TRANSFORMING
COMMUNICATIONS

COMMUNICATIONS INFRASTRUCTURE DIVISION

Intel® Data Plane Development Kit (Intel® DPDK) Overview
Packet Processing on Intel® Architecture

December 2012

Legal Disclaimer

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

All information provided related to future Intel products and plans is preliminary and subject to change at any time, without notice. All dates provided are subject to change without notice. Intel may make changes to specifications and product descriptions at any time, without notice. Intel product plans in this presentation do not constitute Intel plan of record product roadmaps. Please contact your Intel representative to obtain Intel's current plan of record product roadmaps. Intel product plans in this presentation do not constitute Intel plan of record product roadmaps.

Celeron, Intel, Intel logo, Intel Core, Intel Inside, Intel Inside logo, Intel. Leap ahead., Intel. Leap ahead. logo, Intel NetBurst, Intel SpeedStep, Intel XScale, Itanium, Pentium, Pentium Inside, VTune, Xeon, and Xeon Inside are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Intel® Active Management Technology requires the platform to have an Intel® AMT-enabled chipset, network hardware and software, as well as connection with a power source and a corporate network connection. With regard to notebooks, Intel AMT may not be available or certain capabilities may be limited over a host OS-based VPN or when connecting wirelessly, on battery power, sleeping, hibernating or powered off. For more information, see http://www.intel.com/technology/iamt.

Enhanced Intel SpeedStep® Technology for specified units of this processor available Q2/06. See the Processor Spec Finder at http://processorfinder.intel.com or contact your Intel representative for more information

64-bit computing on Intel architecture requires a computer system with a processor, chipset, BIOS, operating system, device drivers and applications enabled for Intel® 64 architecture. Performance will vary depending on your hardware and software configurations. Consult with your system vendor for more information.

No computer system can provide absolute security under all conditions. Intel® Trusted Execution Technology is a security technology under development by Intel and requires for operation a computer system with Intel® Virtualization Technology, an Intel Trusted Execution Technology-enabled processor, chipset, BIOS, Authenticated Code Modules, and an Intel or other compatible measured virtual machine monitor. In addition, Intel Trusted Execution Technology requires the system to contain a TPMv1.2 as defined by the Trusted Computing Group and specific software for some uses. See http://www.intel.com/technology/security/ for more information.

†Hyper-Threading Technology (HT Technology) requires a computer system with an Intel® Pentium® 4 Processor supporting HT Technology and an HT Technology-enabled chipset, BIOS, and operating system. Performance will vary depending on the specific hardware and software you use. See www.intel.com/products/ht/hyperthreading_more.htm for more information including details on which processors support HT Technology. Intel® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, virtual machine monitor (VMM) and, for some uses, certain platform software enabled for it. Functionality, performance or other benefits will vary depending on hardware and software configurations and may require a BIOS update. Software applications may not be compatible with all operating systems. Please check with your application vendor. Intel® AES-NI requires a computer system with an AES-NI enabled processor, as well as non-Intel software to execute the instructions in the correct sequence. AES-NI is available on select Intel® processors. For availability, consult your reseller or system manufacturer. For more information, see Intel® Advanced Encryption Standard Instructions (AES-NI)

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

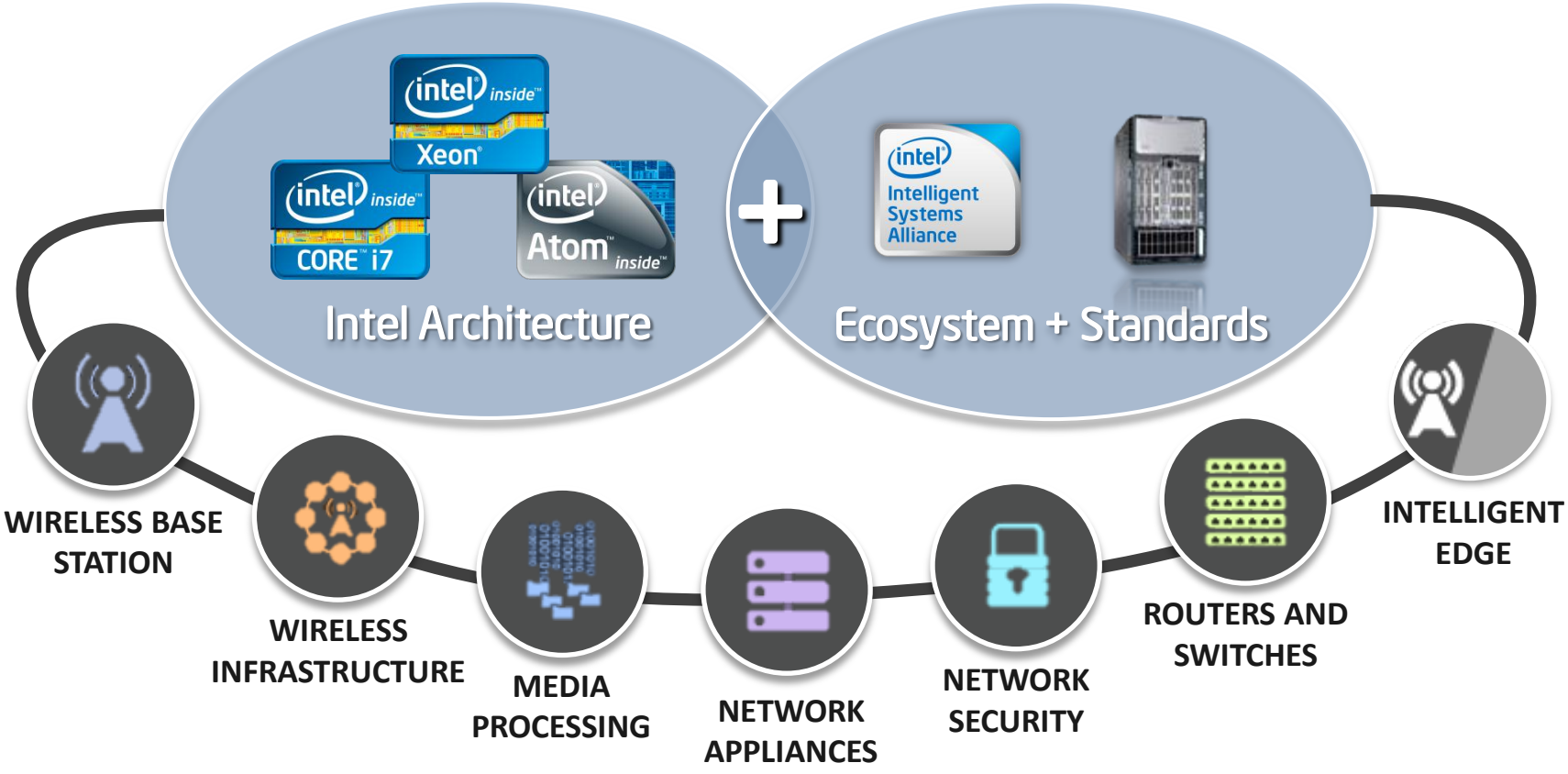
Other vendors are listed by Intel as a convenience to Intel's general customer base, but Intel does not make any representations or warranties whatsoever regarding quality, reliability, functionality, or compatibility of these devices. This list and/or these devices may be subject to change without notice.

Agenda

1. Intel's Packet Processing Motivation and Value Proposition
2. Overview of Intel® DPDK
3. Intel® DPDK Performance Benchmarks
4. Lead Ecosystem Offerings
5. Intel® DPDK Website and Collateral
6. Summary

CID Mission "TRANSFORMING COMMUNICATIONS"

