MARVELL

WHITE PAPER

Network Virtualization: A Data Plane Perspective

David Melman
Uri Safrai
Switching Architecture
Marvell
May 2015

Abstract

Virtualization is the leading technology to provide agile and scalable services in cloud computing networks.

Virtualization is occurring at different areas of the network:

1) Server Virtualization
2) Network Virtualization
3) Network Function Virtualization

Marvell® Prestera® switching devices are based on the flexible *eBridge* architecture, which implements data plane interface virtualization. This enables migration from the legacy “physical” networking paradigm to the virtualized network in the emerging modern data center compute cloud, overlay networks and network function virtualization domains.

This paper outlines the state of art of the virtualization technologies and how the Marvell® eBridge architecture allows Marvell switches to serve as a universal gateway, seamlessly interconnecting different types of virtualization domains and data encapsulations.

Server Virtualization

Server virtualization allows scaling of server resources by partitioning a physical server into multiple independent virtual machines (VMs).

IEEE 802.1 has standardized two approaches for connecting a virtualized server to the network:

- 802.1Qbg Edge Virtual Bridging
- 802.1BR Bridge Port Extension

Although the packet encapsulations differ, in both standards the *Controlling Bridge* is the central entity which performs all forwarding, filtering and policy decisions. The Controlling Bridge views VMs as being attached via a virtual interface.

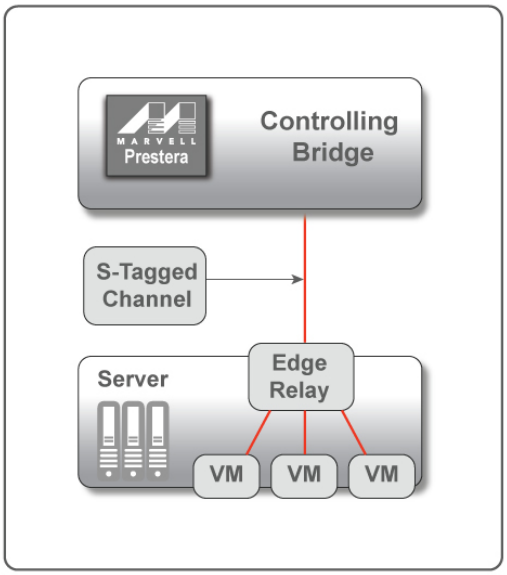
IEEE 802.Qbg Edge Virtual Bridging

IEEE 802.1Qbg defines *VEPA* (Virtual Ethernet Port Aggregation) and a *Multi-channel* S-tagged interface between a Controlling Bridge and the server NIC.

Basic VEPA requires all VM-sourced traffic to be processed by the Controlling Bridge. The Controlling Bridge may need to *hairpin* traffic back to the source port if it is destined to another VM on the same server.

In Multi-channel VEPA, an S-tag is added to the packet indicating the source or target VM. Upon receiving the packet, the Controlling Bridge examines the S-Tag and assigns a source logical interface to the packet. The S-Tag is popped, and the packet is subject to ingress policy, forwarding and filtering rules that are associated with the source VM. When forwarding traffic to a VM, the Controlling Bridge applies egress policy for this VM, and adds an S-Tag to indicate the target VM on the server. In the case of Broadcast, Unknown Unicast and Multicast (BUM) traffic, the switch replicates the packet to the set of VMs in the flood domain, where each packet instance includes the S-Tag associated with the remote VM.

The Marvell eBridge architecture supports the 802.1bg VEPA and Multi-channel VEPA, by flexible mapping of each VM to a Controlling Bridge local virtual interface.



