SCTE CABLE-TEC EXPO 23
CELEBRATING 40 YEARS

2023 Fall Technical Forum
SCTE • CABLELABS • NCTA

Hyperscale Virtual Services Gateway

A Technical Paper prepared for SCTE by

Carl Klatsky
Senior Principal Engineer
Comcast
1800 Arch St., Philadelphia, PA 19103
+1 215 286 8256
carl_klatsky@comcast.com

DeFu Li
Distinguish Engineer
Comcast
1800 Arch St., Philadelphia, PA 19103
+1 267 260 3704
defu_li@comcast.com

Jason Combs
Senior Principal Engineer
Comcast
Jason_combs@comcast.com

Anton Grichina
Senior Software Engineer
Harmonic
anton.grichina@harmonicinc.com

Adam Levy
Director System Architecture
Harmonic
19 Alon Hatavor St., P.O.B 3600
Caesarea Industrial Park 3088900
Israel
adam.levy@harmonicinc.com

Table of Contents

List of Figures

1. Introduction

The Virtual Cable Modem Termination System (VCMTS) and Virtual Services Gateway (VSG) are newer network elements that have quickly become vital to running a large MSO network. The VCMTS disaggregates the traditional Cable Modem Termination System (CMTS) functions across a generalized computing structure. The VSG is in the logical path of the customer data packets enabling usage and quality analysis use cases. The next step is to integrate these two virtualized network elements.

The authors will outline an approach to building a hyper scale VSG that can scale in proportion to future needs. This approach uses a Smart Network Interface Card (NIC) with programmable network function accelerators for the platform's physical connectivity along with a control plane that utilizes the Linux SwitchDev model and Linux Traffic Control (TC) flows to direct traffic between the VCMTS and VSG systems. The approach allows both elements to rapidly scale up by simply increasing the capacity of the switching fabric and adding more compute nodes.

2. System Overview

2.1. Virtual Cable Modem Termination System

VCMTS is a collection of software applications implementing the Converged Cable Access Platform (CCAP) core functions according to the Distribute Access Architecture (DAA) specification, also known as the Modular Head-End Architecture version 2 (MHAv2) [1].

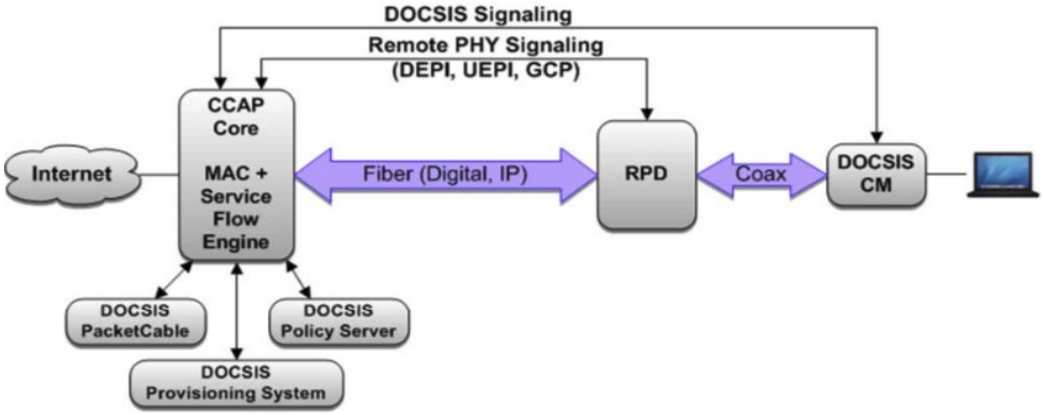

Figure 1 - MHAv2

This suite of software applications is built and deployed on the Edge Cloud Platform. As a software CMTS, all packet processing, encapsulation, encryption, and Data Over Cable System Interface Specification (DOCSIS) protocols are implemented with software running over Commercial Off-The-Shelf (COTS) servers. This approach provides significant power savings, cost reduction, flexible scalability, and performance increases as each generation of Central Processing Units (CPU) becomes more powerful, with very little to no additional development costs. The VCMTS architecture has maintained full independence from hardware from the beginning, while also taking advantage of. opportunities for offloading expensive operations onto the hardware components it runs over.

2.2. Virtual Services Gateway

The Virtual Services Gateway (VSG) provides a new platform where next-generation network services are deployed close to the customer edge. Automation and flexible deployment best practices are used to simplify the creation of new services and augment the capabilities of existing network services. Current state-of-the-art Software-Defined Networking (SDN) and Network-Function-Virtualization (NFV) technologies are used to enable agile implementation and orchestrated management and monitoring.

The VSG is a software application capable of running on general-purpose compute nodes, either virtual or bare metal. It can share platforms with the VCMTS or other virtual elements deployed at the network edge. Initially placed in line with customer traffic, the performance of the VSG is critical to improving the customer experience. It is a low-latency flow-through platform that can forward traffic at a line rate, ensuring the introduction of the VSG does not cause a performance bottleneck.