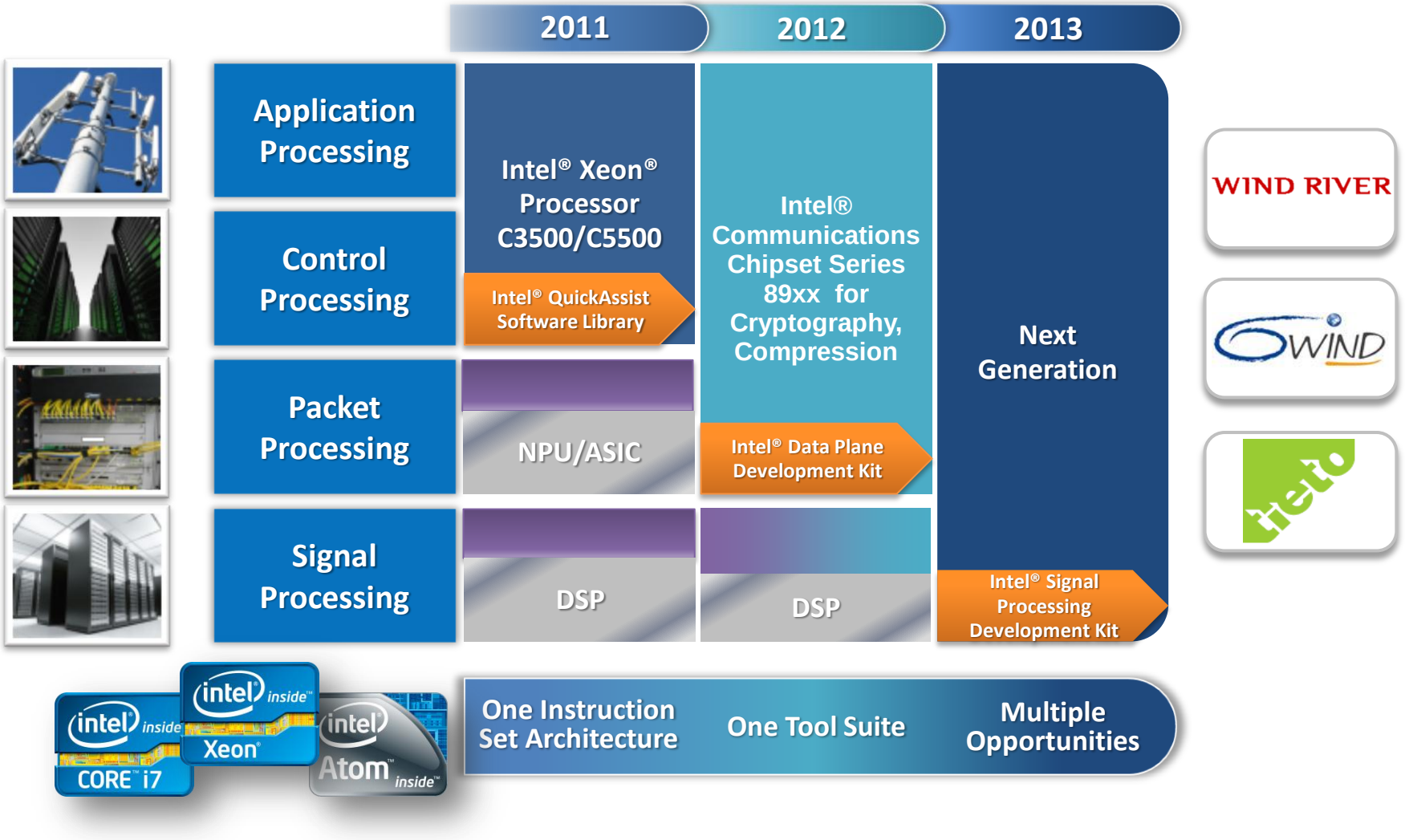
ACCESS NETWORKS | **EDGE/CORE NETWORKS** | **ENTERPRISE NETWORKS**



>10 Years Serving Communications and Networking Segments

4:1 Workload Consolidation Assets

Unleashed by Multi-Core IA and Software



Intel's Data Plane Value Proposition

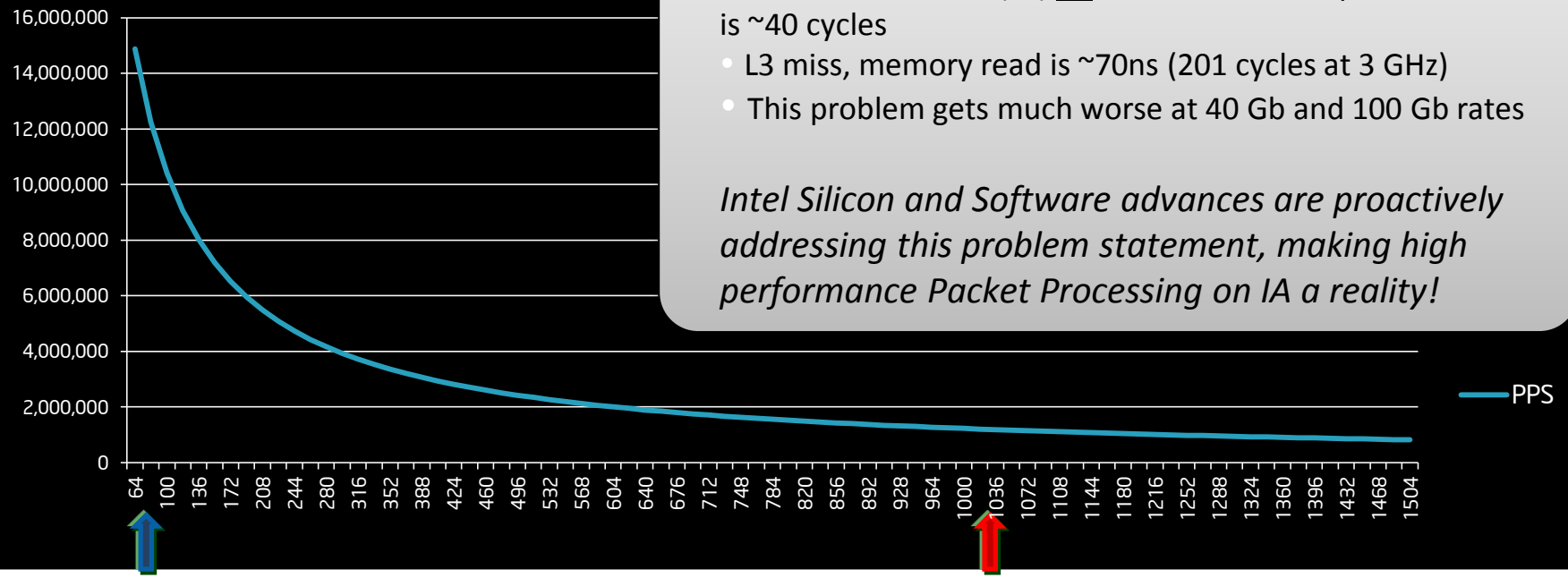
Datacom/Telecom convergence increasing Data Plane processing requirements exponentially

TCO concerns leading customers to seek single architecture design top-to-bottom

Intel addresses TCO and TTM concerns with single architecture, multi-workload IA capability, allied to an industry-leading beat-rate of process and uArchitectural advancements (Tick-Tock Model)

Optimized Data Plane Software solutions will help unleash IA platform potential

The Engineering Problem Statement …



From a CPU perspective:

- A Last-level-cache (L3) hit on Intel® Xeon® processor 5500 is ~40 cycles
- L3 miss, memory read is ~70ns (201 cycles at 3 GHz)
- This problem gets much worse at 40 Gb and 100 Gb rates

Intel Silicon and Software advances are proactively addressing this problem statement, making high performance Packet Processing on IA a reality!

Network Infrastructure Packet Sizes

Packet Size	64 bytes
10G Packets/second	14.88 Million each way
Packet arrival rate	67.2 ns
2 GHz Clock cycles	135 cycles
3 Ghz Clock cycles	201 cycles

Typical Server Packet Sizes

Packet Size	1024 bytes
10G Packets/second	1.2 Million each way
Packet arrival rate	835 ns
2 GHz Clock cycles	1670 cycles
3 Ghz Clock cycles	2505 cycles

Ongoing Silicon Architectural Enhancements

Packet Processing Enhancements:
- Pipeline Depth
- Direct Cache Access
- Integration of Memory Controller
- Integration of High Bandwidth PCIe Gen3
- New AVX Extensions
- Intel® Virtualization Technology (Intel® VT)
- Intel® Data Direct I/O Technology (Intel® DDIO)

Intel® Pentium® 4 Processor Extreme Edition
Supporting Hyper-Threading Technology

Netburst microarchitecture
Intel® Extended Memory 64 Technology, Hyperthreading

Intel® Core™ 2 Microarchitecture

Multi-Core Introduction
Advanced Smart Cache
Wide Dynamic Execution
SSE2/SSE3, Power Management

Intel® Core™ i5 / i7 Microarchitecture

Hyper-Threading, Smart Cache, QuickPath, Integrated Memory Controller, SSE2/SSE3/SSE4 Instructions

