Accelerate Ceph performance via SPDK related techniques

Ziye Yang
Intel Corporation
Oct 2015

Legal Disclaimer

Notice: This document contains information on products in the design phase of development. The information here is subject to change without notice. Do not finalize a design with this information.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at Intel.com, or from the OEM or retailer.

No computer system can be absolutely secure. Intel does not assume any liability for lost or stolen data or systems or any damages resulting from such losses.

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.

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

The products described 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.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

<Use only if applicable> Warning: Altering PC clock or memory frequency and/or voltage may (i) reduce system stability and use life of the system, memory and processor; (ii) cause the processor and other system components to fail; (iii) cause reductions in system performance; (iv) cause additional heat or other damage; and (v) affect system data integrity. Intel assumes no responsibility that the memory, included if used with altered clock frequencies and/or voltages, will be fit for any particular purpose. Check with memory manufacturer for warranty and additional details.

<Use only if applicable> 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.

<Use only if applicable> 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.

<Use only if applicable> Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

<Use only if applicable> 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.

<Use only if applicable> Intel® AMT should be used by a knowledgeable IT administrator and requires enabled systems, software, activation, and connection to a corporate network. Intel AMT functionality on mobile systems may be limited in some situations. Your results will depend on your specific implementation. Learn more by visiting Intel® Active Management Technology.

<Use only if applicable> Intel is a sponsor and member of the Benchmark XPRT Development Community, and was the major developer of the XPRT family of benchmarks. Principled Technologies is the publisher of the XPRT family of benchmarks. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or by visiting www.intel.com/design/literature.htm.

Intel and the Intel logo <Add terms trademarked by Intel and used in this document> are trademarks of Intel Corporation in the U. S. and/or other countries.

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

Agenda

- Recent common requirements for Ceph
- Middle Cache tiering solution
- Building block techniques
  - I/O optimization technique
    - DPDK for storage
  - Data Processing Acceleration techniques
    - ISA-L
- Conclusion

Agenda

- Recent common requirements for Ceph
- Middle Cache tiering solution
- Building blocks techniques
  - I/O optimization technique
    - DPDK for storage
  - Data Processing Acceleration techniques
    - ISA-L
- Conclusion

Recent Common requirements for Ceph

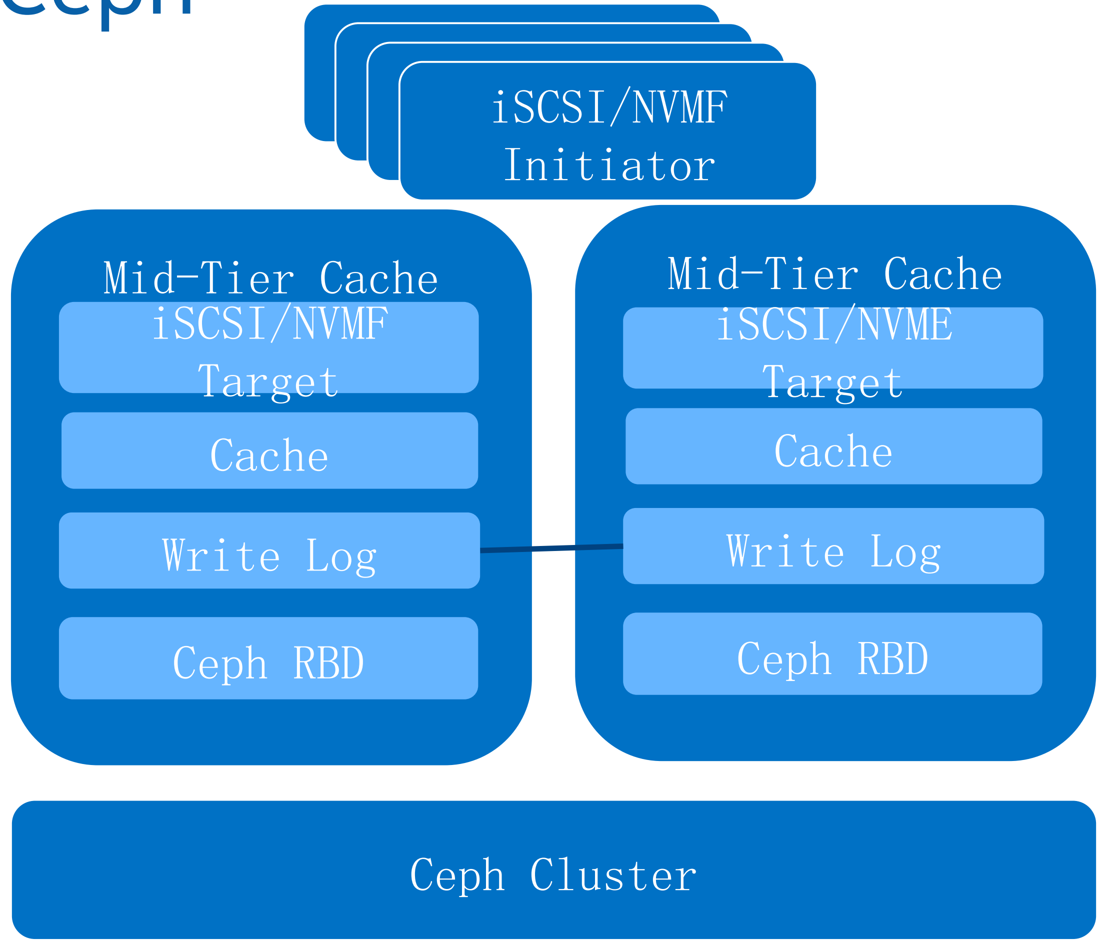
- Legacy protocol support
  - For example, iSCSI interface support for transparently migrating applications from Enterprise storage to cloud storage – Ceph
- High performance requirements for Ceph
  - Low latency for front end applications

