to point or a community of multicast connections.

Common Aspects of PLSB and PBB-TE

PBB-TE is an impending work item at the time of writing this document. This section lists some of the intended mechanisms, because many of its basic data path mechanisms are also utilized by PLSB. PLSB is a control plane capable of controlling and configuring PBB-TE. Manual configured PBB-TE and routed PBB-TE paths can co-exist.

PBB-TE is traffic engineered data plane that does not use any current spanning tree or even necessarily shortest path to establish either point to point data flows or point to multipoint flows. The current specified mechanism for establishing a PBB-TE path requires population of the forwarding tables by configuration.

Configured and Link State Populated Forwarding: In Ethernet, Forwarding tables may be populated by spanning tree mechanisms such as learning or by static configuration. Static configuration exposes the interface represented by the dot-one-D-static subtree of RFC 4188 for management directly into the filter database). PBB-TE uses distinct VLANS where forwarding tables are populated exclusively by configuration (or by other means that is consistent with configuration). PLSB will populate FIBs in a manner that is consistent with configuration as viewed from conventional Spanning tree control. See Figure 1 .

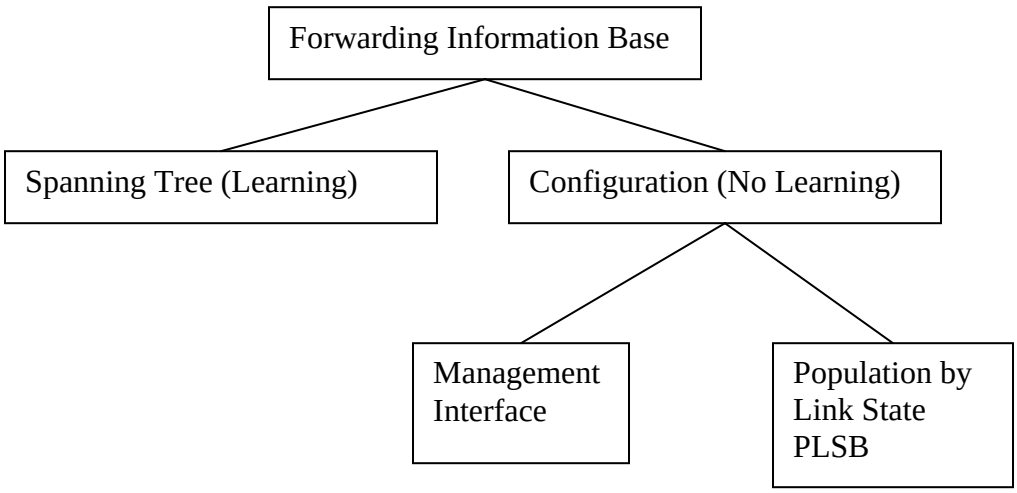

Figure 1 FIB Population

VID Partitioning: The Backbone VLAN (B-VLAN) range utilized by PLSB needs to be partitioned from other B-VLAN spaces. This mechanism is based on the same requirements for PBB-TE. This allows complete decoupling of any PBB spanning tree functions such as Learning and unknown address broadcasts. The shortest path bridging specification and the proposed PBB-TE management mechanism also have an option for VID partitioning via assignment of a reserved MSTID. It is envisioned PLSB would share this mechanism. PLSB mandates symmetric VID usage.

Loop Free Connectivity: PLSB requires the paths to be loop free, just as other Ethernet Control planes do.

Shared Forwarding: Point to point PBB-TE paths allow for the sharing of forwarding paths to a given destination. Many paths that terminate on a destination bridge can share the same B-DA and B-VID but have differing B-SA or I-SID. This reduces the number

of B-DA MACs that are used per destination to one for the auto mesh point to point connectivity. This has desirable properties for certain control planes as we explore later.

Learning: Learning is turned off for PBB-TE and PLSB VLANs.