Intel® Xeon® processor E5-2600 Microarchitecture

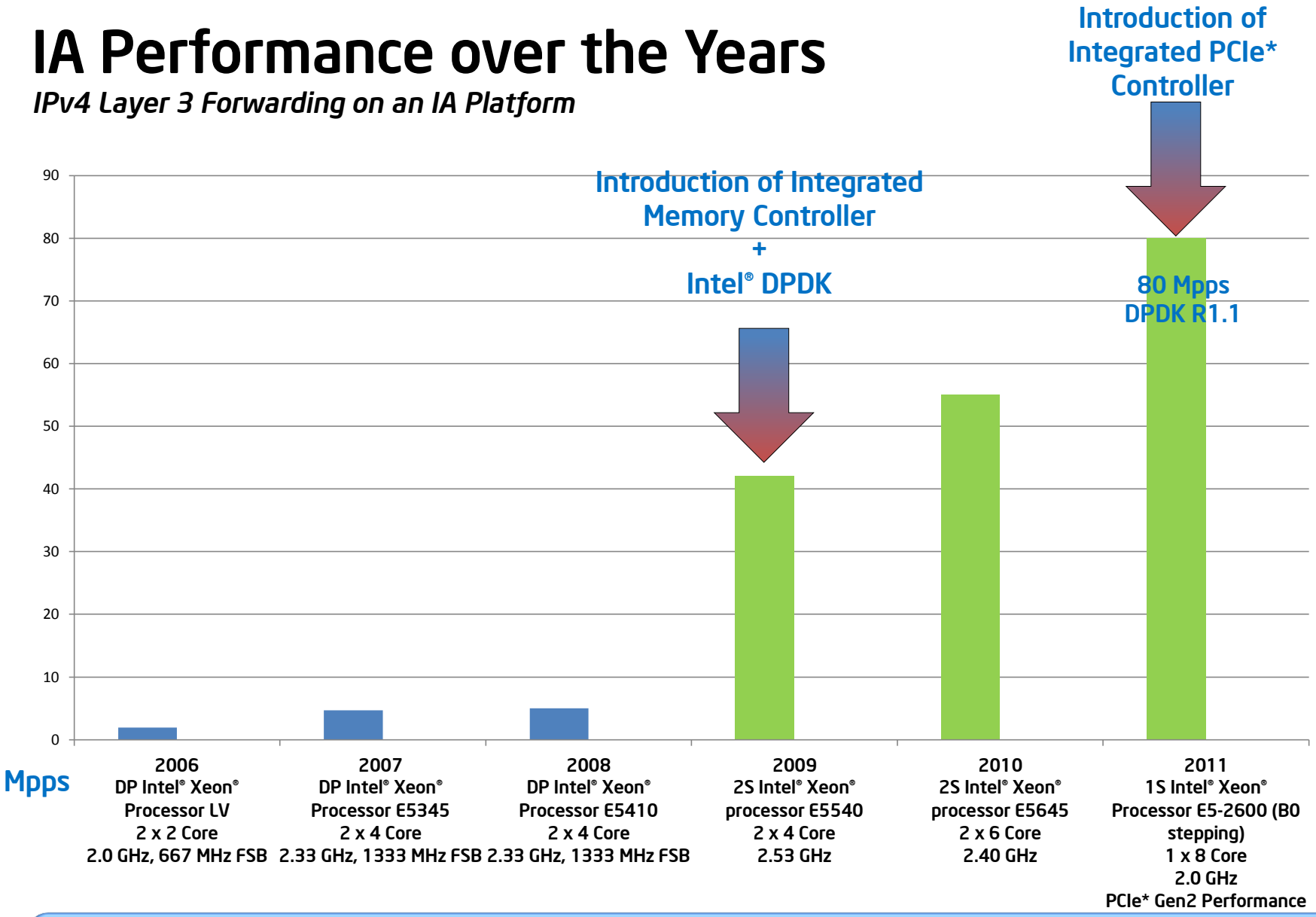
Enhanced Intel® Core Microarchitecture
8C, 6C, 4C, 2C Product Choices
1S, 2S, 4S Configurations
Intel® Hyper-Threading Technology
Integrated Memory Controller(s)
Integrated High BW PCIe gen3
2 QPI Links for 2S Configurations
Up to 20MB of L3 Cache

Beat Rate of Enhancements

Improved Packet Processing Capability

IA Performance over the Years

IPv4 Layer 3 Forwarding on an IA Platform

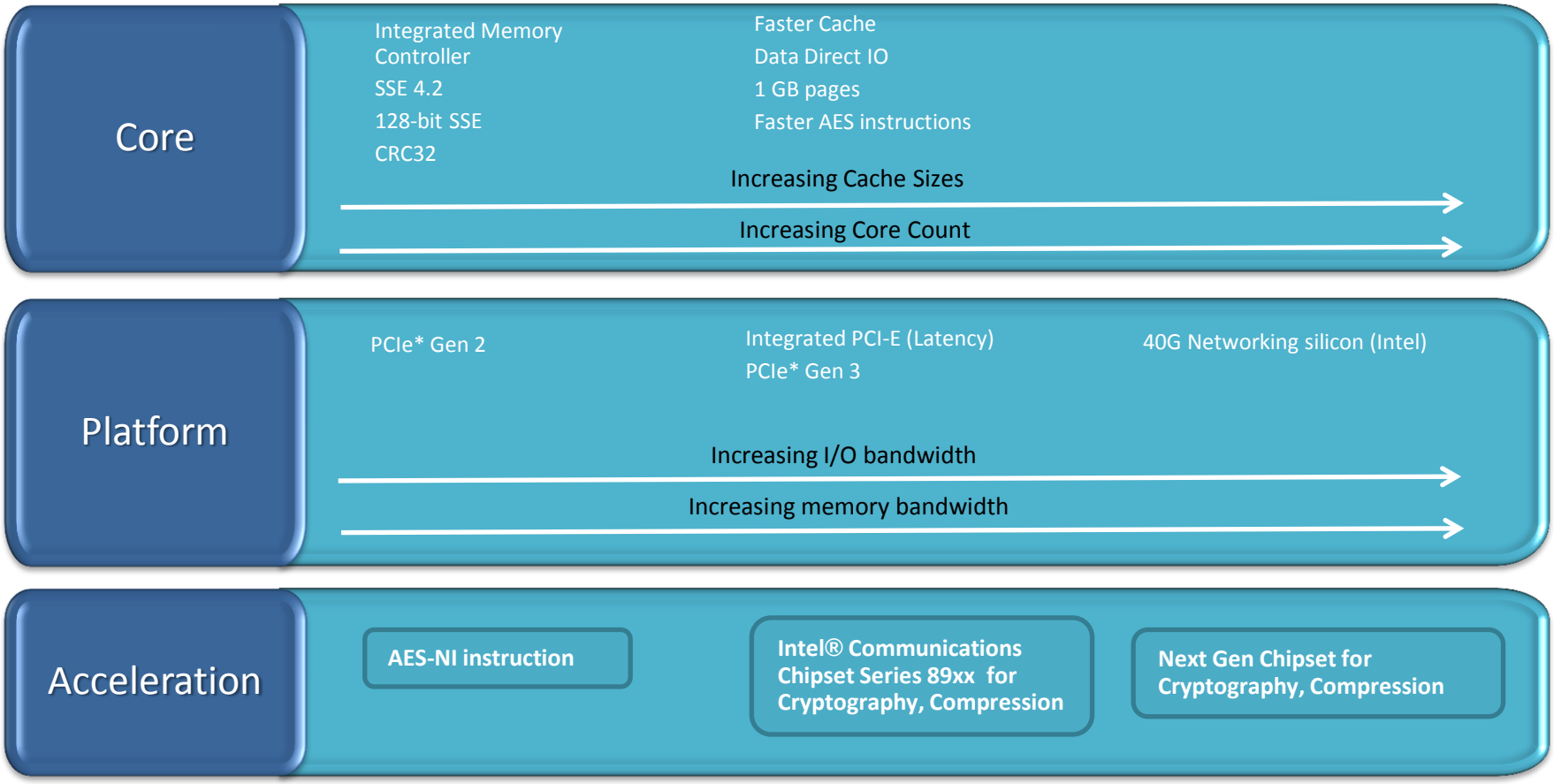


Standard "off-the-shelf" IA platform can deliver huge performance.

Performance jump can be attributed to Core, Memory architecture (iMC) + Intel® DPDK

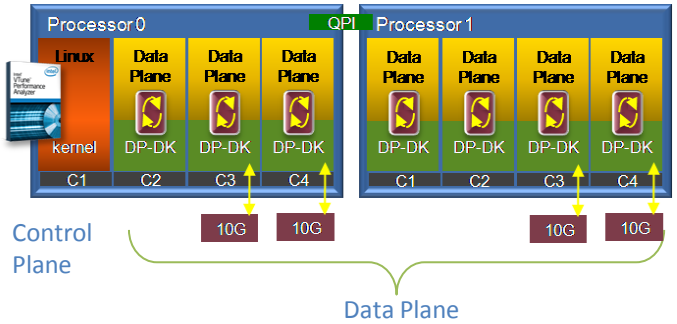
Evolution of Data Plane Support

Software & Hardware performance enhancements over the next 2-3 years



Management, Control & Data Plane Environments

Typically more Data Plane elements than Management/Control Plane

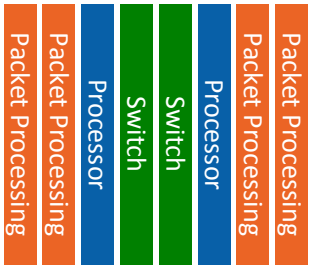


Rack Mount Server, Enterprise Servers

Typically have Control and Data Plane on same board

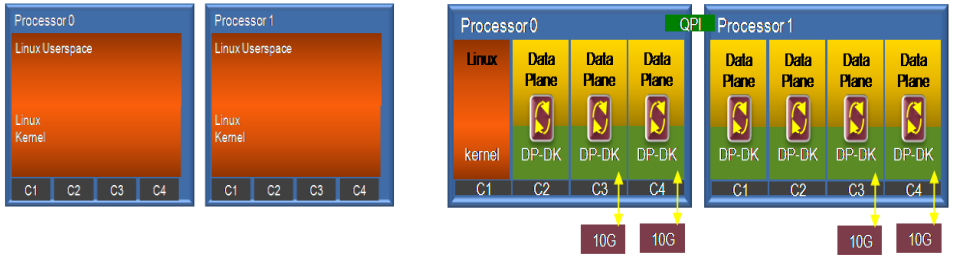
Value Proposition is Consolidation of DP + CP