802.1BR Bridge Port Extension

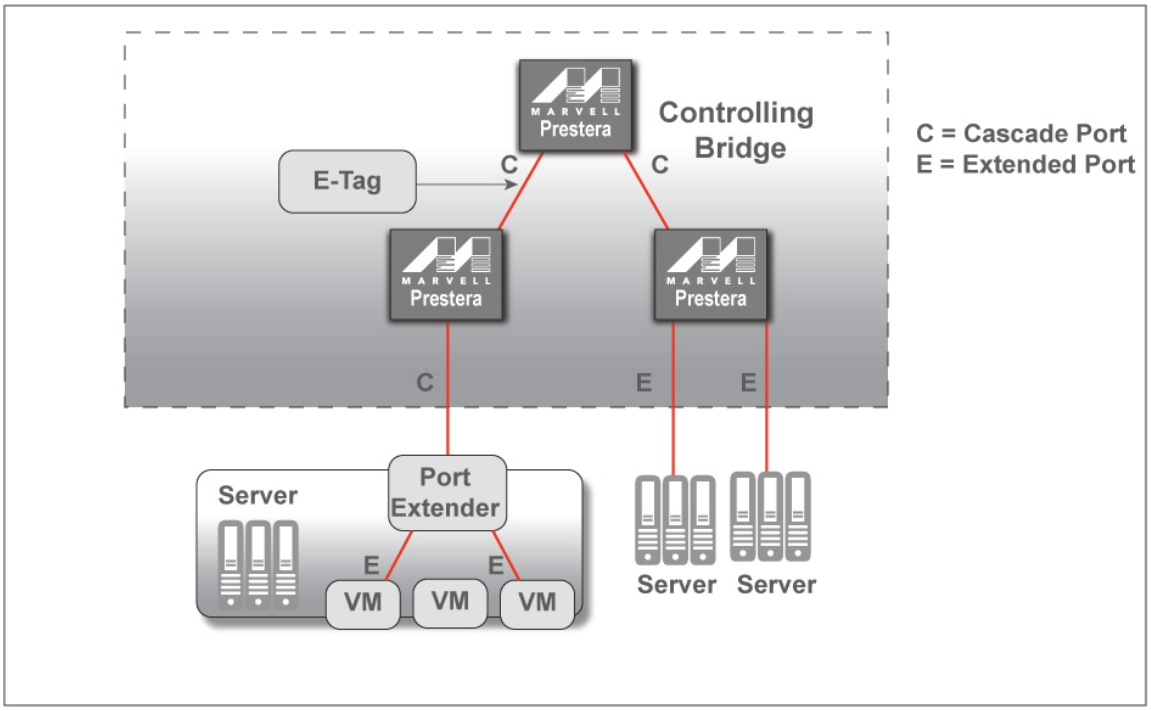
IEEE 802.1BR defines a logical entity called an *Extended Bridge* that is comprised of a *Controlling Bridge* attached to a set of devices called *Port Extenders*. The Port Extenders can be cascaded to interconnect the Controlling Bridge with remote VMs. Logically, adding a Port Extender is similar to adding a line card to a modular switch.

All VM-sourced traffic is sent via the Port Extenders to the Controlling Bridge. The Port Extender implemented in a virtualized server's hypervisor pushes an *E-Tag* identifying the source VM and forwards the packets towards the Controlling Bridge.

On receiving the packet, the Controlling Bridge examines the E-Tag, and assigns a source logical port. The E-Tag is then popped, and the packet is subject to ingress policy, forwarding, and filtering rules that are associated with the source VM.

When the Controlling Bridge transmits traffic to a VM, the switch applies egress policy for this VM, and adds a unicast E-Tag, which indicates the target VM on the server. The intermediate Port Extenders forward the packet towards the server, where the hypervisor Port Extender strips the E-Tag prior to sending the packet to the VM. In case of BUM traffic, the Controlling Bridge pushes a single multicast E-Tag which indicates the multitarget "port group". The downstream Port Extenders replicate the packet to each of its port group members. The hypervisor Port Extender strips the E-Tag prior to sending to its local VMs.

The Marvell eBridge architecture supports the 802.1BR Controlling Bridge and Port Extender standard, by flexible mapping of each VM, remotely or locally attached, to a Controlling Bridge local virtual interface.