Link State Control Planes

There are currently a number of standardization efforts to introduce links state control planes to Ethernet. There are a number of similarities and dissimilarities between the approaches. One of the commonest approaches is to use the IS-IS [IS-IS] routing protocol and strip the IP address identifiers from the protocol. In the place of IP Ethernet MAC addresses can be used.

Shortest Path Bridging

Shortest path bridging [SPB] offers one possibility of bringing Link State routing based on the IS-IS protocol to the generation of Shortest Path Spanning Trees. PLSB has similar goals to SPB but different options enabled by the provider aspects of PBB and PBB-TE. Shortest path bridging uses multiple VIDs to identify the multiple shortest path trees. PLSB has alternatives to the VID provided by the PBBN to be explained later in the paper.

TRILL

The IETF working group activity called TRansparent Interconnection of Lots of Links (TRILL for short) [TRILL], uses IS-IS to compute shortest path spanning trees. While TRILL kicked off some of the early discussion in the Ethernet link state area, TRILL uses a different forwarding paradigm and TRILL does not utilize the PBB encapsulations.

IEEE 802.11s

The IEEE Wireless Mesh networking Group IEEE 802.11s [Mesh] uses, in one variation, a version of Optimized Link State Routing OLSR). OLSR is a different link state system to carry MAC addresses for creating a Wireless mesh. The wireless constraints are quite unique so we do not examine this version of link state routing any further.

Using IS-IS for Provider Link State Bridging (PLSB)

We present the use of IS-IS tailored to Provider Bridges with MAC address called Provider Link State Bridging (PLSB) as having some unique properties outlined in the following sections.

Fundamentally the one property of PBB Ethernet that is exploited is the domain wide uniqueness of the PBB/PBB-TE Ethernet header for objects such as B-MACs and I-SIDs.

The uniqueness of these objects allows the distribution of information in link state and removes a whole layer of signaling that exists in other bridging or routing systems today.

Shortest path for PBB and PBB-TE

Given a set of provider backbone bridges it is possible to create shortest path trees from every PBB to every other PBB using an IS-IS link state database designed to carry Backbone MAC addresses. The Backbone edge Bridges (BEB) and the Backbone Core Bridges (BCB) learn the topology of the network in a standard IS-IS/link state fashion. Once all bridges have learned the topology all functions are performed by computations against the data distributed by IS-IS. Shortest paths for Unicast/Multicast are all computed in this manner and the results populate the bridge's FIB directly.

It is important to understand that we are computing an all pairs shortest path to create simultaneously: a shortest path tree; and an inverse symmetric tree to the shortest path tree. The result is a simultaneous shortest path multicast tree from each root BEB with all edges as leaves of the tree and a congruent multi-point to point unicast tree each the root BEB. This is a very powerful combination of multicast trees and unicast paths that is computed in parallel on the link state database. See Figure 2.

