WHITE PAPER

Intel® Rack Scale Architecture using Intel® Ethernet Multi-host Controller FM10000 Family

Introduction

Hyperscale data centers are being deployed with tens of thousands of servers making operating efficiency a key metric. By optimizing the data center rack design, data center administrators can benefit from improved resource density, more flexible resource provisioning and reduced Total Cost of Ownership (TCO). Several industry initiatives have begun to drive standards for network infrastructure optimization. Facebook founded the Open Compute Project several years ago to define standards around efficient server, storage and data center hardware for scalable computing. In China, Baidu*, Alibaba*, and Tencent* joined forces and started Project Scorpio in order to define efficient data center infrastructure standards. At Intel, we have been working for several years on a complimentary initiative called Intel® Rack Scale Architecture (Intel® RSA) which improves efficiency and deployment flexibility at the rack level while reducing TCO. This white paper will focus on the networking part of Intel RSA and show how our Intel® Ethernet Multi-host Controller FM10000 family can reduce cost and improve performance compared to standard rack solutions.

Intel Rack Scale Architecture

Cloud data centers require an infrastructure that lowers computing costs while providing the following:

- Ease of scalability
- Create applications and provision new services rapidly
- The ability for end users to orchestrate their operations with real-time data and analytics
- The flexibility to address unpredictable workload needs as applications change across compute, network and storage

Intel RSA as depicted in Figure 1 is redefining the cloud platform by providing a logical architecture that disaggregates compute, storage, and network resources while introducing the ability to pool these resources. In addition, it provides a way to simplifying management of compute, storage, and network resources while at the same time enabling the ability to dynamically compose resources based on workload-specific demands.

Table of Contents

Networking resources are a key part of this new architecture. Efficient networking topologies provide high performance, maximum utilization of server and storage resources, and enhanced deployment flexibility. To meet these needs, Intel has chosen the FM10000 family as a key ingredient for Intel RSA networking which will be the focus of this white paper.

Data Center Networking Technology

Ethernet has become the preferred networking technology in data centers. Today, most enterprise data centers are moving to 10Gbps while cloud data centers are starting to migrate from 10Gbps to 25Gbps, 40Gbps, and 100Gbps. Unfortunately, most servers used in these data centers do not include high bandwidth Ethernet interfaces, necessitating the use of Ethernet network interface controllers (NICs) which connect to the processors using high bandwidth PCI Express* interfaces. Today Intel multi-core processors contain 8-and 16-lane PCIe gen3 interfaces that can provide 50Gbps and 100Gbps of data bandwidth. Therefore high-performance and cost effective PCIe network interfaces are key in these data center environments.

Network Virtualization Technology

Administrators need a methodology to isolate multiple tenants or end customers from one another in these large data center 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 these large data centers. Because of this, a wide variety of tunneling or network virtualization overlay (NVO) protocols have been emerging over the years as shown in Table 1.

These protocols provide a way to separate the logical tenant networks 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 forwarded to the correct servers. Since the tenant applications typically exist on virtual machines with unique addresses that may migrate over time to different servers throughout a data center, these protocols must be supported by an orchestration layer that can provide global updates to policy rules and forwarding tables.