The VSG is introduced between the CMTS/VCMTS and Residential Edge Router (RUR) in Headend or Hub sites where there is access to customer traffic near the customer's premises.

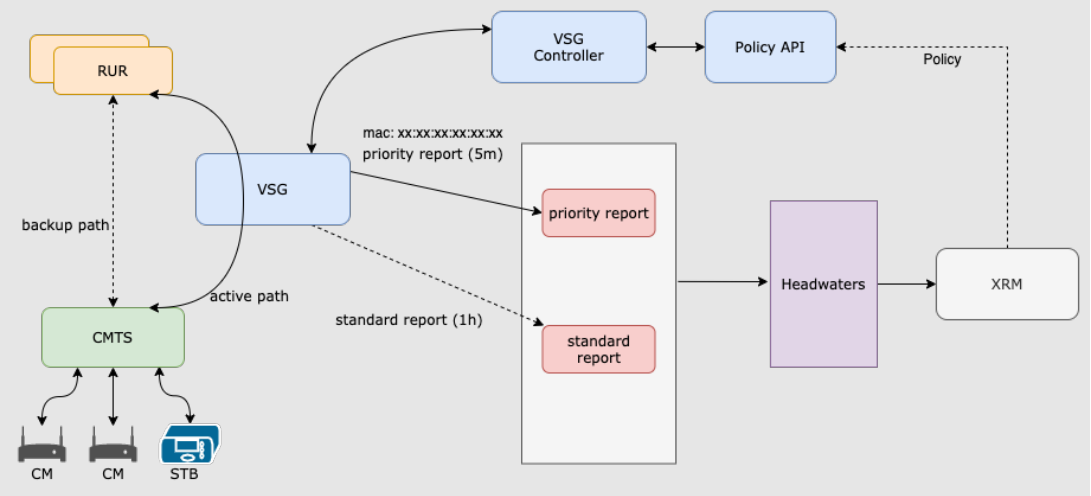

Figure 2 - VSG High Level

Potential use cases for the VSG are Usage Based Billing (UBB) and customer Quality of Experience (QoE) measurements.

For the UBB use case, the VSG platform needs to collect the usage byte information for each customer connected to the network. The VSG platform can be deployed in head-ends where it is able to collect and aggregate upstream and downstream byte usage per IP family per DSCP in regular intervals for each cable modem device. A mediation layer is needed to consume the byte usage, apply any needed business rules, and publish the data to back-office systems such as the Xfinity Resource Manager (XRM). This same data set can be used for capacity planning, usage trend analysis, and other analysis.

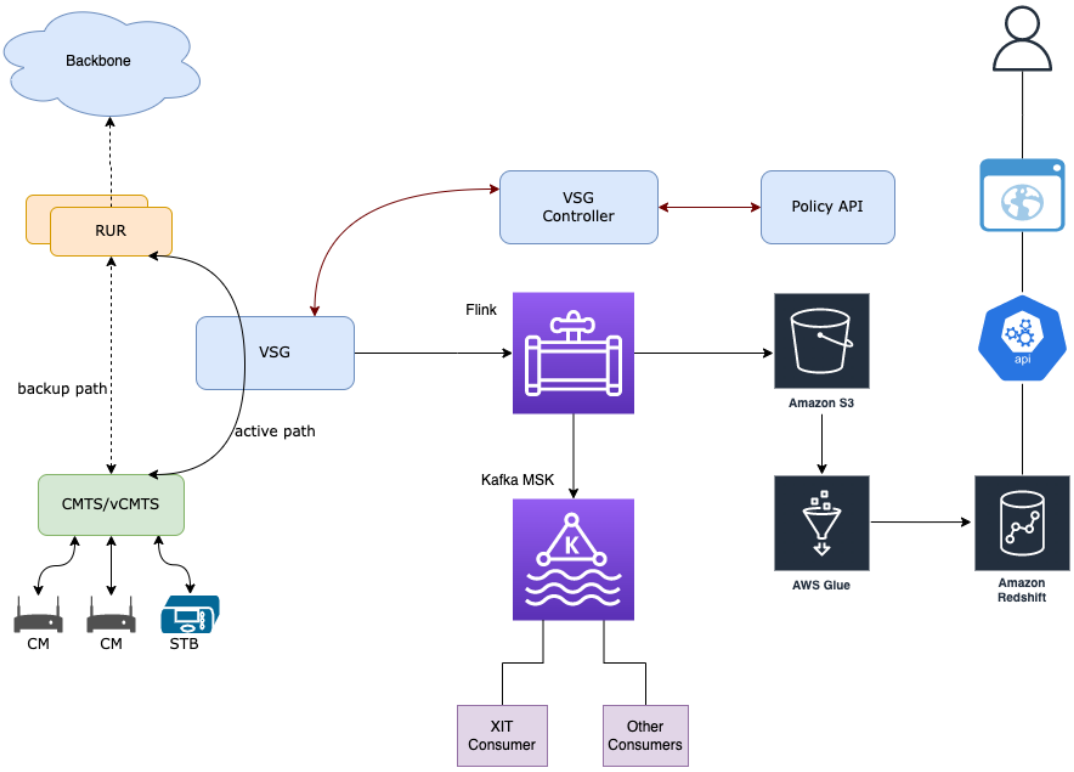

Figure 3 - VSG Reporting Infrastructure

The VSG Quality of Experience use case is to collect and derive the QoE of residential High-Speed Data (HSD) services provided by the MSO. The VSG is intended to be an IP flow-aware system. As such, it is able to analyze key Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) performance and network latency statistics. Statistics are able to be captured for each IP flow, aggregated on a per-customer basis, and streamed to back-office systems. The VSG back-office components are able to tie the collected performance data with other meta-data pulled from various available data sources to calculate the *Quality of Experience* for the HSD customer.