Agenda

- Recent common requirements for Ceph
- Middle Cache tiering solution
- Building block techniques
  - I/O optimization technique
    - DPDK for storage
  - Data Processing Acceleration techniques
    - ISA-L
- Conclusion

Provide Mid-Tier Cache between applications and Ceph

- Protocol support:
  - Aim to target iSCSI/NFS/NVMF protocols.
- High performance
  - Provide local cache in each Mid-Tier node
  - Provide write log for data consistency.
- HA:
  - Replicate to 1+ additional nodes
  - Heartbeat for failed node detection

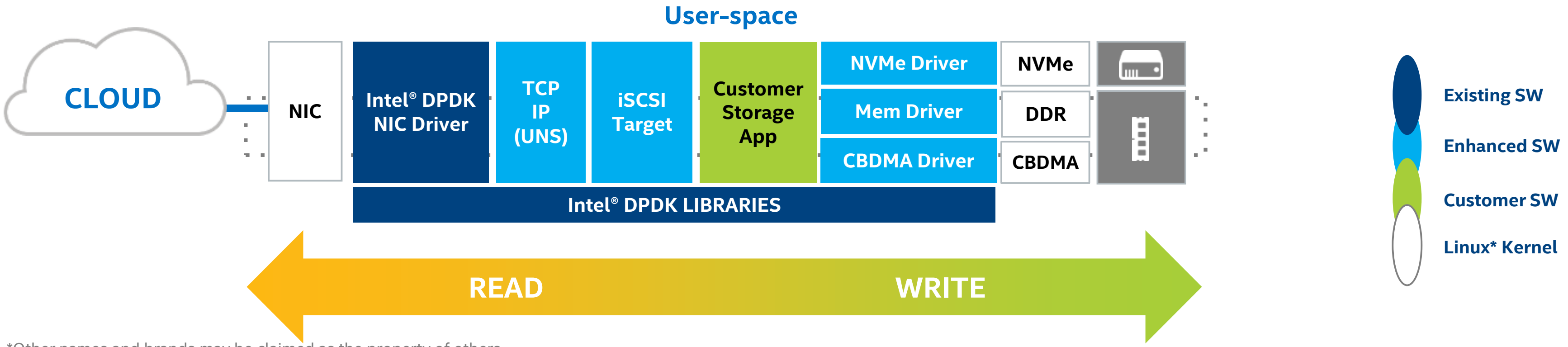


Agenda

- Recent common requirements for Ceph
- Middle Cache tiering solution
- Building blocks techniques
  - I/O optimization technique
    - DPDK for storage
  - Data Processing Acceleration techniques
    - ISA-L
- Conclusion

