November 26, 2006

PROVIDER BACKBONE TRANSPORT OVERVIEW

Paul Bottorff, pbottorf@nortel.com

David Allan, dallan@nortel.com

Introduction

Essential features needed in Ethernet networks to support carrier applications are: scalability, reliability, hard QoS / traffic management, service management and support for TDM services. Current carrier Ethernet standards including the IEEE Provider Backbone Bridging (802.1ah) and Connectivity Fault Management (802.1ag CFM) are addressing scaling, reliability, and service management providing new alternatives for metro networks that are based on Ethernet forwarding, simplicity and cost curves.

This contribution discusses a technology called Provider Backbone Transport (PBT) which can be employed in the service provider domain of a Provider Backbone Bridge Network (802.1ah) to allow configuration of engineered, resilient, SLA driven point-to-point Ethernet trunks facilitating hard QoS and traffic management. These PBT trunks allow carriers to engineer traffic managed circuits which may be monitored, along with the rest of the 802.1ah network, using 802.1ag protocols. Paths generated by PBT may be used to assure route diversity for protection paths as well as network load balancing and support for performance guarantees.

Provider Backbone Transport Overview

In a standard Provider Backbone Bridged Network (802.1ah PBBN) traffic engineering is limited as a consequence of the use of IEEE 802.1Q MSTP control plane protocols which control the population of the bridge filtering tables. The underlying 802.1Q/802.1ad/802.1ah bridge relays however don't have any inherent characteristics which prevent full traffic engineering. For example, the IEEE Shortest Path Bridging (802.1aq) project is improving link utilization by replacing the MSTP control plane with a shortest path spanning tree control plane. Provider Backbone Transport is a method for providing full traffic engineering of point-to-point paths in an 802.1ah network. To do this PBT replaces the MSTP control plane with either a management plane or an external control plane and then populates the bridge filtering tables of the component 802.1ad and 802.1ah bridge relays by creating static filtering table entries (see Figure 1).

The ability of PBT to utilize an external management or control plane is facilitated by 802.1ah because the B-DAs are all managed by the Provider and therefore can all be discovered and identified in the Provider's topology by the external management or control plane. The description of PBT provided here is based on using PBT within a PBBN, however it is possible to extend the use of PBT to other environments where the Provider controls the SA and DA addresses. An example of such an environment is where the Provider and Customer agree on a specific set of MAC address as part of an SLA. The Provider then only accepts frames with a source address as specified in the SLA and destination addresses as specified in the SLA. Though these arrangements are possible they require some form of static agreement on MAC address which are used between the customer and provider.

The external PBT management/control plane is responsible for maintaining and controlling all the topology information to support point-to-point unidirectional Ethernet Switched Paths (ESP) over the PBBN. The PBT topology can co-exist with the existing active MSTP or with the new 802.1aq topology by allocating B-VID spaces to PBT, MSTP, or 802.1aq, or PBT can be stand alone. PBT takes control of a range of B-VIDs from the Backbone Core Bridges (BCB) and Backbone Edge Bridges (BEB) of the PBBN.

The PBT management/control plane forms a topology of B-DA rooted trees. For each <B-DA, B-VID> pair configured by PBT an independent tree is maintained. ESPs are routed by PBT along a tree selected by the B-VID to the destination B-DA. For a single B-DA the number of separate routing trees may be up to the number of PBT reserved B-VIDs. The B-VIDs may be reused for every B-DA, therefore the total number of routing trees maintained by the PBT management/control plane may be up to the number of B-DAs times the number of PBT reserved B-VIDs. The trees maintained by PBT for routing ESPs do not have to be spanning since they only require connectivity to all the source B-MACs which have ESPs to the specific B-DA. Each tree may connect to as many B-SAs as desired with the only limits being implementation imposed table sizes. The PBT management/control plane may use any algorithm desired to select the path for a routing tree thereby providing complete route selection freedom. The PBT management/control plane also manages the bandwidth of all ESPs along each routing tree. For each B-SA which is part of a routing tree maintained by the PBT management/control plane, PBT will maintain a routing tree which provides a co-routed reverse path from the B-DA

to the B-SA. The B-VID used in this reverse ESP does not have to be the same one used for the forward ESP. The reverse ESP is used by CFM management to monitor the ESPs.

