Provider Link State Bridging

David Allan, Peter Ashwood-Smith, Nigel Bragg, and Don Fedyk, Nortel

ABSTRACT

Wide area communications technology has been challenged to virtualize large numbers of Ethernet LAN segments. This is a consequence of a mismatch between the broadcast nature of the LAN segment and the extremely constrained connectivity implied by the p2p connections or tunnels available in the WAN environment, which have been combined to create virtual LAN segments.

PBB-TE has been a practical demonstration of how filtering applied to a broadcast media can result in a connection. This article introduces provider link state bridging (PLSB), which adds a control plane to the PBB data plane in order to extend the techniques for manipulation of Ethernet bridges for unicast paths pioneered by PBB-TE. PLSB solves the problem of large-scale virtualization of LAN segments over an Ethernet infrastructure by applying advances in computation performance to the multicast capabilities inherent in the Ethernet data plane. The result is that the fundamental primitives of connectivity today, the broadcast LAN segment and the connection, can be virtualized in a scalable manner on a common (but substantially larger and better utilized) Ethernet-based infrastructure.

INTRODUCTION

Since its origins in the 1970s, Ethernet has established and grown its now ubiquitous position in both enterprise and service provider networks, while simultaneously becoming the dominant layer 2 for all flavors of WiFi and home networks. This is the consequence of a virtuous cycle of adoption and commoditization. The ubiquity of Ethernet as the network infrastructure of choice for enterprises, the physical layer of choice for service providers, and its introduction as a new provider service offering have continued to drive Ethernet's dominant position in the telecom industry today. Ethernet continues to thrive, gain functionality, and evolve while staying true to many of its design principles. Most recently there has been a sustained and significant effort in the standardization of features to enhance the fit of Ethernet to carrier networks.

The initial IEEE response to the increase in interest of carriers in Ethernet as a provider technology was provider bridging or 802.1ad. The basis of operation was the insertion of a provider tag field into a frame that was already customer tagged. This provided for the coexistence of customer and provider administered tags. 802.1ad was comparatively simple reuse of what had gone before as the basic relay function itself was not modified (only a single tag of the tag stack used in any forwarding domain), but the approach largely inherited enterprise-scale limitations.

Moving beyond these limitations while holding true to the values of reuse and backward compatibility has required two comparatively recent developments that now figure prominently in the evolution of Ethernet.

The first is 802.1ah, provider backbone bridging (PBB) [1]. PBB saw a complete recursion of medium access control (MAC) headers such that customer MAC frames were fully encapsulated in provider frames, and the I-component was introduced. The I-component provided a 24-bit service tag that could be inferred from the customer facing port (port user-network interface, UNI) or customer tag information (tagged UNI). PBB had the effect of radically reducing the number of different MAC addresses known by core bridges, providing for a substantially enlarged service tag field and putting the operator fully in control of all information on which core bridges switched while simultaneously eliminating any requirement for core bridges to examine or otherwise act on the customer MAC.

What PBB did was to completely separate the customer and provider MAC spaces. This not only afforded network operators control over their MAC space and scalability advantages, but it created a situation whereby all bridge controls could be exposed and manipulated. This led to PBB with traffic engineering (PBB-TE) or 802.1Qay.

PBB-TE is a profile of Ethernet behavior whereby provider MAC bridge tables are configured instead of learned. PBB put operators in control of which endpoint identifiers appear in switching tables, and PBB-TE focuses on how those switching tables are used. In that regard it is a comparatively small change to the Ethernet data plane whereby the normal procedure of flooding of frames for which the destination is unknown (and therefore must be learned) is disabled. If the destination MAC address has not already been configured in the filtering database (FDB), as dictated by tools with additional intelligence, the frame has been received in error and should not be forwarded since learning is neither required nor desirable.

PBB-TE produced a highly scalable point-to-point (P2P) connectivity model (for ELINE ser-

0163-6804/08/$25.00 © 2008 IEEE

vices [2]), with operational attributes analogous to synchronous optical nework/digital hierarchy (SONET/SDH) and resiliency being purely a data plane function. While PBB-TE is able to support point-to-multipoint (P2MP) constructs (for ELAN and ETREE services) , efficient and fast resiliency for multicast is hard to achieve with a simple provisioned or even signaled model.

What PBB very importantly demonstrated was how fully featured the Ethernet networking data plane had become, the salient features being:

- VLANs as complete virtual network, providing the ability to mesh the network some 4000 times
- Comprehensive east-west operations, administration, and maintenance (OAM) toolset in the form of 802.1ag fault management and Y.1731 performance and alarm management
- Full unicast, multicast, and broadcast capability
- Tree-based multicast and broadcast replication (as opposed to head-end replication)
- Recursion and layering to enhance scaling
- A data plane fully described by network-global parameters (both addresses and service identifiers), which, when a routing system is applied to it, obviates the need for a separate signaling protocol, as there is no link-local state to configure

For multipoint networks, Ethernet suffers today when compared with the state of the art in link utilization and speed of recovery because of its use of the various spanning tree protocols (STP). Spanning tree was designed in an era where frugality in computing power and state was a virtue. STP and its variants minimize both via the use of message exchange of bridge protocol data units (BPDUs) to converge the network. The advent of PBB has suggested it is time to revisit the compute state-resilience trade-off for Ethernet.