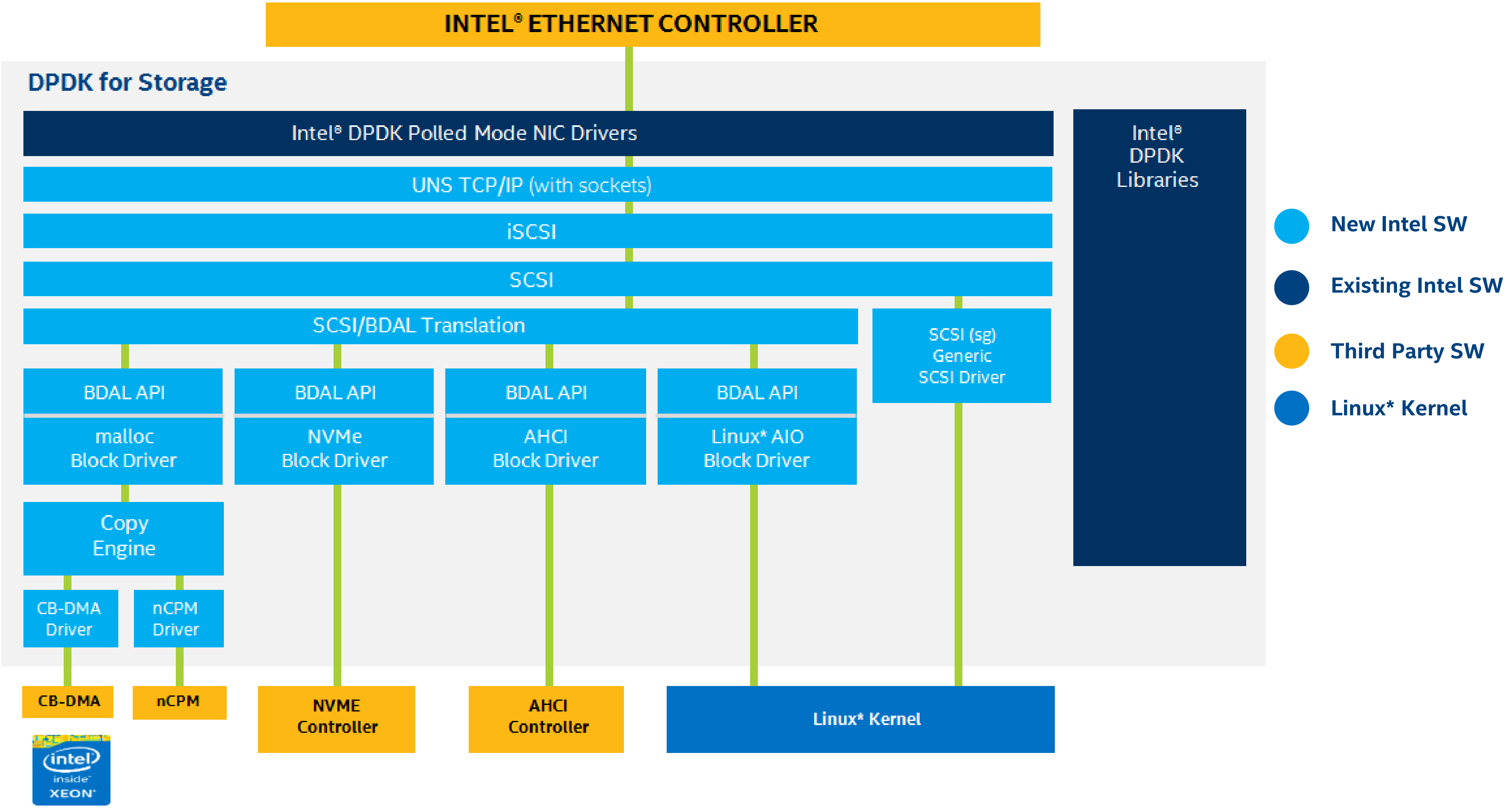
DPDK for Storage **Overview**

- **Uses Intel® DPDK and UNS technology**
  - Optimized user space lockless polling technology in the NIC driver
  - Presents lock-light libraries and network TCP/IP services
- **Provides an enhanced software stack that optimizes iSCSI front end targets**
  - Optimizes packets in user space using lockless polling mechanisms
  - Reference software available for customer application integration to NVMe or other backend
- **Supports Linux* operating systems**
- **Enables a higher system level performance for iSCSI targets**
- **Currently available as reference software**



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

High-Level Block Diagram



INTEL® ETHERNET CONTROLLER

DPDK for Storage

Intel® DPDK Polled Mode NIC Drivers

UNS TCP/IP (with sockets)

iSCSI

SCSI

SCSI/BDAL Translation

SCSI (sg) Generic SCSI Driver

BDAL API

malloc Block Driver

BDAL API

NVMe Block Driver

BDAL API

AHCI Block Driver

BDAL API

Linux* AIO Block Driver

Copy Engine

CB-DMA Driver

nCPM Driver

Intel® DPDK Libraries

CB-DMA

nCPM

NVME Controller

AHCI Controller

Linux* Kernel

● New Intel SW

● Existing Intel SW

● Third Party SW

● Linux* Kernel

Intel DPDK for Storage **Benefits**

Up to 7x Better Performance[+]	▪ vs. Linux*-IO Target (LIO) ▪ Or 1/7 CPU overhead at same performance
Up to 10x fewer cores utilized with the NVMe Driver	▪ Vs. Linux* NVMe Driver
Reduces Total Cost of Ownership	▪ Reduce BOM costs between $80-500 by removing the need for a TOE ▪ Utilize free CPU cycles for other workloads
Free Source Code	▪ Customizable source code available as reference ▪ Evaluation source available upon request
User Space Implementation	▪ Portable/Upgradable and permissive licensing ▪ Requires Software License Agreement for full product use

[+] 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.
Source: Intel Internal Measurements as of 22 August 2014. See back up slides for configuration details.

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 SSE3 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.
For more information go to http://www.intel.com/performance
*Other names and brands may be claimed as the property of others.

Intel DPDK for Storage
Full **Packaging** and **Contents**

Library Package Includes:

- Intel DPDK | UNS | Optimized Storage Stacks as reference software
- User space support code (written in C):
  - POSIX compliant
  - Demo/Usage, Unit test (functional correctness), Basic performance
- API manuals – may include links or copy key papers
- Release.txt (release notes, version, and library serial IDs)
- Linux* Support

Source Agreement

- Source available under Restricted Use License Agreement Confidential (RULAC)
- Source code available under source license agreements (SLA)

Intel® DPDK for Storage:
Case Study1: Performance Comparison with LIO

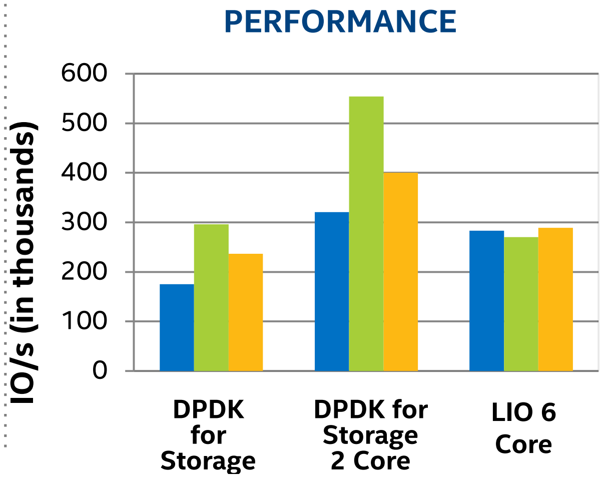
Intel® Xeon® Processor E5-2620v2-iSCSI Read/Write:
4 KB Data (performance per/core)

