DPDK

INTEL® 40GBE ETHERNET CONTROLLER

M Jay & **Helin Zhang** - Intel

DPDK US Summit - San Jose - 2016

About this Document

- The performance measurement and analysis of an embedded platform for communication and security processing can be very challenging due to the diverse applications and workload inherent in the platform. The Internet of Things Group (IoTG) and Network Platform Group(NPG) are dedicated to performing lab measurements which will assist customers in understanding the performance of combinations of Intel® architecture microprocessors and chipsets.
- This document publishes a set of indicative performance data for selected Intel® processors and chipsets. However, the data should be regarded as reference material only and the reader is reminded of the important Disclaimers that appear in this document.
- Intel, Intel Core and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
-

* Other names and brands may be claimed as the property of others.

Disclaimers

- By using this document, in addition to any agreements you have with Intel, you accept the terms set forth below.
- You may not use or facilitate the use of this document in connection with any
- infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.
- INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.
- A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.
- Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.
- The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.
- Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.
- Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm

* Other names and brands may be claimed as the property of others.

Optimization Notice

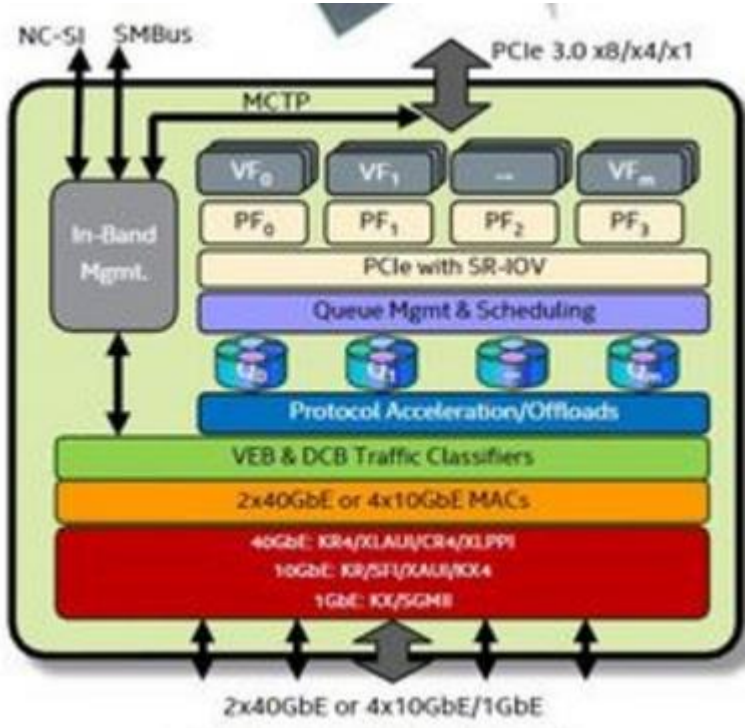
Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel.

Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804

Flexible Packet Processing – XL710

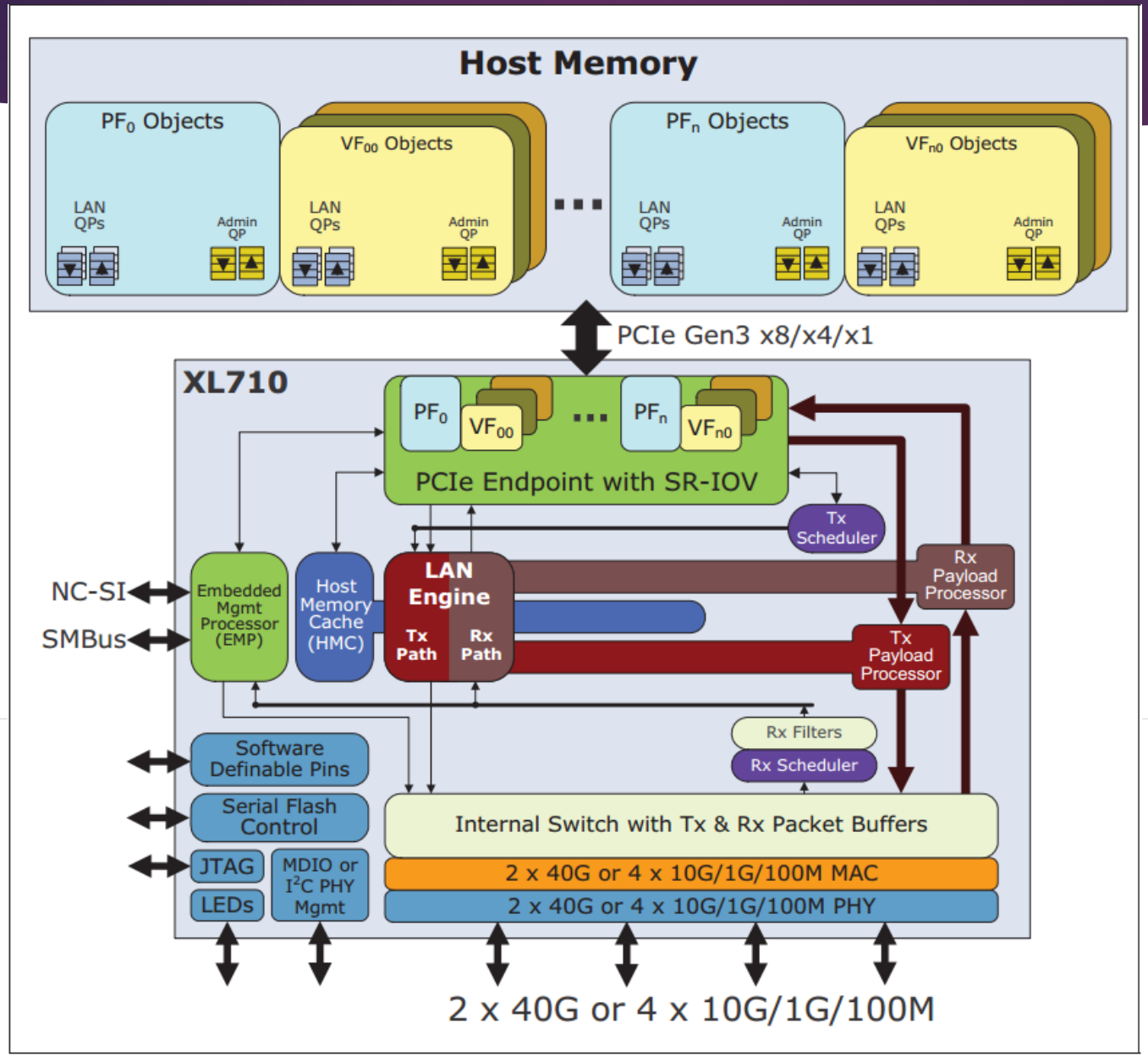
- Server Virtualization – VMDq for Emulated path; SR-IOV for Direct Assignment
- Network virtualization Overlay stateless offloads for VXLAN, NVGRE, VXLAN GRE
- "Flexible" – **Add new features after production** by upgrading firmware
- Intelligent load distribution for high performance traffic flows – **Flow Director**
- Virtual Bridging support that delivers control & management of virtual I/O
  - Both host-side and switch-side