Provider link state bridging (PLSB) is a control plane development leveraging aspects of PBB and PBB-TE. PLSB is currently being proposed for standardization to the IEEE 802 community as part of the 802.1aq (Shortest Path Bridging) project. This project addresses a truly carrier-scale ELAN infrastructure, with a target of 1000 bridges. PLSB enhances PBB with the addition of a layer 2 routing system and loop mitigation in order to produce a better spanning tree for PBB, "better" in the sense that all trees are shortest path, and under fault the only traffic affected is that traversing the failed path. The rest of this article explores the technical details.

Ethernet Shortest Path Bridging

For many years the only networking of Ethernet has been bridging on spanning trees. Spanning trees have the property of confining unicast traffic and multicast traffic to a simple spanning tree. A spanning tree has properties that are true to the roots of Ethernet where a shared infrastructure enables plug-and-play operation of Ethernet.

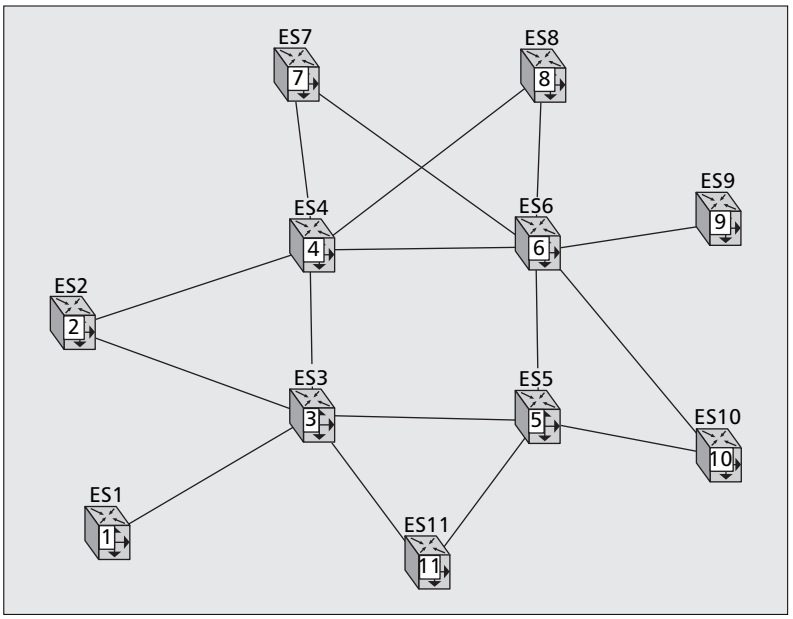

Figure 1. *Sample topology.*

With the development of PBB and PBB-TE [3], several emerging drivers for control plane enhancement arrived. First, the development of a provider address space that is administratively separate from a customer address space allows independence from the customer control plane for the provider network. Next, the large Ethernet switches needed for carrier requirements can afford a more elaborate control plane (note that many larger "Ethernet" switches already support an IP control plane based on link state algorithms). Furthermore, the compute performance delivered by typical embedded processors has increased by orders of magnitude since the initial deployment of both spanning tree protocol and the first IP control planes. Finally, PBB-TE illustrated a break from the conventional use of bridges, VLANs, and the associated spanning tree control planes.

This was the background for the birth of provider link state bridging (PLSB). PLSB combines Intermediate System to Intermediate System (IS-IS) [4], PBB, PBB-TE, and a data plane enhancement in the form of a data plane ingress check [6] to produce an Ethernet mesh solution with a layer 2 distributed control plane for a provider backbone bridged network (PBBN). PLSB is being proposed into 802.1aq SPB as shortest path backbone bridging (SPBB) [5]. The standards work on this is now underway; it is important to emphasize that it is mainly due to the provider scope of PBB that the drivers for a layer 2 link state control plane for Ethernet have become significant.

PBB utilizes backbone VLANs (B-VLANs) as a provider service topology. However, other than defining a B-VLAN, PBB does not define the VLAN attributes. A B-VLAN may be any form of spanning tree. Figure 1 illustrates a sample topology and Fig. 2 a variety of spanning trees on that topology.

Spanning trees enable both unicast and multicast operation. PLSB creates a complete set of shortest path trees, one rooted on every bridge