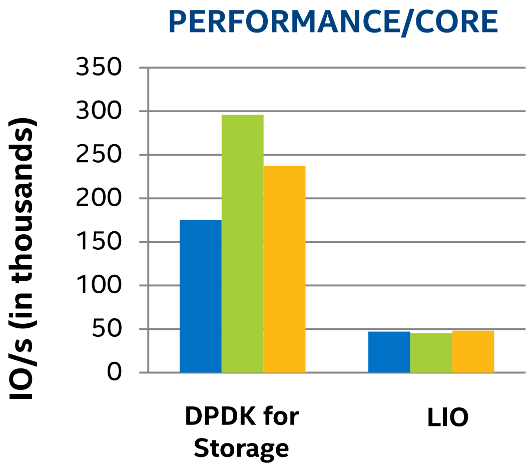
NVM Express Backend

● 4 KB-Random-100% Write

● 4 KB-Random-100% Read

● 4 KB-Random-70% Read 30% Write





Up to 650% increase in max performance per core+

+ 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. Source: Intel Internal Measurements as of 22 August 2014. See back up slide # 10-13 for configuration details.

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 SSE3 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.
For more information go to http://www.intel.com/performance

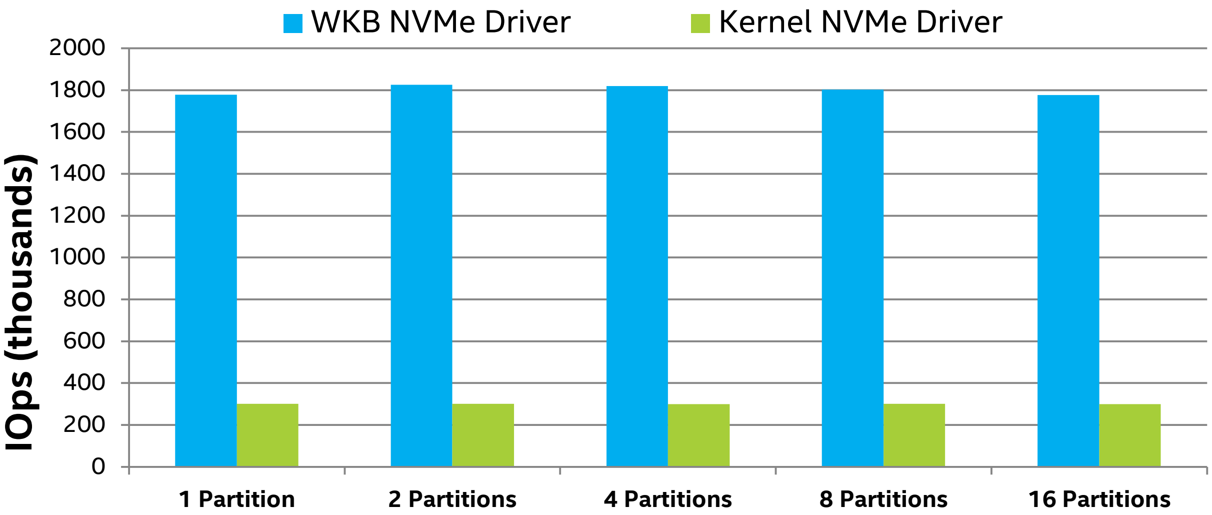
Intel® DPDK for Storage
Case Study2: User space NVME driver(SPDK) Benefit