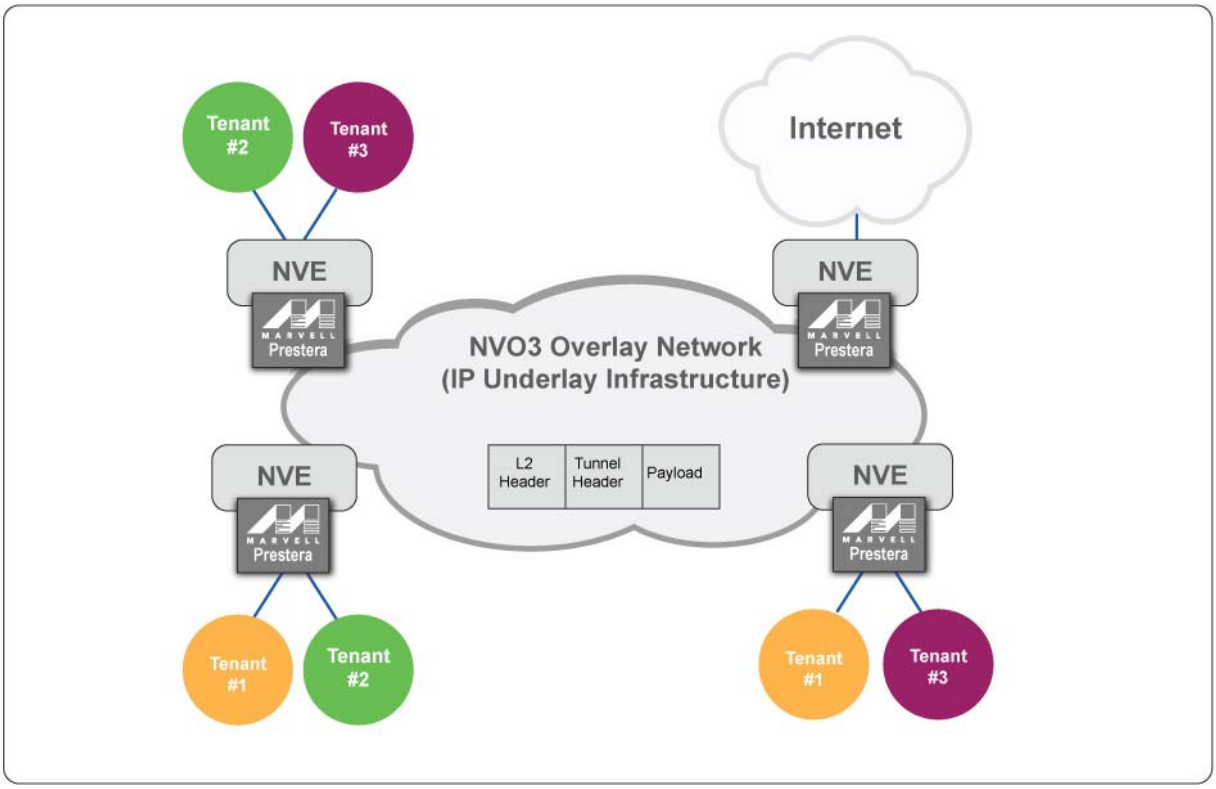


Network Virtualization

A *network overlay* creates a virtual network by decoupling the physical topology from the logical topology, allowing compute and network services to reside anywhere in the network topology and to be dynamically relocated as needed. In particular, network overlays are being deployed today to provide scalable and agile solutions for multi-tenancy services in large data center networks. Key requirements for multi-tenancy services include:

Requirement	Benefit
Traffic isolation between each tenant	Ensures no tenant's traffic is ever leaked to another tenant
Address space isolation between each tenant's address spaces	Enables tenants to possibly have overlapping address spaces.
Address isolation between the tenant address space and the overlay network address space	Enables VMs to be located anywhere in the overlay network and mobility to migrate to any new location, e.g. migrate across the IP subnets in the network.

The IETF NVO3 work group is chartered to define the architecture, control plane and data plane for L2 and L3 services over a virtual overlay network. The NVO3 architecture model is illustrated below.



The overlay network infrastructure is IP-based. The NVE (Network Virtualization Edge), which resides between the tenant and the overlay network, implements the overlay functionality.