QoE is calculated based on multiple network performance parameters, including latency, bandwidth consumption, and packet retransmission. Thresholds are identified from research and data analysis over a period of time.

2.3. SmartNIC

In recent years, the industry term "SmartNIC" has been coined to describe a NIC (Network Interface Card) that includes advanced switching and/or processing offload capabilities. Some SmartNICs even include a CPU as an on-board component. The term SmartNIC is not defined by any standards body such as the Institute of Electrical and Electronics Engineers (IEEE) or the Association for Computing Machinery (ACM), and as such, has differing definitions as presented by the various NIC vendors. With no clear standards-based definition, the selection of a suitable NIC for various networking applications is quite a challenge.

Within the context of this paper, the authors use this definition of a SmartNIC: Any NIC with network switching offload capability to modify, copy, redirect, or forward packets in hardware, regardless of the presence of additional on-board CPU.

3. Challenges & Evolution

3.1. VCMTS & VSG Challenges Requiring SmartNIC Capabiltities

In running the Edge Cloud Platform, the primary business challenge comes in the form of ever-increasing bandwidth and scale.

Service Group (SG) bandwidth capacity continues to increase as part of the ongoing Radio Frequency (RF) plant and access technology upgrades. For example, going from 96 MHz to 192 MHz full band Orthogonal Frequency Division Multiplexing (OFDM) channels, Low-split to Mid-split RF plant, Full-Duplex DOCSIS (FDX), node splits, and so on. These are driven by subscriber growth, network demand increases, and new service offerings for the multi-gigabit symmetric speed-tiers. The VCMTS workloads running on the servers need to scale accordingly.

To attempt to keep up with this growth, each generation of the server rack built for the Edge Cloud Platform has a specific subscriber scale and network bandwidth capacity target. When transitioning to a newer generation server rack build, the number of subscribers is typically doubled, and the network bandwidth supported is more than doubled. For example, in the most recent generation, the change has been from 10 Gigabit per second (Gbps) ethernet connectivity per compute host to 100 Gbps ethernet connectivity—to allow for both the multiplexing of high bandwidth services to individual customers as well as continued general usage growth.

This increase in both subscriber scale and bandwidth utilization presents a problem for maintaining the VSG function. Prior to this generation, standalone bare metal servers were able to scale vertically to support 400 Gbps and thereby keep up with demand. However, this is proving to be unsustainable with the latest generation of the Edge Cloud Platform; this presents a clear need to be able to steer traffic more granularly to multiple VSG servers. Existing NIC solutions have many shortcomings in network traffic steering functions which need to be addressed. Figure 4 shows an example of the host CPU core allocation scheme for the deployment of VCMTS pods. The VCMTS Data Plane (DP) has dedicated CPU cores, while the VCMTS Control Plane shares the remaining CPU cores. The data and control plane traffic relies on the NIC for steering to their targeted Virtual Functions (VF) or Physical Functions (PFs) to reach the intended network interfaces of the VCMTS pods.

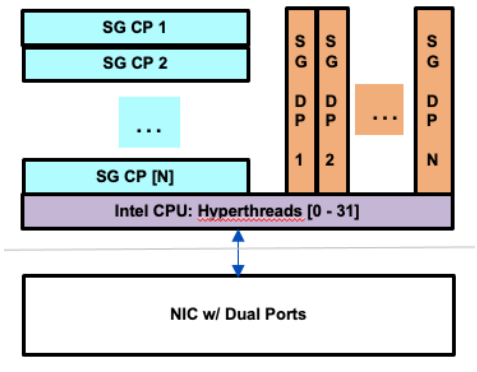

Figure 4 - An Example of VCMTS CPU Core Allocation Model

The proposed solution, detailed in a later section, mirrors and tunnels the traffic from the VCMTS to the VSG utilizing the more capable SwitchDev model. A SmartNIC with programmable network function accelerators provides this ability to address the above challenges without sacrificing VCMTS throughput performance.