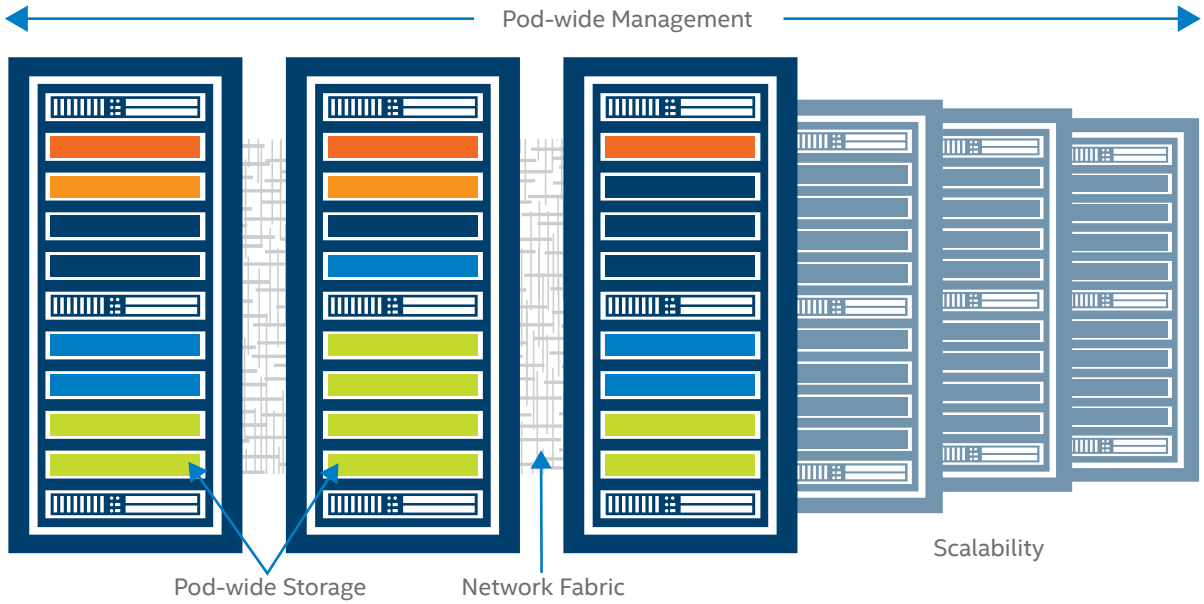

Figure 1. Intel® Rack Scale Architecture.

Table 1. Network Virtualization Technology.

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

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

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

The early protocols such as MPLS and GRE did not comprehend multi-tenant data centers or software defined networking environments, but in some cases have been adopted for these applications. 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. Emerging protocols such as Geneve, VXLAN-GPE, and NSH are adding further tunneling enhancements for network service chaining which is used in network function virtualization (NFV) applications.

Software Defined Infrastructure

The network virtualization protocols described in the last section can add a lot of complexity to how packets are forwarded throughout a data center network. In addition, tenant VMs 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 resources 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 reduces network administrator workload.

As part of the Intel RSA initiative, Intel is developing reference software that can facilitate the discovery and configuration of resources and works with industry standard orchestration layers such as OpenStack. This includes our Intel Open Network Platform (Intel® ONP) software which provides an open Linux*-based multi-host NIC driver for the FM10000 family.

FM10000 Family as a Multi-host Controller

The FM10000 family is a new product category that combines high bandwidth Ethernet controller technology with advanced Ethernet switching technology. As shown in the Figure 2, the FM10000 family contains four 8-lane PCIe gen3 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 Intel Xeon processor interfaces.

The Ethernet interfaces are divided into nine groups with four lanes each. Each lane can be independently configured as 1Gbps, 2.5Gbps, 10Gbps, or 25Gbps. In addition, groups of four lanes can be combined to form 40Gbps or 100Gbps 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

