Scalable Switchless Accelerated Data Plane

The IBM/Silicom Data Plane Implementation uses Intel® FM10K to quadruple bandwidth while delivering a 10x latency drop – and at a significantly lower cost

PREFACE

Cloud data centers operation is becoming more and more demanding. The growing complexity of applications is a challenge for management and provisioning. Modern architecture of applications dramatically augments east-west traffic.

Back in the day, for example, an Apache [1] web server with its PHP web application engine was installed alongside its databases and storage. Users transactions generated network traffic only up to the web front end. But nowadays, each of these components is located in a different tier, so that multiple communications over the network are required to carry out a single user transaction. As a result, each transaction generates 3x-4x as much traffic as it did in the past, creating ***a much heavier traffic load that translates into east-west communication bloat***. Other examples can be found in big data installations and more.

Today's ubiquitous ToR (top-of-rack) architecture [2] was built to provision north-south communications – that is, communications to and from the rack to the edge/core or leaf/spine. In today's reality, however, this architecture is also required to support significant quantities of traffic whose sources and destinations are both inside the rack – a rack that is 3-4 times busier than it ever was before.

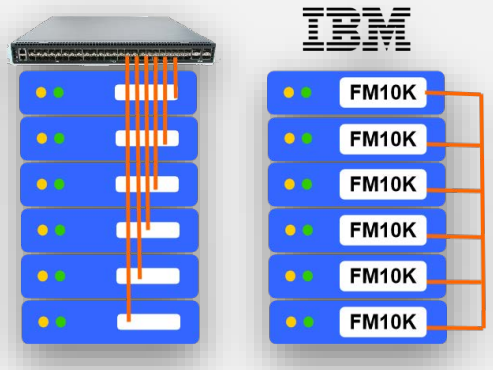

Figure1 - IBM's switchless topology

IBM'S ADVANTAGE

With a record of development leadership for Tier 1 cloud service providers and NFV telco solution providers, the IBM (https://www.ibm.com) team was founded with the mission to create higher-performance data plane and network services provisioning solutions. Their goal was to "do it right," building an innovative new solution from the bottom up that would overcome common hurdles met when employing common open source solutions such as **OVS** [3], or **VPP** [4] projects. As they began development, they brought with them the understanding and insight gained through painful experience with actual use cases.

The exceptional solution created by IBM delivers:

- Increased **bandwidth**
- True **low latency**
- Proven **scalability**
- **Programmability and flexibility**
- Potential network infrastructure **cost reductions**

NO MORE ToR SWITCH

The key differentiating feature of IBM's solution is the elimination of the ToR switch. Instead, IBM's solution uses an array Silicom adapters, based on Intel® FM10K controller, with one adapter residing in each server. Each adapter acts as a data plane comprised of NIC and a switching fabric. Working in tandem with IBM's IOBricks software suite, a complete operation and management envelope is created for each and every Intel® FM10K adapter, resulting in a single unified switching fabric.

Intel® FM10K-based network interface card features an embedded switching and forwarding engine with a filtering engine as well as a packet encapsulation/de-capsulation engine as part of the packet processing pipeline.

Silicom ltd., 14 Atir Yeda St., Kefar Sava, ISRAEL
sales@silicom.co.il www.silicom-usa.com

In addition, as a NIC, Silicom's interface card provides a full range of connectivity options - including 10GbE, 25GbE, 50GbE and 100GbE - alongside two PCIe host interfaces. ***Rounding out its capabilities, the Silicom NIC is able to function as a fully-fledged hardware switch that is tightly coupled with the server.***

BANDWIDTH

To demonstrate the bandwidth advantage delivered by this switchless setup, we carried out the test described herein.

As a baseline, we installed a set of 32 servers on a rack, each connected via a 100GbE NIC to an Arista 7160-32CQ switch. The maximum bandwidth that could flow through this fabric had a hard limit: that is, the limit of the internal backbone of the switch box, which barred at 3.2 Tpbs.