4 KB Random Read Performance: **4 x NVMe Drives**

Single-Core Intel® Xeon® Processor

● DPDK for Storage NVMe Driver

● Kernel NVMe Driver

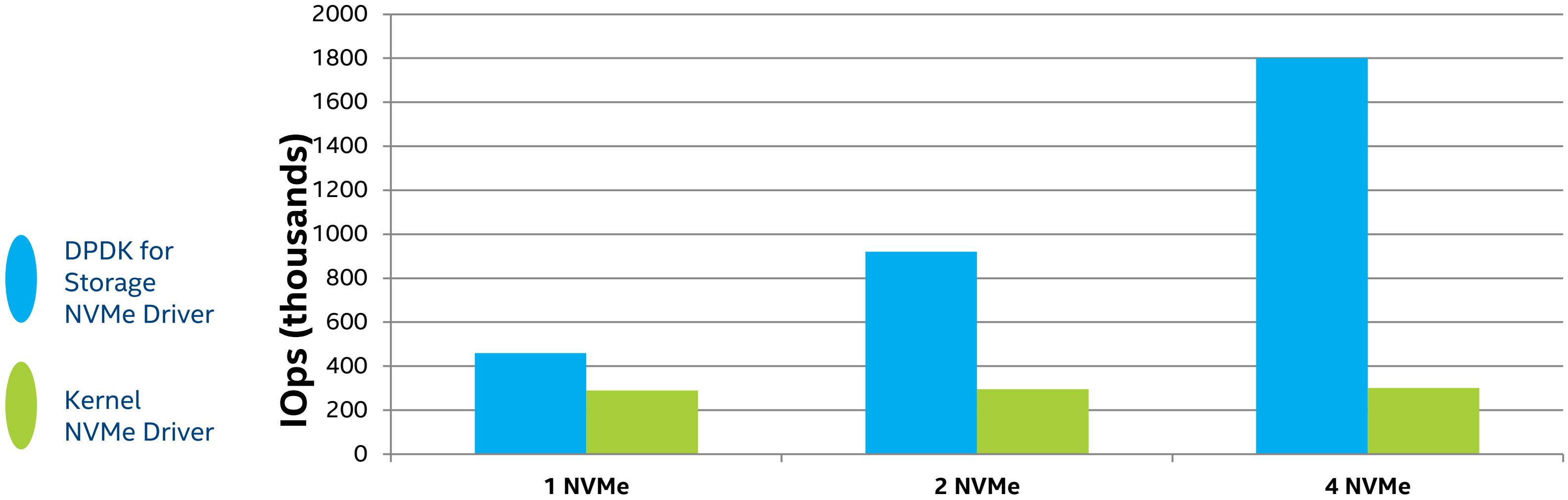


DPDK for Storage NVMe driver delivers up to **6x** performance improvement vs. Kernel NVMe driver with a single-core Intel® Xeon® processor

For test configuration details see slide # 16 and 18

Disclaimer: 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.

4 KB Random Read Performance: **1-4 NVMe Drives** Single-Core Intel® Xeon® Processor



DPDK for Storage NVMe driver scales linearly in performance from 1 to 4 NVMe drives with a single-core Intel® Xeon® processor

For test configuration details see slide # 16 and 18

Disclaimer: 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.

Agenda

- Recent common requirements for Ceph
- Middle Cache tiering solution
- Building blocks techniques
  - I/O optimization technique
    - DPDK for storage
  - Data Processing Acceleration techniques
    - ISA-L
- Conclusion

Benefits of using **Intel ISA-L**

Intel ISA-L enables Storage OEMs to obtain more performance from Intel CPUs and reduce investment in developing their own optimizations

Up to **7X** BANDWIDTH for Hash functions compared to OpenSSL algorithms

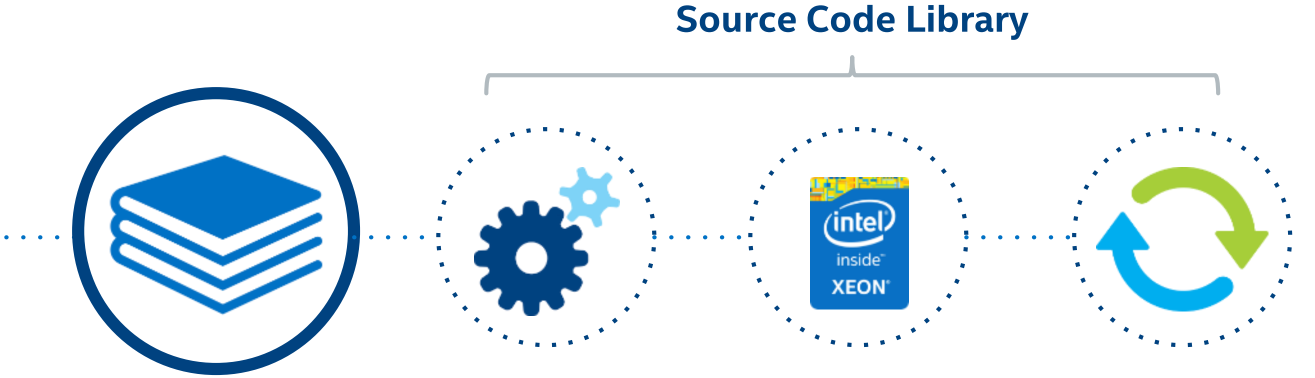
Allows **MAXIMUM** UTILIZATION of additional cores

Up to **4X** BANDWIDTH improvement on compression compared to zlib

FASTER TTM/ **LESS** RESOURCES than developing optimizations from scratch

Allows Intel to DEVELOP **OPTIMIZATIONS** that use new architectural enhancements that are TTM

Intel® ISA-L Packaging and Contents



Source Code Library

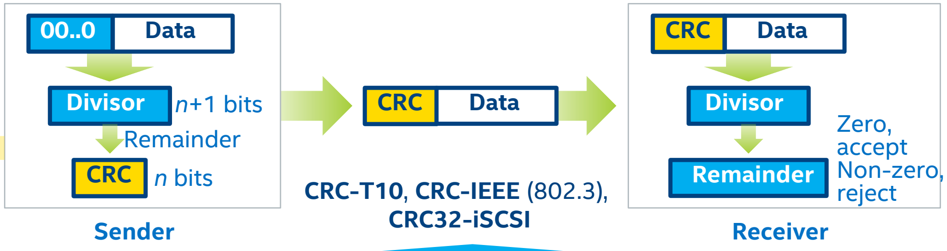
Single Core Low-Level Functions (OS independent functions)

Supports 64-bit, Intel® Xeon® and Atom Processor E5-2600/2400 and Atom C2000 product family forward

