WHITE PAPER

Network Functions Virtualization Using Intel® Ethernet Multi-host Controller FM10000 Family

Introduction

Network service providers are finding it increasingly difficult to keep pace with consumer bandwidth demands while at the same time maintaining reasonable service pricing. They are addressing this by moving away from proprietary network appliances to more open systems using high volume multi-core Intel® processors to reduce system cost and streamline software development efforts. This movement is also embracing Software Defined Infrastructure (SDI) concepts such as Network Functions Virtualization (NFV) which can provide agile services and lower Total Cost of Ownership (TCO) when combined with other industry initiatives such as network virtualization overlays and software defined networking. This white paper will provide an overview of these new industry trends and will also introduce the Intel® Ethernet Multi-host Controller FM10000 family that provides several types of high bandwidth interfaces while also including an advanced set of features for this new NFV environment. In addition, we will provide information on how the FM10000 family can help improve performance while reducing TCO.

The Evolution of Virtual Network Functions

You can find many sources of information on the web describing how service provider revenue is not keeping up with service demand. As shown in the generalized figure below, broadband traffic is growing at a tremendous rate due to factors such as the introduction of new mobile technologies and the significant increase in video content. At the same time service provider revenue is flattening out, driving their need to dramatically reduce the cost of delivering this bandwidth while at the same time providing new features such as improved security. By embracing new SDI initiatives such as NFV, capital equipment costs can be reduced by moving to standard high volume servers and operating costs can be reduced by moving to virtualized, open environments.

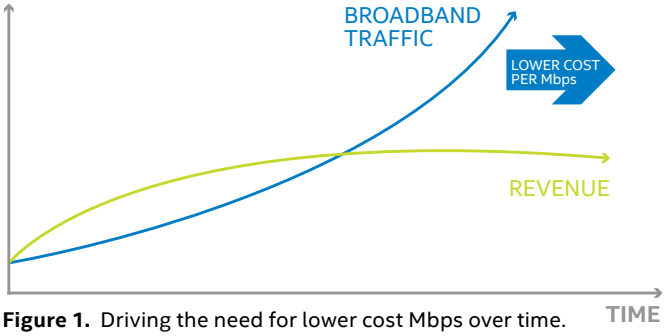

Figure 1. Driving the need for lower cost Mbps over time.

Table of Contents

In today's network environment, service provider equipment is spread across homes, businesses, base stations, central offices, and data centers. A large number of base stations and central offices may utilize traditional network appliance equipment housed in standard proprietary chassis as shown in the left side of the figure below. The specialized processing boards (not shown) that plug into this type of chassis may contain processors, Network Processing Units (NPUs), Field Programmable Gate Arrays (FPGAs), and Application-Specific Integrated Circuits (ASICs) that are used to provide high bandwidth packet processing, forwarding, policing, and billing. These base stations and central offices may also be located in hostile environments requiring expensive industrial temperature range parts and Network Equipment Building System (NEBS) compliance.

With the advent of high performance multi-core processors from Intel along with packet acceleration using Data Plane Development Kit (DPDK), control plane and data plane functions are migrating to Intel processors using Intel® Quick Assist Technology co-processing devices. Many security and network monitoring companies are currently developing two units network appliance systems based on Intel® Architecture. The middle picture in the figure below shows a specialized chassis containing four Intel® Xeon® processor sleds that can handle these high bandwidth workloads without the need for expensive NPUs, ASICs, or FPGAs. In addition, since these applications can be developed using a single software tool chain, development and support expenses can be reduced. By using NFV technology, network functions can be virtualized on these servers, providing a very flexible way to deploy services on demand.

Eventually, Intel Quick Assist Technology will be incorporated into Intel chip sets at which point racks full of standard high volume servers can be used to deploy these virtual functions as shown to the right in the figure below. By using these commodity components and by leveraging the economies developed for hyperscale data centers, costs can be significantly reduced further while providing flexible service allocation. Service providers can move many of their network functions that are currently deployed in thousands of remote locations into a few large data centers connected with high bandwidth links to remote access points. Functions such as firewalls, intrusion detection systems, vRANs, NAT, even functions normally found in set-top boxes can be moved into these data centers as a service provided by a centralized virtual appliance. As a further benefit, since these data centers do not need to endure the extreme temperatures seen