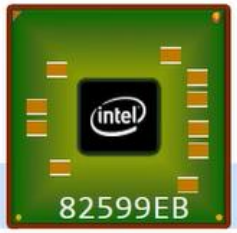
XL710 Internals

Helin

Block Diagram - XL710/X710

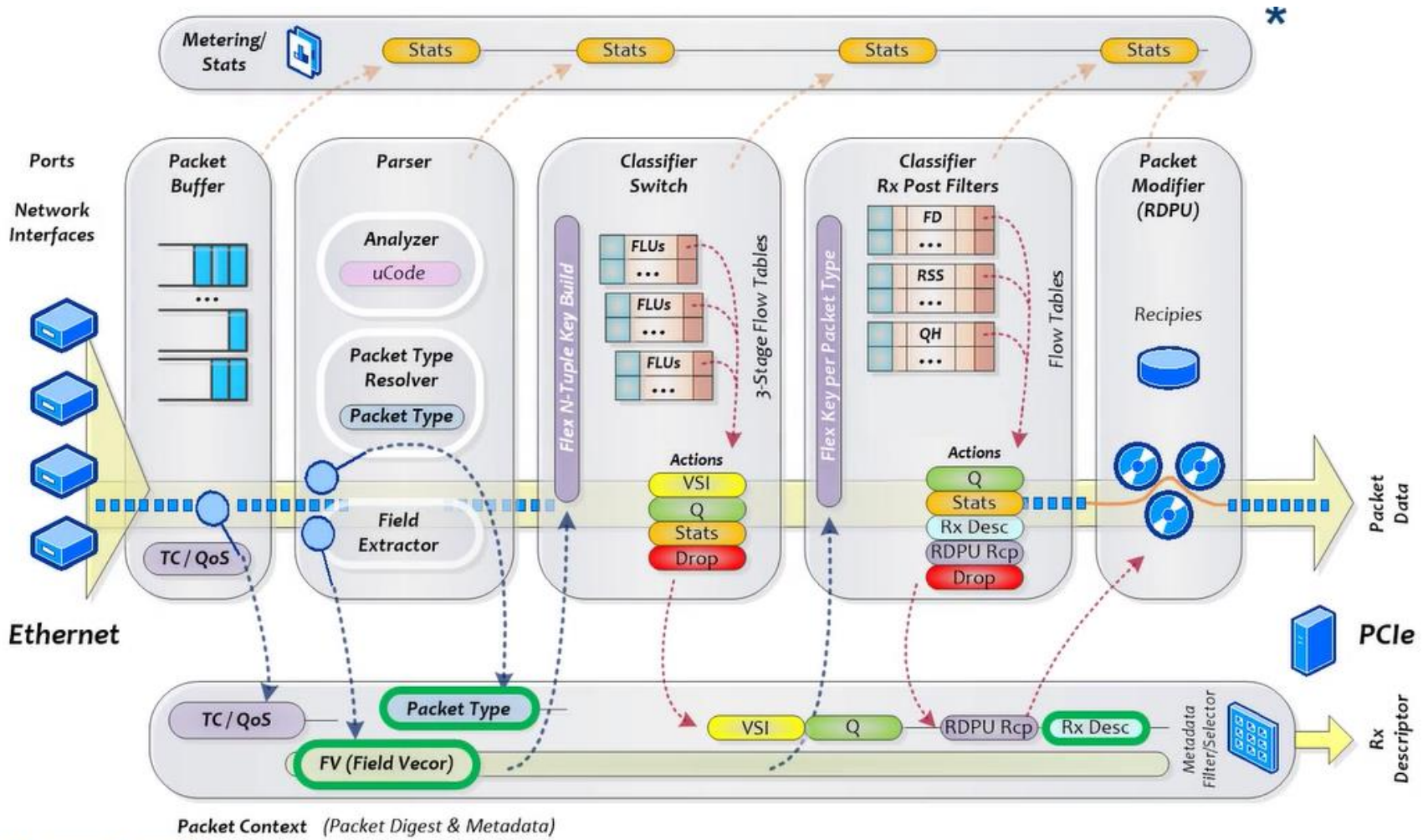


Classification – XL710 Vs 82599



82599EB		XL710
	Hash function	
Toeplitz 40 bytes key		Toeplitz 52 bytes key, Simple XOR, Symmetric (with Simple XOR)
	Hash input set	
static, 5-tuple only		flexible, > 10 fields from a packet can be used
	Flexible payload	
1 word (2 bytes)		up to 8 words from 3 locations within first 480 bytes for L2-L4
	Flow director	
exact and signature match		exact match only, > 10 fields from a packet can be used

NIC Anatomy



* Courtesy of Ronen Chayat

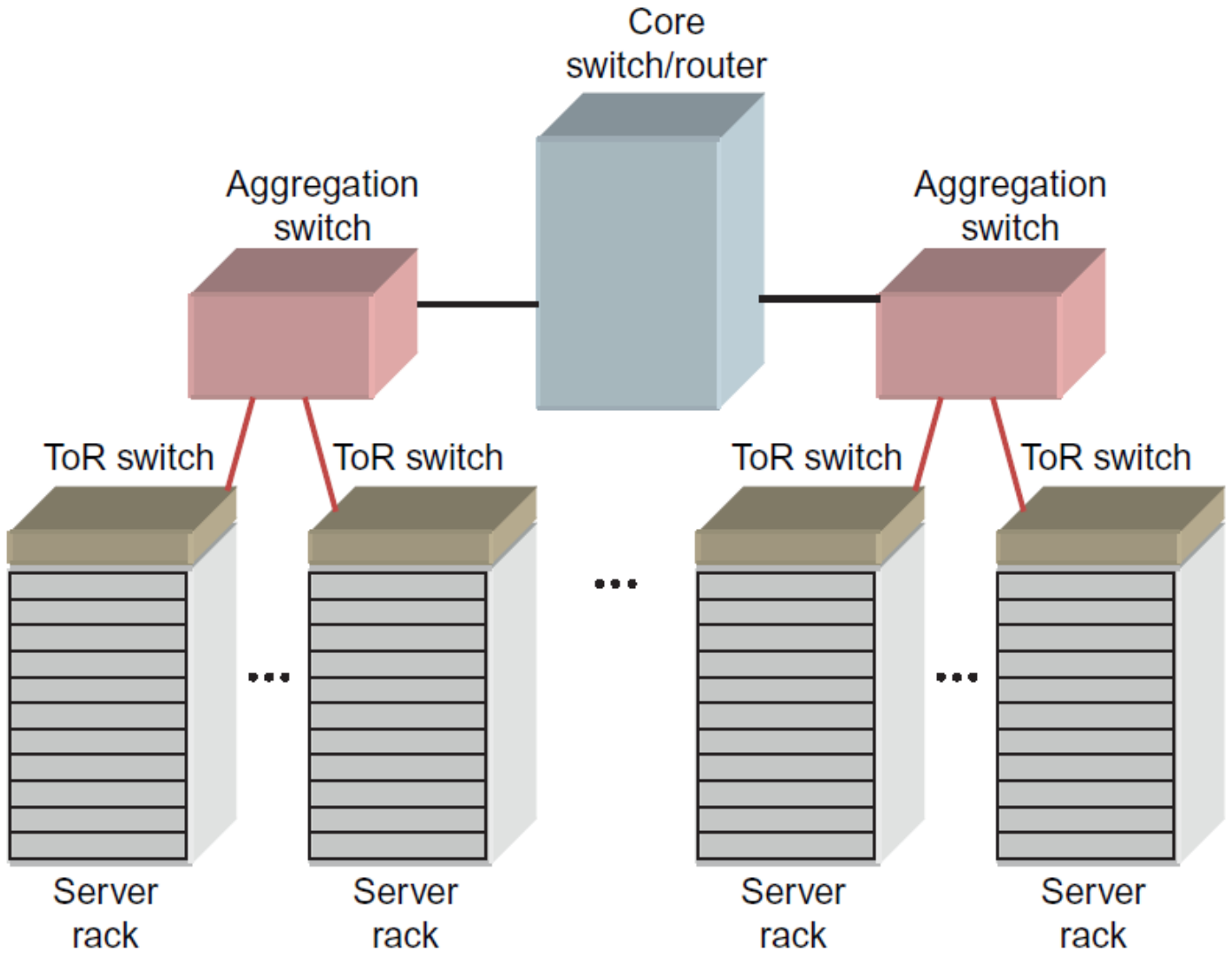
Classification

- **Hash filter (RSS):** load distribution to multiple queues using hash calculated over packet's field selected by input set. Hash signature extracted to Receive Descriptor.
- **Flow Director (FD):** pinning flow to the specific queue, extracting payload's data (up to 8 bytes) to Receive Description.
- **FD can run in "pass-through"** mode. In this mode FD extract data to RXd and then packets are distributed by RSS.
- **Tunnel (Clouds) Filters:** assign tunnelled packets (VXLAN, VXLAN-GPE, GRE, NVGRE) to a queue/VF

Customer Usage Models- Requirements

M Jay

What issues you see With 3-Tier Traditional Data Center Network?



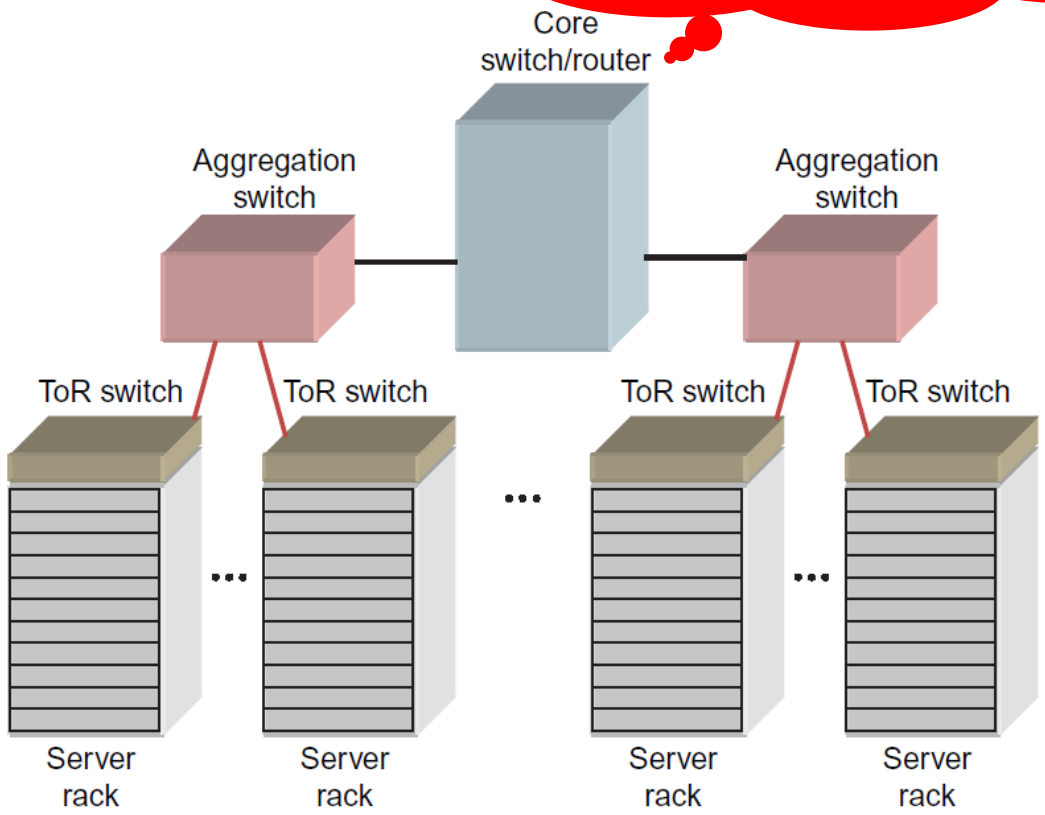
What Scaling Problems Do You See?