Gen Over Gen Function Updates to take advantage of new processor features

Intel® ISA-L **Functions**

XOR (RAID 5), **P+Q** (RAID 6), Reed Solomon Erasure Code



00..0 | Data

Divisor *n*+1 bits

Remainder

CRC *n* bits

Sender

CRC | Data

CRC-T10, CRC-IEEE (802.3), **CRC32-iSCSI**

CRC | Data

Divisor

Remainder

Zero, accept
Non-zero, reject

Receiver

DATA PROTECTION

DATA INTEGRITY

PERFORMANCE OPTIMIZING

CRYPTOGRAPHIC HASHING

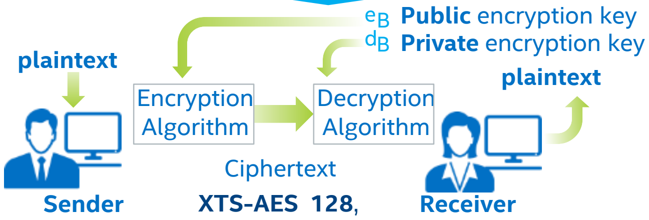
COMPRESSION “DEFLATE”

ENCRYPTION

06d80e7
b0C50bs
49a509t
b49f249
24e8c8o
05x84q4

Multi-Buffer: SHA-1, SHA-256, SHA-512, MD5

IGZIP: Fast Compression



e_B **Public** encryption key

d_B **Private** encryption key

plaintext

Encryption Algorithm

Decryption Algorithm

plaintext

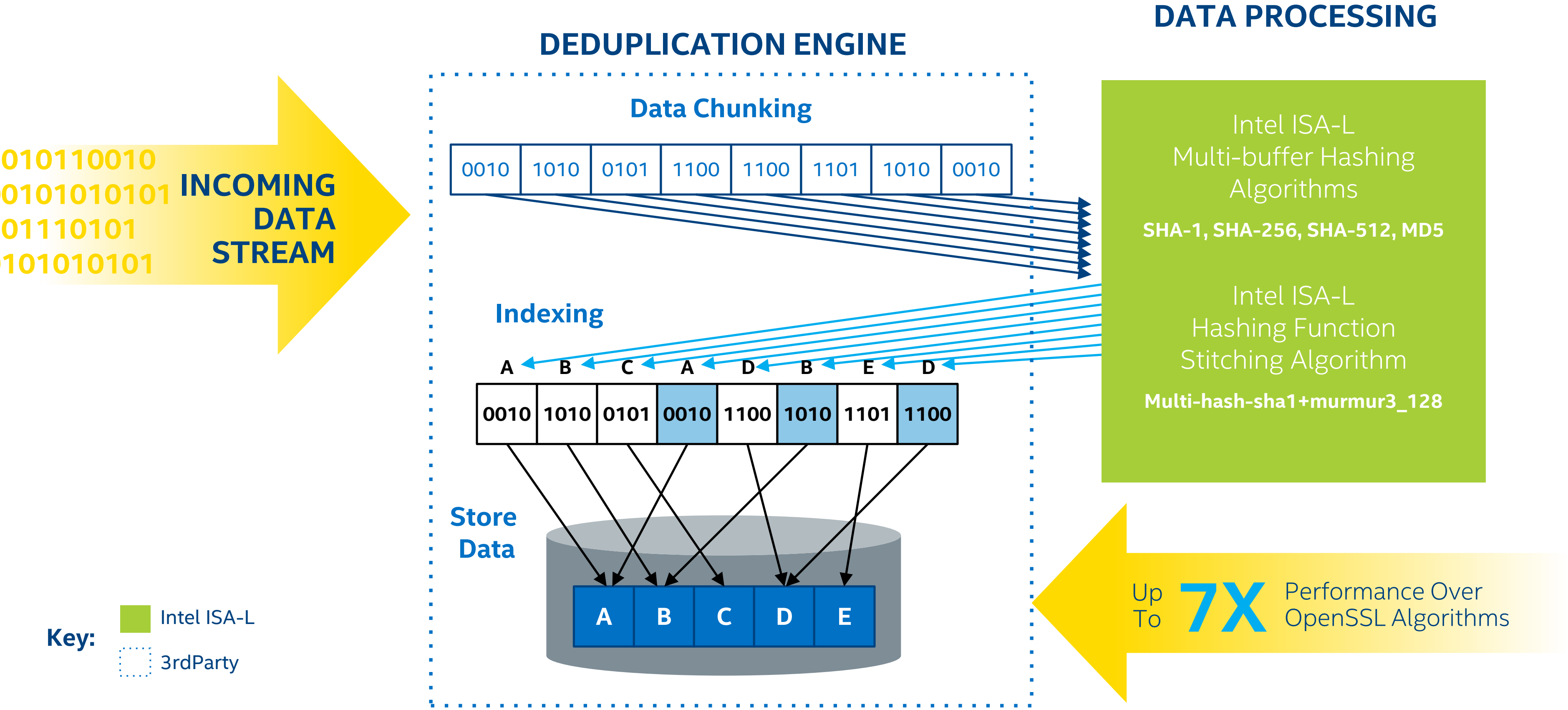
Ciphertext

Sender

Receiver

XTS-AES 128,
XTS-AES 256

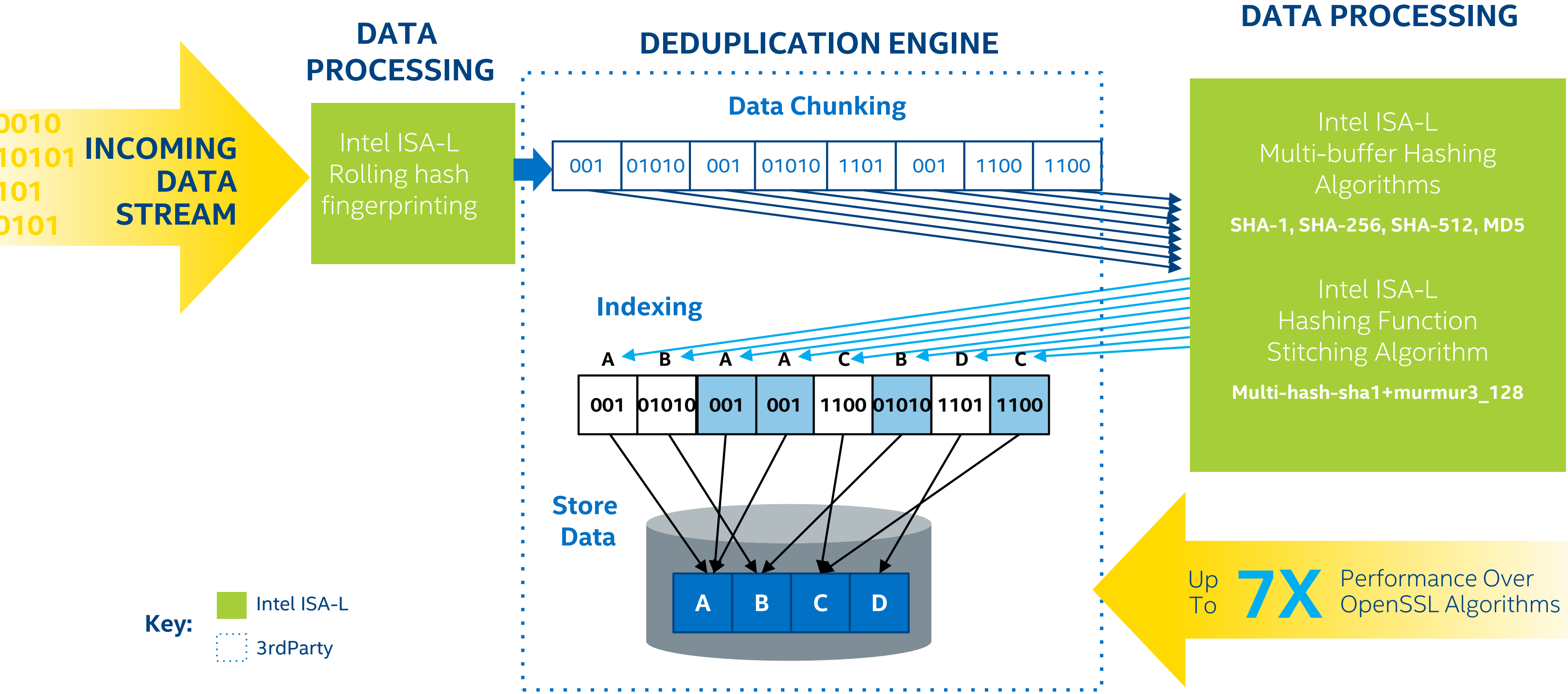
Hashing Usage: Data Deduplication Optimizations (Fix Size)



Key:

- 🟩 Intel ISA-L
- ⬚ 3rdParty

Hashing Usage: Data Deduplication Optimizations (Dynamic Size)