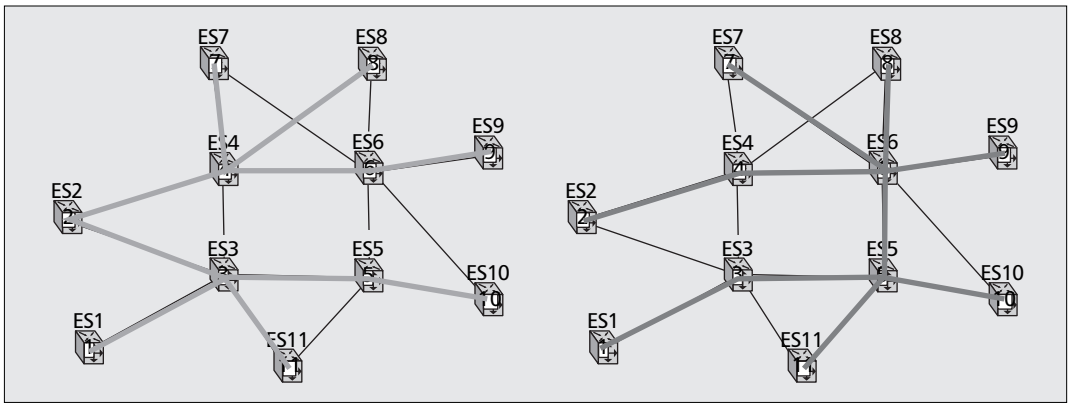

Figure 2. *Sample spanning trees (rooted on ES2 and ES5, respectively).*

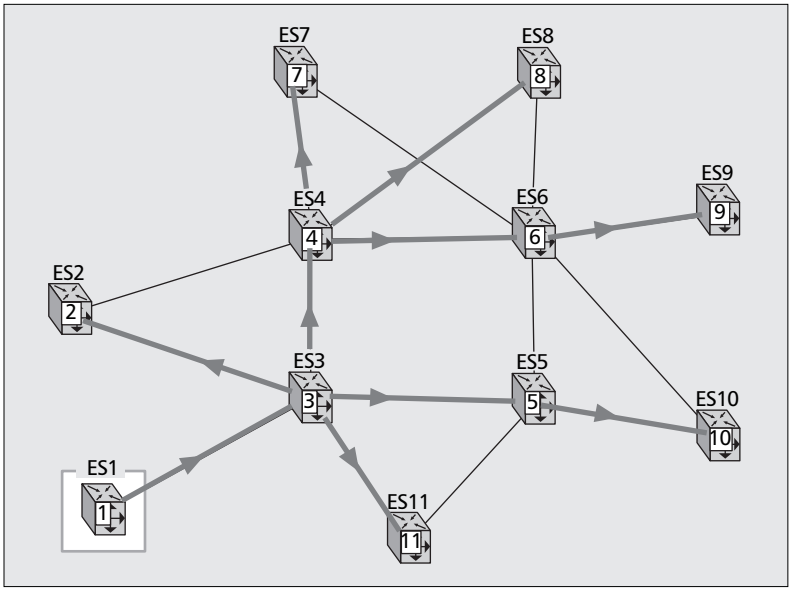

Figure 3. *Shortest path tree from ES1 to all neighbors.*

in the network, all within a single B-VLAN. In order to retain the desired multicast properties, the traditional client bridging functions of flooding and learning onto a PBBN are mapped onto functions performing source-specific multicast implemented at the B-MAC layer of the PBBN. One result is that, similar to PBB-TE, PLSB can fully mesh the network once per B-VID, and this property is exploited for the purposes of edge-based load spreading across multiple paths, each instantiated in a separate B-VID.

PBB defines the necessary constructs and behaviors in the I-component to B-component mappings performed at the edge of the PBBN. PLSB manages the population of the B-component FDB and creates a source-specific multicast B-MAC for the corresponding I-component. PLSB uses source-specific shortest path trees for multicast; therefore, the destination multicast address is required to encode (Source, Group) instead of simply (Arbitrary Source, Group).

PLSB uses standard IS-IS [4] layer 2 processes for network topology discovery and database synchronization. The I-SID to B-MAC configuration of each bridge is also flooded using IS-IS. The algorithm for computing shortest path trees and populating the FDB is then computed locally in its entirety, with no further message exchange required. Although the algorithm is computationally more intensive than traditional layer 3 algorithms, the target size of the provider network combined with modern CPU design make this simple approach highly practical. A description suffices to describe the behavior.

1 Given the current network topology and state reflected in the local IS-IS database, a bridge will first calculate the shortest path from itself to all other bridges. This defines the route the bridge will use for all traffic it itself originates to remote bridges, both unicast and multicast. The unicast routes are also used for loop mitigation, as described later. This computation is used to populate the FDB with nodal unicast B-MACs.

2 Then, to determine its role in forwarding transit traffic, a bridge will compute the shortest path tree seen by every other bridge to determine whether it lies on the shortest path route between any two pairs of bridges.

For each pair on which the bridge lies, the shortest path between it will next determine the intersection of the set of I-SIDs associated with that pair. For each I-SID intersection, the appropriate multicast B-MACs are populated in the FDB. Note that multicast B-MACs are algorithmically generated, leveraging the information of source bridges, I-SIDs, and the B-MACs associated with those I-SIDs.

The result is a fully populated unicast address set between all bridges and per edge/per service multicast trees. The multicast FDB state installed for each I-SID is congruent with the shortest path tree from each root bridge. What this means is when visualizing the resulting connectivity, the complete shortest path tree for each bridge is the prototype for any multicast forwarding originating from that bridge. Figures 3 and 4 illustrate this.

