White Paper

Using Hardware to Improve NFV Performance

Prepared by

Roz Roseboro
Senior Analyst, Heavy Reading
www.heavyreading.com

on behalf of

www.mellanox.com

July 2015

Introduction

Network functions virtualization (NFV) brings a change in the platform used to deliver telecom network functions. Rather than employing tightly-coupled dedicated hardware and software, which tend to be static and difficult to scale, standardized IT servers run virtualized network functions (VNFs) to provide enhanced elasticity and scalability.

The most common implementation will add a hypervisor layer to manage the virtual machines (VMs) that make up the VNFs. In order to meet the needs of communications service providers, this new architecture needs to provide elasticity and scalability without compromising the reliability and performance of dedicated platforms.

Performance is impacted by a variety of factors: virtualization itself, the use of network overlays and the networking protocols of Transmission Control Protocol and Internet Protocol (TCP/IP). Using software acceleration, such as Data Plane Development Kit (DPDK) libraries, is one approach to improving performance. Another is hardware acceleration, which leverages SR-IOV to enable application direct access. Next-generation NICs are able to offload some of the overhead associated with encapsulation techniques, such as VXLAN and GENEVE. Remote Direct Memory Access (RDMA), which introduces the concept of "zero-copy" into the networking domain, is emerging as an alternative to TCP/IP transport.

This white paper is structured as follows:

- **Section II** discusses the impact the move to NFV has in the areas of reliability, elasticity, scalability and performance, including an explanation as to why performance is particularly critical.
- **Section III** explains the different dimensions of server performance. It explains why it is important to consider both raw interface speed, as well as packet throughput, when considering a server platform.
- **Section IV** explores the technologies being used to help improve server performance. It discusses how to overcome the penalties that arise from compute virtualization, overlay networking and TCP/IP.

Move to NFV Leads to New Requirements

The move to NFV requires a shift from purpose-built hardware with tightly coupled network function software to standard IT server platforms. Because servers were not optimized for network functions, they face reliability, elasticity, scalability and performance requirements that differ from those needed in an IT environment.

Reliability

Because the IT servers that run VNFs were not designed for carrier-grade high availability, the focus in NFV moves from five-nines (99.999%) *hardware* availability to five-nines *service* availability. The NFV infrastructure (NFVI), VNF software and management and orchestration (MANO) together will ensure service availability. Indeed, in its NFV Resiliency specification, the European Telecommunications Standards Institute (ETSI) states: "As outlined in the problem description, the key objective is to ensure service continuity, rather than focusing on platform availability."

Elasticity & Scalability

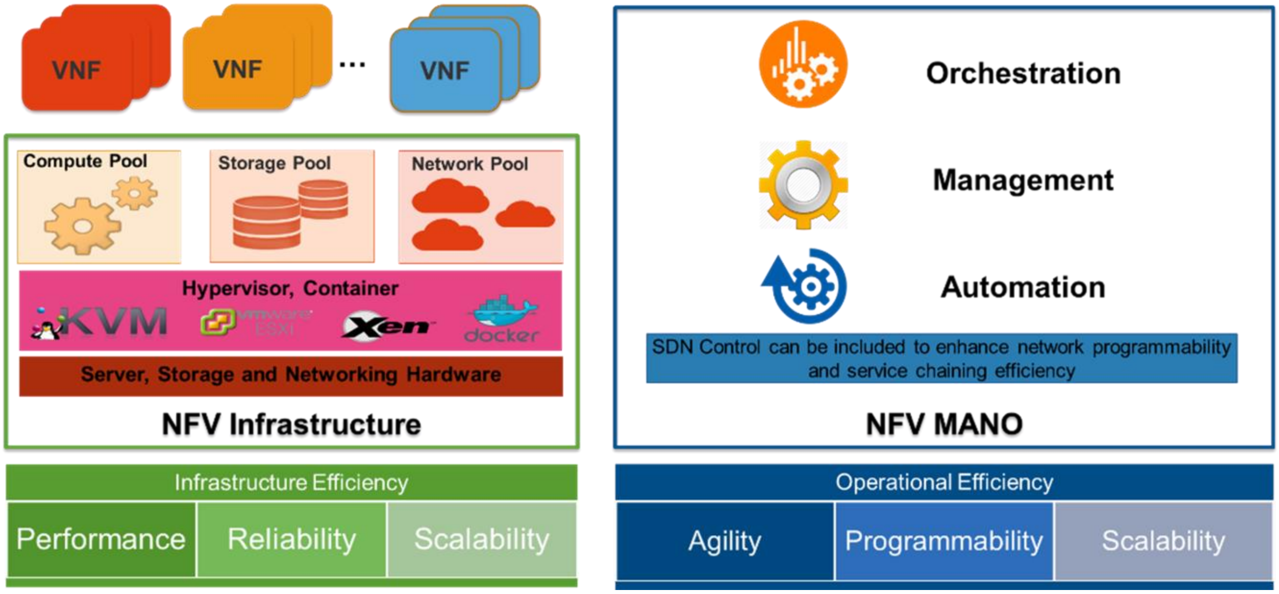
Moving network functions to the cloud is the ultimate goal of NFV. Once in the cloud, service creation can be automated and elastically scaled up and down based on demand. The cloud also allows for resources distributed across various locations to be pooled and managed as a single resource, which helps maximize utilization. This is particularly important to achieve a cost base comparable to hyperscale cloud competitors, as well as match their agility and service reach.