Device Type	Number of devices	Device capacity	Total BW
Arista 7160-32CQ	1	3.2Tbps	3.2Tbps
Intel® FM10K	32	500Gbps	16Tbps

Table 1 - Total capacity comparison

Then, we changed the setup to use a Silicom Intel® FM10K-based adapters (*Figure 1*) for each server. ***In this scenario, the capacity of the fabric as a whole more than quadrupled to 16Tbps*** (*Table 1*). This startling capacity boost derives from the fact that each Intel® FM10K-based adapter offers 600Gpbs of forwarding backbone. Even after reserving a 100Gbps allocation for host interface activities, 500Gbs remain on each adapter, distributed across 5 links of 100GbE each connected in a mesh topology. Together, they form a distributed fabric.

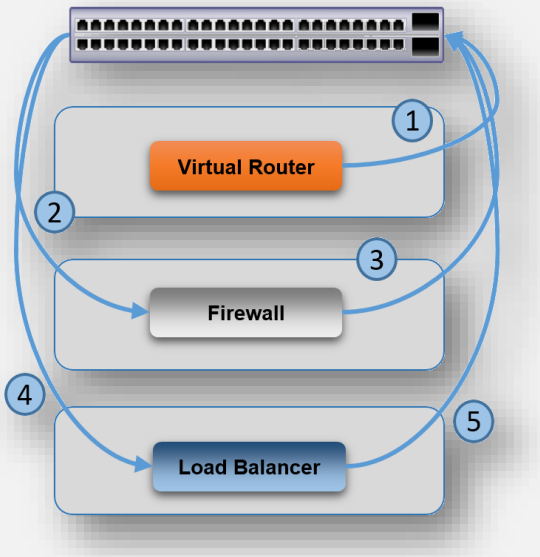

Figure 2 - Top of rack based service chaining

LATENCY

The replacement of ToR topology with a IBM switchless topology also generates a dramatic improvement in network traffic latency, which is another key network performance factor.

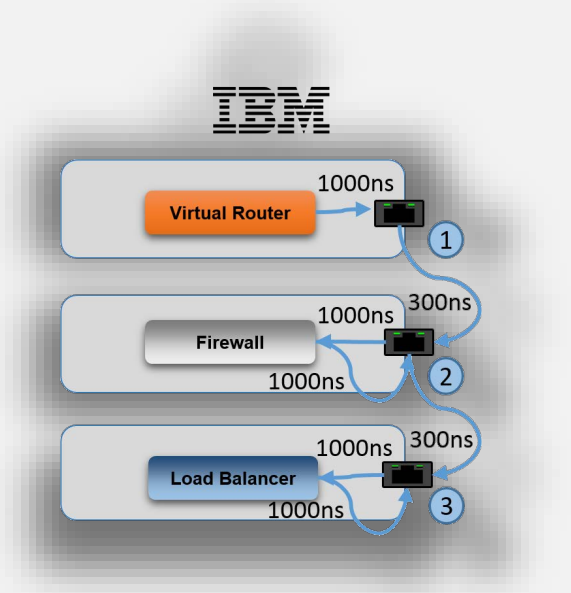

Figure 3 - IBM's switchless fabric with FM10K

This is due to the fact that traffic engineering around top-of-rack architecture performs best when traffic is streamed "vertically" to or from racks and pods, rather than "horizontally" inside racks pods. This vertical architecture was the scenario according to which the virtualized data center paradigm originally evolved, with virtualized components functioning as an overlay on top of the common ToR physical architecture.

Hop	Direction	Latency (uS)		Total
ToR Topology				
1	Host to switch	Router to NIC	1.3	13.3
		Switch	2	
		Software switch	10	
2	Switch to host	NIC to Firewall	1.3	11.3
		Software switch	10	
3	Host to switch	Firewall to NIC	1.3	13.3
		Switch	2	
		Software switch	10	
4	Switch to host	NIC to LB	1.3	11.3
		Software switch	10	
5	Host to switch	Switch	2	12
		Software switch	10	
		Total latency in uS		**61.2**
IBM Switchless Topology				
1	Host to host	Router to NIC	1.3	2.6
		NIC to Firewall	1.3	
2	Host to host	Firewall to NIC	1.3	2.6
		NIC to LB	1.3	
3	Host to host	LB to NIC	1.3	1.3
		Total latency in uS		**6.5**

Table 2 - Latency measurements ToR vs. IBM

However, as explained above, today's networks have an ever-increasing load of intra-rack/pod, inter-VM traffic that the physical infrastructure was not designed to handle. ***As a result, applications running on VMs using a ToR topology suffer from relatively high latency*** (*Figure 2*).