The I-SID-specific multicast trees that originate from the same bridge will be a subset of the prototype, with the pruning of uninterested endpoints a consequence of local computation and FDB population with no need for additional internodal messaging. Since the unicast traffic is populated on these same paths, and the trees are forward and reverse congruent, the unicast traf-

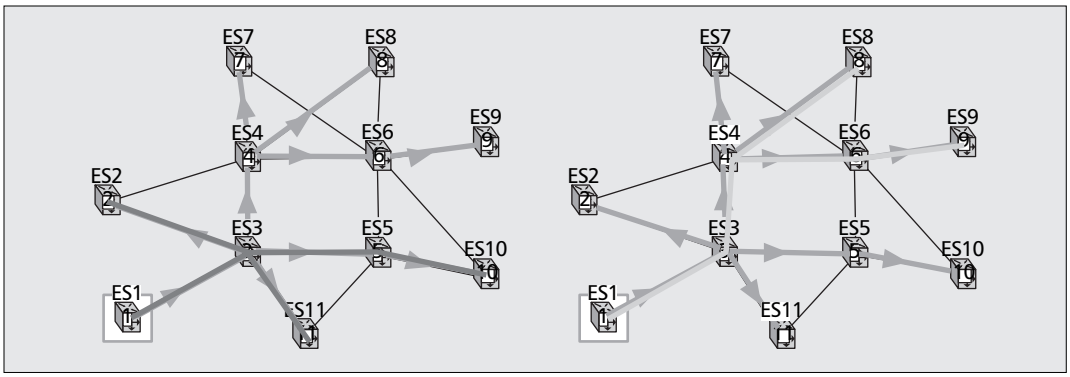

Figure 4. *Multicast trees for ES1 that are perfect subsets of the shortest path tree.*

fic follows these same paths. One additional property is that the unicast forwarding can use common destination B-MAC FDB entries for traffic from all sources, so the unicast forwarding inherits the state efficiency of conventional bridging.

IS-IS [4] is proposed as the preferred layer 2 link state protocol and augmented to carry per backbone edge bridge (BEB) service information. As the service identifiers (I-SIDs) have administrative domain-wide significance, they can be directly associated with the service endpoint B-MACs, and are used by the control plane to compute and directly populate the FDB with efficient per-service per-source multicast trees.

For the PLSB computation, Floyd's "all pairs" vs. the ($n \times$ Dijkstra) algorithm was initially considered (where n is the number of bridges). Ultimately the multiple ($n \times$ Dijkstra) algorithm prevailed due to the reduced complexity. This is because the Dijkstra computation at any given time maintains state in proportion to the circumference of the circle defined by the root under consideration rather than the area, and this gives it a computational edge since time spent traversing the state is diminished accordingly.

It might appear that PLSB is being somewhat intensive in its use of compute resources. It should be recalled that IS-IS in its IP application was first deployed approximately 15 years ago, and since then available embedded compute power has increased by about two orders of magnitude. Our assertion can of course easily be tested on any modern PC/processor.

Loop Prevention and Loop Mitigation

Ethernet forwarding loops are a thing to be avoided. The combination of extremely efficient layer 2 multicast forwarding (tandem replication) and a loop means the effect of a single looping frame may potentially be magnified to the point where it can significantly disrupt a network, especially because in an Ethernet environment the loop time can be only a few hundred microseconds.

In IEEE 802.1aq (SPB), the task group is evaluating alternatives between loop prevention and loop mitigation. PLSB perfectly emulates bridged Ethernet's properties of bidirectional symmetry and unicast/multicast path congruency between any two network elements (NEs) in the network (see "Tie Breaking" below); this means the FDB in a bridge is also able to detect and mitigate loops on a frame by frame basis, as follows.

In a steady state network a frame from a given source B-MAC address in a given B-VLAN will arrive on an interface which is also on the shortest path to that source B-MAC address, and an entry will have been populated into the FDB to indicate this. A frame from a given source B-MAC address arriving on an unexpected interface is an indication of unconverged forwarding, potentially resulting in a loop, and should therefore be discarded.

A simple modification to Ethernet source learning, termed a data plane ingress check [6], is required to simply audit the port of arrival for a given frame with the expected port for the frame's source address as established by the routing system. In terms of implementation, this is equivalent to using the initial "unknown ?" test of the learn/store function to drive discard and can be applied at the ingress to a bridge.

When data plane ingress check is enabled, frames arriving on an unexpected interface (noted as a discrepancy between port of arrival and port in the FDB) are silently discarded by this mechanism. This is an aggressive policy as there may be nonlooping frames arriving on an unexpected interface. However, this does have the benefit of requiring no modifications to the Ethernet PDU and minimal modifications to the implementation of a bridge. This also plays prominently when considering backward compatibility.

The combination of data plane ingress check and consistent tie breaking is not a complete solution to the prevention of looping frames but brings a number of desirable properties to the table:

- For unicast and multicast traffic, ingress checking is performed on a frame-by-frame basis. This does not require unicast port blocking or interruption of connectivity unaffected by topology changes. For multicast, additional prevention techniques are deployed.
- Multicast forwarding is converted to a directed tree.
- A fault or implementation problem on any single bridge cannot produce a loop.

Ethernet forwarding loops are a thing to be avoided. The combination of extremely efficient layer 2 multicast forwarding and a loop means the effect of a single looping frame may potentially be magnified to the point where it can significantly disrupt a network.