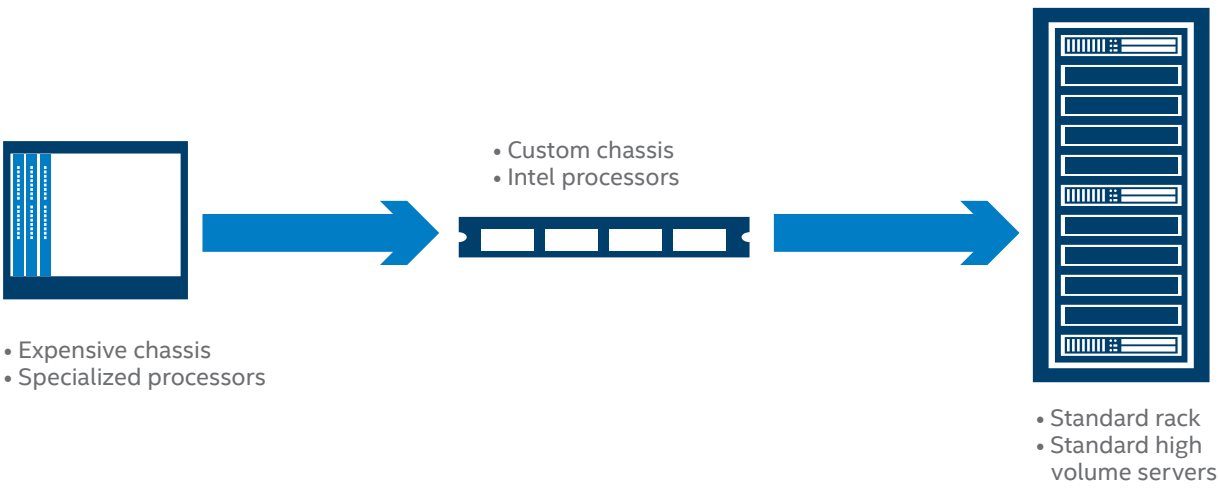

Figure 2. Industry transition to network functions virtualization.

in the central offices or base stations, component costs and total cost of ownership can be further reduced.

Processor Interconnect Technology

In the 1990s, telecom equipment used proprietary switching technologies from a variety of silicon vendors to connect line cards across backplanes. With the advent of standardized systems, Ethernet became the backplane interconnect of choice starting with XAUI which provides a 4-lane 10Gb Ethernet link and later with 40GBASE-KR4 which provides a 4-lane 40G link. Unfortunately, not many processors contain 40GBASE-KR4 interfaces, necessitating the use of Ethernet Network Interface Controllers (NICs) which connect to the processors using high bandwidth PCI Express* (PCIe) interfaces. Today Intel multi-core processors do not contain 40GBASE-KR4 interfaces, but they do contain 8- and 16-lane PCIe* v. 3.0 interfaces that can provide 50Gb and 100Gb of data bandwidth. Since these multi-core processors can process data at rates over 40Gbps, this is the type of data interface bandwidth needed in these new network appliance and NFV applications.

Network Virtualization Technology

Data center administrators and network operators need a methodology to isolate multiple tenants or end customers from one another in these large networks. VLANs were originally used in enterprise networks to isolate various departments, but the 12-bit VLAN ID field does not scale to meet the needs of service providers. Because of this, a wide variety of tunneling or network virtualization overlay (NVO) protocols have emerged over the years as shown in the table below.

The early protocols such as MPLS and GRE did not comprehend multi-tenant data centers or network function virtualization requirements, but in some cases have been adopted for these applications. Large hyperscale data centers have driven the need to isolate a large number of tenants driving the industry to support a new set of protocols such as STT, VXLAN, and NVGRE. In order to minimize processor overhead, Network Interface Controllers (NICs) have been developed that use stateless offloads to support these new protocols.

Network operators need a way to separate the virtual NFV network from the physical network in order to provide agile services. This requires the ability to identify flows, apply policies to these flows, and make sure they are processed by network functions in a predetermined order. These functions typically exist on virtual machines throughout a data center that may migrate over time to different servers. When a flow comes into a data center for processing, header fields are examined to determine the order of services needed and if any special metadata must be transported with this flow. By using one of the emerging protocols listed in the table below, a special tag can be added to the packet headers that can be used to forward them through the correct network functions in the correct order and metadata can be added to aide these network functions in their processing tasks.

Table 1. Data Center Protocols.