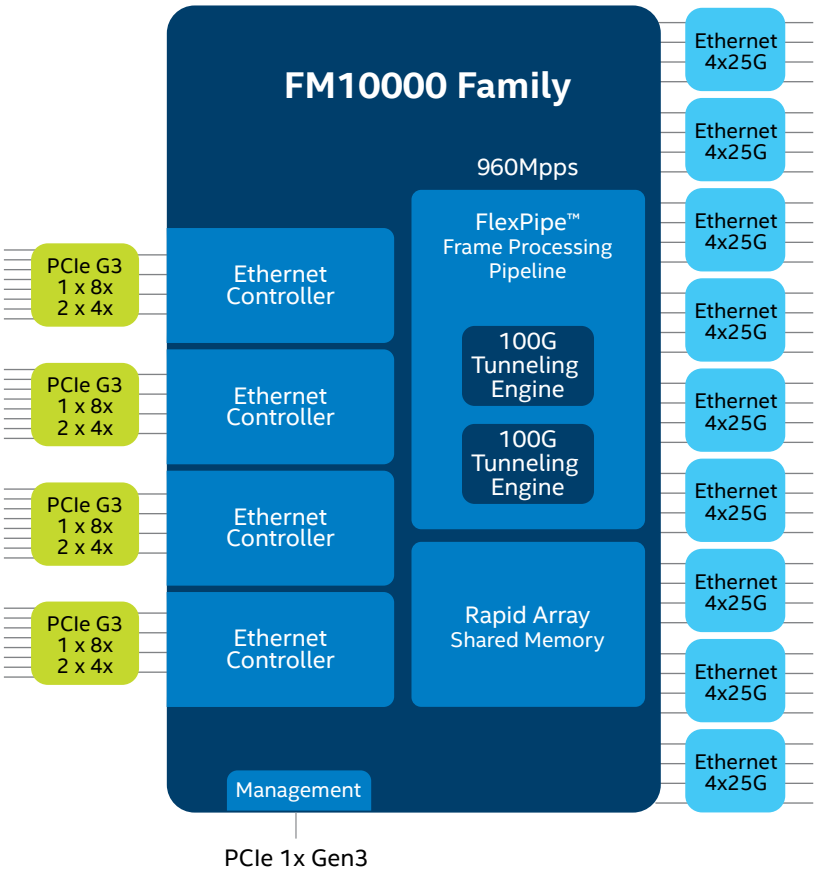

Figure 2. FM10000 family block diagram.

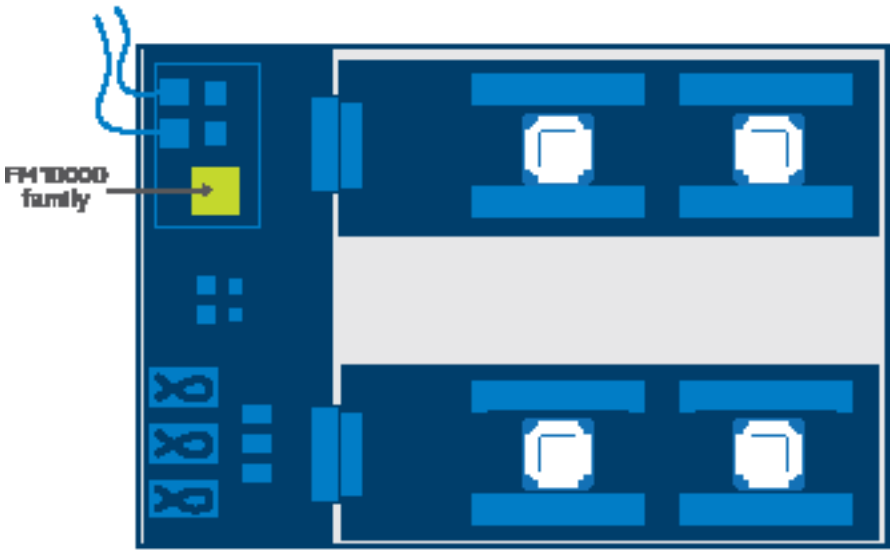

Figure 3. Intel® RSA server shelf using the FM10000 family.

operation, Ethernet to Ethernet latencies as low as 300nS and PCIe to Ethernet latencies as low as 1000nS can be achieved. Frame headers are sent to a flexible frame processing pipeline that can operate up to 960 million packets per second. In addition, the pipeline contains two 100Gbps tunneling engines that can perform encapsulation and de-encapsulation of various tunneling protocols that are shown in Figure 2.

In Intel RSA applications the FM10000 family can be used to connect multiple Intel Xeon processor-based server sleds to the network as shown in Figure 3. In this system, modular Intel Xeon processor sleds are used which connect to the FM10000 family through an 8-lane PCIe gen3 interface. These sleds allow easy upgrade when more powerful Intel Xeon processors become available. The FM10000 family connects off the shelf using 40Gbps or 100Gbps direct attach copper or optical cables to a ToR switch, to other shelves in the rack including storage shelves, or to other parts of the network.

The 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 25Gbps, 40Gbps, or 100Gbps 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 for certain RSA workloads that cannot be found in other products.