Performance

Telecom applications, unlike most IT applications, are rarely best-efforts services. Voice and video applications, for example, are particularly delay sensitive, often requiring less than 100 ms of latency. Performance is further challenged by the need to serve tens of millions of customers in a telecom environment.

When evaluating a virtualized solution, communications service providers are expecting similar performance as its appliance counterpart so that they can maintain a similar device footprint. This means a physical server might need to support 20 Gbit/s throughput for an Evolved Packet Core (EPC) application – significantly higher throughput than is needed for typical IT applications.

Figure 1: NFV Infrastructure & MANO



Source: Mellanox

Service providers' customers demand – and are generally willing to pay for – high-quality services, and use service-level agreements (SLAs) to ensure their requirements are met. These service providers must ensure that performance levels remain the same when they move into virtualized and cloudified environments.

The remainder of this paper will focus on performance and some of the approaches available to enable performance parity in a virtualized environment.

Multiple Dimensions to Server Performance

When considering performance, it is important to recognize that it can be measured multiple ways.

The most fundamental measure is **raw interface throughput**, which is normally measured in gigabits per second (Gbit/s). This describes the total amount of traffic that can be supported on a given input/output (I/O) device – e.g., network interface card (NIC) or adapter. Today, 10 Gbit/s NICs are common, but supporting NFV in the cloud will require 40 or 100 Gbit/s. In the interim, newly standardized 25/50 Gbit/s interfaces that can provide 150 percent more throughput at a competitive price point compared to 10 Gbit/s will be available.

What is of equal importance is the **packet throughput**, which is normally measured in millions of packets per second (Mpps). This describes how fast an I/O system can handle packets. Many VNFs handle large quantities of very small packets, which puts great strain on the system. Ensuring against packet loss is much more complex and difficult when dealing with small packets.

In a virtualized environment, packet throughput is further strained due to the hypervisor layer that manages VMs and containers that can belong to different virtual networks. In addition to the challenges that come just from virtualization is the impact of service chaining. Many VNFs will be comprised of multiple VMs, which may or may not be resident on the same server. Regardless of location, latency accumulates and system overhead increases with each subsequent VM. Most service chains today are implemented as tunnels (e.g., VXLAN, MPLS over GRE/UDP), which adds overhead due to additional headers. Packet handling adds to the overhead for the I/O and CPUs for all servers involved in the service chain. This all increases latency, so in order to achieve deterministic performance, signaling and receiving overhead needs to be minimized.

Overcoming Performance Hit From Virtualization

ETSI acknowledges in its NFV Performance specification that "whenever I/O and memory access are relevant for the overall performance of the VM instances, techniques relying on bypassing the OS, its I/O mechanism and its interruptions may become essential." This section discusses some of those techniques.