EARLIER PROTOCOLS	DESCRIPTION
MPLS (Multi-Protocol Label Switching)	Widely used in IP networks today. Some are adopting for the data center.
GRE (Generic Routing Encapsulation)	Developed to encapsulate a variety of protocols for various needs.

NEW PROTOCOLS	DESCRIPTION
STT (Stateless Transport Tunneling)	Developed by Nicira to tunnel through standard L3 data center networks.
VXLAN (Virtual Extensible LAN)	Developed to isolate L2 data center tenants. Alternative to NVGRE.
NVGRE (Network Virtualization using GRE)	Developed to isolate L2 data center tenants. Alternative to VXLAN.

EMERGING PROTOCOLS	DESCRIPTION
Geneve (Generic Network Virtualization Encapsulation)	Similar to VXLAN and NVGRE but adds metadata transport.
VXLAN-GPE (Generic Protocol Extension for VXLAN)	Extension of VXLAN to support metadata transport.
NSH (Network Service Header)	Design specifically for network service chaining for virtualized functions.

Software Defined Infrastructure

As we described in the last section, network virtualization can add a lot of complexity to how packets are forwarded throughout a data center network or through a chain of virtual functions. In addition, network functions must be setup, configured, moved, or torn down over time, adding to this complexity.

Software Defined Infrastructure (SDI) aspires to provide a centralized control of servers, storage, and networking in order to simplify and dramatically speed up the deployment of resources in these large data center environments. These data center components must have mechanisms to advertise their capabilities to a central orchestration layer, which then can in turn configure them to meet the changing needs of data center clients. In addition, the automation of many tasks such as low-level component configuration, migration and failover will significantly reduce the cost of services generated and also shorten time to revenue.

The FM10000 family as a Network Function Interconnect

The FM10000 family is a new product category that combines high bandwidth Ethernet controller technology with advanced Ethernet switch technology. As shown in the figure to the right, the FM10000 family contains four 8-lane PCIe v. 3.0 interfaces each capable of providing 50Gbps of bandwidth in each direction and can be directly connected to Intel Xeon processors without the need for discrete NICs. Each of these interfaces can also be bifurcated into two 4-lane interfaces allowing the FM10000 family to support up to eight 25Gbps PCIe v. 3.0 interfaces.

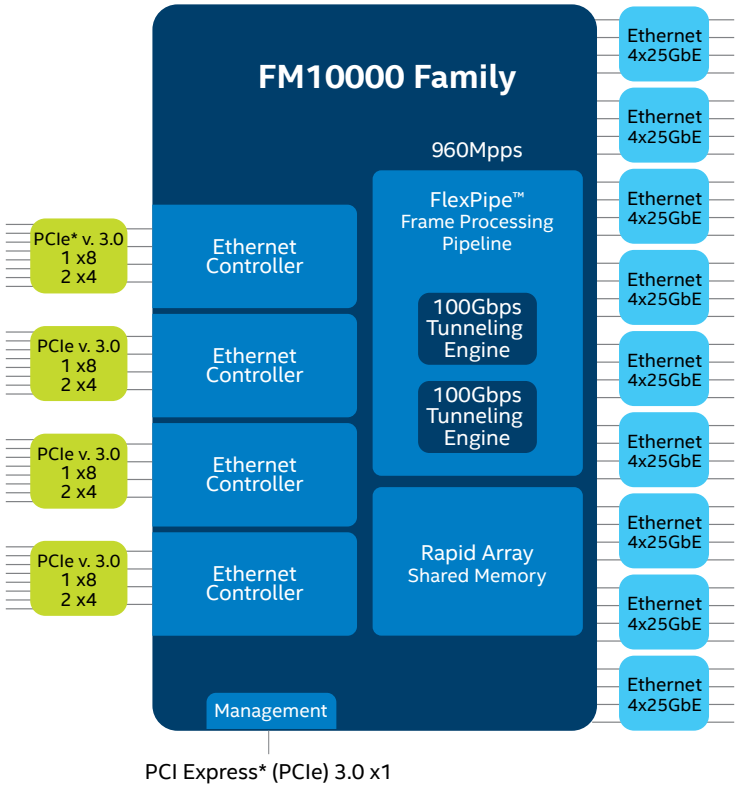

Figure 3. FM10000 family block diagram.