In the VSG system, a SmartNIC is needed to perform the Virtual Local Area Network (VLAN) translation and packet copy functions. As seen in Figure 5 - VSG VLAN Interconnection and Figure 6 - VSG NIC Internal View, the VSG is physically wired to the redundant routers with one 100 Gbps link to each router. Each link supports both upstream and downstream traffic. For the VSG to properly analyze the subscriber QoE, it is critical that the VSG application has some parameter to indicate upstream traffic from downstream traffic; this is achieved by VLAN ID or "VLAN tags" to distinguish the directionality of the traffic as it traverses the VSG. Traffic to and from the VCMTS is marked with a unique VLAN tag. Using the router's patch panel function, the VLAN tagged packets are re-directed through the router between the VCMTS port facing and the VSG-facing port. At the VSG, the VSG NIC makes a copy of the packet to send to the VSG application and changes the VLAN tag of the original packet for transmission back to the RUR and onwards upstream to the Aggregation Routers (AR). From the perspective of the router, it is only aware of the VLAN tag value after it has been modified by the VSG. The router has a layer three sub-interface corresponding to this VLAN tag value and proceeds to route the traffic onto the network as per standard layer three routing functions. The same process occurs in reverse for downstream traffic.

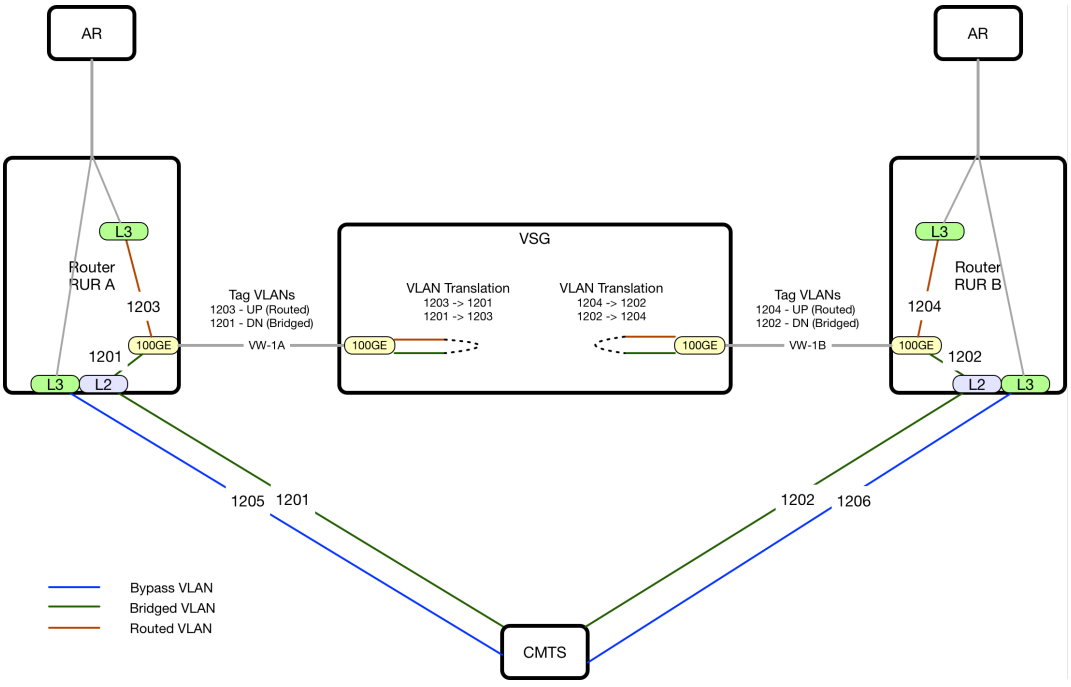

Figure 5 - VSG VLAN Interconnection

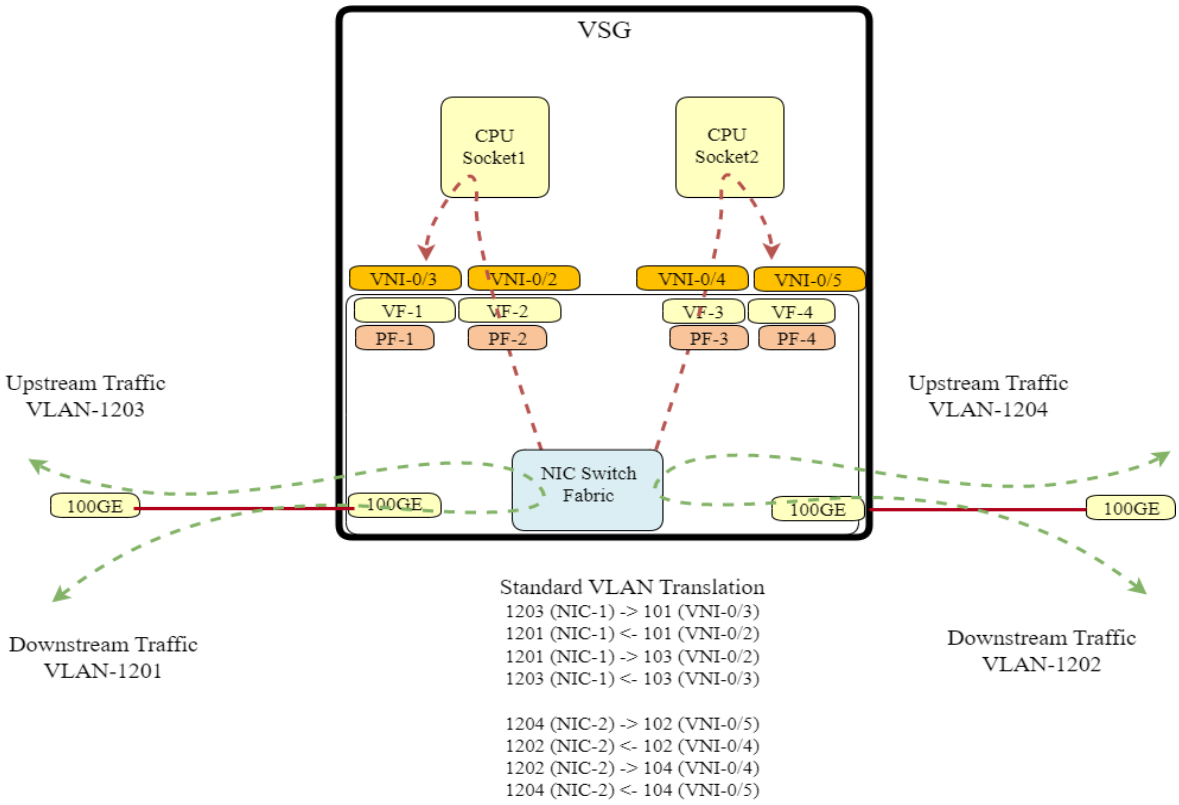

Figure 6 - VSG NIC Internal View

3.2. Evolving the VCMTS

As business needs evolve, the VCMTS needs to evolve to meet those needs. New capabilities are needed, such as increased throughput and improved packet processing offload.

A true SmartNIC with refined traffic steering and increased performance is required—one that supports a higher number of virtualized functions and more elaborate traffic management mechanisms.

Single Root-Input / Output Virtualization (SR-IOV) is an industry standard for NIC virtualization; it allows a physical device to be used simultaneously from different applications or virtual machines. Each user gets one or multiple VFs which are logical Peripheral Component Interconnect express (PCIe) devices with no performance overhead. SR-IOV is standardized and supported by different hardware vendors. SR-IOV can be utilized in a VCMTS.

An advanced traffic management technique is also needed to efficiently re-direct packets through the NIC. SwitchDev mode provides the needed traffic management technique. SwitchDev is an open-source infrastructure construct in the Linux kernel that controls how a SmartNIC processes incoming and outgoing packets. It standardizes a configuration mechanism for implementing SR-IOV. Further, it enables using the NIC as an e-switch and embedded switch within the NIC.