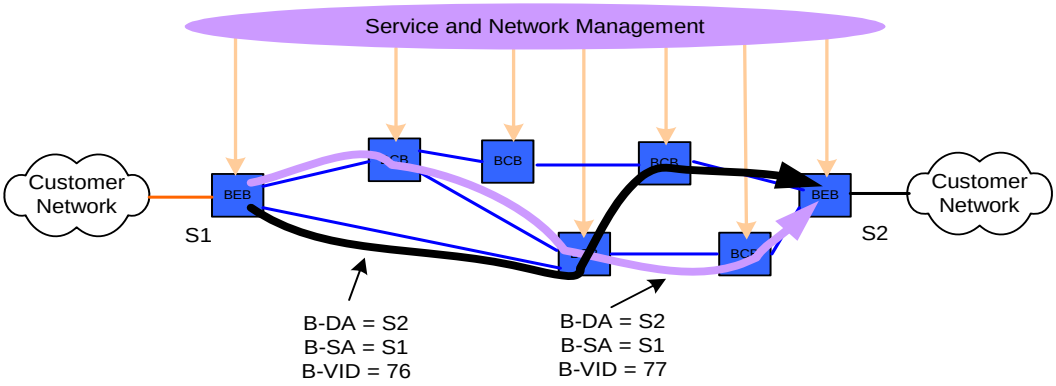

Figure 1 - PBT Network

The relay functions used by PBT are standard bridge forwarding which use a <B-DA, B-VID> tuple rather the B-DA alone for filtering and forwarding decisions. It is possible to treat the combined <B-DA, B-VID> tuple as though it was a single 58 bit address, where 12 bits are the B-VID and 46 bits are the B-DA (allowing for the multicast and local reserved bits in the MAC space). This fact allows PBT to consider the B-VID part of the address as a path selector to the B-DA rather than a B-VLAN ID, allowing up to 4094 unique routing trees to any single B-DA. Typically only a small number of B-VIDs are needed for PBT since it is normally not necessary to have even tens of alternate paths to a single destination. In Figure 1 two paths are configured to reach S2. These two paths are separated by using a different B-VID in combination with the B-DA for a second path.

PBT requires no B-VID translation and operates on most 802.1ah unicast forwarding hardware. PBT allows scaling to an almost infinite number of Ethernet Switched Paths. If the B-VID range delegated is the full 4094 possible values, then each B-DA termination can sink 2^{12} routing trees, and the theoretical network maximum is about 2^{58} ESPs. The fowarding of frames over these ESPs is easily achieved by most existing Ethernet equipment without significantly re-specifying the hardware and management to achieve traffic management.

To make PBT robust, a few aspects of standard 802.1Q bridge forwarding need consideration:

1) Discontinuities in forwarding table configuration for an ESP will result in packets being flooded as "unknown". As there is no loop free topology for the delegated B-VID range, this will result in unbounded flooding, looping and replication. For this reason flooding of packets with unknown destinations must be disabled for the B-VID range allocated to PBT. Similarly, broadcast and multicast traffic that would be flooded must be filtered at the ingress to the relay function. This function is already supported by 802.1Q.

2) B-SA learning is not required, and may interfere with management/control population of the forwarding tables when combined with the potential for configuration errors. For this reason B-SA learning is disabled for the delegated B-VID range.

3) This approach bypasses spanning tree for the delegated B-VID range, so spanning tree or multiple spanning trees are used only for the non-delegated (traditional operation) B-VID range.

4) When used in conjunction with a best effort spanning tree, traffic on Call Admission Control or “CAC'd” paths requires a higher priority than best effort traffic. When engineering the network, a reserve bandwidth must be set aside for best effort traffic to allow for spanning tree reconfiguration.

This approach has a number of useful properties:

The use of a global path identifier (the <B-DA, B-VID> tuples) directly for forwarding and with no translation is inherently more robust than alternatives. Any mis-configuration or forwarding table errors resulting in the deviation of a PDU from the intended path will self identify immediately. There is no possibility of collision with other identifier spaces that can mask the fault. This also suggests that no or minimal changes are required to existing CFM and Y.1731 constructs to successfully instrument Ethernet paths.

The ability to explicitly route and pin paths across the network can be combined with call admission control and 802.1Q class-based queuing in order to provide per path QoS. The call admission control function can be enforced by the external management/control plane without any changes to existing Ethernet bridges.