Issues With Traditional 3-Tier Enterprise Data Center Network

DPDK

1) As Data centers grew larger, aggregation switches ran out of ports, needing to use large swich/routers

2) Frame header processing at very high bandwidth- adds more congestion to the network



3) Since not designed with latency in mind, 3-tier networks do not do a good job of handling east west traffic

4) With Multi cores ramping in performance, ToR switch can't keep up with both storage (without dropping) + network bandwidth.

Larger Switches @ high bandwidth + l3 features => Expensive

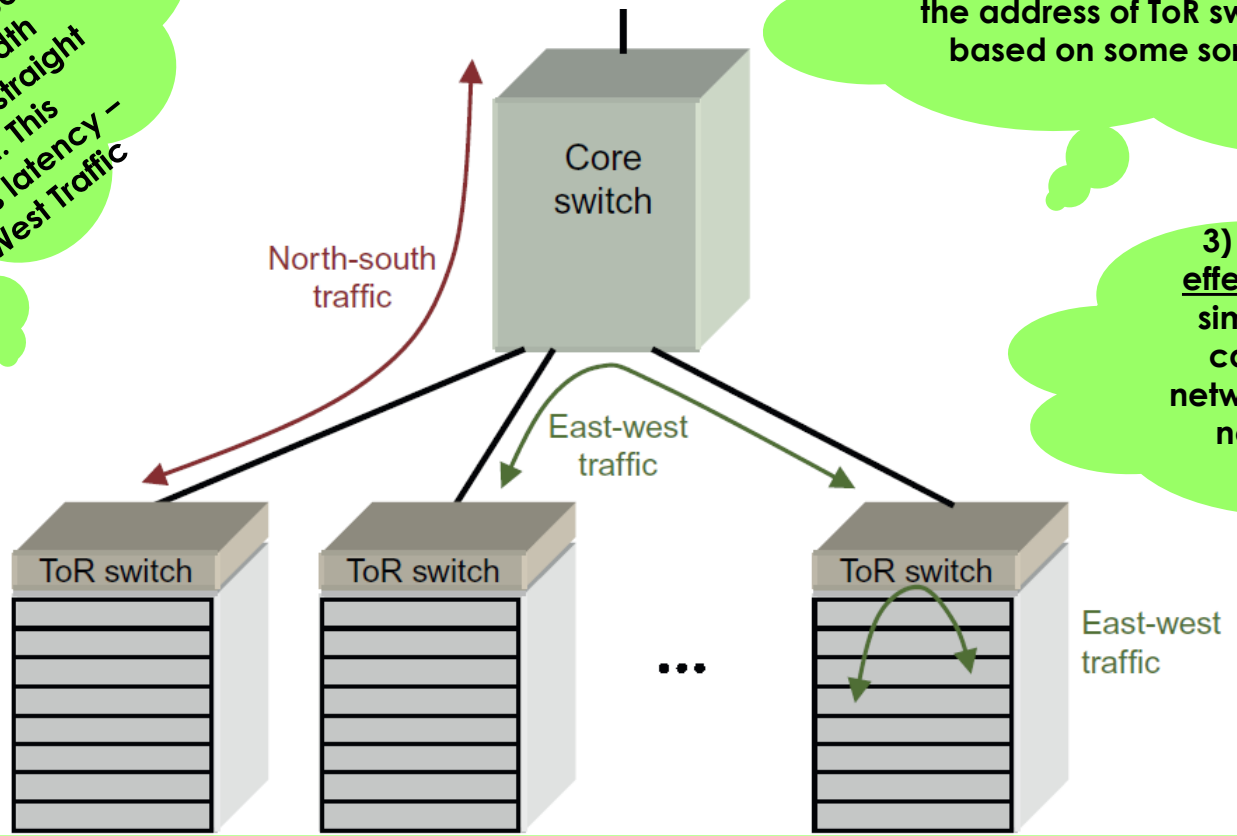
40 Gig Advantages - Flat Data Center Networks

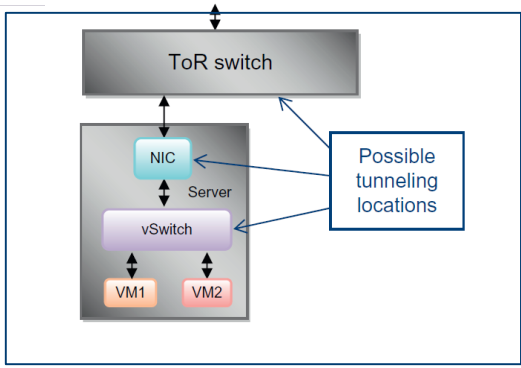
DPDK

1) Low Latency: Flatter network reduces potential congestion hot spots since bandwidth distribution is straight forward. This improves latency – East-West Traffic

2) Smaller Table Sizes with Tunneling Label (Cost effective): Compared to ToR Switches, smaller Table sizes since core switches can simply use tables containing the address of ToR switches in the network based on some sort of tunneling label.

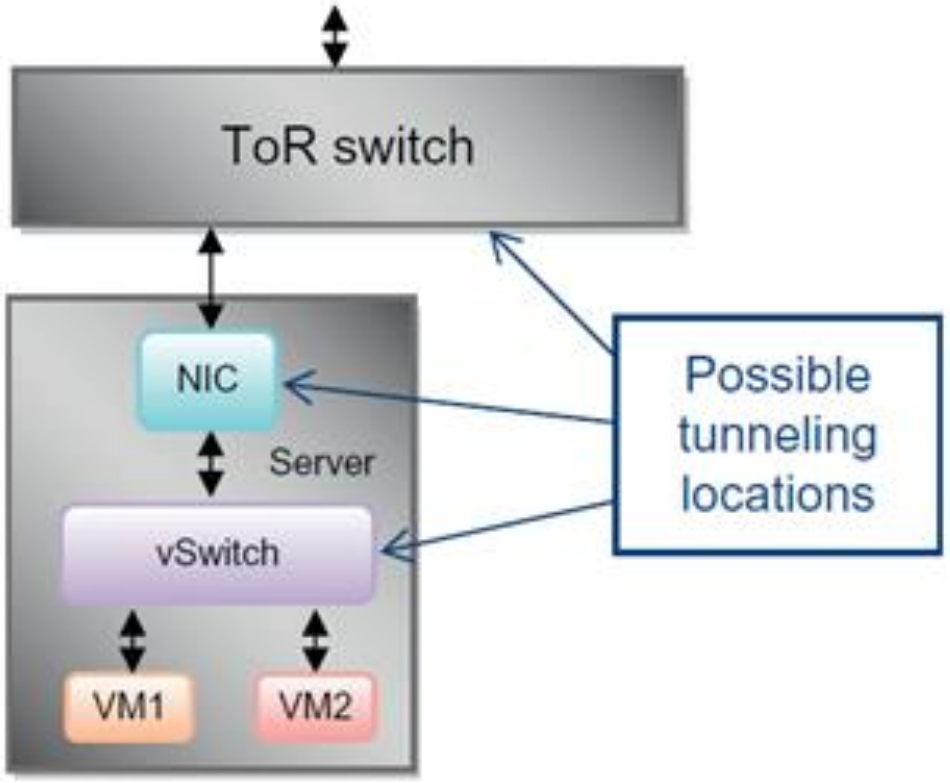
3) Simplify Frame Processing (Cost effective) : Frame processing can be simplified since tunneling functions can be moved to the edge of the network i.e., ToR Switch or vSwitch and not necessarily be done by core switch.