For L2 service, the tenant is provided with a service that is analogous to being connected to an L2 bridged network. If a tenant's packet MAC DA is known unicast, the NVE tunnels the Ethernet frame across the overlay network to the remote NVE where the tenant destination host resides. Tenant BUM traffic is transported to all the remote NVEs attached to the given tenant, using either head-end replication or tandem replication.

For L3 service, the tenant is provided with an IP-only service where the NVE routes the tenant traffic according to the tenant virtual router and forwarder (VRF) and forwards the IP datagram over an overlay tunnel to the remote NVE(s).

In both L2 and L3 overlay services, the underlay network natively routes the IP traffic based on the outer IP tunnel header. The underlay network is unaware of the type of overlay service and payload type.

The Marvell eBridge architecture supports the intelligent and flexible processing required by the NVE to implement L2 and L3 overlay services. This includes support for the leading NVO3 encapsulation modes, VXLAN, VXLAN-GPE, Geneve, and flexibility for proprietary and future standards as well.

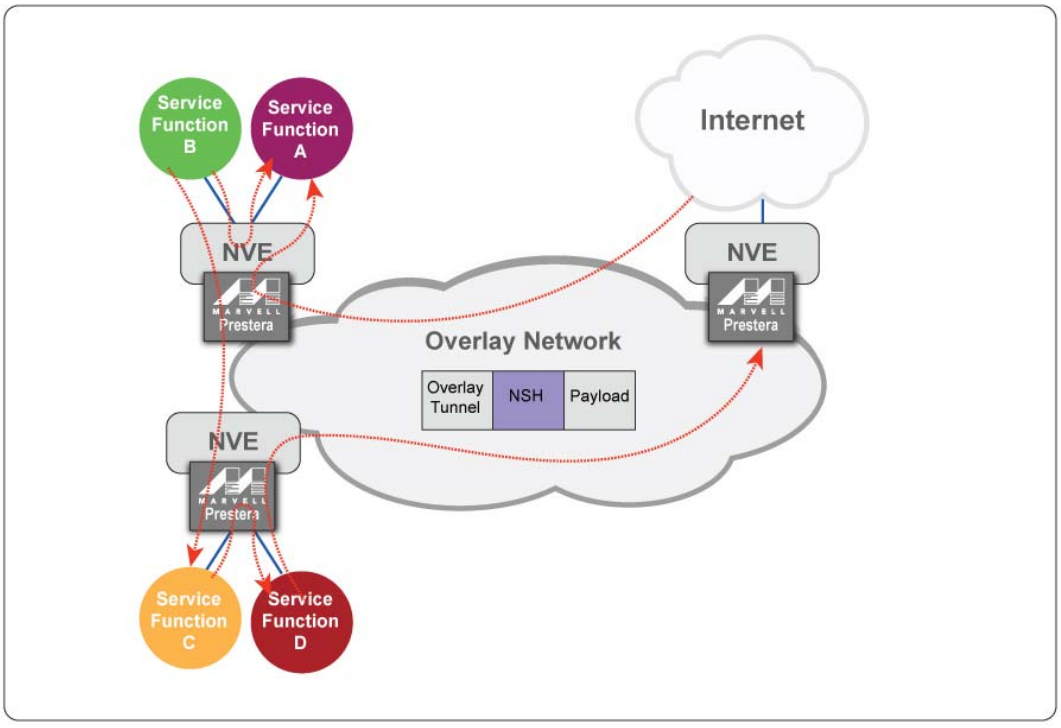
Network Function Virtualization & Service Function Chaining

To improve scaling and reduce OPEX/CAPEX, modern data centers and carrier networks are replacing dedicated network appliances (e.g. firewalls, deep packet inspectors) with virtualized network service functions running as an application in the server or VMs. This technology is called Network Functions Virtualization (NFV).

A “service function chain” (aka VNF Forwarding Graph) is an ordered set of network service functions that are required for a given packet flow, (e.g. Firewall ➔ Deep Packet Inspector ➔ Load Balancer). Packets are steered along the service path using an overlay encapsulation (e.g. VXLAN-GPE) between service functions. This allows the service functions to be located anywhere in the network (e.g. in server VMs), independent of the network topology.

