Introduction to Port Extension (IEEE P802.1Qbh)

Joe Pelissier
Cisco Systems
3 West Plumeria Drive
San Jose, CA 95134
USA
+1.503.628.0801
jopeliss@cisco.com

Rene Raeber
Cisco Systems
Mail Stop WLSN01/3/
Swing Building, 3rd and 4th floor
Richtistrasse 7
Wallisellen, Zurich 8304
Switzerland
+41 79 446 7737
rraeber@cisco.com

Abstract – Data centers today are experiencing a rapid proliferation of switches as a direct result of the deployment of high-density server platforms. Given the layered switching architecture found in today's data centers, many of the switches are performing a simple aggregation function; that is, the majority of traffic is moving between downlinks and uplinks. However, despite this simple function, these switches contribute to a significant portion of the capital expenditure and ongoing administrative and management costs of the data center. The IEEE is standardizing a new technology, Bridge Port Extension, which replaces these aggregating switches with a Port Extender that extends the ports of the switch in the next higher layer. This technology has the potential to reduce significantly the number of managed switches in the data center as well as reduce the upfront capital expenditure costs.

I. Introduction and Motivation

Data centers today are experiencing a dramatic increase in the number of installed Ethernet switches as a direct result of the deployment of high-density servers and blade servers. In addition, deployment of virtualization technology within servers has resulted in an even further increase in the number of installed switches. These switches are typically embedded within the actual server itself. It is important to note, however, that virtualization is not the sole source of this explosion in switch proliferation, but it has added significantly to this phenomenon.

The growth in switch deployment has resulted in the corresponding growth of associated costs.

A new technology has is being standardized by the IEEE referred to as "Bridge Port Extension". The Port Extension technology introduces a new device called a "Port Extender" that effectively acts as additional ports for the switch to which it is connected. The switch to which the Port Extender is connected is referred to as the "Controlling Switch". The Controlling Switch and a set of Port Extenders connected to it form a single logical switch, referred to as an Extended Switch. Network administrators to not mange Port Extenders directly; Port Extenders are managed as part of the combined Extended Switch. To enable the Controlling Switch to act as the central point of management of the Extended Switch, the Bridge Port Extension standard specifies a Port Extension Control and Status Protocol (PE CSP). The Controlling Switch uses PE CSP to control the attached Port Extenders. This paper introduces Bridge Port Extension and the Port Extender Control and Status Protocol.

II. Bridge Port Extension Overview

Bridge Port Extension provides the capability to combine distributed network

components into a single Extended Switch. These components consist of:

- A Controlling Switch
- Distributed Port Extenders (that may be cascaded)
- A protocol enabling control of the Port Extenders by the Controlling Switch

Figure 1 illustrates a network utilizing Port Extension.

The top switch in this figure represents the Controlling Switch. One of the ports on this switch connects to a Port Extender. In Upstream Port is the port of a Port Extender that connects to a Controlling Switch, or to a higher level Port Extender within a cascade. This Port Extender connects to four additional devices below it. Two of these devices are additional Port Extenders. The other two devices are conventional switches. The Controlling Switch and the three Port Extenders shown in this figure combine to form a single Extended Switch. Cascade Ports are ports of the Port Extenders that connect to lower level Port Extenders within a cascade. In a cascade of Port Extenders, the Cascade Ports of Port Extenders in one layer connect to the Upstream Ports of the Port Extenders in the next lower layer of the cascade. Thus, the topology formed by a cascade of Port Extenders is a loop-free tree.

Figure 2 illustrates the logical network achieved by the physical network illustrated in figure 1. Note that the three Port Extenders logically become ports of the Controlling Switch.

A Controlling Switch detects the connection of a Port Extender utilizing Link-Layer Discovery Protocol [1].

S-channels deliver unicast traffic. An S-channel is a point-to-point channel identified by a Service VLAN Identifier (S-VID). The S-VID may be expressed explicitly in the frame using a Service VLAN Tag (STag), or may be expressed implicitly (i.e. no STag present) using defaults assigned to the receiving ports.