The Ethernet interfaces are divided into nine groups with four lanes each. Each lane can be independently configured as 1GbE, 2.5GbE, 10GbE, or 25GbE. In addition, groups of four lanes can be combined to form 40GbE or 100GbE ports. Inside the chip, all frames are treated like Ethernet frames and they are stored in a low-latency single output queue shared memory structure. By using cut-through operation, Ethernet to Ethernet latencies of 300 nS and PCIe to Ethernet latencies of 1000 nS can be achieved. Frame headers are sent to a flexible frame processing pipeline that can operate up to 960M frames per second. In addition, the pipeline contains two 100Gbps tunneling engines that can perform encapsulation and de-encapsulation of various tunneling protocols which will be described later in this paper.

There are several use cases for the FM10000 family in NFV applications.

Multi-socket Network Appliances

In multi-socket network appliance boards, it can be difficult to direct high bandwidth traffic to each socket when using individual high-bandwidth NICs. The FM10000 family effectively combines four 50Gbps NICs plus a switch into a single device providing up to 200Gbps of total bandwidth in a four socket server board as shown in the figure on the next page.

This can be complemented by two or more 100Gbps Ethernet interfaces to the external network. Unlike individual NICs, the FM10000 family can efficiently distribute flows across the sockets and cores as detailed below. With flexible Ethernet interfaces, the FM10000 family can provide a variety of port speeds and port counts in addition to the ability to cluster multiple boards into a larger appliance using mesh and/or ring interconnect topologies.

SDI Adapter Card

In some cases, network appliance vendors have existing high performance server blades to which they want to add 100GbE interfaces using the FM10000 family. In these cases, the FM10000 can be deployed in a PCIe card form factor allowing easy adaptation to existing server boards as shown in Figure 5. Since the FM10000 family has 8-lane PCIe interfaces, a 16-lane PCIe edge card connector can be bifurcated through BIOS changes in the processor to accept two separate 8-lane interfaces. The PCIe card can include two QSFP28 connectors allowing two 100GbE, two 40GbE, or eight 25GbE or 10GbE ports using breakout cables.

Telecom Data Centers

Telecom data centers are migrating to centralized NFV data centers that along with SDN can help service providers reduce cost and speed deployment. Rack Scale Architecture is one way to achieve these goals. The idea is to leverage the economies of scale from standard high-volume servers and to use SDI to reduce the cost of NFV deployments. Figure 6 shows an example of an Intel Xeon processor-based server shelf used in RSA applications.

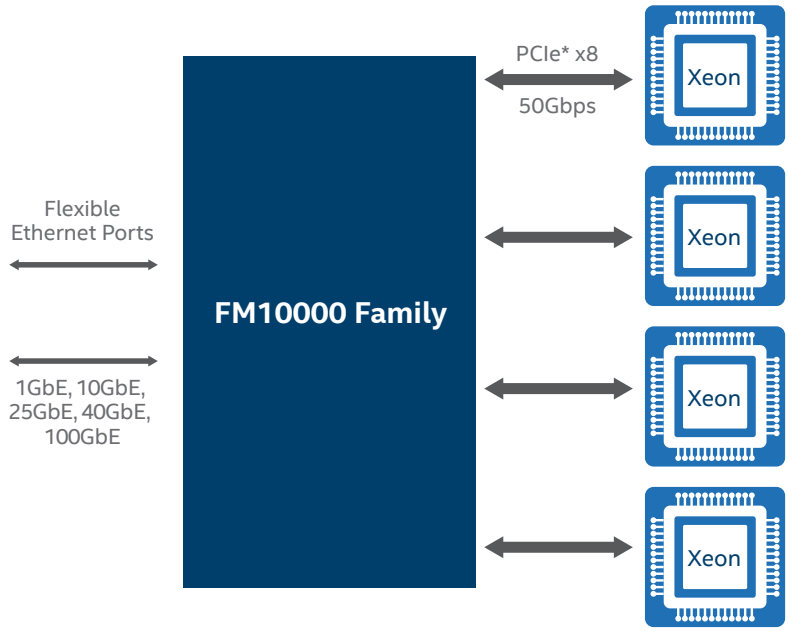

Figure 4. Multi-socket Network Appliance.