The IETF Service Function Chaining work group is standardizing a service layer call Network Service Header (NSH.) NSH is transport independent, that is, it can reside over any type of overlay encapsulation (e.g. VXLAN-GPE, Geneve, GRE) or directly over Ethernet L2 header.

NSH contains the service path ID and optionally packet metadata. The service path ID represents the ordered set of service functions that the packet must visit. The metadata may be any type of data that may be useful to the service functions along the path.



The Marvell eBridge architecture supports the ability to classify and impose the Network Service Header (NSH), and encapsulate with any overlay tunnel header. At overlay tunnel termination points, the NSH header can be examined and a new overlay tunnel can be applied to transport the packet to the next service function. At the end of the service function chain, the NSH and overlay tunnel can be removed, and L2/L3 forwarding is based on the original packet payload.

Marvell eBridge Architecture

The legacy switching paradigm is based on the concept of physical interfaces, where incoming traffic is associated with the ingress physical interface, and outgoing traffic is forwarded to an egress physical interface.

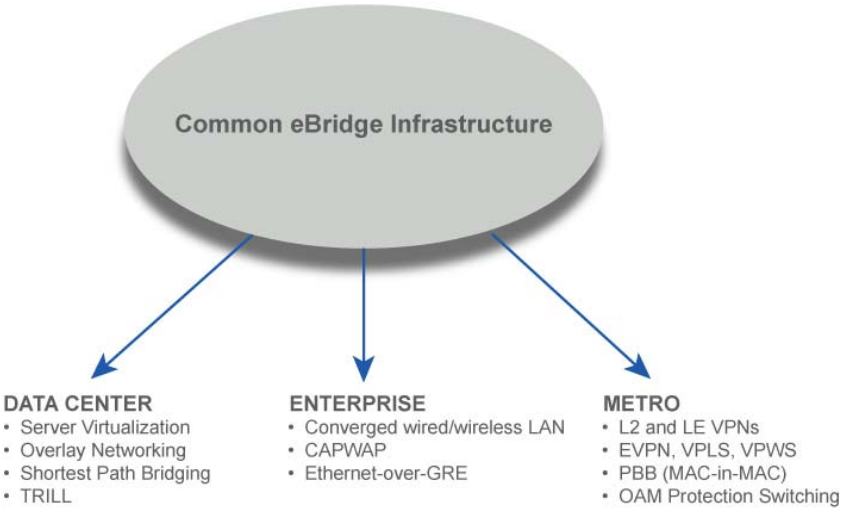
This paradigm worked well in basic layer-2 networks. However, in modern data centers, carrier networks, and converged wired/wireless networks, there are additional challenges, such as cross-tenant isolation, tunnel overlay encapsulations, learning and forwarding over virtual interfaces and per virtual interface attributes.

Today, system architects must design a complex software virtualization layer to hide the physical layer limitations from the end-user software. This software virtualization layer enables the application software to configure traffic flows, services, VMs and other logical entities.

The eBridge architecture provides a common scalable infrastructure that has been proven to meet the new challenges of virtualization. This architecture uses opaque handles to identify virtual interfaces (ePorts) and virtual switching domains (eVLANs). This unique implementation brings the hardware configuration driver view close to the application software view, thus significantly reducing the size and complexity of the required software virtualization layer, and with it, the development and debug efforts and resulting time-to-market of the system.

While supporting virtual interfaces, the eBridge architecture fully supports legacy network paradigms based on physical interfaces. In the legacy network paradigm, an ePort maps 1:1 with a physical port, and an eVLAN maps 1:1 with a VLAN bridge domain. Prestera devices integrating the eBridge architecture are fully backward-compatible with legacy packet-processors' processing pipe and feature set.

The eBridge architecture has been successfully applied to many of the emerging networking standards associated in different market segments, as illustrated in the figure below.



eBridge ePorts

The eBridge architecture utilizes, at the packet processor data plane level, an entity called *ePort* to represent a virtual interface. For example, an ePort may represent a physical port, Port-VLAN interface, 802.1BR E-Tag interface, VXLAN-GPE tunnel, Geneve tunnel, MPLS pseudowire, MAC-in-MAC tunnel, TRILL tunnel, etc. Ingressed packets are classified and assigned a "source ePort." Forwarding engines assign a "target ePort" indicating the virtual interface through which the packet is egressed. The ePort entity, source or target, is completely decoupled from the physical interface the packet is received on / transmitted to.