The Controlling Switch creates an S-channel between itself and the control entity

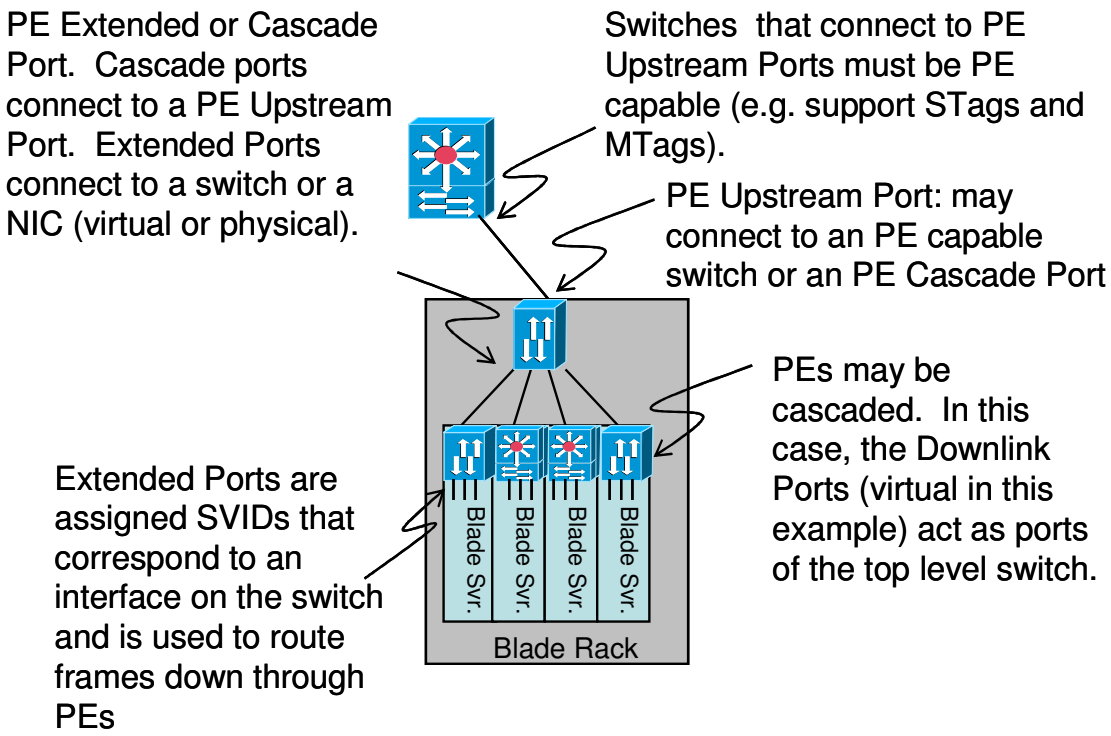

Figure 1 – Port Extension Anatomy

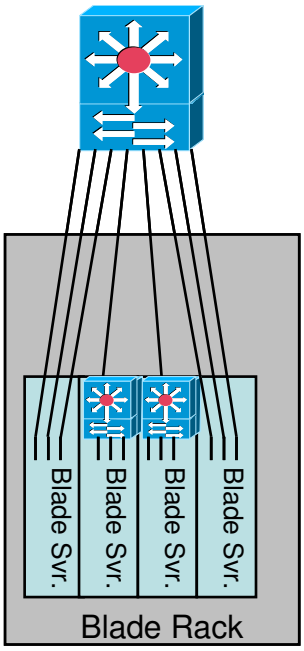

Figure 2 – Equivalent Logical Network

within each Port Extender. This channel carries the PE CSP. In addition, The Controlling Switch creates an S-channel for each downlink port on the Port Extender. Since there is a one to one correspondence between S-channels and the downlink port on a Port Extender, the Controlling Switch is able to determine the ingress Extended Port from the S-channel indicated by the STag carried in the frame. In addition, the Controlling Switch directs a frame to a particular Extended Port by placing it on the corresponding S-channel.

