UNLOCKING GENERAL PURPOSES PROCESSORS TO BOOST PACKET PROCESSING PERFORMANCE

Liang Cunming
Platform Solution Architect
Data Center / Network Platforms Group
Legal Notices & Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

© 2017 Intel Corporation. Intel, the Intel logo, and Intel Xeon are trademarks of Intel Corporation in the U.S. and/or other countries. *Other names and brands may be claimed as property of others.
Tick-Tock Development Model:
Sustained Microprocessor Innovation Leadership

Innovation delivers new microarchitecture with Skylake
Skylake-SP Server CPU Overview

14nm Process Technology

Skylake: 6th gen Core processor
IPC increase vs. Broadwell

Intel® Hyper-Threading Technology (2 threads/core)

Intel® Turbo Boost Technology

Power Management:
Per Core P-State (PCPS)
Uncore Frequency Scaling (UFS)
Energy Efficient Turbo (EET)

Integrated Voltage Regulator

Power Management Enhancements (HWPC)

Integrated Fabric: Intel® Omni-Path Architecture

IO Enhancements

Memory Enhancements

Mesh Interconnect (SCF)

Non-Inclusive Cache Hierarchy:
SNC: Sub-NUMA Clustering Mode

Intel® AVX-512
32 DP FLOPs/Cycle/Core

New Feature
Enhanced Feature
## Skylake Core Micro-Architecture

<table>
<thead>
<tr>
<th></th>
<th>Sandy Bridge</th>
<th>Haswell</th>
<th>Skylake</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out of Order Window</td>
<td>168</td>
<td>192</td>
<td>224</td>
</tr>
<tr>
<td>In-flight Loads</td>
<td>64</td>
<td>72</td>
<td>72</td>
</tr>
<tr>
<td>In-flight Stores</td>
<td>36</td>
<td>42</td>
<td>56</td>
</tr>
<tr>
<td>Scheduler Entries</td>
<td>54</td>
<td>60</td>
<td>97</td>
</tr>
<tr>
<td>Integer Register File</td>
<td>160</td>
<td>168</td>
<td>180</td>
</tr>
<tr>
<td>FP Register File</td>
<td>144</td>
<td>168</td>
<td>168</td>
</tr>
<tr>
<td>Allocation Queue</td>
<td>28/thread</td>
<td>56</td>
<td>64/thread</td>
</tr>
</tbody>
</table>

Extracting more parallelism each generation, ~10% IPC improvement
Cycle Per Packet Improvements

System configuration is the same as the one used in DPDK layer 3 forwarding test covered in this presentation.
Skylake-SP Scalable Coherent Fabric Overview

Mesh Improves Scalability with Higher Bandwidth and Reduced Latencies
Loaded Memory Access Latency

**Memory** Load Line enables deterministic packet processing at peak levels

- Network Function Virtualization requires deterministic throughput as VMs are added
- Memory controller design and two additional memory channels yield a significant improvement in the loaded latency

(*) Source as of May 2017: Intel internal measurements of BW/latency on platform with Skylake-SP H0 28C internal sample, Core=turbo, CLM=turbo, UPI=10.4, SNC1, 6x32GB DDR4-2400/2667 per CPU, 1 DPC, and platform with E5-2699 v4, Turbo enabled, 4x32GB DDR4-2400, RHEL 7.0. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to [http://www.intel.com/performance](http://www.intel.com/performance) *Other names and brands may be claimed as the property of others.
PCIe Bandwidth

**PCI Express** platform performance increases up to 2x

- Mesh to I/O improvement, three MS2PCI mesh stops
- Additional Gen 3 x16 PCI E interface, three in total – resulting in up to 82GB/Bytes per socket
- Improvement in Data Directed I/O architecture, separation of RX and TX data

*Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to [http://www.intel.com/performance/datacenter](http://www.intel.com/performance/datacenter). Configurations: see next slide*
Translating Core, Memory and I/O Performance to Packet Processing

Data Plane Development Kit

Linux* Foundation Project

• More than 20 key open source projects build on DPDK libraries, including MoonGen*, mTCP*, Ostinato*, Lagopus*, Fast Data (FD.io), Open vSwitch*, OPNFV*, and OpenStack*

SKL-SP Optimizations

• Large MLC enables packet processing application foot print to remain close to the core

*Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to http://www.intel.com/performance/datacenter. Configurations: see next slide *