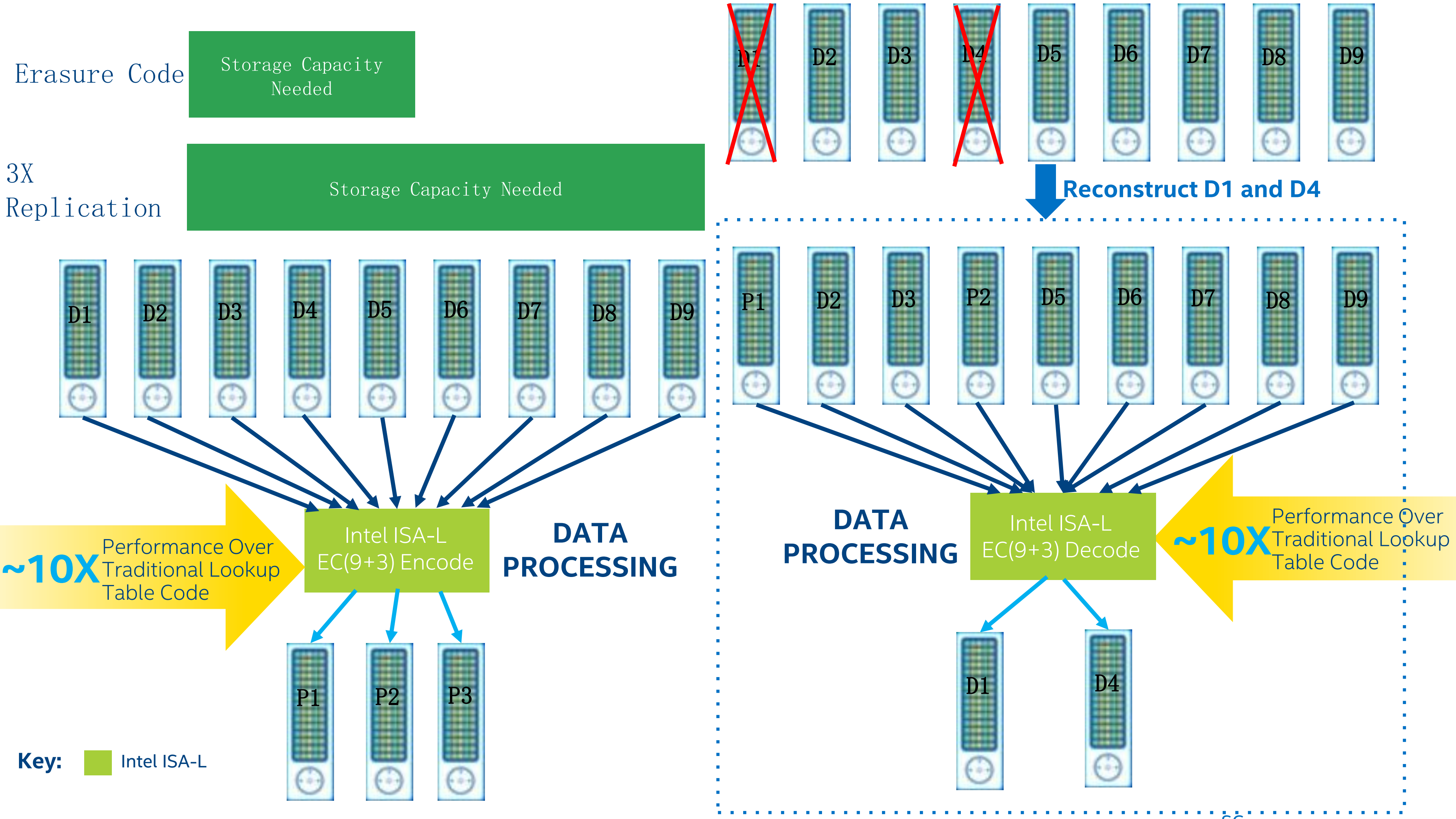
0010
10101
101
0101
INCOMING DATA STREAM
DATA PROCESSING
Intel ISA-L Rolling hash fingerprinting
DEDUPLICATION ENGINE
Data Chunking
001
01010
001
01010
1101
001
1100
1100
Indexing
A
B
A
A
C
B
D
C
001
01010
001
001
1100
01010
1101
1100
Store Data
A
B
C
D
DATA PROCESSING
Intel ISA-L Multi-buffer Hashing Algorithms
SHA-1, SHA-256, SHA-512, MD5
Intel ISA-L Hashing Function Stitching Algorithm
Multi-hash-sha1+murmur3_128
Up To 7X Performance Over OpenSSL Algorithms

Key:

🟩	Intel ISA-L
⬚	3rdParty

Intel ISA-L provides a solution to deploy Erasure Code (EC) with better performance, so that data replication can be done faster with half the space of other methods.

- **Support any Matrixes: Vandermonde Reed-Solomon EC, Cauchy Reed-Solomon EC**
- **Support the different EC strategies: Local Reloadable Code EC, Regeneration Code EC, Hitchhiker Code EC**



Erasure Code — Storage Capacity Needed

3X Replication — Storage Capacity Needed

Reconstruct D1 and D4

DATA PROCESSING

Intel ISA-L EC(9+3) Encode

~10X Performance Over Traditional Lookup Table Code

DATA PROCESSING

Intel ISA-L EC(9+3) Decode

~10X Performance Over Traditional Lookup Table Code

Key: 🟩 Intel ISA-L

Source: "Erasure Code and Intel® Intelligent Storage Acceleration Library"
http://www.intel.com/content/www/us/en/storage/erasure-code-isa-l-solution-video.html

Solving Real-World Problems: Qihoo 360

DEPLOYED **INTEL-ISAL-BASED** HDFS Raid for **INTEL-XEON-BASED** Cold Storage

EC Encode SPEEDS **45X FASTER** THAN JAVA VRS

EC Decode SPEEDS **36X FASTER** THAN JAVA VRS

REDUCED COSTS BY **25%~30%**

Source: Case Study "Intel and Qihoo 360 Internet Portal Datacenter - Big Data Storage Optimization Case Study”
https://software.intel.com/en-us/articles/intel-and-qihoo-360-internet-portal-datacenter-big-data-storage-optimization-case-study

Solving Real-World Problems: Alibaba

DEPLOYED **INTEL-ISAL-BASED** Sheepdog Erasure Code for **INTEL-ATOM-C2000-BASED** Cold Storage

Data Recovery SPEEDS

4X FASTER

THAN Sheepdog ZFEC

5X

CPU utilization reduction

Source: Case Study "Lambert: Achieve High Durability, Low Cost & Flexibility at Same Time, Open source storage engine for exabyte data in Alibaba"
http://events.linuxfoundation.org/sites/events/files/slides/LFVault2015_Alibaba.pdf

Conclusion

- In this presentation, we introduce the storage optimization techniques provided by Intel for accelerating the Ceph performance:
  - I/O optimization technique: DPDK for storage (SPDK)
  - Data Processing Acceleration: ISA-L
- These kinds of building block techniques can help customers to accelerate the Ceph performance on IA platform

Q & A ?

intel®