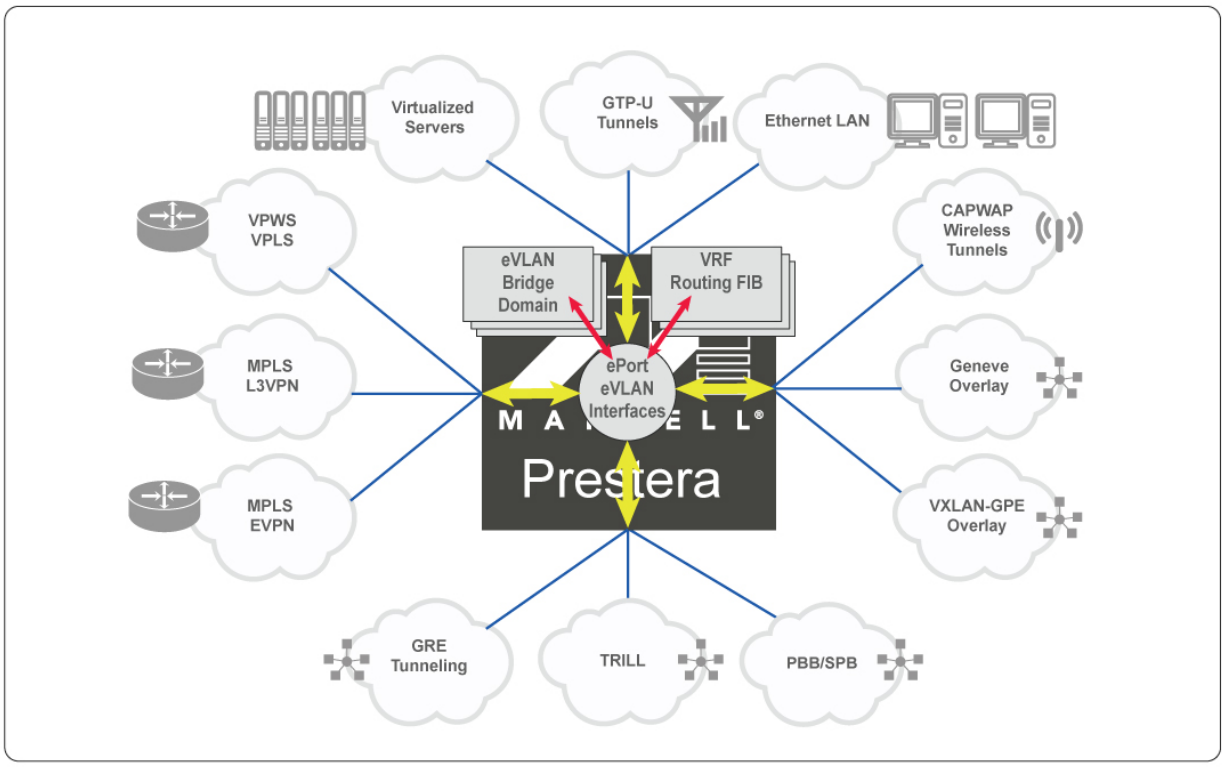
eBridge eVLANs

The eBridge architecture utilizes, at the packet processor data plane level, an entity called *eVLAN* to represent a virtual switching domain. In concept, this is similar to IEEE 802.1Q VLAN, but can be extended beyond 4K VLAN-ID range to support a large number of switching domains (e.g. per tenant) independent of the packet VLAN-ID.

Marvell eBridge Universal Gateway

Different network domains are virtualized using different technologies, e.g. server virtualization using 802.1BR, data center virtualization using VXLAN-GPE overlay, data center interconnect (DCI) using VPLS, branch office WAN connection using GRE tunnels, etc. To allow connectivity between these domains, traffic with dissimilar encapsulations must be bridged or routed from one domain to another.

