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.

Copyright © Intel Corporation 2016. All rights reserved.

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

* Other names and brands may be claimed as the property of others.
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
Helin
## Classification – XL710 Vs 82599

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

* Courtesy of Ronen Chayat
- **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
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

1) As data centers grow larger, aggregation switches ran out of ports. Needing to use large switch/ 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

1) Low Latency: Flatter network reduces potential congestion hot spots since bandwidth 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

<table>
<thead>
<tr>
<th>Tunneling at Vswitch</th>
<th>Tunneling at NIC</th>
<th>Tunneling at ToR Switch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
VEPA – Consistent Treatment Of All Network Traffic

VEPA – An Overview

VEPA – XL710

VEPA enabled switch

Cascaded virtual interface (CVSI)

VSI0

VSI1

VSI2

VSI3

VSIa

PF0

VF0

VF1

VFn

VM0

VM1

VM2

VM3

VMn

SW

SW VEPA

VMM

XL710

Uplink Port
VXLAN – Packet Flow

VXLAN encapsulated frame

VXLAN header

1. Host-A
   - MAC-A: MAC-A
   - IP-A: 10.1.1.100

2. VTEP-1
   - MAC-1: IP-1: 165.123.1.1

3. Router-1
   - MAC-2: IP-2: 165.123.1.2
   - MAC-3: IP-3: 140.123.1.2

4. VTEP-2
   - MAC-4: IP-4: 140.123.1.1

5. Host-B
   - MAC-B: IP-B: 10.1.1.101

Outer MAC
Outer IP
Outer UDP
VXLAN header
Inner MAC
DA, SA, VLAN
Data
CRC
4-bytes

VXLAN Network

UDP
VXLAN VNI: 10
S-MAC: MAC-A
D-MAC: MAC-B
S-IP: IP-A
D-IP: IP-B

VXLAN VNI: 10 (Tenant Blue)
Question: What is UDP Source Port Used For?

VTEP-A
- Dest IP = VTEP-B
- Dest IP from VNI/DMAC/VLAN
- VNI from port
- DMAC/VLAN = VM3

VTEP-B
- VM3
- VM4

VM1
VM2

Data center layer 3 network
### VXLAN vs NVGRE

<table>
<thead>
<tr>
<th>VXLAN</th>
<th>NVGRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>UDP + VXLAN header</td>
<td>Only GRE header</td>
</tr>
<tr>
<td>Inner L2 header contains VLAN tag</td>
<td>No VLAN tag in inner L2 Tag</td>
</tr>
<tr>
<td>UDP Port for Hash</td>
<td>Reserved 8 bits (Random for uniform distribution) + VSID for Hash</td>
</tr>
</tbody>
</table>

**NVGRE encapsulated frame**

- Outer MAC DA, SA, VLAN
- Outer IP DA, SA
- GRE header
- Inner MAC DA, SA
- Data
- CRC 4-bytes

**GRE header**

- 0x200 + Version
- Protocol: 0x6558
- VSID: 24-bits
- Reserved: 8-bits
# 40 GbE - Step by Step Walk Through