Port Extenders may be cascaded. The S-channel associated with the downstream port of the Port Extender to which the new Port Extender is attached transports PE CSP. The Controlling switch creates additional S-channels for each downlink port on the new Port Extender. The process repeats for each Port Extender attached in the cascade.

The Controlling Switch also creates M-channels through a single Port Extender or a Cascade of Port Extenders. M-channels are a point to multipoint channel starting at the Controlling Switch and ending at one or more Extended Ports. These channels transport multicast frames. Port Extenders replicate frames to each downlink port that is part of the M-channel. Like S-channels, an identifier is included in each multicast frame to identify the M-channel to which it belongs. The frame carries the M-channel Identifier (MCID) using an M-channel tag (MTag).

III. Protocol Overview

The Controlling Switch utilizes the PE CSP to perform the following functions:

- Establish initial communication with the Port Extender
- Establish the S-channels to the Port Extender Downlink Ports
- Adding S-channels to intermediate Port Extenders within a cascade
- Deleting S-channels
- Adding and deleting M-channels
- Obtain traffic statistics from the Port Extenders

The PE CSP operates over a simple transport protocol referred to as Edge Control Protocol (ECP). PE CSP packages commands and responses into Protocol Data Units (PDU). On transmission, ECP transmits the PE CSP PDU and waits for an acknowledgement. ECP retransmits the PDU if it receives no acknowledgement. On reception, ECP passes the PDU to PE CSP, and sends an acknowledgement.

PE CSP uses a credit based flow control mechanism to enable processing of multiple commands. During initial communication establishment, each endpoint provides the other a credit limit. The transmission of a command consumes a credit and the reception of a response replenishes one credit. Once the credit reaches zero, a device may not issue PE CSP commands until the peer endpoint

replenishes credit by responding to outstanding commands.

IV. Establishing Initial Communication

Both a Controlling Switch and a Port Extender initiate PE CSP communication using the CSP Open Command. This command provides the credit limit and the number of S-channels and M-channels supported. A device may not send additional PC CSP commands until it has received the response to its CSP Open command and has received the CSP Open command from its peer.

V. Establishing S-channels

The Port Extender initiates the establishment of S-channels. It does so by sending an S-channel Create command to the Controlling Switch. This command includes an index of the port associated with the requested S-channel. The Controlling Switch includes in its response the SVID used to identify the S-channel. The Port Extender adds a STag containing this SVID to all frames received on this port. In addition, the Port Extender forwards all frames received from the Controlling Switch containing this SVID in the STag to this port.

VI. Establishing S-channels within a Cascade

The S-channel Create command establishes an S-channel at the Controlling Switch and the final Port Extender. If intervening Port Extenders exist within a cascade, then the Controlling Switch must also establish the S-channel within these intervening Port Extenders. To accomplish this, the Port Extender issues an S-channel Register command to each intervening Port Extender. In this command, the Controlling Switch provides a list of SVID, port identifier pairs. Each pair represents an additional SVID associated with the corresponding port. Upon receiving this command, the Port Extender updates its forwarding tables to forward frames with the SVID to the corresponding port thus creating the S-channel though the Port Extender. The Controlling Switch may use a single S-channel register to register multiple S-channels within a Port Extender.

VII. Deleting S-channels

The Controlling Switch utilizes the S-channel Deregister command to remove an S-channel from a Port Extender. A list of SVIDs within the command indicates the S-channels to delete.

VIII. Adding and Deleting M-channels

The Controlling Switch utilizes the M-channel Register command to create M-channels though one Port Extender or a cascade of Port Extenders. In the case of a cascade, the Controlling Switch sends an M-channel Register command to each Port Extender in the cascade that forms part of the path of the M-channel.