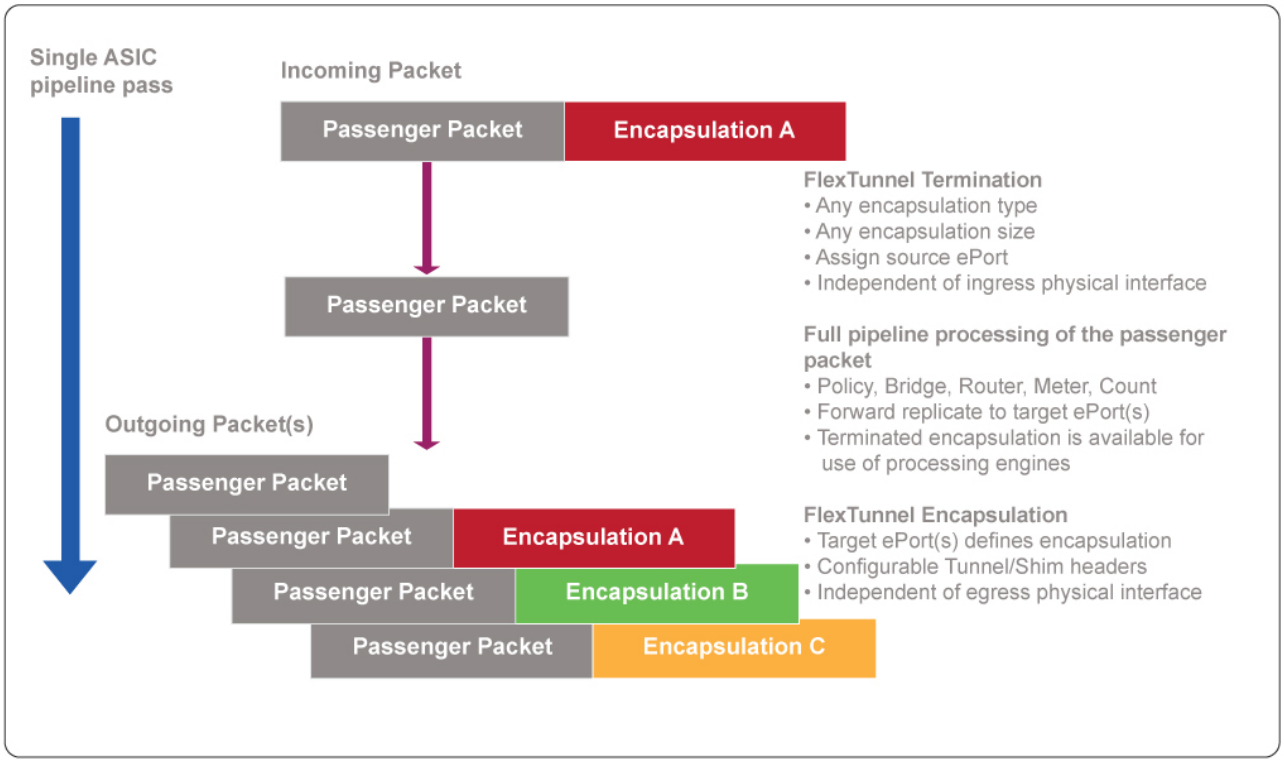
The eBridge architecture allows Marvell Prestera switches to serve as a universal gateway, seamlessly *stitching* between different virtualization domains and their respective data encapsulations.



The Prestera switching devices perform any-to-any stitching in a single pipeline pass, including BUM traffic replication, full passenger and/or encapsulation-based packet processing, such as policy TCAM rules, policing and metering, bridging, routing and more, with no performance impact.

The eBridge architecture supports single pass any-to-any stitching as follows:

1. The incoming packet encapsulation is classified, tunnel-terminated and assigned a source ePort representing the ingress virtual interface.
2. The payload is subject to the full pipeline processing. This includes policy TCAM rules, bridging based on the eVLAN bridge domain, routing based on the VRF, metering, counting, etc.
3. Unicast traffic is assigned a target ePort representing the egress virtual interface.
4. Multitarget traffic (BUM, routed IP multicast) is replicated to a set of target ePorts, each representing a different egress encapsulation, independent of the underlying physical interface.