The technologies leveraged in these techniques include the following:

- **Single Root I/O Virtualization (SR-IOV)** is a technology that allows a single physical NIC to be shared by multiple VMs.
- **Data Plane Development Kit (DPDK)** is a set of libraries and drivers that enable enhanced packet processing.
- **Open vSwitch (OVS)** is an open-source virtual switch used to route traffic between VMs. It generally resides in kernel space of the CPU.
- **Embedded switch (eSwitch)** is an emerging technology that resides on the NIC and accelerates switching of both inter- and intra-host packets. It shares the forwarding table with the virtual switch, and it is anticipated that eSwitch could eventually replace virtual switching in some cases.

Overcoming the Compute Virtualization Penalty

As mentioned in the previous section, it could take multiple VMs to create a single service, and those VMs reside in user space on the same or different servers. In a hypervisor-managed virtualized environment (as opposed to bare metal), traffic between VMs goes through a hypervisor into kernel space where the virtual switch resides. If going to the same host, traffic goes back through the hypervisor and into user space; if going to a different host, it is put into a buffer where encapsulation, checksum and CRC calculations are performed. A copy is made in the kernel, then another copy is made onto the I/O system, where outer packet encapsulation, checksum and CRC calculations are performed. Traffic is then sent across the network to another server, where the process runs in reverse until the traffic reaches the VM in user space. Due to the numerous interrupts and store-and-forward mechanisms, packet performance is dramatically reduced. In some cases, less than 1 Mpps is achieved on a 10 Gbit/s link, which can, in theory, translate to 15 Mpps small packet performance.

Software & Hardware Assist Options

Multiple options are being proposed to address the performance issue described above. Two popular choices are software assist and hardware offload. Each has its benefits and trade-offs.

Software-assist can address the overhead that comes from the interrupt that occurs in a push model. By linking with DPDK libraries and application programing interfaces (APIs), applications continuously poll for new packets, rather than interrupt processing each time new packets arrive. A key benefit of this approach is that it significantly improves processing performance while maintaining hardware independence and eliminates PCI bandwidth overhead. However, unless other technologies are employed, packet processing will still take place in user space, so overhead to the CPU is not reduced. This packet processing overhead can be further exacerbated when NFV moves toward cloud-native, micro-service type of VNFs.

With hardware-assist, SR-IOV is used to enable application direct access. VMs leverage direct memory access (DMA) to eliminate the need to copy traffic into buffers. This approach can be used in conjunction with DPDK libraries as in the software-assist model to gain the efficiencies of switching from push-to-poll mode. In implementations that leverage DPDK, the virtual switch is often moved out of the kernel and into user space. In this approach, many of the switching functions typically performed in the kernel can be accelerated on the NIC card using eSwitch. While the PCIe load is increased in this approach, offloading the switching from the CPU to the NIC results in accelerated processing since CPU resources can be assigned to tasks other than packet processing.

The importance of packet processing techniques becomes apparent when considering VNFs. Not all VNFs share the same requirements – e.g., some require only packet termination, while others require packet processing as well. The latter scenario results in much more East/West traffic, putting additional load on the vSwitches and NICs, making I/O a bottleneck to performance. By using SR-IOV along with accelerated vSwitch (AVS), I/O performance can be significantly improved, allowing the VNF software to run closer to the native software performance.

Affirmed Networks, a leading virtualized mobile solutions provider, has investigated the impact of packet forwarding architecture/vSwitch and NIC throughput on the overall performance of its Mobile Content Cloud (MCC) software suite – a virtualized

EPC solution designed to run in a cloud environment. With regard to packet forwarding, Affirmed says that the MCC software can scale out on demand to saturate a server's maximum I/O capacity, which equals the throughput of all NICs, often 10 Gbit/s or 20 Gbit/s for single-NIC and dual-NIC servers. The introduction of traditional OVS can curtail performance down to 10-20 percent of server I/O capacity. DPDK accelerated OVS can bump actual performance to 80 percent of server I/O capacity, while SR-IOV gets to nearly 100 percent native line rate.