- Customers have come up with solutions to add additional 3rd party boards for Data Plane
- IA can fill both roles



Ecosystem or Proprietary: Data Plane

Intel Provided: Intel® DPDK



Processor Blade
Control Plane

Packet Processing Blade
Data Plane

AdvancedTCA*

Typically Control and Data Plane on different boards

Value Proposition single blade for multiple purposes

- Normally many more Packet Processing Blades
- IA allows a single Blade Architecture for both now

Value proposition of Intel® DPDK is workload consolidation:
Provides framework and performance for NPU workloads on IA cores

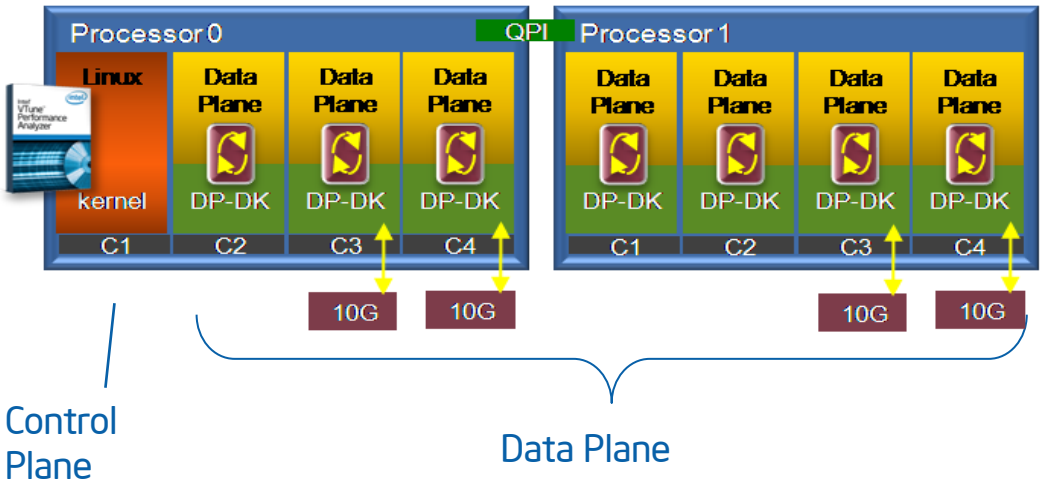
Agenda

1. Intel's Packet Processing Motivation and Value Proposition
2. Overview of Intel® DPDK
3. Intel® DPDK Performance Benchmarks
4. Lead Ecosystem Offerings
5. Intel® DPDK Website and Collateral
6. Summary

Goals of this Overview Section

- Look into the Intel® DPDK architecture and see how it's designed to squeeze the best packet processing performance out of an IA-based platform
- Be able to articulate the most common performance bottlenecks for packet processing software on IA
- Understand the optimization tricks to remove them

The Intel® DPDK Philosophy



Intel® DPDK Fundamentals

- Implements a run to completion model or pipeline model
- No scheduler - all devices accessed by polling
- Supports 32-bit and 64-bit with/without NUMA
- Scales from Intel® Atom™ to Intel® Xeon® processors
- Number of Cores and Processors not limited
- Optimal packet allocation across DRAM channels

- **Must run on any IA CPU**
  - **From Intel® Atom™ processor to the latest Intel® Xeon® processor family**
  - Essential to the IA value proposition
- **Focus on the fast-path**
  - Sending large number of packets to the Linux Kernel /GPOS will bog the system down

Provide software examples that address common network performance deficits

- Best practices for software architecture
- Tips for data structure design and storage
- Help the compiler generate optimum code
- Address the challenges of achieving 80 Mpps per CPU Socket

Intel® Data Plane Development Kit (Intel® DPDK)

Intel® DPDK embeds optimizations for the IA platform:

- **Data Plane Libraries and Optimized NIC Drivers in Linux User Space**

  Queue & Buffer Management, Packet Flow Classification, Poll-Mode NIC Drivers (1/10GbE), and more!
  Simple API Interface, Uses standard tool chain (gcc/icc, gdb, profiling tools)

- **Run-time Environment**

  Low overhead, run-to-completion model optimized for fastest possible data plane performance

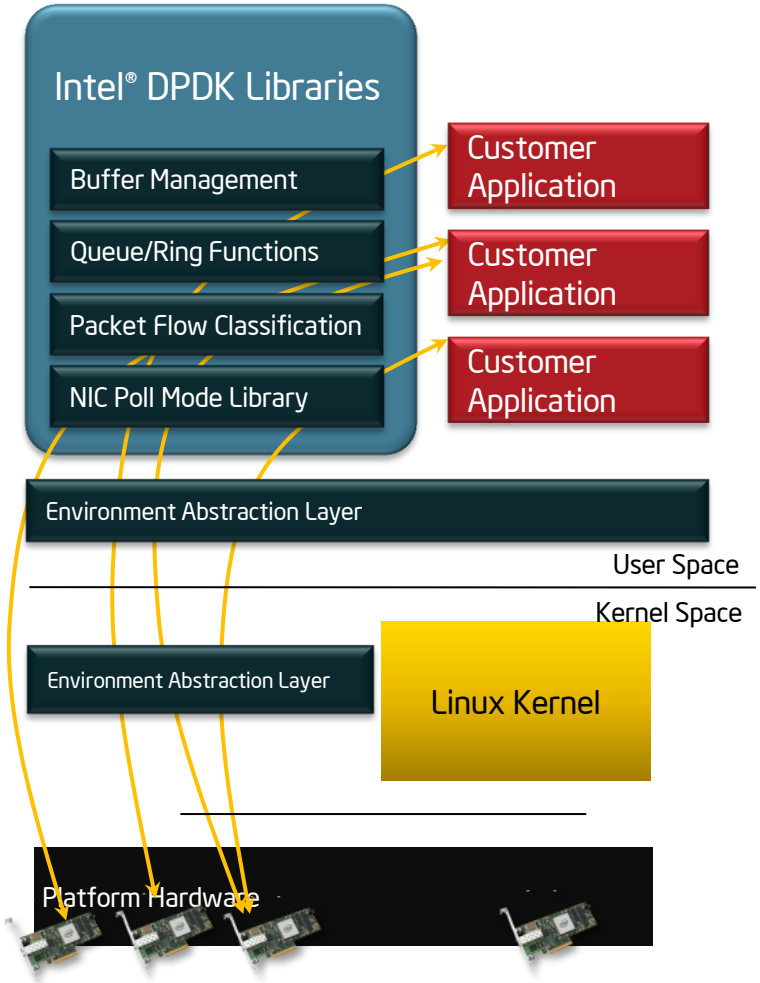
- **Environment Abstraction Layer and Boot Code**

  Primarily platform-specific boot guidelines and initialization code, eases application porting effort

- **BSD-licensed & source downloadable from Intel and leading ecopartners**

  Provided under a very flexible BSD licensing model

  Offered as a free, unsupported standalone solution by Intel or as part of commercial solutions and offerings from leading ecopartners



The Intel® DPDK is a great starting point for customers and the industry in general – delivering breakthrough packet processing performance.