Low Latency, High Quality Network + Simplified Core Switch => Cost Effective

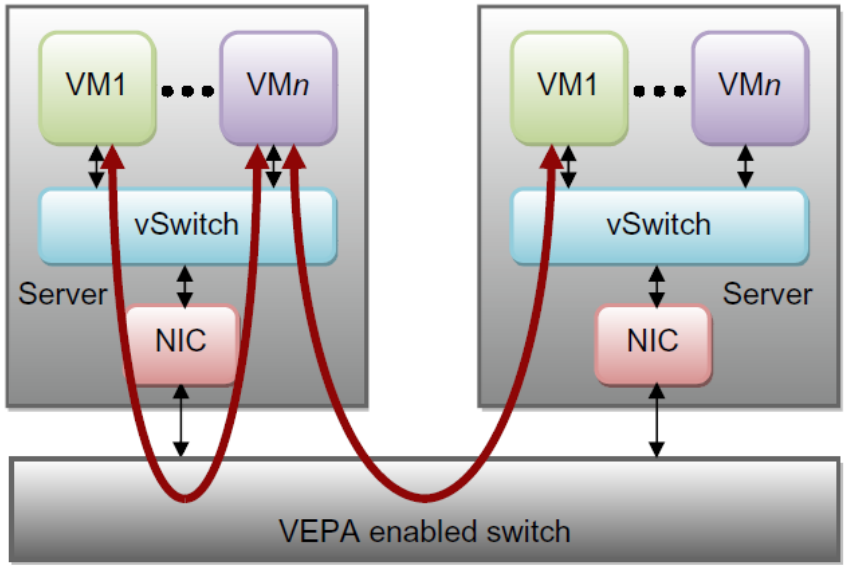
Possible Tunneling Locations



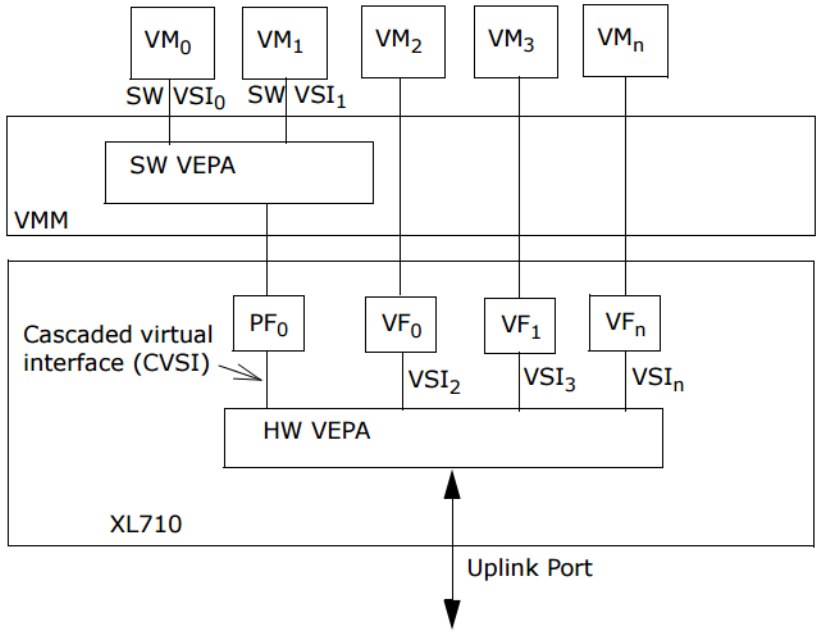
Tunneling at Vswitch	Tunneling at NIC	Tunneling at ToR Switch