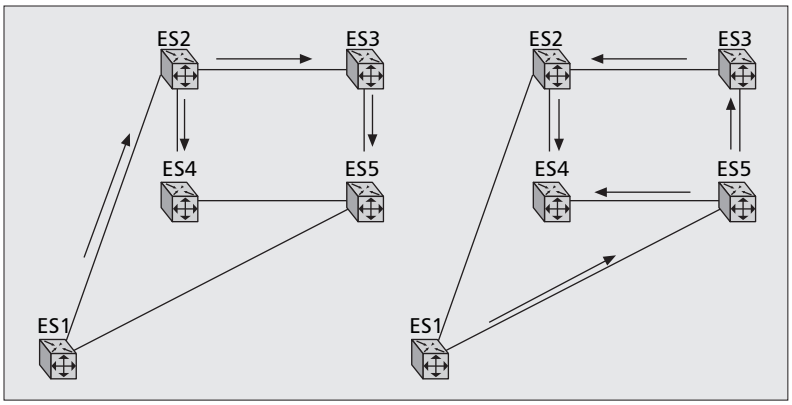

Figure 5. *Consistent tie breaking prevents cycle formation.*

- If a loop did form for whatever reason, the loop's only valid source is itself; it will not admit further traffic, so an exponential increase in looping traffic cannot occur. This is because the ingress check ensures that a node has only one valid ingress for either unicast or multicast traffic, so if a loop has formed, traffic from outside the loop cannot leak into it.

The data plane ingress check and consistent tie breaking augmented with a local neighbor synchronization mechanism for multicast FDB changes offer authoritative protection against the formation of transient loops. These can be either an artifact of network operations or due to faults and implementation issues.

Although it can be demonstrated that no single link metric change in isolation can produce a loop when the data plane ingress check is implemented, examples can be found where the combination of two topology changes and the resultant unsynchronized views of topology can cause a loop. To mitigate this for multicast traffic, the PLSB strategy is to perform the unicast tree computation immediately when a topology change is received and block (delete from the FDB) those multicast trees for which the route to the source has changed. Restoration of the full FDB then awaits topology synchronization with neighbors; meanwhile, traffic unaffected by the topology change is forwarded without interruption.

Tie Breaking

A key component of the routing system and the overall loop-free robustness of the network was ensuring consistent resolution of tie breaking in the presence of equal cost paths.

The fundamental loop avoidance technique is the data plane ingress check (see above), which requires that the port through which a frame enters a bridge is the same port as would be used to reach the source of the frame. At the network level, this translates into the requirement that forward and reverse paths between any pair of bridges must follow exactly the same path. For consistent behavior, PLSB required a tie breaking technique that produces the same decision when executed at every bridge in the network.

Analysis of the behavior of the data plane ingress check also suggested that the presence of equal cost paths, and correct and consistent symmetric resolution of them was a key component in overall network robustness under transient loss of synchronization caused by topology changes.

In the example of Fig. 5, tie breaking will always resolve the path between ES2 and ES5 as being via ES3. If this were not true, a single topology change (which made the shortest path to ES1 via ES5 instead of ES2) could activate an ES5-ES2 path via ES4, forming the loop ES2-ES3-ES5-ES4-ES2.

The tie breaking algorithm proposed for PLSB is deterministic and symmetric. The actual algorithm employed can easily be described as picking the path with the lowest *path identifier*, where this path identifier is defined as a sorted list of the bridge *identifiers* forming the path. This is based on the IS-IS sys-id, and is independent of any port B-MACs advertised by the BEB, all of which share the set of equal cost paths rooted on the BEB. If port B-MACs were used as identifiers, computation complexity could increase from O(Nodes) to O(ports)

This resulted in a tie breaking algorithm with the interesting property that any segment of a shortest path was also the shortest path between the segment endpoints. This in turn had the highly desirable property of minimizing the amount of state maintained at interim steps in the path computation (essentially allowing on-the-fly tie breaking at very high speeds).

Equal Cost Multipath Trees

One artifact of the tie breaking algorithm is that the bookends of the path ranking scheme (lowest path identifier and highest path identifier), while not authoritatively diverse, did have a significant amount of diversity. This can be exploited to improve overall network efficiency. If both the lowest path identifier and the highest path identifier tie breaking paths were retained as output of the all pairs path computation, the result was two complete and significantly diverse shortest path trees that individually preserved symmetrical congruency and could be generated from a single set of link metrics. Each can now be assigned to a different B-VID and the FDBs populated accordingly (Fig. 6).

The result is that each B-VID could be viewed as a virtual multipoint link in a LAG. Then either flow-based (implicit) statistical load spreading can be applied at the edge, or explicit I-SID to B-VID assignment could be utilized. The implicit load spreading offers efficiencies when only a very small number of communities of interest are supported by the network, while explicit load spreading has more desirable properties when large numbers of communities of interest are supported, because of the OAM benefits of deterministic assignment of service to B-VID.

A topic of research is additional multipath/tie breaking mechanisms with an eye to even greater load spreading without resorting to the configuration of additional per link metrics.