While with SR-IOV, each application is connected directly to the network via a VF, with SwitchDev the VF is connected to the eSwitch in the NIC. The SwitchDev mode enables more advanced schemes of communication between applications over VFs. The NIC becomes a configurable switch with extended functionality. It is possible to forward packets between “ports” according to a defined configuration. The

NIC utilizes two forwarding layers from the port to the e-switch and from the e-switch to the VFs. This additional e-switch layer makes the system more complex, but it also allows for greater flexibility. It also forces multi-tenant security by enforcing the e-switch steering to be configured from the host.

One example of a host-based infrastructure which uses SwitchDev is Open vSwitch (OVS). OVS is a mature virtual switching software widely used by telecom operators. OVS provides control plane and data plane functionality, and with SwitchDev integration, its data plane could be offloaded to the SmartNIC, which tremendously increases throughput while minimizing CPU load.

A SmartNIC is able to support OVS offloading via SwitchDev. The dynamic nature of OVS – learning from packets sent to the software – is superfluous for the VCMTS application as it already has all the information about all connected hosts. However, OVS provides a standard implementation and consistent behavior, and is therefore worth considering.

A SmartNIC enables many new opportunities, and a VCMTS implementation with a SmartNIC should try to balance between hardware offload opportunities and reduced complexity for best-in-class product stability and performance.

3.3. Evolving the VSG

In a typical deployment, a VSG needs the capability to perform 100 Gbps line rate packet copy and VLAN translation functions. A suitable VSG NIC did not include any additional onboard CPU for generalized compute functions, but as originally envisioned, the VSG does not need that functionality. The needed VLAN translation is shown above in Figure 5 - VSG VLAN Interconnection. The NIC needs to change the VLAN ID in all received packets based on the directionality of the packet. The NIC also needs to provide a packet copy function to send a copy of all received packets to the VSG software application.

To integrate a VSG more closely with the VCMTS and to do so at scale, the required packet copy and VLAN translation functionality will need to split across the two systems. The VCMTS SmartNIC will need to perform the packet copy function, plus an additional function to tunnel the copied packets to the VSG using the Virtual eXtensible LAN (VXLAN) [2] tunneling protocol. The VSG will then need to receive and decapsulate the tunneled packets, translate the VLAN ID as needed, and forward the packet to the VSG software application. A SmartNIC with hardware offload support for the VXLAN protocol decapsulation function is needed. This SmartNIC also needs to provide a line-rate VLAN translation function with hardware offload support.