VEPA – Consistent Treatment Of All Network Traffic

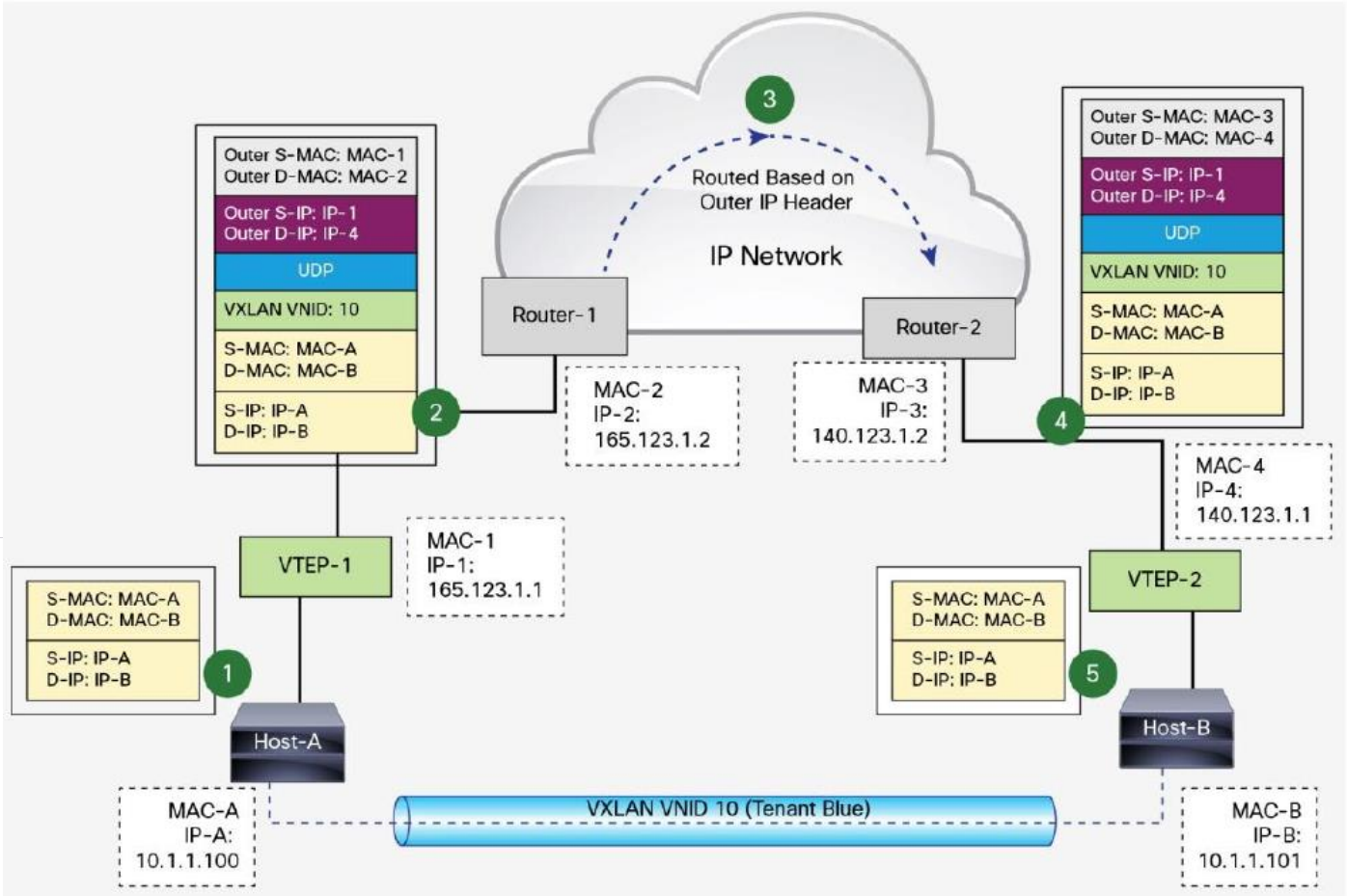
VEPA – An Overview



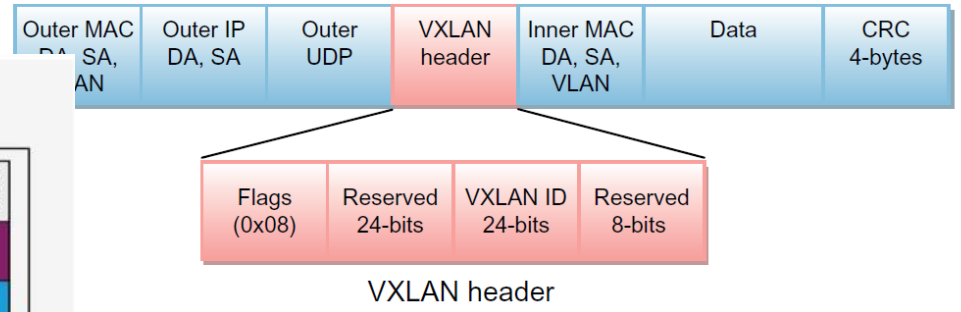
VEPA – XL710



VXLAN – Packet Flow

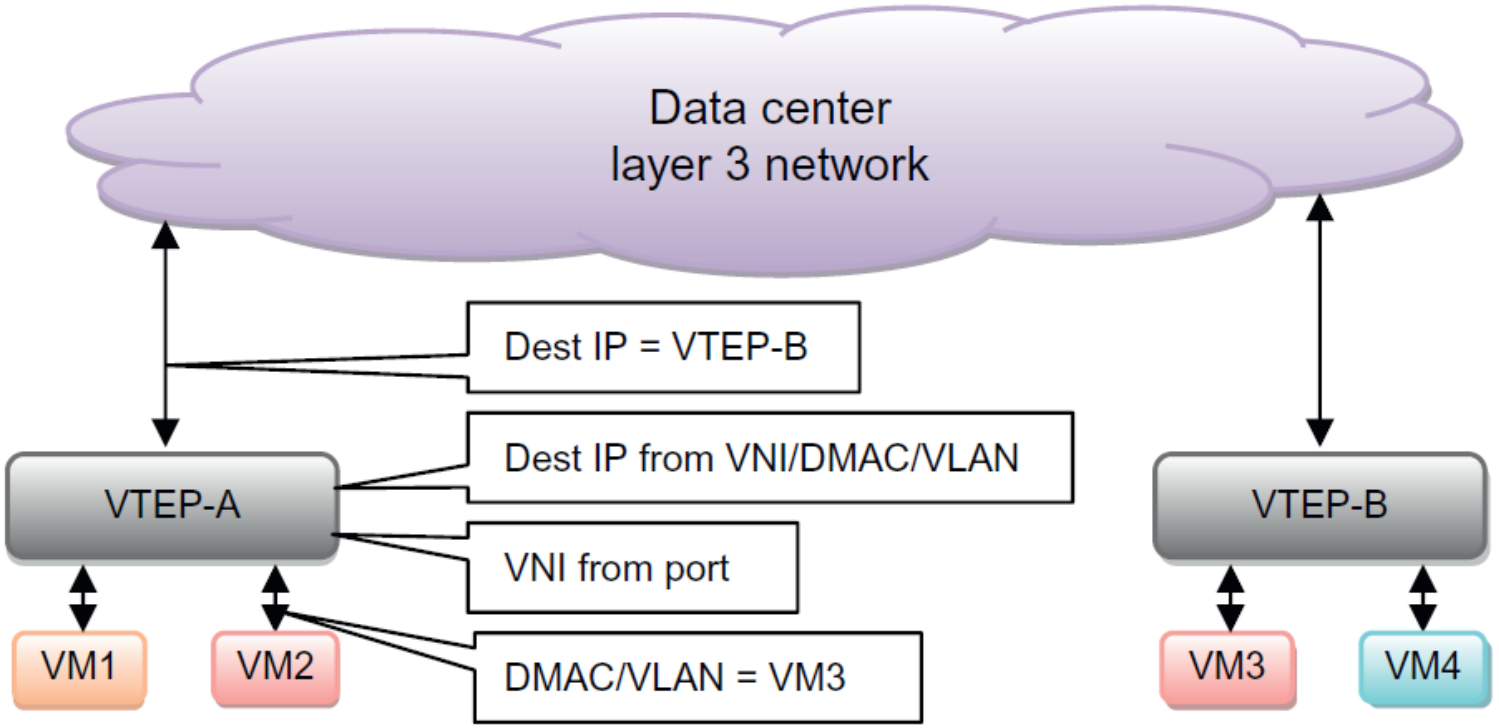


VXLAN encapsulated frame

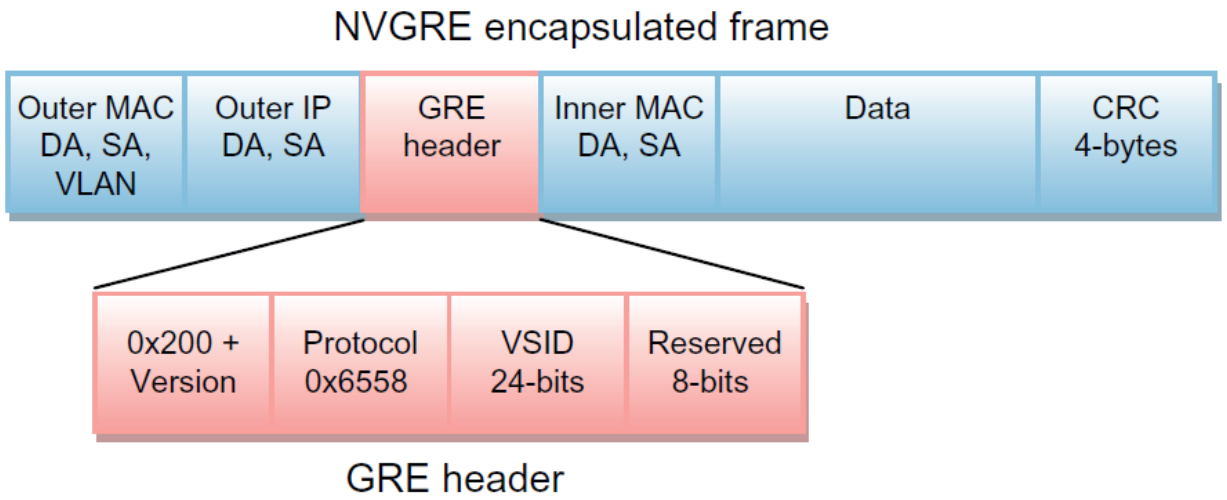


VXLAN header

Question: What is UDP Source Port Used For?



NVGRE



VXLAN	NVGRE
UDP + VXLAN header	Only GRE header
Inner L2 header contains VLAN tag	No VLAN tag in inner L2 Tag
UDP Port for Hash	Reserved 8 bits (Random for uniform distribution) + VSID for Hash

40 GbE - Step by Step Walk Through

Description	Requirement	Reference
What is important in my h/w Platform?	Ensure all the 4 memory channels are populated. AND use –n 4 in the command line also * Note: This is one important element to affect the performance	use "dmidecode -t memory" to check the memory status. **Since this is very important please procure additional memory and populate all the memory channels**
Where the NIC should be plugged in? And Why?	Use PCIe Gen3 slots, such as Gen3 x8 or Gen3 x16 NUMA considerations	Because PCIe Gen2 slots can't provide enough bandwidth for 2x10G and above.
What needs to be updated in NIC?	Make Sure each NIC has flashed the latest version of NVM/firmware.	Go do downloadcenter.intel.com and search for XL710 NVM Update. **It takes you here:** https://downloadcenter.intel.com/search?keyword=NVM+Update+Utility+for+Intel%C2%AE+Ethernet+Converged+Network+Adapter+XL710+%26+X710+Series

40 GbE - Step by Step Walk Through

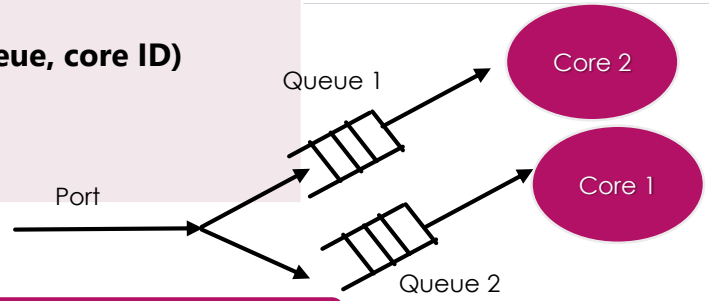
Description	Requirement	Reference
BIOS settings	Refer BIOS Settings	
Linux System Essentials	Real Time Nature of the Process, cgroup	
Huge Page	1) Size of the FIB Table, 2) Locality challenges of packets	TLB Miss, Page Walk
Scheduler	Isolcpus option under title Grub Parameters - Essential Requirement	

BIOS Setting

Menu (Advanced)	BIOS Setting	Required Setting	BIOS default
CPU Configuration ->Advanced Power Management Configuration			
	Power Technology	Disable	Custom
-> CPU P State Control	ESIT (P-States)	Disable	Enable
-> CPU P State Control	Turbo Mode	Disable	Enable
-> CPU P State Control	P-State Coordination	HW_ALL	HW_ALL
-> CPU C State Control	Turbo Mode	Disable	Enable
-> CPU C State Control	CPU C3 Report	Disable	Enable
-> CPU C State Control	CPU C6 Report	Disable	Enable
-> CPU C State Control	Package C State Limit	[C6 (Retention)]	[C6 (Retention)]
-> CPU C State Control	Enhanced Halt State(C1E)	Disable	Enable
Chipset Configuration			
-> North Bridge -> QPI Configuration	Isoc Mode	Disable	Disable
	COD Enable	Disable	Auto
	Early Snoop	Disable	Auto
-> North Bridge -> Memory Configuration	Enforce POR	Disable	Auto
	Memory Frequency	2133	Auto
	DRAM RAPL Baseline	Disable	Auto
-> North Bridge -> IIO Configuration	Intel VT for Directed I/O (VT-d)	Disable	Enable
PCIe/PCI/PnP Configuration	ASPM	Disable	Disable

40 GbE - Step by Step Walk Through

#	Description	Requirement	Reference
9	For Intel® 40 Gig NICs, special configurations should be set before compiling it. This is very Important.	For at least DPDK release 1.8, 2.0 and 2.1, in <dpdk_folder>/config/common_linuxapp [this step is not needed from R16.07] CONFIG_RTE_PCI_CONFIG=y and CONFIG_RTE_PCI_EXTENEDED_TAG=on	This helps increase the efficiency of PCIe by increasing the number of outstanding transactions from 36 to 256.
10	**Phase 1:** Running l3fwd application & command to run for testing 2 x 10 G Please only run l3fwd, to start with, to have a baseline performance for comparison purpose.	With 2 core, 2 Threads, 2 Ports (with only 1 Queue/port) **Note:** Please do not run full application. Run l3fwd to benchmark your platform and configuration.	./l3fwd -c 0x3fc00 -n 4 -w 05:00.0 -w 05:00.1 -- -p 0x3 --config '(0,0,10),(1,0,11)' ***Note config (port, queue, core ID) is the format above**
11	In l2fwd, #define NB_MBUF 16384 [This increases buffer count to 16K – from 8K] – in the file examples/l2fwd/main.c	Change in examples/l2fwd/main.c the values of RTE_TEST_RX_DESC_DEFAULT and RTE_TEST_TX_DESC_DEFAULT both to 1024.	L2fwd. After making the changes, Save. Build l2fwd with make.
12	**Phase 2:** Running l3fwd application & command to run for testing 4 x 10 G	With 4 core, 4 Threads, 4 Ports (with only 1 Queue/port) – Single port x 40 Gig configuration	./l3fwd -c 0x3fc00 -n 4 -- -p 0xf --config '(0,0,10),(1,0,11),(2,0,12),(3,0,13) ***Note config (port, queue, core ID) is the format above**