The table below lists some common data center stitching use cases supported by Marvell eBridge architecture:

Use Cases	eBridge Support
Inter-VXLAN Routing	The incoming VXLAN traffic is tunnel terminated; the tenant IP packet is routed using VRFs and its L2 header is updated and the resulting Ethernet packet is re-encapsulated in a new VXLAN tunnel.
802.1BR Controlling Bridge to VXLAN overlay	E-tagged traffic is received from the server, bridged in the eVLAN domain, and egressed as VXLAN tunneled traffic over the IP core network. Similarly, VXLAN traffic is received from the IP core, bridged within the eVLAN domain and forwarded with E-Tags to the respective server VM(s).
Data Center Interconnect (DCI): VXLAN ←→VPLS/EVPN	VPLS/EVPN traffic received from the MPLS core network is tunnel-terminated; the passenger packet is bridged in the eVLAN domain and egressed as VXLAN tunneled traffic over the IP core network Similarly, VXLAN traffic received from the IP core network is tunnel-terminated; the passenger packet is bridged in the eVLAN domain and egressed as VPLS/EVPN tunneled traffic over the MPLS core network. Split-horizon filtering prevents traffic received from one core network from being looped back to the same core network.

About the Authors

David Melman
Marvell Switching Architect

David Melman is a 20-year veteran of the networking industry. For the past 15 years, Melman has been a switch architect at Marvell, involved in the definition of the Marvell Prestera® family of packet-processing devices. He is an active participant in the Internet Engineering Task Force (IETF), co-author of the IETF draft Generic Protocol Extension for VXLAN and contributor to the IETF draft *Network Service Header*.

Uri Safrai
Marvell Software Solution Architect

Uri Safrai has more than 17 years of networking experience. Prior to his current position, Safrai worked at Galileo Technology until its acquisition by Marvell in 2001. Since then he has held variety of technological positions at Marvell including his former role as a switch architect of the Prestera® line of packet processors, where Safrai was involved in the definition and micro-architecture of networking features, protocols and various ASIC engines and mechanisms. He also led the definition of the Prestera® eBridge architecture. Since 2010, Safrai represents Marvell at the Metro Ethernet Forum (MEF), and recently joined the Open Networking Foundation (ONF) Chipmakers Advisory Board (CAB).

Acronyms

BUM	Broadcast, Unknown Unicast, Multicast
CAPWAP	Control And Provisioning of Wireless Access Points
DCI	Data Center Interconnect
EVPN	Ethernet Virtual Private Network
GENEVE	Generic Network Virtualization Encapsulation
GRE	Generic Routing Encapsulation
MPLS	Multiprotocol Label Switching
NFV	Network Function Virtualization
NSH	Network Service Header
NVE	Network Virtualization Edge
PBB	Provider Backbone Bridging
SFC	Service Function Chaining
TRILL	Transparent Interconnect of Lots of Links
VEPA	Virtual Ethernet Port Aggregator
VLAN	Virtual Local Area Network
VM	Virtual Machine
VNF	Virtual Network Function
VPLS	Virtual Private LAN Service
VPWS	Virtual Private Wire Service
VXLAN-GPE	Virtual Extensible LAN - Generic Protocol Encapsulation

References

An Architecture for Overlay Networks (NVO3)
https://datatracker.ietf.org/doc/draft-ietf-nvo3-arch/

Generic Protocol Extension for VXLAN
https://datatracker.ietf.org/doc/draft-quinn-vxlan-gpe/
Geneve: Generic Network Virtualization Encapsulation
https://datatracker.ietf.org/doc/draft-ietf-nvo3-geneve/

Network Service Header
https://datatracker.ietf.org/doc/draft-ietf-sfc-nsh/

802.1BR - Bridge Port Extension
http://www.ieee802.org/1/pages/802.1br.html

802.1Qbg - Edge Virtual Bridging
http://www.ieee802.org/1/pages/802.1bg.html