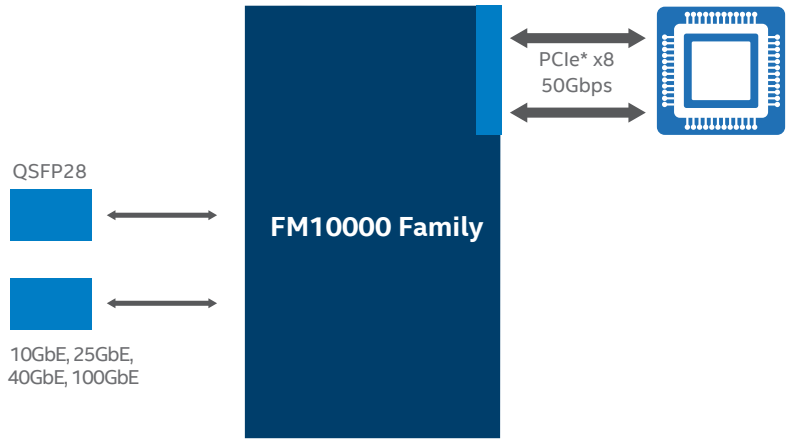

Figure 5. Network Appliance.

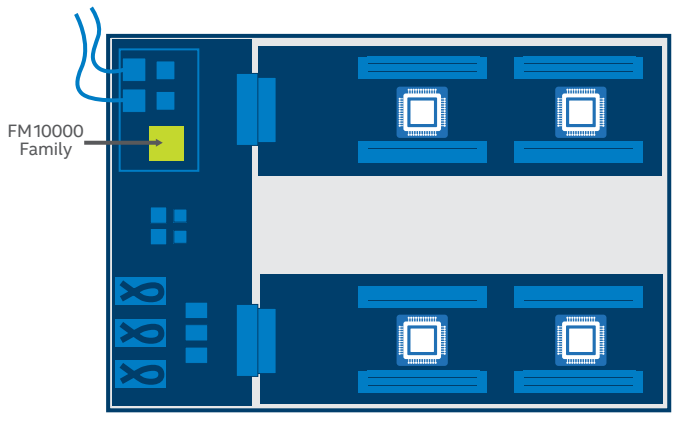

Figure 6. RSA server shelf using the FM10000 family.

In this system, modular Intel Xeon processor sleds are used which connect to the FM10000 family through an 8-lane PCIe* v. 3.0 interface. These sleds allow easy upgrade when more powerful Intel Xeon processors become available. The FM10000 family connects out of the shelf using 40GbE or 100GbE direct attach copper or optical cables to other shelves in the rack including storage shelves, or to other parts of the network. This shelf design is very similar to the dedicated 2U network appliances being developed today, but since these types of server shelves will be used in hyperscale data centers and use standard high-volume servers sleds, costs can be significantly reduced.

FM10000 Family Advantages

The FM10000 family is a unique product in the market which integrates multiple high bandwidth NICs into a single package and aggregates traffic to 25GbE, 40GbE, or 100GbE uplinks. But since the FM10000 family also integrates full Ethernet switching capability along with a frame processing pipeline, it can provide many more advanced features in network appliance and NFV applications that cannot be found in other products.

Advantages in Today's Systems

Today many OEMs and ODMs are building network appliance systems using IA processors to provide application layer, control plane, and data plane processing functions. Many of these system designers have chosen the FM10000 family as a key part of their design since it provides several immediate advantages.

The FM10000 family has the high bandwidth capabilities required in network appliance applications. It provides 50Gbps PCIe interfaces to each Intel Xeon processor removing the bottlenecks found in many discrete NIC-based solutions. In addition multiple 1GbE, 10GbE, 25GbE, 40GbE, and 100GbE ports are available providing flexible interfaces to the network. Since the FM10000 family can achieve overall packet processing levels up to 640Gbps or 960Mpps, unlike NIC-based solutions, multiple FM10000 family Ethernet ports can be used to cluster several network appliance shelves together to provide very high performance solutions.

The FM10000 family provides multiple integrated NICs along with full Ethernet switching capability. It supports simple aggregation of traffic from multiple Intel Xeon processors to 100Gbps uplinks, and also supports low latency switching of east-west traffic between Intel Xeon processors—both improving application performance and removing congestion from the attached network. In addition, the FM10000 family contains advanced hash-based load balancing mechanisms to efficiently spread incoming high bandwidth flows across the Intel Xeon processing resources.

Advantages in NFV

Intel continues to enhance the FM10000 family software and hardware so that it will have a robust feature set to support virtual network functions in either dedicated network appliance systems or rack scale architecture NFV implementations.