Example Ethernet Switched Paths

Figure 2 shows an example Provider Backbone Bridged Network running Provider Backbone Transport over the PBBN core. The PBBN B-VID space has been partitioned between MSTP and PBT with B-VIDs 7 and 8 allocated to PBT. PBT has taken over the port forwarding state machines in the Backbone Edge Bridges and Backbone Core Bridges for the allocated B-VID range and controls frame forwarding by adding static entries to the filtering databases of the switches within the PBBN core. MSTP operates normally in parallel to PBT on the other B-VIDs.

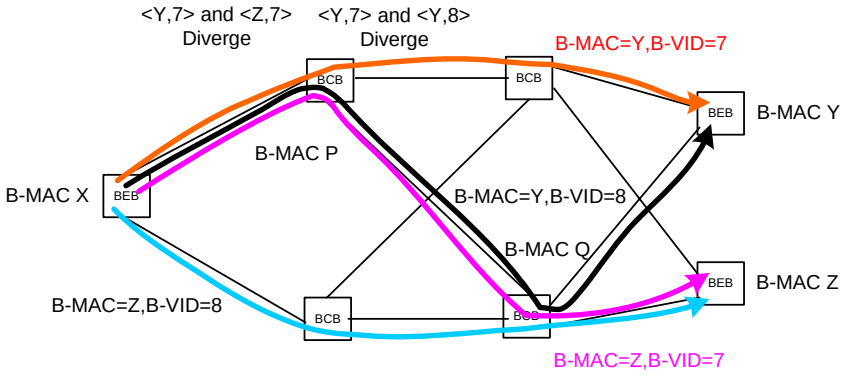

Figure 2 - PBBN with B-VIDs 7 and 8 allocated to PBT operation

The B-VIDs allocated to PBT are not treated as VLAN IDs, instead they are individual instance identifiers for one of a maximum of ‘n’, where ‘n’ is the number of allocated B-VIDs, or possible routing trees to the destination B-DA address. B-VIDs in the allocated range may be re-used for many routing trees within the PBBN as long as the <B-DA, B-VID> tuple is unique. This PBT addressing arrangement results in a 58 bit globally unique destination address that may be shared by any number of B-SAs.

The example in Figure 2 illustrates the complete route freedom of configured forwarding in bridges. In the example a total of 4 ESPs use 2 B-VIDs to forward traffic to 2 B-DA terminations. At node 'P' above, despite collisions in both the B-DA and the B-VID space, the forwarding properly resolves because the switch uses the B-DA and B-VID together to establish route uniqueness. At node ‘P’ the red and purple ESPs diverge even though they have the same B-VID because they are addressed to different B-DAs. Likewise at node ‘P’ the red and black ESPs diverge even though they have the same B-DAs because they have different route selector B-VIDs.

Changes Required to 802.1Q to Support PBT Forwarding

Three changes are necessary to 802.1Q for PBT forwarding support. These three modifications are:

- The B-VID address space must be divided between MSTP (or SPB) and PBT.
- It must be possible to discard unknown and group addresses rather than flooding them on PBT B-VIDs
- The PBT port state must be set to forwarding and not learning

PBT may operate by taking over the entire control plane or may operate in parallel to MSTP. When PBT takes over the entire bridge the entire spanning tree function is removed from the bridge and PBT controls the use of all B-VIDs, port states, and filter database. When PBT operates in parallel to MSTP the B-VID space is split between MSTP and PBT. When operating in parallel PBT controls a set of B-VIDs allocated to it and the associated port states while MSTP retains control of the other B-VIDs. The procedure for splitting the B-VID space requires extensions to the existing management interfaces.

- The MSTI list (12.12.1) is restricted to listing MSTIDs that are to be handled by MSTP.

- Define a special MSTID called the PBTID (use 0xFFE) which identifies PBT rather than a MSTI. An MSTID not in the MSTI list indicates some protocol other than MSTP that may run in parallel to MSTP (802.1Q-2005 12.12.1 and 8.6.2).
- The FID to MSTID Allocation Table (802.1Q-2005 12.12.2) is allowed to contain MSTIDs that are not in the MSTI List (existing restriction is removed).
- The MST Configuration Table (802.1Q-2005 12.12.3 showing the allocation of B-VIDs to MSTIDs) is allowed to contain "MSTIDs" that are not in the MSTI list and "MSTIDs" that are not statically configured in the FID to MSTID Allocation Table.
- The PBTID code of hex FFE in the MST Configuration Table means "B-VID is not allocated to an MSTI and is available to PBT for use as a route selector".
- PBTIDs not in the MSTI list indicate another protocol that might actually run at the same time as MSTP.
- Use a Configuration Identifer Format Selector of 1 to distinguish a Configuration Digest that is conveying an MST Configuration Table extended as above.