Intel® DPDK Libraries and Drivers

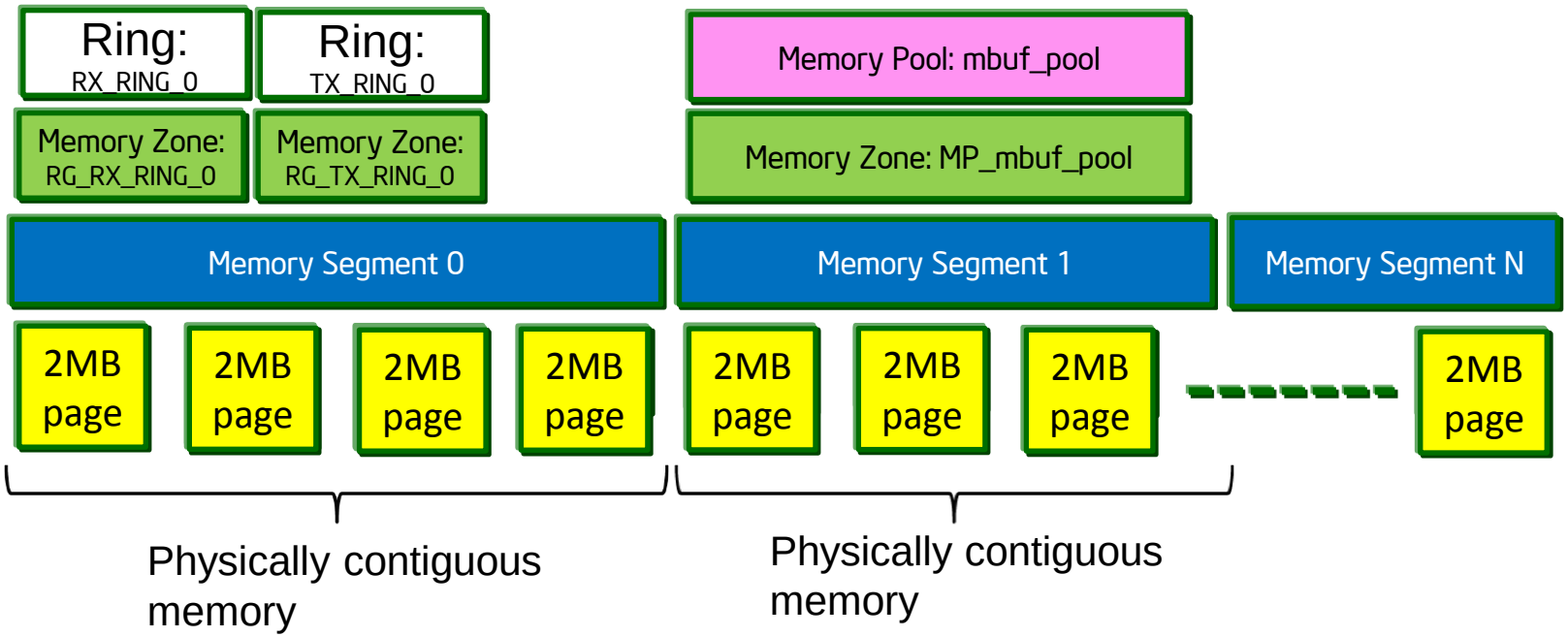
- **Memory Manager:** Responsible for allocating pools of objects in memory. A pool is created in huge page memory space and uses a ring to store free objects. It also provides an alignment helper to ensure that objects are padded to spread them equally on all DRAM channels.
- **Buffer Manager:** Reduces by a significant amount the time the operating system spends allocating and de-allocating buffers. The Intel® DPDK pre-allocates fixed size buffers which are stored in memory pools.
- **Queue Manager::** Implements safe lockless queues, instead of using spinlocks, that allow different software components to process packets, while avoiding unnecessary wait times.
- **Flow Classification:** Provides an efficient mechanism which incorporates Intel® Streaming SIMD Extensions (Intel® SSE) to produce a hash based on tuple information so that packets may be placed into flows quickly for processing, thus greatly improving throughput.
- **Poll Mode Drivers:** The Intel® DPDK includes Poll Mode Drivers for 1 GbE and 10 GbE Ethernet* controllers which are designed to work without asynchronous, interrupt-based signaling mechanisms, which greatly speeds up the packet pipeline.

Component Overviews

- EAL Memory Management Overview
- Queue/Ring Overview
- Buffer Management Overview
- Flow Classification Overview

Memory Usage

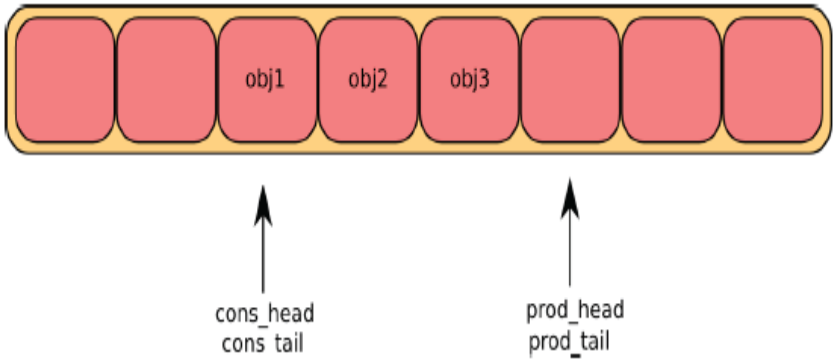
- Basic unit for runtime object allocation is the memory zone
- Zones contain rings, pools, LPM routing tables, or any other performance-critical structures
- Always backed by Huge Page (2 MB/1 GB page) memory



Queue/Ring Management API

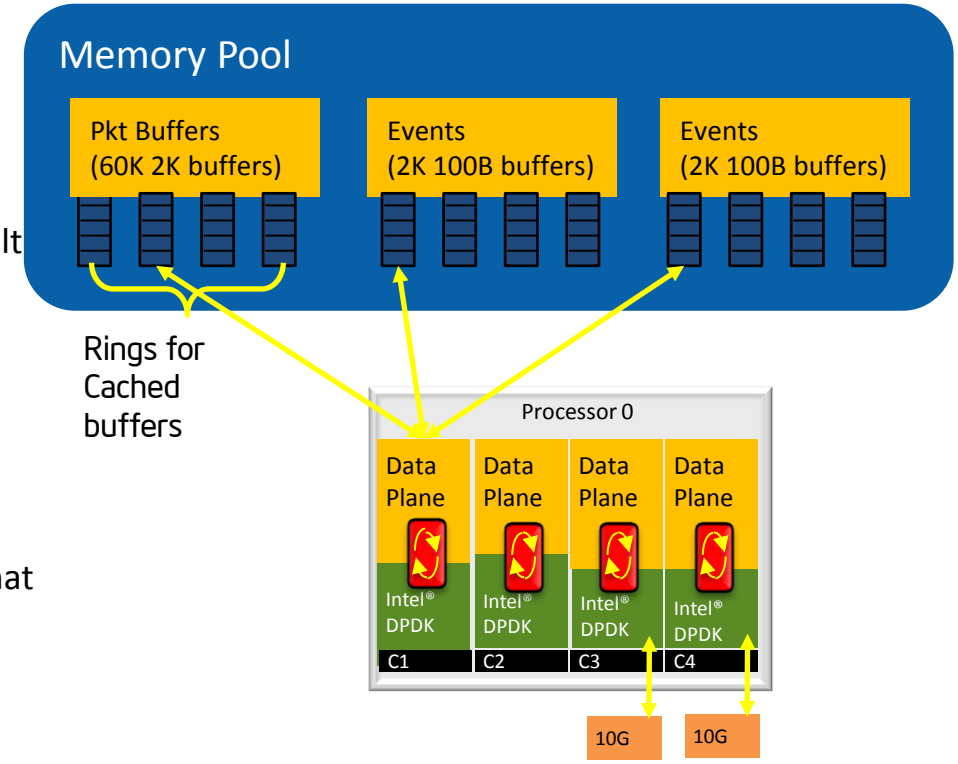
- **Effectively a FIFO implementation in software**
  - Lockless implementations for single or multi-producer, single or multi- consumer enqueue/dequeue
  - Supports bulk enqueue/dequeue to support packet-bunching
  - Implements high & low watermark thresholds for back-pressure/flow control

- **Essential to optimizing throughput**
  - Used to decouple stages of a pipeline



The Buffer Management API (mempool)

- **The Buffer Manager allocates memory from the EAL and creates pools with fixed element sizes.**
  - Typical usage is packet buffers, descriptor ring buffers, etc.
  - Intent is to speed up runtime allocation/de-allocation
  - Does not support runtime resizing of pools
- **Multi-producer/multi-consumer safe**
  - Pools are based on Intel® DPDK rings so are multi-producer and multi-consumer safe
  - No locking; use CAS instructions
  - Pools can also be used in multi-process environments
- **Optimized for performance**
  - Cache alignment
  - Per core buffer caches for each buffer pool so that allocation/freeing can be done without using shared variables
  - Bulk allocation/freeing support

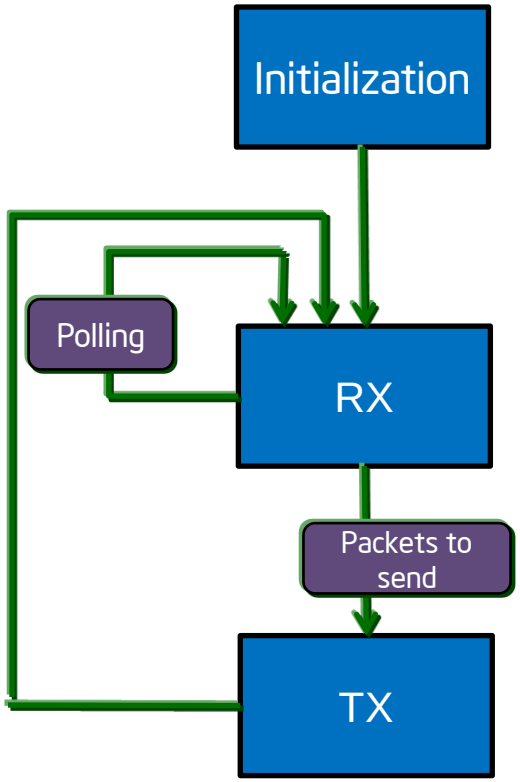


Flow Classification API

- **The Intel® DPDK provides a Flow Classification API**
  - Not expecting every customer to use it – more of a showcase to demonstrate how to do optimization for IA
  - Classification is something that is very customer-specific – Each customer/segment has different needs
    - Router implementations typically use “longest-prefix-match”
    - Security implementations need to identify individual flows and can use flow classification

- **The Flow classification API is designed to take advantage of current and future hardware-based flow classification capabilities**
  - Intel® 82599 10GbE Ethernet Controller implements flow classification (limited in number of flows)
  - Future Chipsets are expected to implement an extensive classifier

30,000 ft Overview of Packet Flow



1. **Initialization**
   - Initialize memory zones and pools
   - Initialize devices and device queues
   - Start the packet forwarding application
2. **Packet Reception (RX)**
   - Poll devices' RX queues and receive packets in bursts
   - Allocate new RX buffers from per queue memory pools to stuff into descriptors
3. **Packet Transmission (TX)**
   - Transmit the received packets from RX
   - Free the buffers used to store the packets

Overcoming The Challenge of Achieving 80 Mpps (and More…) Per CPU Socket

The system can't keep up with the amount of interrupts for packet Rx

Switch from an interrupt-driven network device driver to a polled-mode driver.