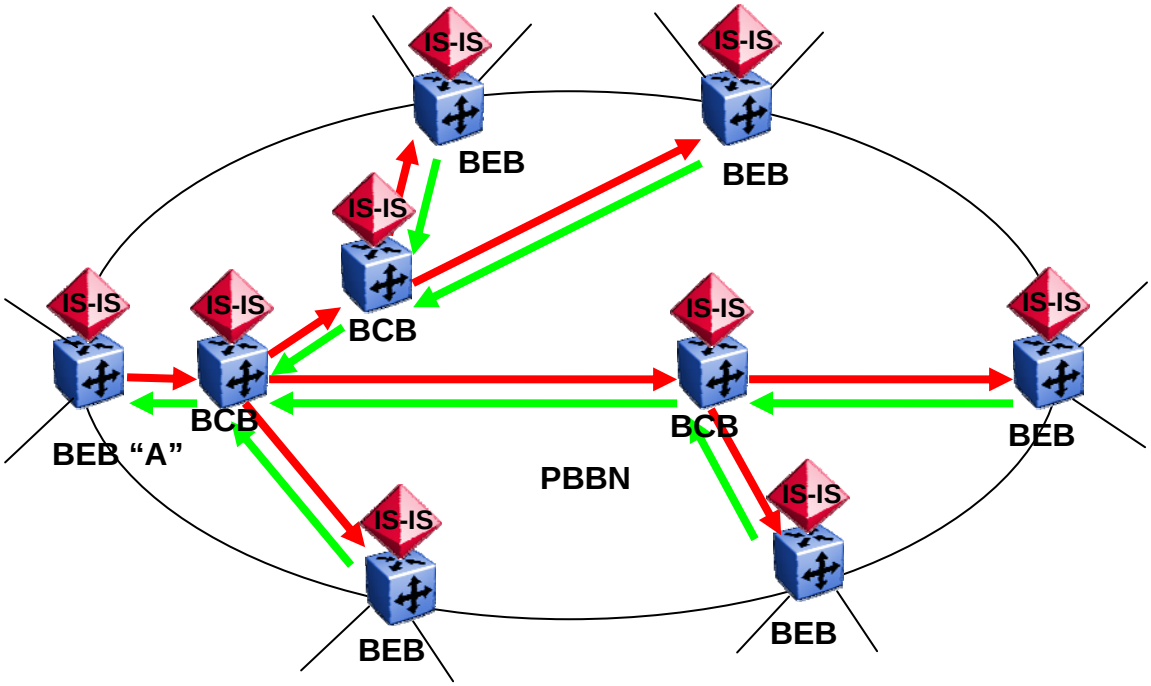

Figure 2 Multicast and Unicast Paths computed by link state for BEB A.

In Figure 2 the Backbone Edge Bridge A (BEB A):

- Is the root of a shortest path multicast tree to all other bridges which uses PBB encapsulation. This uses one Multicast PBB MAC Address and one VLAN ID (VID).
- Is also the source of a Point to point PBB-TE path to every destination that followed the shortest path multicast tree but uses a single VID+DA allocated from BEB"A" for every leaf. When the paths are computed for every possible bridge this uses one PBB-TE MAC address per destination and one VID.
- The fact that these provider addresses are only visible in the domain of the provider allows for these B-MAC addresses to be allocated out of the local address space. A simple policy can be set up to choose addresses from a Base address that can be uniquely assigned

Design Characteristics

The fundamental tenets of this system are simple efficient link state computed shortest path Spanning tree combined with shortest path unicast. The following sections detail proposals of how to make this operationally practical.

Loop Prevention and Loop Mitigation

The converged link state topology allows distributed computation of loop free multicast and unicast paths. Computing trees in a distributed routing system requires a mechanism to address transient loops, ideally in as non-intrusive a manner as possible. Transient loops may be created when the set of bridges in a PBBN have an unsynchronized view of topology during the period of time that link state advertisements propagate and the subsequent calculations are performed. Three things may happen in forwarding, non-optimal paths, broken paths and/or temporary loops may form. Interestingly there is a tradeoff between purposely interrupting forwarding or trying to maintain forwarding during convergence. Loops should be prevented or if they cannot be prevented there must be a mechanism to mitigate the effect of the loops.

Fortunately the combination of bi-directional symmetry and unicast/multicast path congruency between any two network elements (NE's) in the network means that a consistent FIB entry holds sufficient information to detect and mitigate looping frames. Note that having this information is not new, but having this information populated by link state database allows greater accuracy for equal cost paths. This combination is unique to Ethernet networks and holds true for the PBBN.

Due to the symmetry, a network a frame **from** a given B-MAC address will arrive on an interface which is also on the shortest path **to** that same given B-MAC address. The information in the forwarding database can be used to determine if this condition is true. Therefore, any frame from a given B-MAC address arriving on an unexpected interface is an indication of a problem and potentially a loop.