Using these rules we allocate PBT B-VIDs to the hex FFE FID, and allocate the hex FFE FID to the hex FFE MSTID. This configuration then allows the B-VID space to be split between MSTP and PBT with a single port state for the PBT B-VIDs and a port state for each MSTI allowing both to operate independently from each other.

The second modification to 802.1Q to support PBT frame forwarding is to discard unknown and group addresses rather than flooding them. To implement this we may us the all Group Addresses, for which no more specific Static Filtering Entry exists (802.1Q-2005 8.8.1 bullet a) filter to discard all unspecified group addresses. To support discarding unknown unicast addresses we need to add a new Static Filtering type for all unicast addresses, for which no more specific static filtering entry exists (change 802.1Q-2005 8.8.1 bullet a). By setting both of these static filters to discard on match all unknown addresses will be discarded.

The third modification to 802.1Q is to enable forwarding on the PBT ports while disabling MAC address learning. All VIDs allocated to PBT have a port state at each bridge port who's state is forced to forwarding=on and learning=off (change 802.1Q-2005 8.4).

Supporting CFM CCMs for PBT

In a PBBN PBT is used to replace B-VLANs to support engineered trunks. To manage these PBT trunks standard 802.1ag CCMs, which are addressed using a group address, may not be used since they will be discarded by the PBT trunk. To support CCM in on a PBT trunk a provision must be added for unicast addressed CC messages. These messages will then be forwarded just link PBT data and will follow the same path defined for PBT forwarding. The ITU-T Y.1731 specification already supports unicast CC messages. This functionality would need to be added to 802.1Q for PBT CC support.

In addition to CCMs unicast LB CFM messages are also used to monitor links. These are already supported by 802.1ag and do not require any additions.

Protection switching is also an important management feature of PBT. PBT may provide 1:1 protection configuring an alternate path with a different B-VID or B-MAC. Both the working and protection paths are constantly monitored using unicast CCMs. On a failure of the working path the data stream is transferred to the protection path by switching the B-VID or B-MAC.

Scaling Considerations

The number of PBT trunks which may be supported within a single PBBN is limited only by the limits of the MAC address space and the ability of the provisioning and management system. The combined <B-MAC, B-VID> address used by PBT provides up to 58 bits of addressing which may be used to define PBT trunks. An entry must exist in the filtering database for each switch which the PBT trunk passes through. This limits the number of PBT trunks which can be supported by a single switch to the number of entries supported in the filtering database. Since PBT trunks are routed through the PBBN by the provisioning system, the load on the filtering database in any given switch and the data load on any given switch may be distributed over as many switches as needed to support the desired number of trunks. A PBBN could theoretically support up to the entire 2**58 PBT trunks if enough switches were available and if the provisioning database could reach this size.

Another scaling consideration for PBT is the load generated by constant CCMs used to monitor the trunks and perform protection switching. For CCMs generated at the fastest rate of 10 msec, allowing 50 msec protection switching, the total load generated per trunk is ~6.4 Kbytes/sec (or 100 frame of 64 byte each). This represents about 0.64 % of a 10 Mbit

trunk. Given that PBT trunks will normally run over 10 Gbit links and typically be 10 Mbits and more the bandwidth utilization for management would never exceed around 0.5% of the link bandwidth and would typically be less.

In addition to bandwidth utilization by PBT trunk monitoring the end systems will have to handle the CFM frame load. For instance a single end station terminating 1000 PBT trunks would have a frame load of 200,000 CCM frames received by both the working and protection paths. Though this is a lot of processing it is well within the ability of today's network processors. To scale the network beyond the CCM frame processing load of the BEBs, it is always possible to distribute trunks over as many chassis as desired to support indefinite scaling with today's switches.

Another aspect of scaling which is indirectly related to PBT is exhaustion of the I-SID address space for identification of services. The number of services can be scaled using the peer E-NNI provided by 802.1ah for interconnection of PBBNs. The I-SID service address is translated at a peer the E-NNI providing indefinite scaling of the service address space. PBT trunks extended over peer E-NNIs will have the PBT route selector (B-TAG) stripped and regenerated at the next PBBN allowing each PBBN to have an independent route for the PBT trunk.