The M-channel Register command contains an MCID and a list of ports. Upon receiving this command, the Port Extender configures its forwarding logic to forward frames received from the Controlling Switch with the corresponding MCID to each of the listed ports, replicating frames as necessary.

The Controlling Switch also uses the M-channel Register command to add additional ports to or to remove ports from and existing M-channel. To accomplish this, the Controlling Switch simply sends an M-channel Register command containing the MCID of the M-channel and a list of the ports that are to remain in the M-channel. Upon receiving this command, the Port Extender removes any ports from the M-channel that were not included in the list and adds any new ports.

To delete an entire M-channel, the Controlling Switch sends an M-channel Register command containing the MCID of the

channel and no entire in the port list. Upon receiving this command, the Port Extender removes the entire M-channel from its forwarding logic.

IX. Obtaining Traffic Statistics

To obtain traffic statistics from an Extended Port, the Controlling Switch sends a Get Statistics command to the Port Extender. The Controlling Switch includes in this command the SVID corresponding to the port from which statistics are to be gathered. In response, the Port Extender returns the values of the statistical counters for that port. There has been no proposal in the IEEE of counted events; however, the list will likely include the usual statistics such as frames and octets received and transmitted, number of frames received with errors, etc.

X. Port Extension and Virtualization

Port Extension provides unique benefits in server virtualization environments in addition to the benefits provided in traditional data center networks.

A key capability of server virtualization is the ability to move an active virtual server from one physical machine to another (referred to as virtual machine migration). Virtualization software commonly provides this capability.

The network complicates the migration process. Each virtual machine requires certain characteristics of the switch port that connects it to the fabric. This could include parameters such as flow control, congestion notification, VLAN assignment, access control lists, etc. The term "port profile" commonly refers to this collection of parameters.

There is no standard process for transferring a port profile from one switch to another in synchronization with the migration of a virtual machine. Virtual machine migration often involves a manual step of pre-provisioning the switch port within the target physical server. Thus, the efficiency of the migration is negatively impacted.

If the migration is to take place between two ports of a given switch, it is quite trivial for the switch to transfer the port profile simultaneously. Immediately after a migration, a virtual machine typically broadcasts a frame to "announce" its new location and to allow switches in the network to update their filtering databases (typical a RARP frame serves this purpose). Since the virtual machine has moved to a new port on the same switch, upon reception of the announcement frame, the switch may simply move the port profile from the old switch port to the new one. Since no coordination is required between switches, no standard protocols or procedures are required.

Unfortunately, virtual machines in today's data centers nearly never migrate between ports on a given physical server. Recall that each physical server contains its own embedded switch. It rarely makes sense to migrate a virtual machine within a given server (such a migration is essentially meaningless). Thus, migrations are nearly always between physical servers, and by definition, between switches.

However, if a Port Extender replaces the embedded switch within each physical server, and these Port Extenders connect to a common Controlling Switch, then a single Extended Switch connects all of the virtual machines within all of the virtual servers. As a result, it is possible for migration between the physical servers without manual pre-provisioning of the port profiles.

XI. Summary

Because of high-density server technology, modern data centers are experiencing a dramatic increase in the number of deployed switches. This results in increased capital expenditure and management costs while stretching the scalability limits of the management applications. Many of these

switches perform little frame relay other than between adjacent layers in the network architecture. Port Extension allows the removal of these switches from the network resulting in significantly fewer acquired and managed switches. In addition, port extension extends the reach of the more capable higher layer switches to the edge of the network. This results in superior use of the capabilities of the higher layer switches. The Port Extension Control Protocol enables the Controlling Switch to operate as the single point of management for the entire Extended Switch.

XII. References

[1] IEEE Computer Society, *IEEE Std 802.1AB™-2009, IEEE Standard for Local and metropolitan area network – Station and Media Access Control Connectivity Discovery,* 3 Park Avenue, New York, NY 10016-5997, USA, 17 September 2009.