It is possible to consider exploiting edge-based multipath to offer service-level call admis-

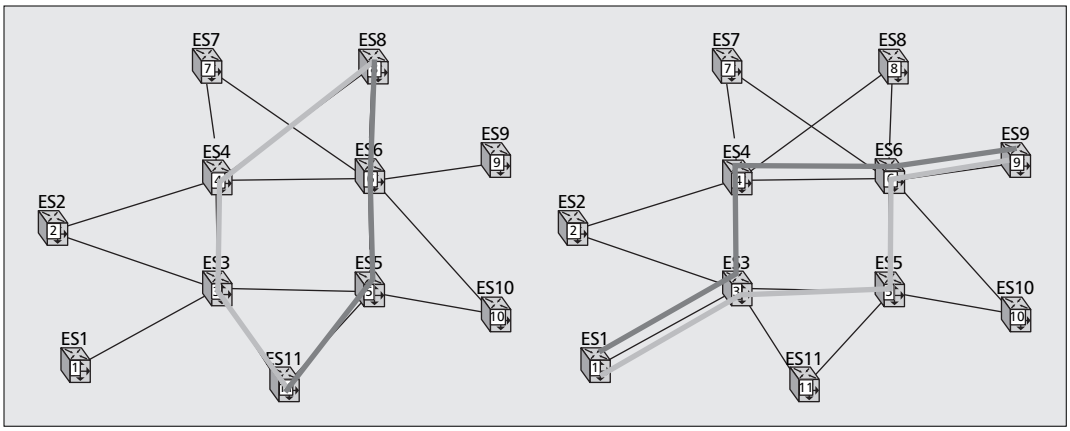

Figure 6. *Example equal cost paths: ES8-ES11 and ES1-ES9.*

sion control (CAC) in the network when explicit I-SID to B-VID assignment is used. An example would be to track the offered load on each multipath variation and do a best fit assignment of the customer traffic matrix to a multipath B-VID for new service uptake.

Resilience

PLSB inherits the resilience properties of a routed system. Its key virtue is that the only internodal system or subsystem that requires synchronization is the routing system. All of the required convergence elements for complete FDB generation are derived from nodal computation applied to the IS-IS database. The primary beneficiary is multicast convergence, which does not require any subsequent signaling, pruning, or other forms of care and feeding after the routing system has converged and the FDB updates have been installed.

Recently proposed is the use of multicast in the control plane. This is not necessarily a new concept but one that comes naturally to a technology that robustly instantiates large numbers of virtual broadcast domains. This will largely eliminate hop-by-hop control plane propagation of link state packets (LSPs) from the overall recovery budget. When a failure occurs local to a node, it is permitted to use an I-SID delegated to the control plane for the initial flood of the LSP advertising the change. The LSP will only suffer data plane latency as it is multicast to all nodes on the portions of the I-SID multicast tree unaffected by the failure. Normal IS-IS replication procedures then take over to provide for completeness and robustness. The net result is that network convergence times will become independent of network diameter and approach nodal convergence times.

ETREE and ELINE

Ethernet bridging in the form of a PBBN has an interesting property in that the scoping of flooding and learning dictates customer connectivity. This is directly used to scope membership for a given community of interest (i.e., broadcast domain) for ELAN services but conceptually it can be extended further.

PLSB permits multicast attributes to be associated with I-SID advertisements in IS-IS. These attributes are *source* and *sink* encoded as two Boolean flags. For ELAN, all clients for a given community of interest are both source and sink. They will originate and terminate flooding, and therefore client layer bridging can "learn" connectivity.

ELINE is the simple degenerate case; no multicast/flooding or learning is required of a P2P service, so no multicast is associated with an ELINE I-SID. Both attribute bits are turned off.

For ETREE, the community of interest can be divided into two behaviors denoted by group membership. The membership is designated by I-SIDs, and PLSB shares learned client MAC information between the two I-SIDs in a common I-component bridge. The use of multicast attributes makes it comparatively simple to instantiate an ETREE with multiple root UNIs [2].

On the first I-SID, the leaf behavior is source and the root behavior is sink. On the second I-SID, the root behavior is source and sink, and the leaf behavior is sink. The result is a "split horizon" in the leaf bridging function. The leaves only learn about, and therefore can only communicate with, the roots. The roots see both the set of leaves and the other roots, and can communicate with both leaves and roots. This is a useful feature for applications such as broadband aggregation and backhaul. Broadband customers have layer 2 access to broadband network gateways, but cannot learn about or communicate with each other. It is easy to envision other variations of combining multicast attributes.

Scalability

PLSB builds on the scaling enhancements embodied in PBB. The addition of the I-component means the data plane supports some 2^{24} service instances in a PBBN, and although not commonly implemented, the ability to nest 802.1ad S-tags inside the B-MAC header in theory extends this to 2^{36} when the concept of tagged and port UNIs is carried forward into 802.1ah.

PLSB, like PBB, only operates on the B-MAC layer. Client MAC address tables are therefore confined to the I-components around the edge of the network, yielding a huge gain in

PLSB, like PBB, only operates on the B-MAC layer. Client MAC address tables are therefore confined to the I-components around the edge of the network yielding a huge gain in scale. The industry interest in using PBB to "front-end" other Ethernet emulation technologies attests to the scaling value this brings.

PLSB multicast follows a logical tree structure which follows exactly the physical topology of the network, unlike proposed logical MPLS ring structures for H-VPLS which are by no means optimal when compared to the physical topology.

scale. The industry interest in using PBB to "front-end" other Ethernet emulation technologies attests to the scaling value this brings.

PLSB does require the use of per-service source-specific multicast trees whose identifiers are encoded in the multicast DA. The actual MAC encoding uses the *local* bit such that the OUI information is available to encode the source or root, and the lower 24 bits encode the I-SID directly. This imposes a theoretical limit of 2^{20} bridges in a PBBN. It should be noted that the root (multicast address) is decoupled from the unicast source address, a design property that can be exploited in the future.