Use 2 Cores

System Configuration

Hardware		
Motherboard		Supermicro* X10DRX
CPU	Product	Intel ® Xeon® Processor E5-2658 v4
	Speed(MHz)	2300
	Number of CPUs(per socket)	14Cores/28 Threads/socket
	Stepping	M0
	LLCCache	35840K
	Max TDP(W)	105W
Memory	Vendor	Samsung*
	Type	DDR4-2400 RDIMM
	Configured Speed(MT/s)	2400
	Part Number	36ASF2G72PZ-2G3A3
	Size per DIMM	16GB
	Channel	1 DIMM/Channel, 4 Channel per Socket
BIOS	Vendor	American Megatrends Inc.*
	Version	Version 2.0 Release date 12/17/2015
OS	Vendor	Fedora 23
	Version	4.2.3-300.fc23.x86_64

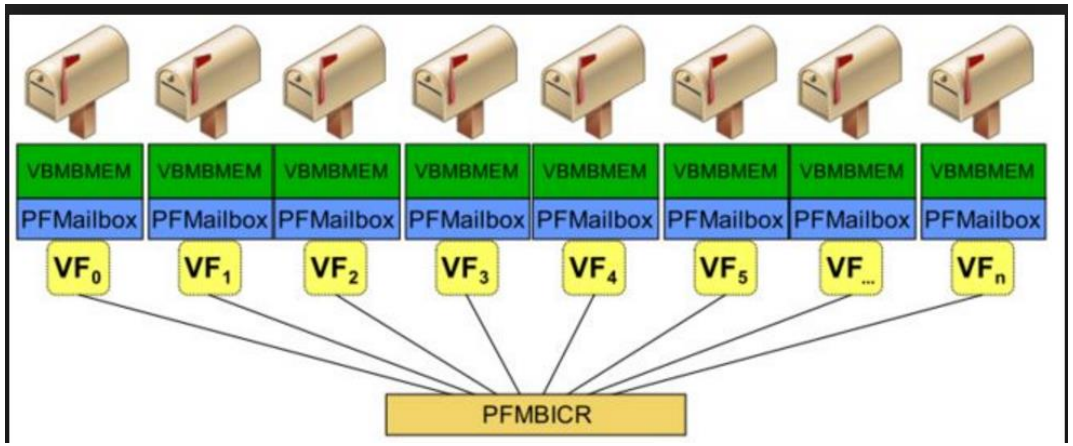
BIOS Tuning Settings

Menu (Advanced)	BIOS Setting	Required Settings for Performance	BIOS Default
CPU Configuration -> Advanced Power Management Configuration			
	Power Technology	Disable	Custom
	Energy Performance Tuning	Disable	Enable
	Energy Performance BIAS Setting	Performance	Enable
	Energy Efficient Turbo	Disable	Enable
-> CPU P State Control	EIST (P-States)	Disable	Enable
-> CPU P State Control	Turbo Mode	Disable	Enable
-> CPU P State Control	P-State Coordination	HW_ALL	HW_ALL
-> CPU C State Control	Package C State Limit	[C0/C1 State]	[C6 (Retention)]
-> CPU C State Control	CPU C3 Report	Disable	Enable
-> CPU C State Control	CPU C6 Report	Disable	Enable
-> CPU C State Control	Enhanced Halt State (C1E)	Disable	Enable
Chipset Configuration			
-> North Bridge -> IIO Configuration	EV DFX Features	Enable	Disable
	Intel VT for Directed I/O (VT-d)	Disable	Enable
->North Bridge -> IOAT Configuration	Enable IOAT	Enable	Enable
	No Snoop	Disable	Disable
	Relaxed Ordering	Disable	Disable
-> North Bridge -> QPI Configuration			
	Link L0 P	Disable	Enable
	Link L1	Disable	Enable
	COD Enable	Disable	Auto
	Early Snoop	Disable	Auto
	Isoc Mode	Disable	Disable
-> North Bridge ->Memory Configuration	Enforce POR	Disable	Auto
	Memory Frequency	2400	Auto
	DRAM RAPL Baseline	Disable	Auto
	A7 Mode	Enable	Enable
-> South Bridge	EHCI Hand-off	Disable	Auto
	USB3.0 Support	Disable	Enable
PCIe/PCI/PnP Configuration	ASPM	Disable	Enable
	Maximum Payload	AUTO	AUTO
	Maximum Read Payload	AUTO	AUTO
	Onboard LAN 1 OPROM	Disable	PXE

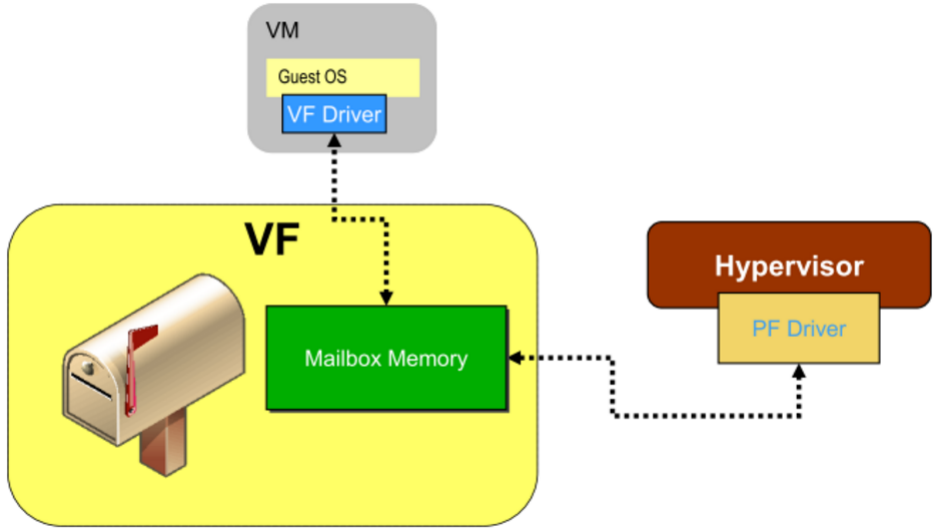
Latency & Throughput – How To Improve?

- Latency Hiding – Prefetch
- Throughput - Bulk

Admin Queues – DOs and Don'ts



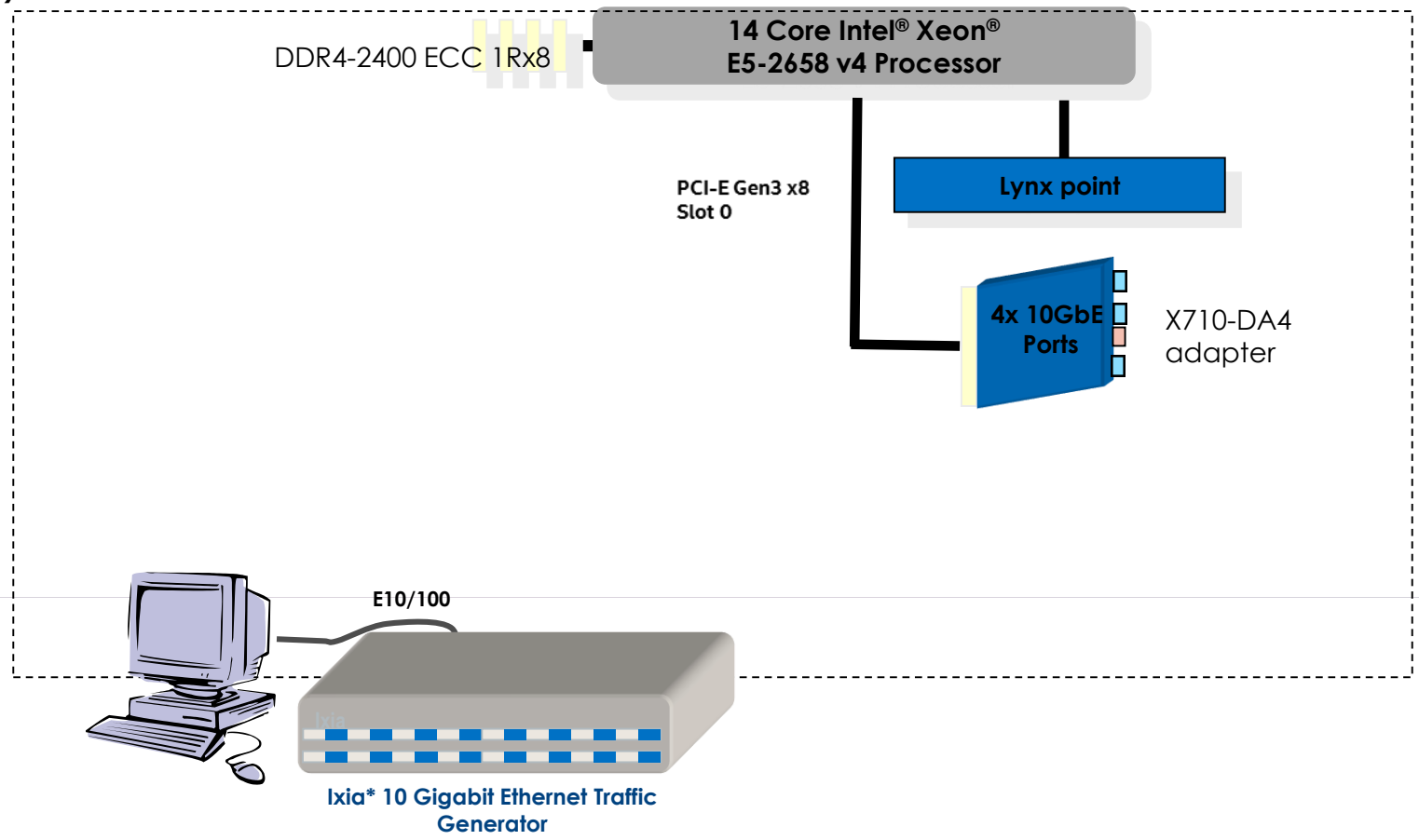
- XL710 Admin Queue Versus 82599 Mail Box
- Run time changing MTU? - Think Again. Why?
- Run time Resetting VFs from PF?