Traffic Aggregation

Figure 4 shows a 1Gbps link running a database application. As you can see, the link is fully utilized when there are traffic bursts, but the overall bandwidth utilization is much lower than the peak traffic (as indicated in figure 4). The bursty nature or data center traffic depends on the application, but will

also be seen when using higher performance Intel Xeon processor-based servers.

With a traditional star architecture within the rack where each server has a NIC connection to the ToR, every NIC and cable must be able to accommodate the potential peak bandwidth required for the application. When using the FM10000 family to aggregate the bandwidth from four servers to a ToR, 2:1 oversubscription can be used since the probability of more than two servers bursting traffic at the same time is quite low. Table 2 shows a comparison of a traditional ToR switch using discrete NICs with one using the FM10000 family as an aggregation point connected to a ToR in a rack-scale network.

Table 2 shows that the FM10000 family switch module will cost less than four discrete NICs due to silicon integration. In addition, 2:1 oversubscription can be used with the FM10000 family which reduces both the cable count and the number of ToR ports required, further reducing the cost per server. This allows data center administrators to eliminate 50 percent of the ToR switches throughout the data center.

Server Clustering

Another way to reduce the number of ToR ports required and/or improve application performance is to use direct connections between server shelves in addition to the ToR connections as shown in Figure 5. In this figure, only three, two unit server shelves are shown for simplicity.

The implementation on the left in Figure 5 provides a standard north-south data path between the servers on the shelf and the ToR switch through the FM10000 family which is acting as an aggregation point using oversub-

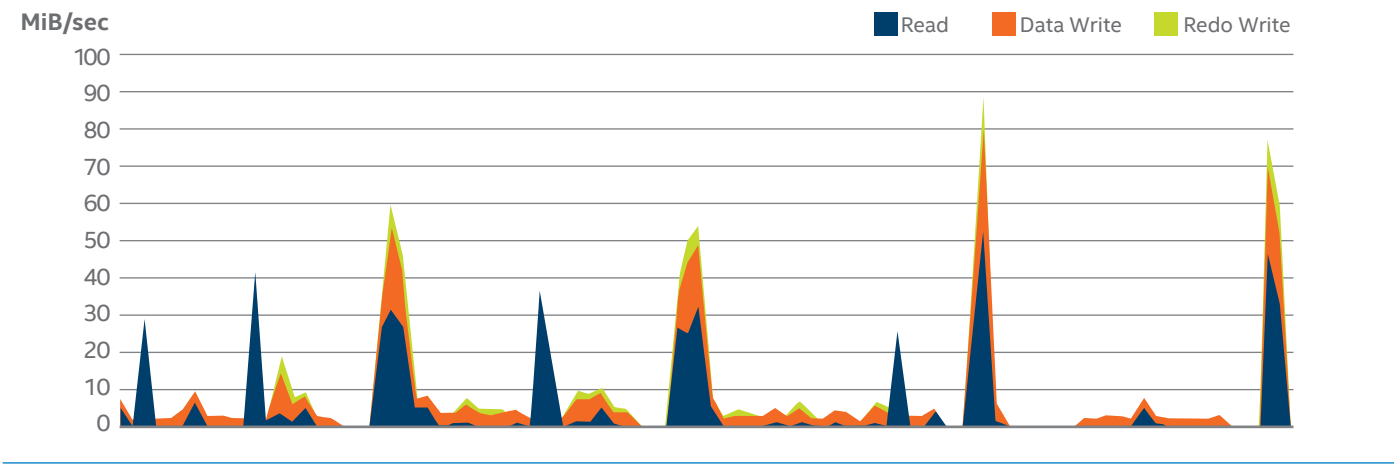

Figure 4. Database throughput.

Table 2. Data center cost savings using the FM10000 family.