*Other names and brands may be claimed as property of others
Packet Processing Problem Statement

From a CPU perspective:
- Last-level-cache (L3) hit ~40 cycles
- L3 miss, memory read is ~70ns (140 cycles at 2GHz)
- Added security complexity
- Harder to address at 100Gb rates
Terabit Throughput Level with Unmodified SW

Breaking the Software Defined Network Services Barrier
1 Terabit Services on dual Intel® Xeon® Server !!! with DPDK, Fortville-25, Lewisburg

Intel® XEON® CPUs (Skylake-SP)
- Per socket have 48 lanes of PCIe Gen3
- 2x 280Gbps of packet I/O per socket

Intel® XEON® CPUs (E5 v3/v4)
- Per socket have 40 lanes of PCIe Gen3
- 2x 160Gbps of packet I/O per socket
Unlocking Platform Capability by DPDK

**DPDK Fundamentals**

- Implements run-to-completion and pipeline models
- No scheduler - all devices accessed by polling
- Supports 32-bit and 64-bit OSs, with and without NUMA
- Scales from Intel® Atom® to Intel® Xeon® processors
- Number of cores and processors is not limited
- Optimal packet allocation across DRAM channels
- Use of 2M & 1G hugepages and cache aligned structures
- Uses bulk concepts - processing ‘n’ packets simultaneously
- Open source and BSD licensed

**Network Functions (Cloud, Enterprise, Telco)**

<table>
<thead>
<tr>
<th>Core libraries</th>
<th>Packet classification</th>
<th>Accelerated SW libraries</th>
<th>Stats</th>
<th>QoS</th>
<th>Packet Framework</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core functions such as memory management, software rings, timers etc.</td>
<td>Software libraries for hash/exact match, LPM, ACL etc.</td>
<td>Common functions such as IP fragmentation reassembly, reordering etc.</td>
<td>Libraries for collecting and reporting statistics.</td>
<td>Libraries for QoS scheduling and metering/policing</td>
<td>Libraries for creating complex pipelines in software.</td>
</tr>
</tbody>
</table>

**PMDs for physical and virtual Ethernet devices**

- ETHDEV
- CRYPTODEV
- EVENTDEV
- SECURITY
- COMPRESS
- RAW

**Generic devices w/o specific type**
Bridging Various Accelerators
seamless interface to accelerators

**DPDK Framework**

- Generic APIs
- Application is abstracted from the underlying SW and HW with DPDK
- Preserve Platform and Application software investment
- Optimized platform software ingredients (e.g. vSwitch) to take advantage of HW and SW ingredients
- Flexible and outstanding performing data plane
Community Ecosystem

A fully open source software project with a strong development community
Boosts Open Source Projects

vSwitches/vRouters
- VPP
- BESS
- Lagopus
- Open vSwitch
- tf.tungstenfabric
- CloudRouter

DPDK in OS Distros
- redhat
  Version 7.1 +
- ubuntu
  Version 10.1 +
- WIND
  Version 6 +
- CentOS
  Version 15.10 +
- fedora
  Version 22 +

Storage
- Storage Performance Development Kit
  + Many more

Packet Generators
- OTRex
- Pktgen
- Ostinato
- MoonGen

TCP/IP Stacks
- mTCP
- TLDK & VPP
- Seastar
- LWIP DPDK
Enriches Research & Innovation

mTCP [NSDI '14]  mOS [NSDI '17]
IX [OSDI '14]      FTMB [SIGCOMM '15]
MoonGen [IMC '15]  MICA [NSDI '14]
NetBricks [OSDI '16] BlindBox [SIGCOMM '15]
SoftFlow [ATC '16]  Flowtune [NSDI '17]
APUNet [NSDI '17]   NFP [SIGCOMM '17]
StatelessNF [NSDI '17]  VigNAT [SIGCOMM '17]
STYX [SOCC '17]  NFP [SIGCOMM '17]
Future: Toward Cloud-Native Network Functions

- **Primary Constructs**
  - DevOps/Continuous delivery/Micro services/Containers

- **Unique Considerations of Network Functions**
  - Data plane packet processing requires an optimized architecture
  - Domain specific protocol is absent
  - Intergenerational transforming & compatibility
Summary

- Powerful Multi-Core Scalable Architecture Processor
- Unlock Packet Processing Capability by DPDK
- Seamless Interface to Various Accelerators
- Fantastic Ecosystem for Innovation