FUNCTIONAL PERFORMANCE MEASUREMENT FOR COMMUNICATIONS: LAYER 3 FORWARDING USING 10GBE AND 40GBE

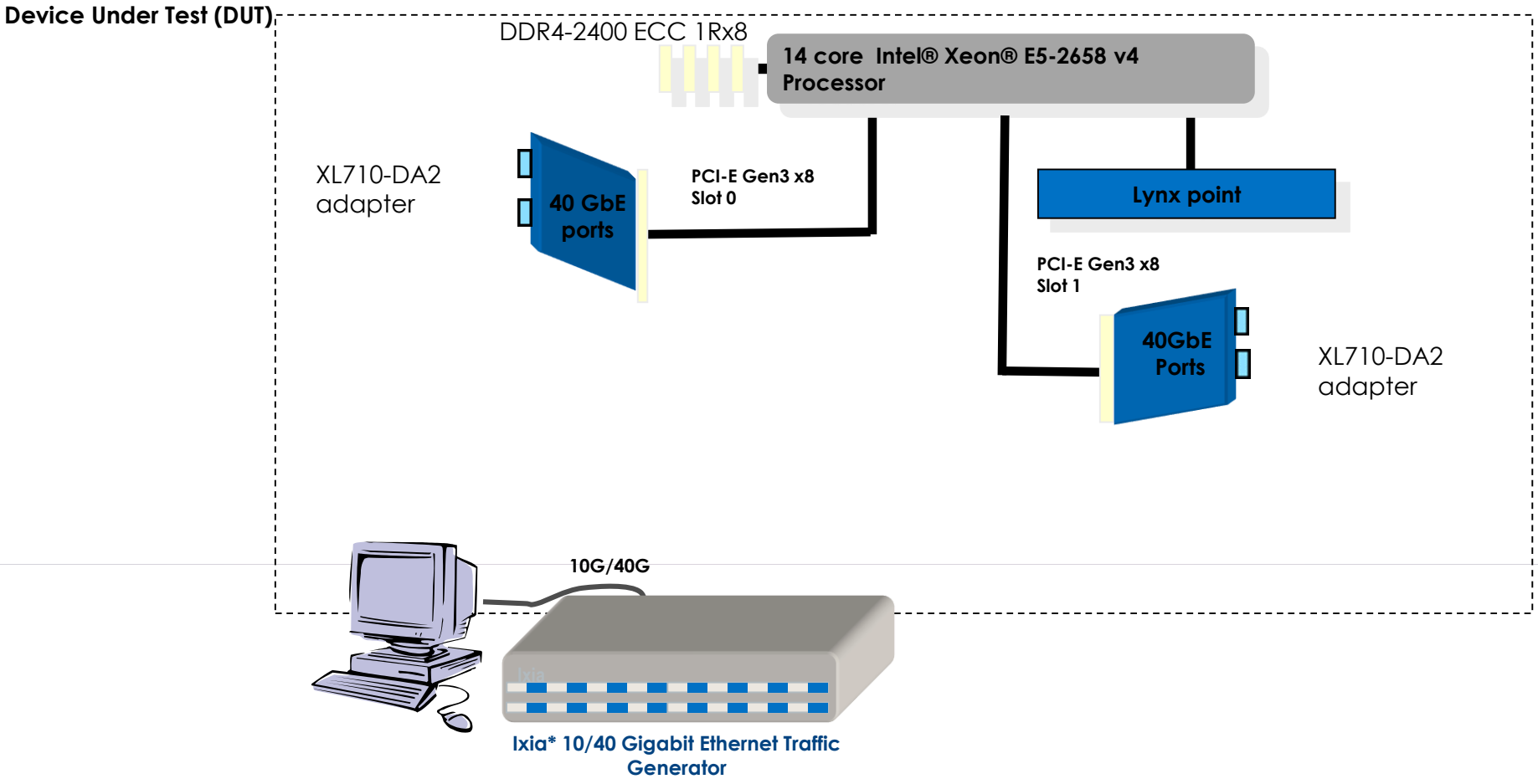
Test Setup for 10G Cards

Device Under Test (DUT)



* Other names and brands may be claimed as the property of others.

Test Setup for 40G Cards



* Other names and brands may be claimed as the property of others.

Test Setup -Cont.

DUT:

- Intel® Xeon® E5-2658 v4 processor,35MB L3 cache
- Super Micro* Platform (X10DRX)
- DDR4 2400 MHz, 4 x 1Rx4 registered ECC 16GB (total 64GB), 4 memory channels per socket Configuration, 1 DIMM per channel
- 1 x Intel X710-DA4-FH PCI-E Gen3X8 Quad Port Ethernet Controller (NVM: 5p04)
- 2 x Intel XL710-DA2 PCI-E Gen3x8 Dual Port 40GbE Ethernet Controller (NVM: 5p04)

IXIA* Traffic Parameters:

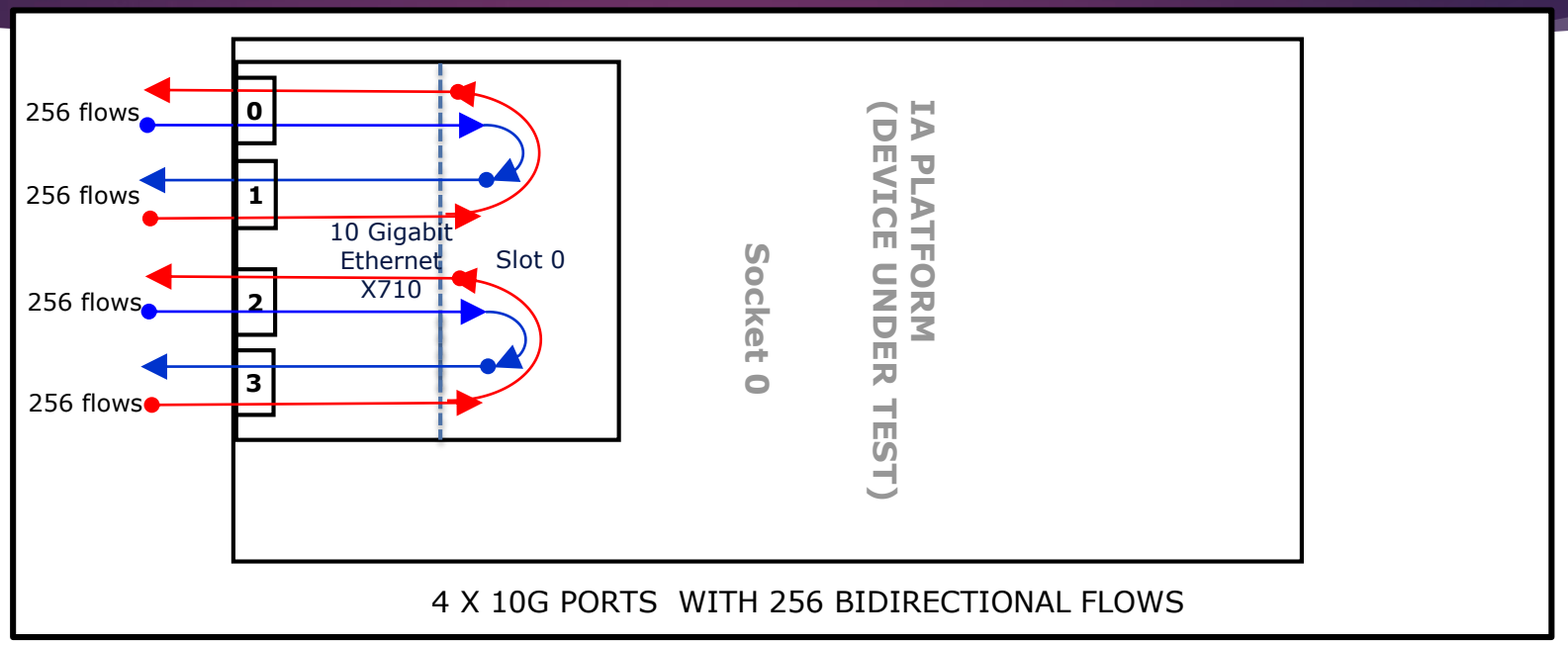
- Acceptable Frame Loss: 0.00001%
- Resolution: 0.1
- Traffic Duration: 20 Seconds

Software:

- BIOS version: Version: 2.0 & Date: 12/17/2015
- Operating system: Fedora 23
- Kernel version: 4.2.3-300.fc23.x86_64
- IxNetwork* : 7.40 EA
- DPDK version: 16.04
- DPDK L3fwd example application on Linux user space (LPM for route lookup)
  - .hw_ip_checksum = 0, /**< IP checksum offload enabled */
  - #define RTE_TEST_RX_DESC_DEFAULT 1024
  - #define RTE_TEST_TX_DESC_DEFAULT 1024

Flow Traffic Configuration

4 x10G Ports

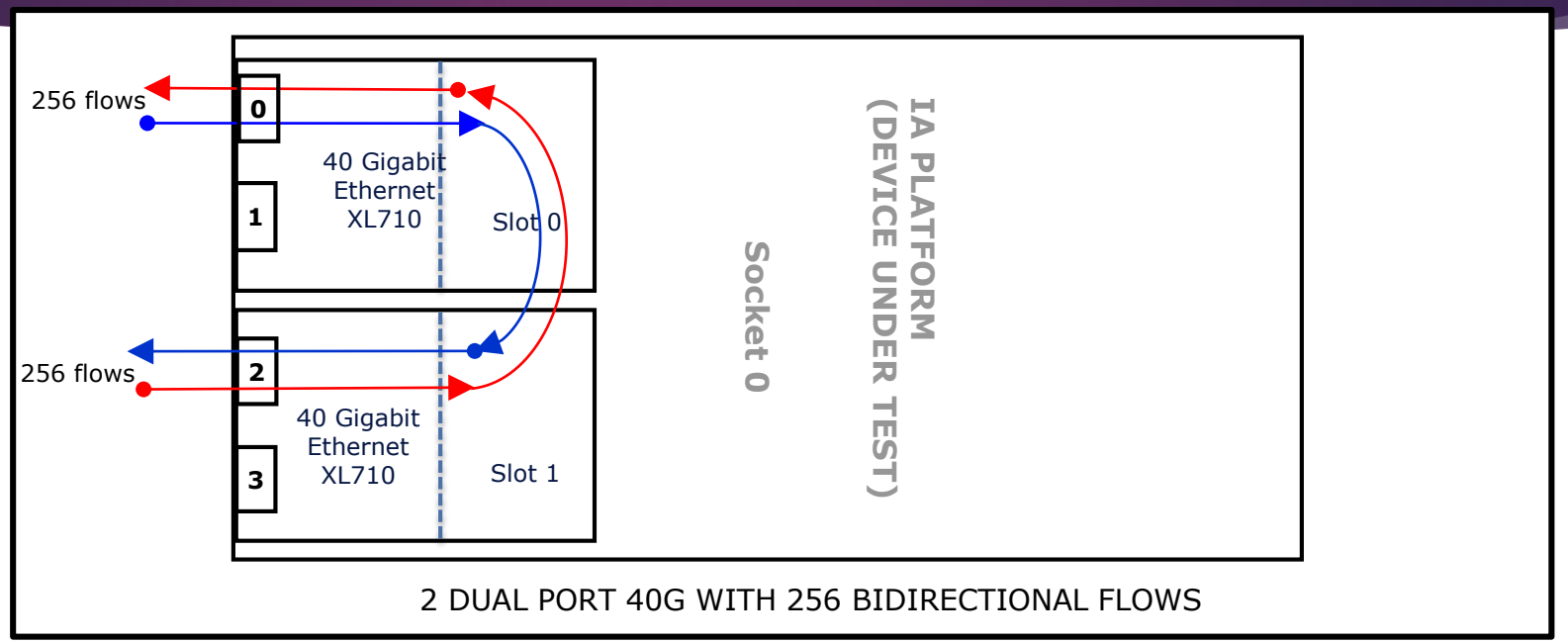


2 port configuration with 256 bi-directional flows per port

- Port 0 -> Port 1
- Port 1 -> Port 0
- Port 2 -> Port 3
- Port 3 -> Port 2

Flow Traffic Configuration

2 x40G Ports



2 port configuration with 256 bi-directional flows per port

- Port 0 -> Port 1
- Port 1 -> Port 0

Polling Affinity for Ethernet Queues- 4x10G ports

- **2 ports – (1 Core/1 Thread /1Queue)**
  - CPU1 (Core 1 SMT 0) polls port 0
  - CPU1 (Core 1 SMT 0) polls port 1
  - CPU1 (Core 1 SMT 0) polls port 2
  - CPU1 (Core 1 SMT 0) polls port 3

- **2 ports - (1 Core / 2 Threads/1 Queue**)
  - CPU1 (Core 1 SMT 0) polls port 0
  - CPU2 (Core 15 SMT 1) polls port 1
  - CPU1 (Core 1 SMT 0) polls port 2
  - CPU2 (Core 15 SMT 1) polls port 3

- **2 ports - (2 Core / 2 Threads/1 Queue)**
  - CPU1 (Core 1 SMT 0) polls port 0
  - CPU1 (Core 2 SMT 0) polls port 1
  - CPU1 (Core 1 SMT 0) polls port 2
  - CPU1 (Core 2 SMT 0) polls port 3

Each polling core has 100% CPU Utilization.
Remaining cores are IDLE

Polling Affinity for Ethernet Queues-2x40G ports

- **2 ports –** (1 Core / 1 Thread/2 Queues)
  CPU1 (Core 1 SMT 0) polls port 0 queue 0
  CPU1 (Core 1 SMT 0) polls port 0 queue 1
  CPU1 (Core 1 SMT 0) polls port 1 queue 0
  CPU1 (Core 1 SMT 0 polls port 1 queue 1

- **2 ports –** (1 Core / 2 Thread/2 Queues)
  CPU1 (Core 1 SMT 0) polls port 0 queue 0
  CPU2 (Core 15 SMT 1) polls port 0 queue 1
  CPU1 (Core 1 SMT 0) polls port 1 queue 0
  CPU2 (Core 15 SMT 1) polls port 1 queue 1

- **2 ports –** (2 Core / 2 Thread/2 Queues)
  CPU1 (Core 1 SMT 0) polls port 0 queue 0
  CPU1 (Core 2 SMT 0) polls port 0 queue 1
  CPU1 (Core 1 SMT 0) polls port 1 queue 0
  CPU1 (Core 2 SMT 0) polls port 1 queue 1

- **2 ports –**(2 Core / 4 Thread/2 Queues)
  CPU1 (Core 1 SMT 0) polls port 0 queue 0
  CPU2 (Core 15 SMT 1) polls port 0 queue 1
  CPU1 (Core 2 SMT 0) polls port 1 queue 0
  CPU2 (Core 16 SMT 1) polls port 1 queue 1

- **2 ports –**(4 Core / 4 Thread/2 Queues)
  CPU1 (Core 1 SMT 0) polls port 0 queue 0
  CPU1 (Core 2 SMT 0) polls port 0 queue 1
  CPU1(Core 3 SMT 0) polls port 1 queue 0
  CPU1(Core 4 SMT 0) polls port 1 queue 1

Each polling core has 100% CPU Utilization.
Remaining cores are IDLE

Reference & Acknowledgements

- Cloud Networking – Understanding Cloud-Based Data Center Networks – Gary Lee.
- http://cat.intel.com Get NDA performance foils here.
- **DPDK Cook Book on Vtune – M Jay**
  - https://software.intel.com/en-us/articles/profile-dpdk-code-with-intel-vtune-amplifier
- DST 2016: v-ISG-Fortville: Explaining Fortville Features Enabled with DPDK Rel 16.04 – Hash and Flow Director Filters, Native MPLS (Virtual) – Andrey Chilikin, Eoin Walsh.
- CISCO White Paper January 2016 – VXLAN Best Practices
- Intel® XL710/X710 Data Sheet
- George for Performance setup
- Rashmin foils for Virtualization
- http://blog.jgriffiths.org/?p=929 Deep Dive: How does NSX Distributed Router Work

Questions?

M Jay
Muthurajan.Jayakumar@intel.com
Helin Zhang
Helin.Zhang@intel.com

Comparing XL710/X710 to Prior NIC 82599

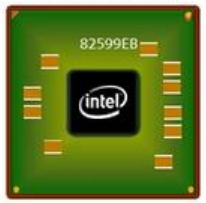
DPDK

Fortville family (XL710/X710)

- Low power single chip design for PCI Express* 3.0
- **Support for standard and custom network headers**
- Intelligent load balance for high performance traffic flows
- Network virtualization Overlay stateless offloads for VXLAN, NVGRE, Geneve, VXLAN GPE, NSH, MPLS
- **Flexible pipeline processing – add new features after production by upgrading firmware**

Power Efficiency Improvements

Comparing Controller Typical Power



2 x 10GbE
5.2 watts[1]
Typical Power

1 x 40GbE
3.3 watts[2]
Typical Power

Source as of Aug 2014: 1: 82599 Datasheet rev 2.0 Table 11.5 for 2x10GbE Twinax Typical Power [W]
2: XL710 Data sheet rev 1.21 Table 14-7 Typical Active Power 1x40GbE Power [W]

30% UP TO 30% Reduction TYPICAL POWER

65% UP TO 65% Reduction In GIGABIT PER WATT

2x Increase in TOTAL BANDWIDTH

Port density

GENEVE & NSH added after the chip is released - Flexibility !