Using a similar mechanism employed for Ethernet source learning, a “reverse path forwarding check” (RPFC) can be created. RPFC audits the incoming frame’s source MAC address at the port of arrival. The FIB contains unique entries for unicast and group multicast MAC addresses which permits distinct treatment to be applied to each type of traffic. When RPFC is enabled for a class of traffic (unicast or multicast), frames arriving on an unexpected interface are silently discarded. When the multicast tree from a BEB corresponds to or is a strict subset of the unicast tree to a BEB we have achieved the equivalent of the Extended Reverse Path Forwarding model [Metcalfe].

For PLSB unicast forwarding, a transient loop is not strictly speaking catastrophic. Frames may be temporarily buffered in a loop or silently discarded. This suggests that RPFC could safely be disabled on point to point links during periods of network change and re-enabled after some period of convergence. One reason for disabling the RPFC is to prevent it aggressively discarding frames making the tradeoff mentioned earlier. RPFC could also be used periodically as a sanity check on unicast paths.

For multicast forwarding, a transient loop could result in unbounded replication, a situation to be avoided. Therefore, RPFC should be enabled at all times for the processing of frames with multicast group addresses. Multipoint or hubbed segments are a special case where both unicast and multicast traffic are replicated. RPFC is enabled for both unicast and multicast destination addressed frames in this case to avoid inadvertent generation of multiple copies of a frame. In this case, RPFC will prune the traffic such that only the frame on the shortest path between the BEBs will not be filtered.

One advantage of link state topology creating shortest path trees over the use of Spanning tree protocols creating minimal spanning trees, is that frequently a network change will not modify the path taken by a given shortest path tree. Therefore forwarding along that shortest path tree continues uninterrupted.

Optimal Multicast Service Trees

By default, we created a shortest path multicast tree from every source node to every destination node. While we computed the shortest path tree to all bridges, we have not populated the FIB along the paths. A default Multicast address (B-MMAC) that is derived uniquely, is computed as part of the Provider Local address space. One possibility, to derive a unique B-MMAC, is to make it a function of a unique node identifier such as each root bridge’s IS-IS identifier. These multicast addresses are installed in the FIB as the equivalent of a common spanning tree to talk to all nodes.

We also are required to populate the partial sub trees which include just the BEBs needed to support each specific multicast service. The I-SID can be used for this purpose in the following manner. The I-SID designates a "service" community of interest for any set of edge ports. A multicast service is the set of all customer ports that support that I-SID. We use the link state control plane to distribute this community of interest of in the form of all the sets I-SIDs. On any node when computing multicast trees, in order to create a

sub tree, a Multicast address (B-MMAC) that is a function of the I-SID value and the root bridge IS-IS identifier is computed out of the Provider Local address space. This allows each group to have a unique unambiguous multicast address. Since all bridges have computed the shortest path trees for all members it is a simple matter to populate the FIB for the set of destination multicast MAC addresses that are required to support the I-SID. In order to send a multicast address to the group the root bridge encapsulates the PBB frame using the B-MMAC corresponding to the I-SID. While I-SID drives the community the forwarding paradigm is still based on VID and B-DA. In this manner only bridges involved in the forwarding of traffic for a service will ever see traffic for that service. Also another advantage is there is no “signaling” of B-MMACs since all computations can be performed by local calculations on the distributed IS-IS link state data.

Equal Cost Paths

One issue is that when these algorithms are applied to mesh networks, there may be multiple paths with equal cost. The computation of trees is symmetric. The algorithms are deterministic and repeatable. In the case of multiple equal cost paths, multiple trees may be computed. The trees are distinguished by using different B-VIDs for each "topology". In order to be repeatable a unique tie breaker is chosen for each tree. Typically a small number of B-VIDs would satisfy most cases of networks equal cost routes.

PLSB impacts on Scalability

PLSB uses the IS-IS link state protocol to create unicast any to any forwarding and Multicast trees. IS-IS, like other link state protocols, typically scales well up to several 100s of nodes. Note this number relates only to the number Backbone Edge and Core Bridges in any one domain. Therefore, PLSB based on IS-IS does not pose a scalability issue. Scalability of the solution is primarily determined by the amount of forwarding capacity, port fan-out and forwarding memory of the individual provider bridges. Scalability and convergence time is improved by PLSB over other types of solutions.