The use of per-service source-specific trees does increase the number of multicast MAC addresses used in the network.[1] Individual addresses tend to be sparse in the scenario where the network supports large numbers of small communities of interest. A valuable consequence of the "all pairs" computation is that bridges which implement multiple FDBs (a common architecture on larger switches where each line card has its own forwarding tables) have sufficient information to personalize and substantially prune the contents of each FDB, because any specific multicast address need only be installed on the interface through which a frame from that source may legitimately enter.

Unlike architectures such as H-VPLS, PLSB does not employ tandem learning points (with attendant customer MAC scaling issues) to increase multicast efficiency. PLSB leverages PBB encapsulation to hide customers' MAC addresses service edge to service edge, and never has to look at customer MAC information for any reason at tandem points. PLSB meets the idealized criteria of multicast efficiency whereby only one copy of a frame traverses any given link, all frames take a shortest path, and no copies are delivered to uninterested recipients, while being oblivious to the contents of the customer MAC.

PLSB multicast follows a logical tree structure that exactly follows the physical topology of the network, unlike proposed logical MPLS ring structures for H-VPLS, which are by no means optimal compared to the physical topology.

Operational Simplicity

Ethernet's popularity is in no small part a result of the light to zero configuration demands of most devices. PLSB continues this trend by employing the IS-IS layer 2 link state protocol, since IS-IS requires no additional L3 addressing/encapsulation and resultant configuration. IS-IS is also a general-purpose, well designed, and thoroughly field-proven link state protocol that is easily extensible. A well designed PLSB implementation with intelligent defaults should require only I-component configuration at the BEBs, while the rest of the network "takes care of itself."

With respect to the I-component configuration, PLSB combines all the functions of automation of VPN single-touch provisioning, per service connectivity establishment, and endpoint discovery into the IS-IS protocol. The normal mode of operation is that a UNI port is associated with an I-SID, via either configuration or user authentication and registration. IS-IS floods the I-SID information to the network peers, and when the network converges, the new UNI port has been grafted onto the connectivity meshing the existing I-SID endpoints without having touched any of the other I-SID endpoints.

Backwards Compatibility

The IEEE always mandates backward compatibility in standardization projects. PLSB has a number of backward-compatible aspects that can be considered in both the concrete and abstract senses:

- PLSB, since it leverages PBB, does not involve any changes to Ethernet frame formats, addressing, or VLAN semantics.
- A PLSB UNI is backward compatible with 802.1ah, which again is backward compatible with 802.1ad and 802.1Q.
- Similar to a spanning tree, in a given PLSB B-VID there is only a single symmetric shortest path from any source to any destination for both unicast and multicast.
- The use of the IS-IS protocol will permit backward control compatibility for many years to come.

Furthermore, Ethernet networks today are constructed of physical symmetric links, and when virtualizing Ethernet as PLSB does, it is very desirable to preserve the behavior and properties as much as possible.

PLSB for PBB preserves what is termed "the symmetric congruency of Ethernet forwarding." What this means is that both unicast and multicast frames follow the same path in both directions. If the traffic on a link on a PLSB network is examined, the appearance of the traffic transiting that link is the same as that of any other Ethernet network. Traffic from a given source will be observed going in one direction, traffic to that source will be observed going in the other direction. In the context of PBB this will be true of both observed B-MAC and encapsulated C-MAC headers.

A number of good things ensue from preserving this property, which is completely aligned with PBB:

- No race conditions or misordering between the flooding of unknown C-MACs and the forwarding of frames on learned C-MACs paths.
- Customer layer OAM frames properly fate share with the connectivity traversed by customer traffic at the B-MAC layer.
- The B-MAC layer itself is a closed and complete system from the perspective of OAM. For transactions between maintenance endpoints (MEPs) and transactions terminating at maintenance intermediate points (MIPs), symmetrical connectivity will exist in both directions in a properly functioning network. Therefore, bidirectional OAM transactions have a return path.
- The likelihood of asymmetric failures (which impact client STP convergence) is minimized as the components in common in both directions are maximized compared to unique routing of paths in each direction.

PLSB for PBB also has a number of properties that are ultimately complementary with

[1] *Aggregate trees were considered and found to introduce significant complexity for an insignificant reduction in overall state.*

other transparent LAN services (TLS) solutions, and this permits PLSB to overlay networks that offer TLS (e.g., VPLS):

- TLS services appear to PLSB as switched LAN segments that employ flooding and learning; the IS-IS control plane has existing models of such structures. PLSB uses active filtering dictated by the control plane's knowledge of topology. When learning encounters a filtering boundary, such as when PLSB is interconnected with QinQ [7] or VPLS, active filtering is the master process, and the learning service will simply align its connectivity with the filtering as imposed by the PLSB control plane.
- PLSB, being Ethernet, can utilize any Ethernet link, be it shared or P2P.
- PLSB is able to utilize parallel paths across the network. This greatly simplifies the task of providing efficient redundancy, as the role of any other technology offering transit bandwidth can be simplified to that of one or more switching hubs. PLSB has no expectations that the transit has any control or blocking capabilities; it is transparent.