The out-of-the-box Linux* Scheduler causes too much overhead to task switch

Bind a single software thread to a logical core. Use CPU core isolation and thread affinities for 1:1 mapping of SW threads to HW threads.

Memory and PCIe* access is really, really slow compared to CPU operation

Process a bunch of packets (e.g. 4 packets at a time) to minimize external memory and PCIe bandwidth. Avoid read-modify-write transactions in favour of single write, and multiple reads in favour of single read.

Data doesn't seem to be near the CPU when it needs it (and so it waits)

For memory access, use HW or SW controlled prefetching and align data structures to cache line size (64 Byte) to minimize external memory and PCIe* bandwidth, as all external memory accesses are in cache line increments; for PCIe access, use Direct Data IO (available on Intel® Xeon® processor E5 Product Family) to read data directly into cache.

Access to shared data structures is a bottleneck in the application

Figure out clever access schemes that reduce the amount of sharing (e.g. use lockless queues for message passing as semaphores/spinlocks are costly).

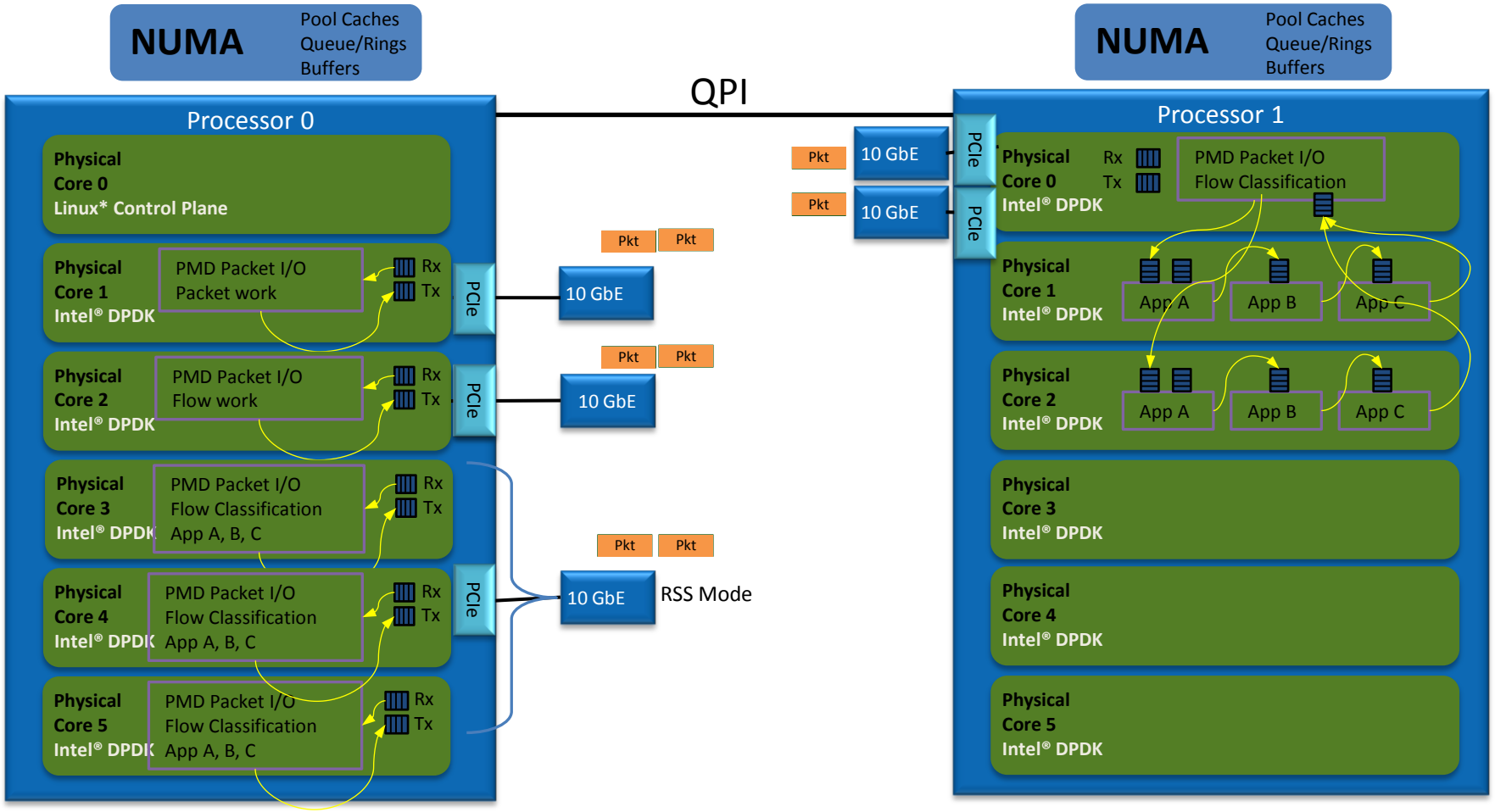
Intel® PTU Tool indicates that page tables are constantly evicted (D-TLB Thrashing)

Use 2MB or 1G Huge Pages in Linux* to reduce TLB misses.

The Challenge can be Overcome with Smart Programming and Hardware assists!!

PCIe* connectivity and core usage

Using run-to-completion or pipeline software models



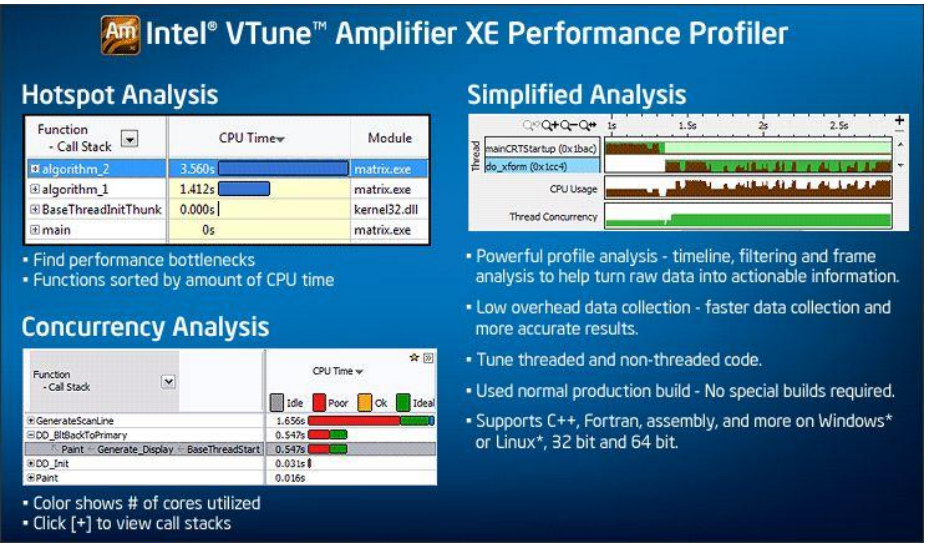
Run to Completion model

- I/O and Application workload can be handled on a single core
- I/O can be scaled over multiple cores