<table>
<thead>
<tr>
<th>Description</th>
<th>Requirement</th>
<th>Reference</th>
</tr>
</thead>
</table>
| 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.  
* Note: This is one important element to affect the performance of the system. 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+Seri es](https://downloadcenter.intel.com/search?keyword=NVM+Update+Utility+for+Intel%C2%AE+Ethernet+Converged+Network+Adapter+XL710+%26+X710+Seri es) |
<table>
<thead>
<tr>
<th>Description</th>
<th>Requirement</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIOS settings</td>
<td>Refer BIOS Settings</td>
<td></td>
</tr>
<tr>
<td>Linux System Essentials</td>
<td>Real Time Nature of the Process, cgroup</td>
<td></td>
</tr>
<tr>
<td>Huge Page</td>
<td>1) Size of the FIB Table, 2) Locality challenges of packets</td>
<td>TLB Miss, Page Walk</td>
</tr>
<tr>
<td>Scheduler</td>
<td>Isolcpus option under title Grub Parameters - Essential Requirement</td>
<td></td>
</tr>
<tr>
<td>Menu (Advanced)</td>
<td>BIOS Setting</td>
<td>Required Setting</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>---------------------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td><strong>BIOS Setting</strong></td>
<td><strong>Required Setting</strong></td>
<td><strong>BIOS default</strong></td>
</tr>
<tr>
<td><strong>Menu (Advanced)</strong></td>
<td><strong>BIOS Setting</strong></td>
<td><strong>Required Setting</strong></td>
</tr>
<tr>
<td><strong>CPU Configuration -&gt; Advanced</strong></td>
<td><strong>Power Technology</strong></td>
<td><strong>Disable</strong></td>
</tr>
<tr>
<td><strong>Power Management Configuration</strong></td>
<td><strong>ESIT (P-States)</strong></td>
<td><strong>Disable</strong></td>
</tr>
<tr>
<td><strong>CPU P State Control</strong></td>
<td><strong>Disable</strong></td>
<td><strong>Enable</strong></td>
</tr>
<tr>
<td><strong>CPU P State Control</strong></td>
<td><strong>Enable</strong></td>
<td><strong>Disable</strong></td>
</tr>
<tr>
<td><strong>CPU P State Control</strong></td>
<td><strong>HW_ALL</strong></td>
<td><strong>HW_ALL</strong></td>
</tr>
<tr>
<td><strong>CPU C State Control</strong></td>
<td><strong>Turbo Mode</strong></td>
<td><strong>Enable</strong></td>
</tr>
<tr>
<td><strong>CPU C State Control</strong></td>
<td><strong>CPU C3 Report</strong></td>
<td><strong>Disable</strong></td>
</tr>
<tr>
<td><strong>CPU C State Control</strong></td>
<td><strong>CPU C6 Report</strong></td>
<td><strong>Disable</strong></td>
</tr>
<tr>
<td><strong>CPU C State Control</strong></td>
<td><strong>Package C State Limit</strong></td>
<td><strong>[C6 (Retention)]</strong></td>
</tr>
<tr>
<td><strong>CPU C State Control</strong></td>
<td><strong>Enhanced Halt State(C1E)</strong></td>
<td><strong>Disable</strong></td>
</tr>
<tr>
<td><strong>Chipset Configuration</strong></td>
<td><strong>Isoc Mode</strong></td>
<td><strong>Disable</strong></td>
</tr>
<tr>
<td><strong>-&gt; North Bridge -&gt; QPI Configuration</strong></td>
<td><strong>COD Enable</strong></td>
<td><strong>Disable</strong></td>
</tr>
<tr>
<td><strong>-&gt; North Bridge -&gt; Memory Configuration</strong></td>
<td><strong>Early Snoop</strong></td>
<td><strong>Disable</strong></td>
</tr>
<tr>
<td><strong>-&gt; North Bridge -&gt; IIO Configuration</strong></td>
<td><strong>Enforce POR</strong></td>
<td><strong>Disable</strong></td>
</tr>
<tr>
<td><strong>PCIe/PCI/PnP Configuration</strong></td>
<td><strong>Memory Frequency</strong></td>
<td><strong>2133</strong></td>
</tr>
<tr>
<td><strong>-&gt; North Bridge -&gt; IIO Configuration</strong></td>
<td><strong>DRAM RAPL Baseline</strong></td>
<td><strong>Disable</strong></td>
</tr>
<tr>
<td><strong>PCIe/PCI/PnP Configuration</strong></td>
<td><strong>Intel VT for Directed I/O (VT-d)</strong></td>
<td><strong>Disable</strong></td>
</tr>
<tr>
<td><strong>PCIe/PCI/PnP Configuration</strong></td>
<td><strong>ASPM</strong></td>
<td><strong>Disable</strong></td>
</tr>
</tbody>
</table>
## Description

For Intel® 40 Gig NICs, special configurations should be set before compiling it. This is very important.

### Requirement

For at least DPDK release 1.8, 2.0 and 2.1, in `<dpdk_folder>/config/common_linuxapp`

- `CONFIG_RTE_PCI_CONFIG=y`
- `CONFIG_RTE_PCI_EXTENDED_TAG=on`

This helps increase the efficiency of PCIe by increasing the number of outstanding transactions from 36 to 256.

### Step 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.

#### 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*

### Step 2: Running l3fwd application & command to run for testing 4 x 10 G

- 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

Use 2 Cores
# System Configuration

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

Other names and brands may be claimed as the property of others.
### BIOS Tuning Settings

#### CPU Configuration

<table>
<thead>
<tr>
<th>Menu (Advanced)</th>
<th>BIOS Setting</th>
<th>Required Settings for Performance</th>
<th>BIOS Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced Power Management Configuration</td>
<td>Power Technology</td>
<td>Disable</td>
<td>Custom</td>
</tr>
<tr>
<td></td>
<td>Energy Performance Tuning</td>
<td>Disable</td>
<td>Enable</td>
</tr>
<tr>
<td></td>
<td>Energy Performance BIAS Setting</td>
<td>Performance</td>
<td>Enable</td>
</tr>
<tr>
<td></td>
<td>Energy Efficient Turbo</td>
<td>Disable</td>
<td>Enable</td>
</tr>
<tr>
<td>-&gt; CPU P State Control</td>
<td>EIST (P-States)</td>
<td>Enable</td>
<td></td>
</tr>
<tr>
<td>-&gt; CPU P State Control</td>
<td>Turbo Mode</td>
<td>Disable</td>
<td>Enable</td>
</tr>
<tr>
<td>-&gt; CPU P State Control</td>
<td>P-State Coordination</td>
<td>HW_ALL</td>
<td></td>
</tr>
<tr>
<td>-&gt; CPU C State Control</td>
<td>Package C State Limit</td>
<td>[C0/C1 State]</td>
<td>[C6 (Retention)]</td>
</tr>
<tr>
<td>-&gt; CPU C State Control</td>
<td>CPU C3 Report</td>
<td>Disable</td>
<td>Enable</td>
</tr>
<tr>
<td>-&gt; CPU C State Control</td>
<td>CPU C6 Report</td>
<td>Disable</td>
<td>Enable</td>
</tr>
<tr>
<td>-&gt; CPU C State Control</td>
<td>Enhanced Halt State (C1E)</td>
<td>Disable</td>
<td>Enable</td>
</tr>
</tbody>
</table>