4. Integrated VCMTS-VSG Solution

4.1. Integrated System

The core of the VCMTS is a collection of software components deployed in a highly available cluster. Kubernetes is an integral part of the system, playing a key role in the management, monitoring, and provisioning of various applications. Kubernetes isolates components from each other into multiple pod instances, enabling horizontal scaling.

Packet processing applications based on the DPDK library, such as the VCMTS, must have direct access to the network to minimize latency and increase throughput without intermediate software components. The SmartNIC provides two mechanisms for network direct access.

The first mechanism is SR-IOV, which allows the splitting of a single physical PCIe device into multiple virtual PCIe devices. Each packet processing application pod is allocated with one or more VFs, used as a networking interface.

The second network direct access mechanism provided by the SmartNIC is traffic steering, the ability to maintain steering rules that determine which traffic will be forwarded to each VF.

NIC traffic steering is either vendor-specific or based on standard APIs, such as rte_flow, SwitchDev, or OVS. SwitchDev is an infrastructure construct provided by recent Linux kernel versions which enables traffic steering configuration via the "tc" utility. The result of "tc" commands are the configured traffic steering rules. A traffic steering rule consists of a matcher, which matches specific packets and actions, which defines what the NIC will do with the matched packets. There are two classes of steering rules: receive and transmit rules. Receive rules define which packets will be sent from the NIC to a specific VF; transmit rules define how to process packets sent from the VF to the NIC.

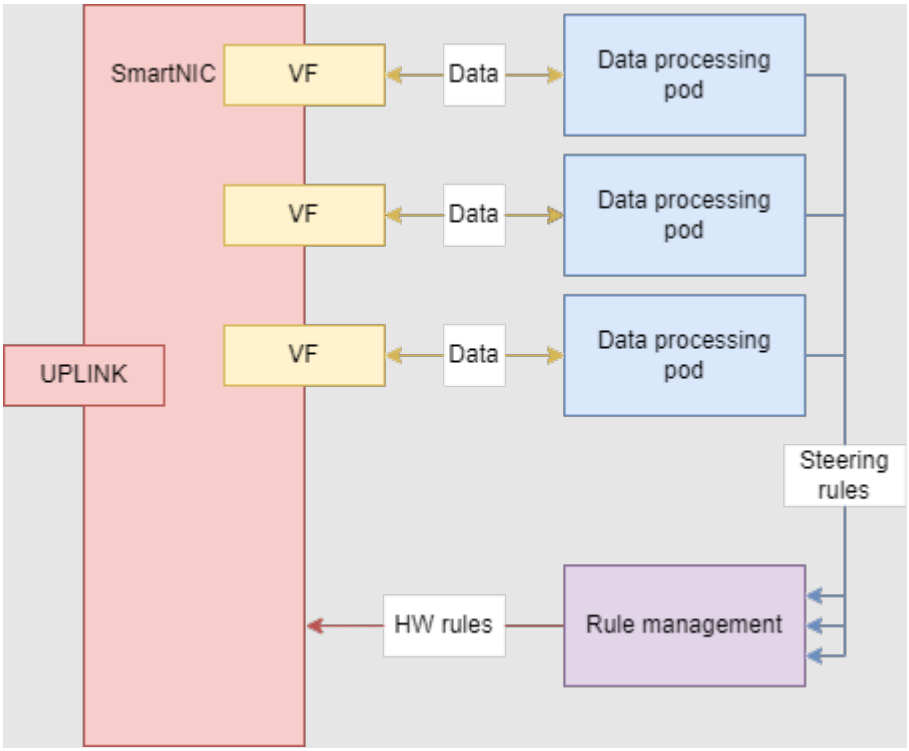

Figure 7 - Traffic Steering in VCMTS

With the development of the next-generation Edge Cloud Platform architecture solution, one of the new requirements has been to support any scale. Instead of relying on the VSG server being in a specific network location and mirroring specific ports, the VCMTS now takes over the mirroring of all traffic towards the VSG server instances, thus reducing the dependency on the network topology almost completely.

The VSG application requires all downstream and upstream traffic to be copied, encapsulated, and sent to the VSG server. Steering rules are used to separate, mirror, and encapsulate traffic into a VXLAN tunnel toward the VSG server.