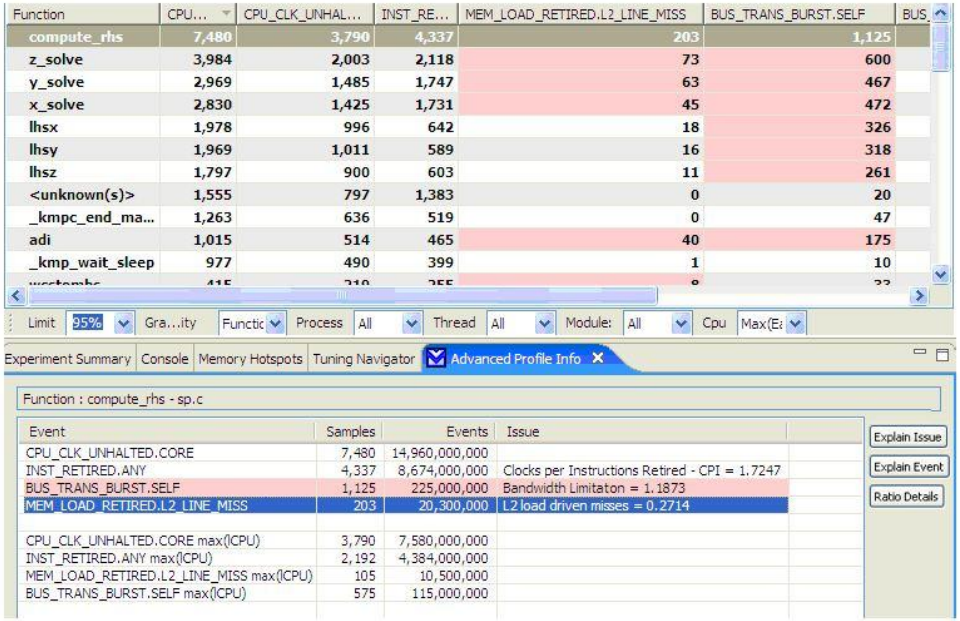
Pipeline model

- I/O application disperses packets to other cores
- Application work performed on other cores

Tools for Optimizing Intel® DPDK



Intel® VTune™ Amplifier XE profiles performance



Intel® Performance Tuning Utility (Intel® PTU) offers specific tuning advice

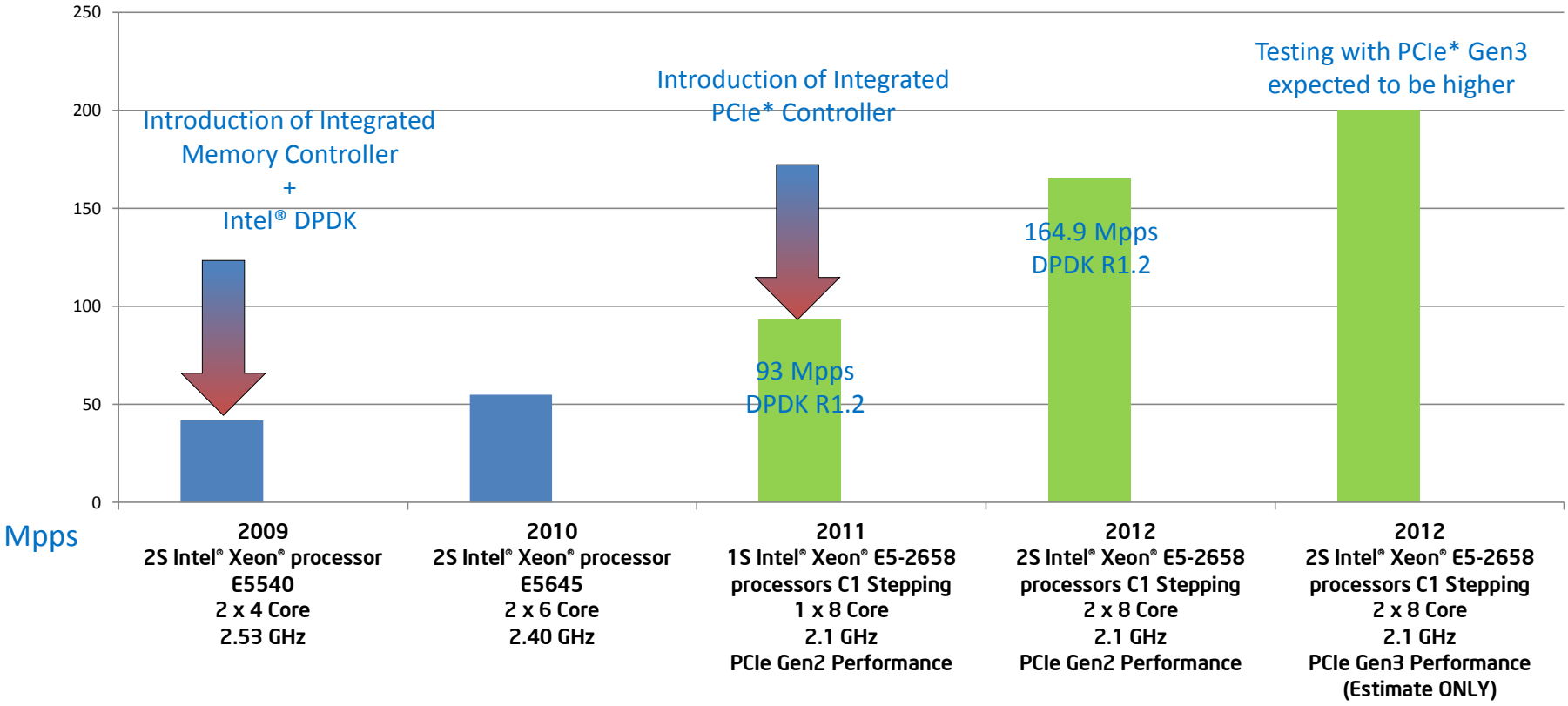
Download both tools at whatif.intel.com

Agenda

1. Intel's Packet Processing Motivation and Value Proposition
2. Overview of Intel® DPDK
3. Intel® DPDK Performance Benchmarks
4. Lead Ecosystem Offerings
5. Intel® DPDK Website and Collateral
6. Summary

Performance Going Forward

IPv4 Layer 3 Forwarding on an IA Platform



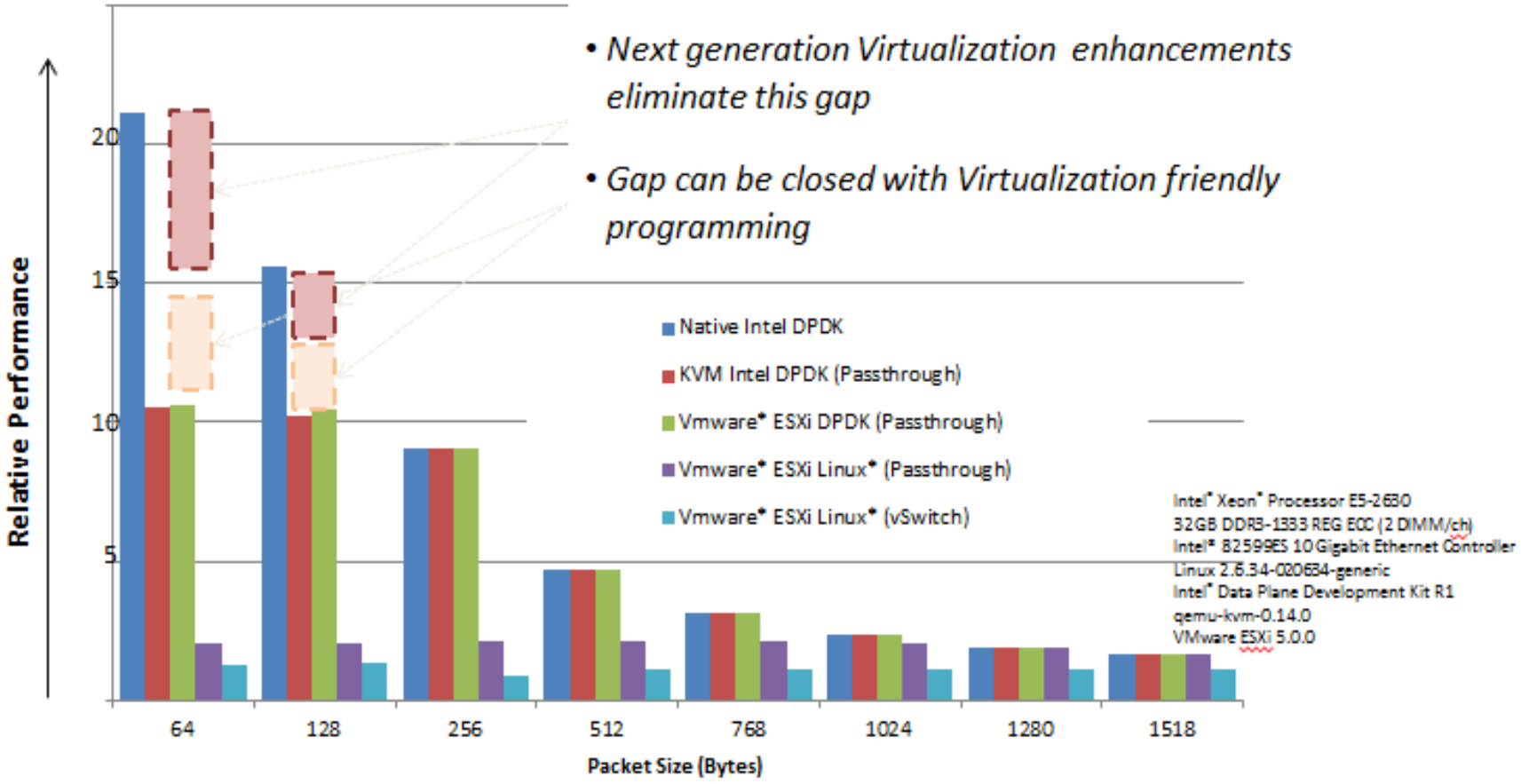
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may 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.