For example, the analysis and "ToR vs. IBM" comparison provided in *Table* 2 demonstrates a **61.2uS** latency penalty for traffic running through a ToR (Arista 7160-32CQ, in this setup, with 2uS minimal latency).

In contrast, a setup running similar services over IBM's data plane accelerated with the Silicom Intel® FM10K adapter (*Figure 3*) reveals **6.5 uS** latency, which is a dramatic ***10x reduction in total latency***. As demonstrated in *Table 2*, not only has the latency improved ten-fold, ***but the number of hops has declined by 40%.***

PROLIFERATION OF SMART NICS: AN ASSET

As network architecture moves away from top of rack architecture towards the deployment of multiple NICs, and as NICs become smarter, network architects are able to shift an increasing list of fabric-implementing activities to be carried out independently by the NICs themselves.

Multipath and Redundancy

Redundant multiple paths from VM-to-VM or from node-to-node are inherent in topologies in which data planes are disaggregated and implemented through NIC interconnectivity. Since intelligent management of this resource, as implemented by IBM, strengthens redundancy, it naturally strengthens network resiliency.

Programmability

The potential for smart and efficient routing within a resilient network environment enables the network administrator to specify ***application-specific data paths***, thereby improving the performance of critical applications or applications that are sensitive to network inconsistencies. For example, this feature could be utilized to enhance the performance of a video streaming application that is sensitive to the *busy neighbor* effect.

Pod Scale Up

Implementation of smart and efficient routing can enable a dramatic increase in the number of nodes (servers) in a pod. Using IBM's patent-pending intelligent routing solution, the number of servers in a pod is scalable up to 768 on basic setup and up to 4096 on high-end setup, with no requirement for a ToR switch.

Cabling

ToR architecture cabling within a pod is a challenge in unto itself. Switchless architecture cabling is much simpler – and allows gradual, easy-to-implement scaling. IBM holds a number of patented methodologies for simplified, cost-effective switchless cabling.

ROI

The Intel® FM10K-based adapter combines 100Gbps NIC capability with a high capacity switching fabric, and is offered at a price point close to a standard 100G NIC. In addition, the elimination of the need for expensive ToR switches is a huge relief for capital and budgets.

Multi-Host FM10K Adapter

The Intel® FM10K-based network adapters made by Silicom arrive in several different link speed configurations [5]. In multi-host configuration, a single FM10K chipset extends the network PCIe reach towards two CPU sockets, ***avoiding the internal QPI bus penalty.*** In this way, the use of the Silicom NIC further reduces latency.

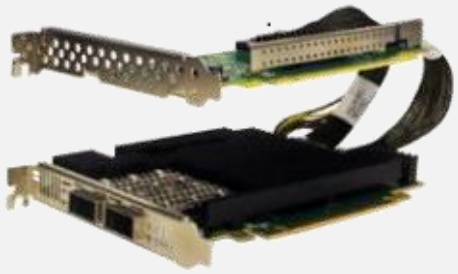

Figure 4 – Multi-host FM10K based adapter

SUMMARY

The deployment of IBM's comprehensive data plane implementation, leveraged by the use of Silicom Intel® FM10K adapters, delivers **dramatic benefits** in terms of **increased bandwidth** and **reduced latency**, while also providing a full set of **routing and programmability features** that can be utilized to further enhance network performance. This architecture cost-effectively overcomes a long list of deficiencies associated with the use of ToR-based switches in a distributed, multi-tier network.

REFERENCES

[1] https://httpd.apache.org/

[2] http://www.cisco.com/c/en/us/products/collateral/switches/nexus-5000-series-switches/white_paper_c11-522337.html

[3] http://openvswitch.org/

[4] https://wiki.fd.io/view/VPP

[5] http://www.silicom-usa.com/cats/server-adapters/networking-adapters/100-gigabit-ethernet-networking-server-adapters/