Finally, PLSB and IS-IS already acknowledge the existence of switched segments. LAN segments have existed since the dawn of Internet time, and are a "well understood" and solved problem from the point of view of control plane architecture; and the modeling of LAN segments in link state routing systems is a known and implemented technology.

PLSB, PBB AND PBB-TE

PLSB can be run side by side on the same infrastructure as both PBB, using other spanning tree control planes, and PBB-TE. It operates on a different set of B-VIDs (nominally requiring only two) than the other modes of behavior, but shares all other aspects of the network including the OAM toolset and associated administration.

It is comparatively easy to envision migration from PBB using other control planes to PLSB operation by loosely synchronized migration of services from the VID assigned within one instance of PBB to a VID assigned to PLSB. Only loose synchronization is required because both topologies are active during the migration, and only the VID selection made at a source determines that any frame travels over one topology or the other but not both. Augmentation of PLSB with PBB-TE is also possible where it is desirable to move traffic off the shortest path, with migration being handled by the same VID assignment technique. In this way engineering of ELINE is easily achieved on a common technology base, and as PLSB is simply PBB-TE exploited by a routing system, it is easy to envision other arrangements.

CONCLUSION

PLSB is the product of applying contemporary levels of computing power and memory to link state routing applied to driving Ethernet unicast and multicast connectivity.

PLSB is able to take advantage of a link state control plane to combine the functions of spanning tree convergence, flooding and learning, and registration protocol exchange into a single FDB population step, and as flooding is eliminated from the equation, PLSB can do this while making the most efficient use of a meshed network environment.

The combination of the scaling of Ethernet identifiers embodied in PBB combined with the efficiencies of full mesh utilization pioneered with PBB-TE and the single step convergence permit the number and geographic scope of virtualized LANs to achieve new levels of scalability, efficiency, and overall operational simplicity.

The combination of the scaling of Ethernet identifiers embodied in PBB combined with the efficiencies of full mesh utilization pioneered with PBB-TE and the single step convergence permit the number and geographic scope of virtualized LANs to achieve new levels of scalability, efficiency and overall operational simplicity.

REFERENCES

[1] IEEE 802.1ah D4.1, "IEEE Draft Standard for Local and Metropolitan Networks, Virtual Bridged Local Area Networks, Amendment 6: Provider Backbone Bridges," Feb. 2008.
[2]Metro Ethernet Forum, Tech. Spec. MEF D00057_006, "Ethernet Services Definitions — Phase 2," Approved Draft 6, Feb. 2008.
[3] D. Allan and D. Fedyk, "Ethernet Data Plane Evolution for Provider Networks" *IEEE Commun. Mag.*, vol. 46, no. 3, 2008.
[4] ISO/IEC 10589, Information Technology — Telecommunications and Information Exchange Between Systems — Intermediate System to Intermediate System Intra-Domain Routing Information Exchange Protocol for Use in Conjunction with the Protocol for Providing the Connectionless-Mode Network Service (ISO 8473)," 2nd ed., 2002.
[5] IEEE 802.1aq D0.4, "IEEE Draft Standard for Local and Metropolitan Networks, Virtual Bridged Local Area Networks, Amendment 9: Shortest Path Bridging," Feb. 2008.
[6] Y. Dalal and R. Metcalfe, "Reverse Path Forwarding of Broadcast Packets," *Commun. ACM*, vol. 21, no. 12, 1978.
[7] IEEE Std. 802.1Q, "IEEE Standard for Local and Metropolitan Networks, Virtual Bridged Local Area Networks," 2005.

BIOGRAPHIES

DAVID ALLAN (dallan@nortel.com) is a distinguished member of technical staff at Nortel. He has been active in data telecommunications standards for the past 12 years including WG chair roles in the DSL Forum and IETF. He has been active for over 25 years as an architect, design engineer, and developer of real-time systems in diverse areas of technology ranging from process control and avionics to financial transaction processing. His current role at Nortel is focused on carrier infrastructure based on carrier Ethernet. He has a B.Eng. (1978) from Carleton University in Ottawa.

PETER ASHWOOD-SMITH (petera@nortel.com) is a Nortel Fellow with B.S and M.S. degrees in computer science from the University of Toronto. He has worked on the design, standardization, implementation, deployment, and support of many modern routing protocols (e.g., early label switched networks including ATM, MPLS, and GMPLS). His interests and research are now tending in the direction of computation-based/non-label-switched protocols.

NIGEL BRAGG (nbragg@nortel.com) is a Nortel Fellow with degrees from Trinity College, Cambridge, and Southampton University. He has spent 20 years in the telecommunications industry, initially in a contract product development environment, and for the last 11 years with Nortel. He has contributed to a wide variety of projects during that time, ranging from design leadership on a high-performance voice switching system to many aspects of data switching and routing, to automated power optimization in optical line systems. For the last four years he has focused on carrier packet transport, and is a co-inventor of PBT and PLSB

DON FEDYK (dwfedyk@nortel.com) is an authority on routing system design for both connectionless and path oriented routing. He is an active contributor to several IETF WGs in the routing area including MPLS and CCAMP. In the past couple of years he has been focused on data and control planes for provider Ethernet. He received his B.S. and M.S. degrees in electrical engineering from the University of Waterloo, Ontario, Canada.