PCIe* Gen 3 will offer better performance and more options

In the case of IPv4 Layer 3 forwarding, we are still I/O limited – i.e. cores capability is not maxed out!!

Intel® DPDK Native and Virtualized Forwarding Performance

- *Performance Degradation at 64 byte and 128 byte packets due to IOTLB eviction*
- *Next generation Virtualization enhancements eliminate this gap*
- *Gap can be closed with Virtualization friendly programming*



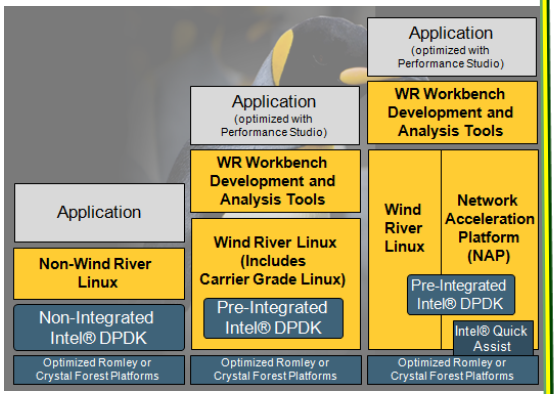
Source: Intel® Corporation, Communications Storage Infrastructure Group

Agenda

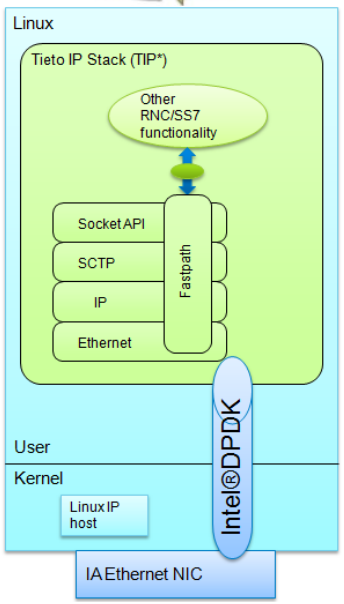
1. Intel's Packet Processing Motivation and Value Proposition
2. Overview of Intel® DPDK
3. Intel® DPDK Performance Benchmarks
4. Lead Ecosystem Offerings
5. Intel® DPDK Website and Collateral
6. Summary

Intel® DPDK Go To Market Options

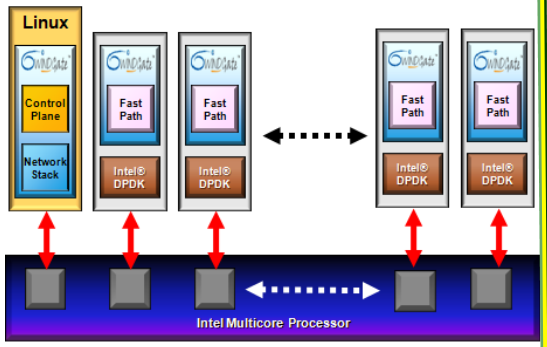
WIND RIVER



tieto

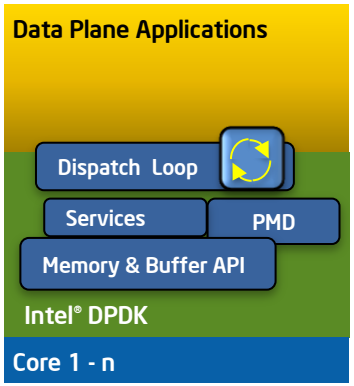


6WIND



intel

Intel® DPDK free, unsupported standalone package for integration with proprietary customer stacks



Ecosystem Provided: Intel® DPDK Integrated into Commercial Solutions and Intel® DPDK services offerings
For more information about ecosystem solutions, visit www.intel.com/go/dpdk

Intel Provided: Intel® DPDK – Free, Unsupported Package

Agenda

1. Intel's Packet Processing Motivation and Value Proposition
2. Overview of Intel® DPDK
3. Intel® DPDK Performance Benchmarks
4. Lead Ecosystem Offerings
5. Intel® DPDK Website and Collateral
6. Summary

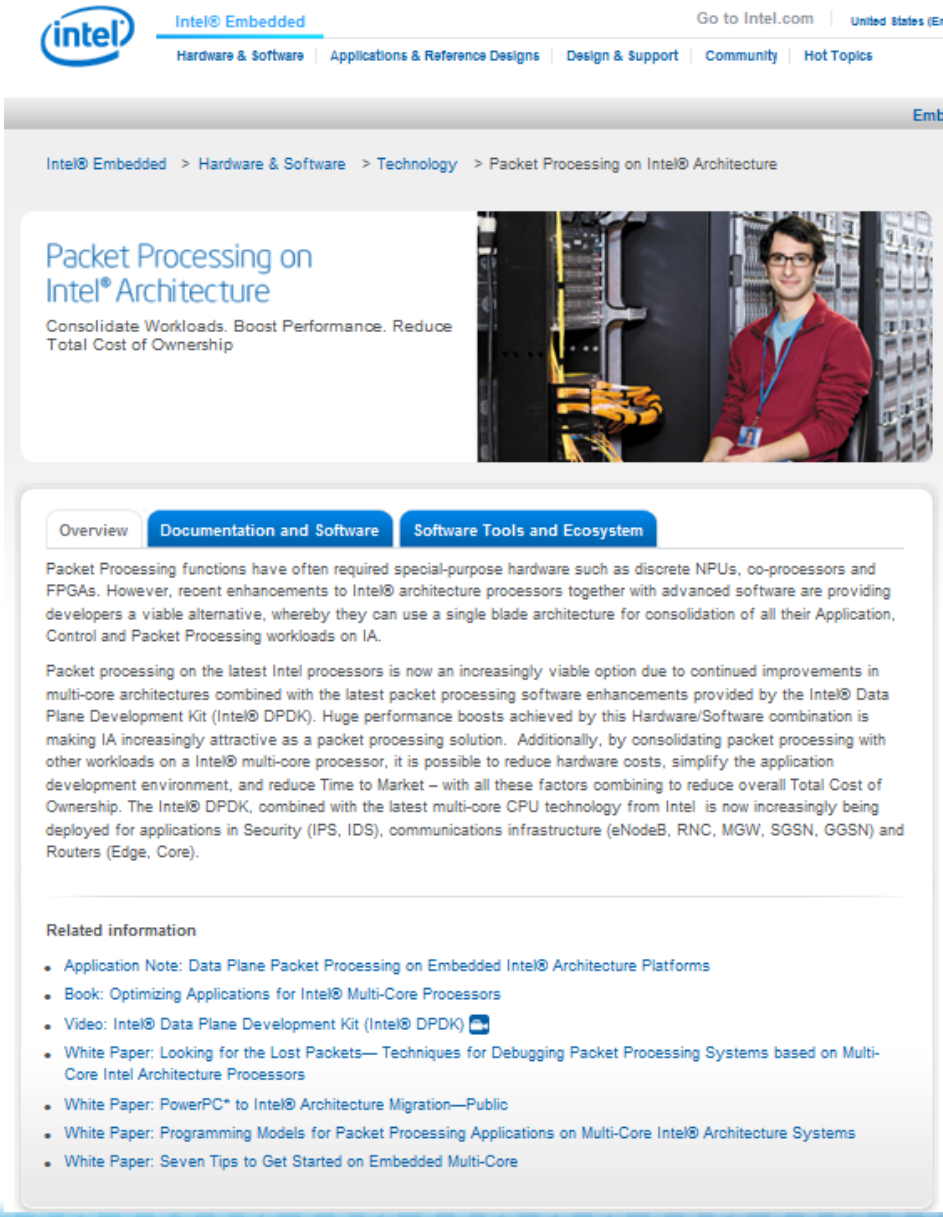
Intel® DPDK

www.intel.com/go/dpdk

Your One-Stop-Shop for:

- Documentation and articles, white papers, pod casts
- Ecosystem information and articles

Examples
See the Video: "Intel® Data Plane Development Kit (Intel® DPDK)". Found on EDC site at www.intel.com/go/dpdk under Video: Intel® Data Plane Development Kit (Intel® DPDK)



Agenda

1. Intel's Packet Processing Motivation and Value Proposition
2. Overview of Intel® DPDK
3. Intel® DPDK Performance Benchmarks
4. Lead Ecosystem Offerings
5. Intel® DPDK Website and Collateral
6. Summary

Summary

Intel® DPDK enables multi-workload/single architecture potential by making IA extremely competitive for packet processing workloads

Distribution of enabling software under flexible and cost-free licensing model enabling maximum customer usability

Fully featured and supported IA Data Plane software solutions via Intel's lead Ecosystem partners

Intelligent Systems