#### Chipset Configuration

<table>
<thead>
<tr>
<th>Menu (Advanced)</th>
<th>BIOS Setting</th>
<th>Required Settings for Performance</th>
<th>BIOS Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>-&gt; North Bridge -&gt; I/O Configuration</td>
<td>EV DFX Features</td>
<td>Enable</td>
<td>Disable</td>
</tr>
<tr>
<td></td>
<td>Intel VT for Directed I/O (VT-d)</td>
<td>Disable</td>
<td>Enable</td>
</tr>
<tr>
<td>-&gt; North Bridge -&gt; IOAT Configuration</td>
<td>Enable IOAT</td>
<td>Enable</td>
<td>Enable</td>
</tr>
<tr>
<td></td>
<td>No Snoop</td>
<td>Disable</td>
<td>Disable</td>
</tr>
<tr>
<td></td>
<td>Relaxed Ordering</td>
<td>Disable</td>
<td>Disable</td>
</tr>
<tr>
<td>-&gt; North Bridge -&gt; QPI Configuration</td>
<td>Link L0 P</td>
<td>Disable</td>
<td>Enable</td>
</tr>
<tr>
<td></td>
<td>Link L1</td>
<td>Disable</td>
<td>Enable</td>
</tr>
<tr>
<td></td>
<td>COD Enable</td>
<td>Disable</td>
<td>Auto</td>
</tr>
<tr>
<td></td>
<td>Early Snoop</td>
<td>Disable</td>
<td>Auto</td>
</tr>
<tr>
<td></td>
<td>Isoc Mode</td>
<td>Disable</td>
<td>Disable</td>
</tr>
<tr>
<td></td>
<td>Enforce POR</td>
<td>Disable</td>
<td>Auto</td>
</tr>
<tr>
<td></td>
<td>Memory Frequency</td>
<td>2400</td>
<td>Auto</td>
</tr>
<tr>
<td>-&gt; North Bridge -&gt; Memory Configuration</td>
<td>DRAM RAPL Baseline</td>
<td>Disable</td>
<td>Auto</td>
</tr>
<tr>
<td></td>
<td>A/M Mode</td>
<td>Enable</td>
<td>Enable</td>
</tr>
<tr>
<td>-&gt; South Bridge</td>
<td>EHCI Hand-off</td>
<td>Disable</td>
<td>Auto</td>
</tr>
<tr>
<td></td>
<td>USB 2.0 Support</td>
<td>Disable</td>
<td>Enable</td>
</tr>
<tr>
<td></td>
<td>ASPM</td>
<td>Disable</td>
<td>Enable</td>
</tr>
<tr>
<td></td>
<td>Maximum Payload</td>
<td>AUTO</td>
<td>AUTO</td>
</tr>
<tr>
<td></td>
<td>Maximum Read Payload</td>
<td>AUTO</td>
<td>AUTO</td>
</tr>
<tr>
<td></td>
<td>Onboard LAN 1 OPTROM</td>
<td>Disable</td>
<td>PXE</td>
</tr>
</tbody>
</table>

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

14 Core Intel® Xeon® E5-2658 v4 Processor

DDR4-2400 ECC 1Rx8

PCI-E Gen3 x8 Slot 0

Lynx point

4x 10GbE Ports

X710-DA4 adapter

E10/100

Ixia* 10 Gigabit Ethernet Traffic Generator

* Other names and brands may be claimed as the property of others.
Test Setup for 40G Cards

Device Under Test (DUT):

- DDR4-2400 ECC 1Rx8
- 14 core Intel® Xeon® E5-2658 v4 Processor
- DDR4-2400 ECC 1Rx8
- PCI-E Gen3 x8 Slot 0
- PCI-E Gen3 x8 Slot 1
- XL710-DA2 adapter
- Lynx point
- 40GbE Ports
- XL710-DA2 adapter

Ixia* 10/40 Gigabit Ethernet Traffic Generator

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

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

4 x 10G 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

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

2 x 40G Ports

2 port configuration with 256 bi-directional flows per port

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

* Other names and brands may be claimed as the property of others.
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 – (2 Core / 2 Threads/1 Queue)**
  - 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

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

* Other names and brands may be claimed as the property of others.
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
  - CPU1 (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.

* Other names and brands may be claimed as the property of others.
• Cloud Networking – Understanding Cloud-Based Data Center Networks – Gary Lee.
• **DPDK Cook Book on Vtune – M Jay**
• 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

* Other names and brands may be claimed as the property of others.
Comparing XL710/X710 to Prior NIC 82599

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 watts¹ Typical Power
- 1 x 40GbE 3.3 watts² Typical Power

30%
UP TO 30% Reduction TYPICAL POWER

65%
UP TO 65% Reduction in GIGABIT PER WATT

2x
Increase in TOTAL BANDWIDTH

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