In addition, Affirmed has found that its MCC software generates about equal amount of East/West traffic among its different modules in the EPC data path as North/South traffic coming into the EPC from a service provider gateway router (the effective throughput of an EPC cluster). This means that to support effective throughput of X, the server I/O capacity needs to be at least 2X. For example, to stay competitive and support 20 Gbit/s of cluster capacity, MCC software will consume 40 Gbit/s of server I/O. This can be supported in a variety of configurations, with the server footprint increasing and compute resource utilization decreasing as you move down this list:

- One server with one 40 Gbit/s NIC
- Two servers each with two 10 Gbit/s NICs
- Four servers each with one 10 Gbit/s NIC

Overcoming the Overlay Networking Penalty

Overlay networking technologies, such as VXLAN, NVGRE and the latest, GENEVE, add a new header and CRC to encapsulate traffic. This results in even more stress being placed on CPU resources.

Traditional NICs only see this outer header that directs a packet from physical source server to physical destination server, and not the inner one with the route information to direct packets to the actual VM or container that runs a VNF. Many of the current generation of NICs can only offload the processing of packet checksum/CRC calculation, encapsulation and decapsulation of the outer packet, resulting in increased CPU workload to handle inner packet checksum/CRC calculation.

However, NICs with overlay offload capability can handle checksum/CRC calculation in the NIC hardware, resulting in significant throughput enhancement to almost bare-metal performance at significantly reduced CPU load for packet handling and increased efficiency of the cloud infrastructure. Moreover, eSwitch in the NIC is able to perform both outer and inner packet encapsulation and decapsulation, and route traffic based on the inner packet. This can further enhance throughput and enable more deterministic latency at an even lower CPU overhead.

Overcoming the TCP/IP Penalty

The communications protocol used in networking, TCP/IP, requires state to be maintained. TCP/IP operates over an assumed lossy environment and relies on dropped packets as an implicit congestion notification mechanism. The resulting sender side timeouts, retransmission and out-of-order packets make offloading TCP much more complex, so rather than offload, it is instead usually handled in CPU.

RDMA is a non-proprietary transport protocol that directly addresses two key limitations of current compute and networking architectures, namely overhead created by data copies between user/application and kernel memories, and latencies introduced by the TCP/IP protocol. RDMA (either over Ethernet or InfiniBand) was designed to provide a reliable in-order sequence of **messages** while TCP was designed to provide a reliable in-order sequence of **bytes**. It has been widely adopted by computing clusters that require high performance.

RDMA is fundamentally an accelerated I/O delivery mechanism. It introduces the concept of "zero-copy" data placement, which allows specially designed RDMA NICs on both ends of a transaction (also called an R-NIC) to transfer data directly from the user memory of the source server to the user memory of the destination server bypassing the operating system (OS) kernel.

RDMA can be run as a transport protocol over Ethernet and IP in place of TCP/IP, allowing traffic to be offloaded onto a host channel adapter (HCA) and reducing the overhead on the CPU. This process works in a similar manner as the DPDK-enabled hardware offload described earlier, except that Accelio (open source messaging and remote procedure call libraries) is used instead.

Conclusion

The move away from dedicated hardware platforms to standardized IT servers is among the most significant changes that come with NFV. Not only were these servers not designed to provide the reliability, elasticity and scalability that communications service providers require, technologies such as compute virtualization, overlays and TCP/IP also add performance penalties.

In order to overcome these performance issues, software and hardware acceleration techniques can be used. Leveraging DPDK allows for a reduction in CPU overhead, while SR-IOV provides VMs a way to share a single physical NIC. In addition, new NICs can help address issues with overlay networking, while RDMA is emerging as a viable alternative to TCP/IP for transport.

About Mellanox

Mellanox provides an extensive portfolio of Ethernet and InfiniBand adapters. It was the first to market with a 40 Gbit/s NIC, and has recently launched NICs supporting 100 Gbit/s and the new 25/50 Gbit/s specification. Mellanox's adapters support SR-IOV and OpenFlow-enabled eSwitch functionality to provide the hardware acceleration needed to overcome the performance challenges that come with virtualization and VXLAN, NVGRE, GENEVE and MPLS overlay networking. They also support RDMA over converged Ethernet to overcome the performance penalty inherent in TCP/IP.