The FM10000 family frame processing pipeline has dedicated tunneling engines that are designed to encapsulate or de-encapsulate GRE, NVGRE, VXLAN, and GENEVE headers for use in multi-tenant data center environments. The FM10000 tunneling engine can also support VXLAN-GPE along with the encapsulation and de-encapsulation of network service headers (NSH) for use in NFV applications. By utilizing an on-board TCAM, the FM10000 family can classify packets to encapsulate them with the proper NSH header information. The FM10000 family devices can also use the TCAM to inspect incoming NSH headers in order to make proper forwarding decisions. In cases where forwarding is between PCIe ports on a given the FM10000 family device, forwarding can be done directly without involving and external switch. This low latency bypass path can improve performance while also reducing congestion within the attached network.

Systems providing virtualized network functions typically host each function on a dedicated VM. The VMs are interconnected using a vSwitch that runs on the host. To satisfy the needs of these NFV applications, the vSwitch requirements are becoming more complex, including:

- The need to operate at high bandwidth up to 40Gbps per host
- The need to make forwarding decisions based on the inspection of a variety of header fields
- The need to apply policies based on flow identification
- The need to add, remove, and inspect service chaining headers
- The need to efficiently distribute flows across multiple cores within a processor

Unlike standard multi-host NIC devices, the FM10000 family can provide a variety of vSwitch hardware acceleration features that can improve system performance and/or free up processor resources for other purposes. As an example, integrated Ternary Content Addressable Memory (TCAMs) in the FM10000 family can be used to accelerate vSwitch applications.

By accelerating high bandwidth switching functions using the FM10000 family, virtual switching on the server shelf is no longer restricted to a given processor but switching can also occur between processors using a common forwarding engine with up to 640Gbps of unidirectional bandwidth. In addition, the FM10000 family frame processing pipeline can be configured to present a variety of different header fields to the TCAM to make forwarding decisions at up to 960M packets per second. A TCAM match can spawn a variety of actions on a given packet including routing, policy enforcement, statistics gathering, and QoS enablement. Since a host can communicate through the control plane processor to update forwarding tables in the FM10000 family, once a flow is identified, the host could instruct the FM10000 family to no longer send it packets belonging to this flow and instead forward them in a low latency bypass mode. This not only provides low bump-in-the-wire latency, but frees up additional PCIe bandwidth and processor resources.

The FM10000 family contains two bi-directional 100Gbps tunneling engines that are designed for encapsulation and de-encapsulation of various industry standard tunneling headers including the new NSH header being driven by Cisco. Tunneling is a complex process that can quickly bog down a vSwitch when operating on high bandwidth flows. By accelerating this functionality using the FM10000 family, overall system performance can be improved. The FM10000 family can also forward frames based on the inspection of various tunneling headers as described earlier.

DPDK and NFV

DPDK libraries have established themselves in the industry as an ideal way to accelerate packet processing performance using standard Intel Architecture devices. Intel is currently working on DPDK acceleration enhancements which can take advantage of the FM10000 family frame processing pipeline to improve system performance. In addition, Intel will release a version of Flow Director for the FM10000 family that learns the source processor core of a given egress flow and can direct ingress traffic back to the same core based on this information. This is much more efficient than using techniques such as Receive-Side Scaling (RSS) which uses a stateless hash-based mechanism to distribute flows to processor cores that may not be associated with the given flow.

Conclusions

Network appliance systems must continue to provide increased performance while at the same time maintain or even reduce cost of ownership. Intel's SDI vision is to deliver advanced networking products that can meet the needs of virtualized compute, storage, and networking. To achieve these goals, service providers are moving to standard high-volume servers to improve agility, offer new services, and reduce costs. They are also using NVO, NFV, and SDN to build high performing networks that are flexible and scalable to support multiple tenants. Intel developed the FM10000 family to help enable SDI and it provides an ideal network connectivity solution for both dedicated network appliance systems and the emerging rack-based NFV systems by providing virtualized networking features. The FM10000 family provides high-bandwidth low-latency connectivity between Intel processors and the network, and by providing the ability to accelerate vSwitch functions to hardware, the FM10000 family can help improve system performance and/or reduce system cost by freeing up processor resources.

For more information on how the Intel Ethernet Multi-host Controller FM10000 family can improve your NFV designs, contact your Intel field sales representative.
Please visit www.intel.com/ethernet

Printed in USA 1115/MBR/HBD/PDF Please Recycle 333020-001US