Performance under failure scenarios

PLSB ends up with unique properties with respect to failure scenarios. It does not use port blocking but relies on instead on RPFC. RPFC is selectively applied to multicast traffic and broadcast segments on a frame by frame basis. RPFC may be applied to unicast traffic as well as a operations policy. The net result is that under failure scenarios, no disruption occurs for those paths unaffected by the failure, and re-converging multicast forwarding is minimally impacted.

PLSB is built on PBB and supports the broadcast of PBB MAC addresses rather than relying on learning. This relies on Link state resiliency within the PBB network cloud. When a customer end system changes the attachment point to the network, its MAC information is simply relearned at the new attachment point via normal MAC learning.

Network changes within the B-MAC layer do not affect the C-MAC to B-MAC bindings at the edge. A failure in the Backbone core network does not affect customer MAC tables. Solutions for 802.1ad to 802.1ah interconnect will equally apply to PLSB.

Constructing ELAN and ETREE

ELAN and ETREE are similar services. The difference between ELAN and ETREE is one of connectivity policy. In the application of link state in PLSB the implementation of this policy can be applied to how Ethernet clients flood and learn B-MAC forwarding. This means policy only need be applied to per service backbone multicast connectivity, and unicast connectivity can be shared across all services.

The implementation of this policy is straightforward in PLSB. Two attributes are associated with service advertisements in the IS-IS routing system. When a node is configured to participate in a service it can be a source, a sink or both.

For ELAN service, all sites are simply set to "both source and sink", so all devices observe flooded traffic and will populate their forwarding tables correctly.

The ETREE service is implemented using two I-SIDs; one identifies the source community and the other identifies the sink community. The result is intra community connectivity (i.e. sink to sink or source to source) is obstructed (cannot be learned) and only inter-community connectivity is possible (i.e. source to sink or sink to source). A common application of this is hub a spoke (branch office/head office). This results in an efficient and optimum distribution of traffic.

Conclusions

PLSB builds upon a number of recent developments in 802 standards to provide shortest path forwarding in the PBBN to complement the scalability improvements embodied in PBB (802.1ah).

This results in a substantially more scalable Provider Ethernet solution. When PBBN, PLSB and PBB-TE are combined, this offers a comprehensive and optimal solution to supporting all manner of Ethernet services, and numerous options for interconnecting legacy L2 and L3 devices over a common Ethernet based network.

Acknowledgements

The authors would like to thank Dave Allan, Peter Ashwood-Smith, Nigel Bragg and Paul Unbehagen for significant input and valuable comments.

References:

[PBB] Paul Bottorff, Steve Haddock, editors, “IEEE 802.1ah - Provider Backbone Bridges”, Draft 3.3, December 2006, work in progress.

[IS-IS] "Intermediate System to Intermediate System Intra-Domain Routeing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)", ISO DP 10589, Second Edition, November 15, 2002.

[SPB] Tony Jeffree, editor, “IEEE 802.1aq – Shortest Path Bridging”, Draft 0.3, May 9 2006, work in progress.

[Metcalfe] Yogen K. Dalal and Robert M. Metcalfe, "Reverse Path Forwarding of Broadcast Packets", Xerox Corporation and Stanford University, Communications of the ACM , Vol 21, 12, (December 1978) 1040-1048.

[TRILL] See http://www.ietf.org/html.charters/trill-charter.html

[Mesh] 802.11 Working Group of the LAN/MAN Committee, “IEEE P802.11s™/D1.00 Draft Amendment to Standard for Information Technology - Telecommunications and Information Exchange Between Systems -LAN/MAN Specific Requirements - Part 11: Wireless Medium Access Control (MAC) and physical layer (PHY) specifications: Amendment: ESS Mesh Networking”,

Authors

Don Fedyk, dwfedyk@nortel.com
Paul Bottorff, pbottorf@nortel.com