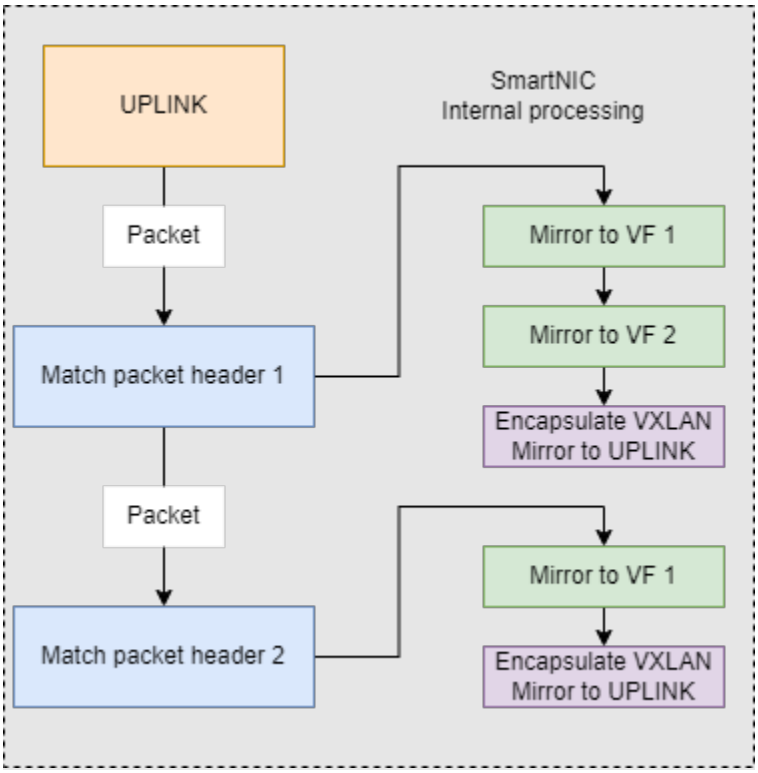

Figure 8 - VSG Replication & Encapsulation Rules in VCMTS

The mirroring and encapsulation could be implemented with software on the CPU. The memory and computing required for mirroring and encapsulating the entire VCMTS traffic is extremely costly, even when run on modern CPUs, with a potentially considerable performance hit.

With SmartNIC offloading enabled, there is no CPU performance penalty. The potential performance improvement is a clear driver towards implementing packet processing applications with SmartNIC HW offloading.

Potential disadvantages with SmartNIC usage are few but should be mentioned:

- Encountering vendor-specific behavior is probable, even when using standard APIs or open-source implementations. Configuring the same feature with identical API calls may result in different behavior across multiple NIC vendor implementations.
- The ability to debug issues is limited, a natural consequence of using hardware offloads. Debugging hardware is harder than debugging a software-only implementation.

4.2. Ability to Scale the System

The deployment philosophy is to “build it once and deploy it at any scale.” Sizing and scaling are part of the deployment plan to match the number of VCMTS pods that can be served by a single VSG bare metal server. If more VSG capacity is needed, we simply add more VSG bare metal servers and connect it to the switch fabric.

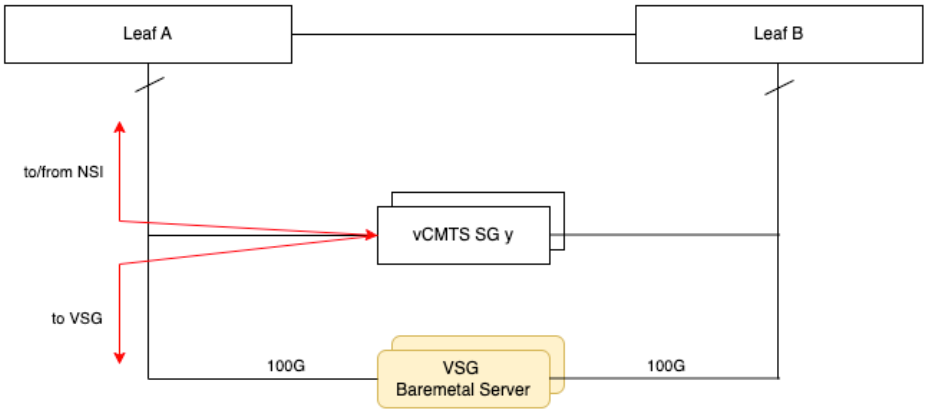

Figure 9 - Baremetal VSG Deployment View

The VCMTS pod to the VSG bare metal server mapping can be dynamically provisioned by the service orchestration layer. Together, the VCMTS pods and the steering operator are able to manage the SmartNIC traffic steering rules and mirroring functions. The ability to steer and mirror VCMTS traffic is, in essence, the ability to partition the service; the system scale is no longer bound by a single VSG bare metal server's limitation. Therefore, the system scales horizontally.

4.3. Future Plans for VSG Containerization

The Edge Cloud Platform can have a tremendous amount of computing and network resources. The next level of sophistication is to build the system such that it is elastic and can auto-scale. These aspects enable the systems to automatically adjust the provisioned network and compute resources to match the current demand.

Future ideas for integrating the VSG with the VCMTS may involve "containerizing" the VSG application to run on the Edge Cloud Platform. The VSG application function may be disaggregated into its constituent functions, potentially with each function running in its own container. There are several possibilities on how to distribute the VSG container workload across the control and data compute nodes of the edge cloud. The key requirements for this solution are:

- The ability to create and destroy the VSG application containers automatically and/or on request, scaling the system as demand grows and subsides.
- The ability to copy and forward data plane packets to the appropriate VSG application container
- The ability to copy and forward VSG control plane packets amongst the VSG application containers

One potential solution is shown in Figure 10 - Containerized VSG & VCMTS Deployment View. In this solution, the VSG pod supports all the needed VSG application containers running using available compute from a VCMTS "worker" node. The VSG pod supports both the VCMTS pod co-resident on the same compute and a VCMTS pod running on separate compute nodes. A variation on this solution is to only copy the packet headers instead of the full packet. This potential solution will reduce the overall link input/output bandwidth needed to support the packet copy function.