	25 GBPS PER SERVER		50 GBPS PER SERVER	
Solution type	FM10000 family	NIC	FM10000 family	NIC
Cable type	50Gbps	25Gbps	100Gbps	50Gbps
Oversubscription	2 : 1	1 : 1	2 : 1	1 : 1
Number of modules	1	4	1	4
Number of cables	1	4	1	4
ToR ports	1	4	1	4
Cost per module	3.0x	1.0x	3.0x	1.2x
Cost per cable	2.0x	1.0x	4.0x	2.0x
ToR cost per port	2.0x	1.0x	4.0x	2.0x
TOTAL COST	**1.0x**	**1.4x**	**1.25x**	**2.0x**

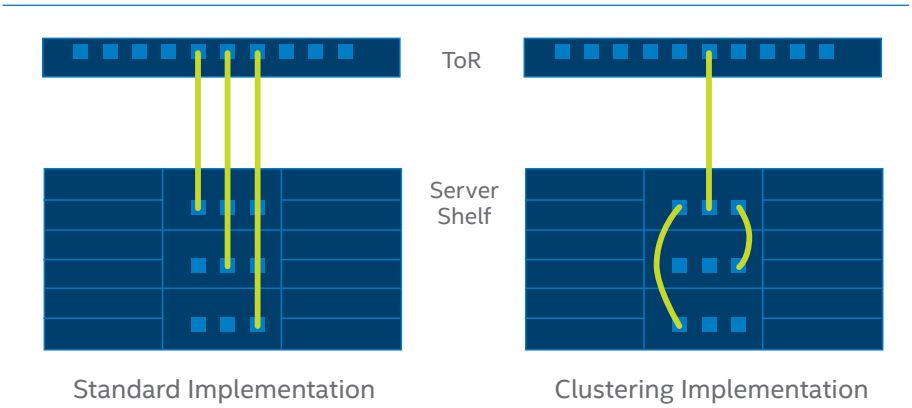

Figure 5. Server clustering using the FM10000 family.

scription as described above. The implementation on the right in Figure 5 provides high-bandwidth low-latency clustering between server shelves as well as the option for one, two, or three uplink ports to the ToR. This can be used for high performance server clustering applications such as big data analytics and network functions virtualization.

In some use cases, this configuration can also be used increase the oversubscription, further reducing the number of ToR ports required. For example, in the clustering implementation above, 12 servers can share a 100Gbps link to a ToR. Or more uplinks can be used to provide 200Gbps or 300Gbps for 12 servers while also maintaining the local cluster interconnections.

vSwitch Acceleration

Unlike standard 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, competing products do not contain integrated Ternary Content Addressable Memory (TCAM) which can give the FM10000 family a speed advantage in applications compared to traditional vSwitch implementations.

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 configure to present a variety of different header fields to the TCAM to make forwarding decisions at up to 960 million frames per second. A TCAM match can spawn a variety of actions on a given packet including routing, policy enforcement, statistics gathering, and QoS enablement.

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 frame processing pipeline has dedicated tunneling engines that are designed to encapsulate or de-encapsulate GRE, NVGRE, VXLAN, Geneve, and NSH headers for use in multi-tenant data center environments. By utilizing on-board TCAMs, the FM10000 family will be able to classify packets to encapsulate them with the proper network virtualization header information. The FM10000 family devices can also use the TCAMs to inspect incoming network virtualization 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 the ToR switch. This low latency bypass path can improve performance.

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 Intel® Ethernet 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. By offloading the functions described in this section to the locally attached FM10000 family multi-host NIC, fewer processor resources are required, reducing cost.

Conclusions

Hyperscale data centers are evolving and in some cases embracing new infrastructure standards such as the Open Compute Platform and Project Scorpio. The FM10000 family is a key ingredient in Intel's complimentary rack scale architecture initiative and can provide networking solutions that are lower cost than traditional server networking approaches by aggregating traffic through a single high bandwidth link. In addition, the FM10000 family contains integrated Ethernet controllers providing high-bandwidth low-latency connectivity between processors and the network, and by providing the ability to accelerate vSwitch functions into 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 rack scale designs, contact your Intel field sales representative. Please visit www.intel.com/ethernet

Printed in USA 1015/MBR/HBD/PDF ♻ Please Recycle 333019-001US