The key takeaway for these solutions is the capability of the SmartNIC to modify, copy, redirect, or forward packets both across and within the supported compute nodes. The SmartNIC in use also provides

packet encryption using hardware offload, so it is possible to encrypt the VXLAN tunneled traffic between the VCMTS and VSG if desired.

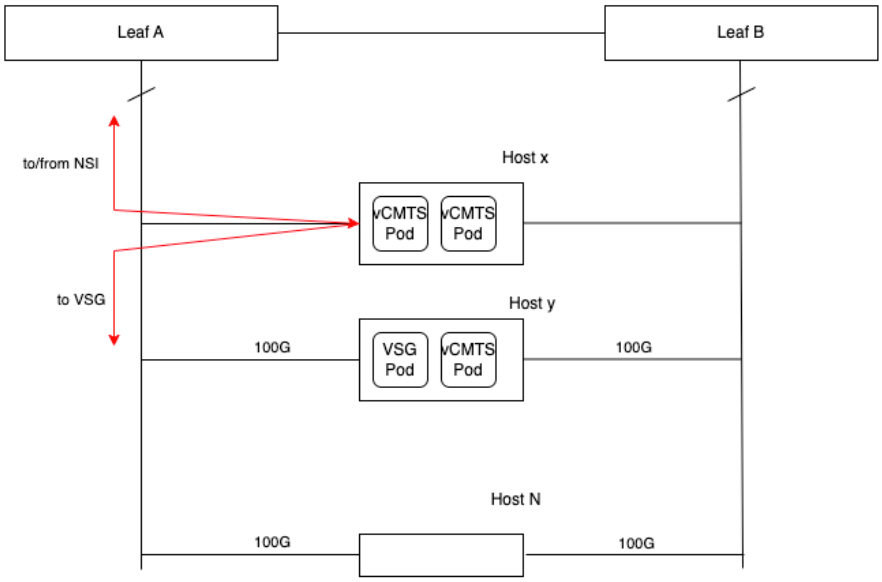

Figure 10 - Containerized VSG & VCMTS Deployment View

5. Conclusion

The VCMTS and the VSG platforms may be developed to achieve new functionality to meet their respective goals. They are able to scale to a point where the next step must be taken together to continue to grow, which is the creation of an Edge Cloud Platform. This platform is modeled after a traditional elastic cloud infrastructure but modified to the unique requirements of the edge and access networks. As such, it must be capable of hyper-scaling automatically and horizontally while maintaining the performance and demands of all the services provided by its individual components.

In this paper, the authors have laid out:

- The function and architecture of the vCMTS and VSG systems
- Introduced the SmartNIC and its capabilities
- Describe challenges that may require SmartNIC functionality
- A proposed architecture for implementing a horizontally scaled solution
- Additional future improvements to move toward and improve the Edge Cloud Platform

The edge cloud is the future of the access network. The authors believe the potential solution documented in this paper is not only the right solution for hyper-scaling the VSG but has defined a roadmap for incorporating other edge and access services into that same platform seamlessly and indefinitely scalable.

Abbreviations

ACM	Association for Compute Machinery
AR	Aggregation Router
COTS	Commercial-Off-The-Shelf
CCAP	Converged Cable Access Platform
CPU	Central Processing Unit
DAA	Distributed Access Architecture
DOCSIS	Data Over Cable System Interface Specification
DP	Data Plane
FDX	Full Duplex DOCSIS
Gbps	Gigabit per second
HSD	High Speed-Data
IEEE	Institute of Electrical and Electronics Engineers
LAN	Local Area Network
M-CMTS	Modular-Cable Modem Termination System
MHA	Modular Head-end Architecture
NFV	Network Function Virtualization
NIC	Network Interface Card
OFDM	Orthogonal Frequency Division Multiplexing
OVS	Open vSwitch
PCIe	Peripheral Component Interconnect express
PF	Physical Function
QAT	Quick Assist Technology
QOE	Quality of Experience
RF	Radio Frequency
RUR	Residential U-ring Router
SDN	Software Defined Network
SG	Service Group
SR-IOV	Single Root - Input / Output Virtualization
TC	Traffic Control
TCP	Transmission Control Protocol
UBB	Usage Based Billing
UDP	User Datagram Protocol
VCMTS	Virtual Cable Modem Termination System
VF	Virtual Function
VSG	Virtual Services Gateway
VXLAN	Virtual eXtensible LAN
XRM	Xfinity Resource Manager

Bibliography & References

[1] Data-Over-Cable Service Interface Specifications MHAv2 Remote Downstream External PHY Interface Specification, CM-SP-R-DEPI-I16-210804, August 4, 2021, Cable Television Laboratories, Inc.

[2] IETF RFC 7348 https://datatracker.ietf.org/doc/html/rfc7348