Accelerating Data Movement with Intel® Data Streaming Accelerator (Intel® DSA) on 4th Gen Intel® Xeon® Scalable Processor

[Presenter name]

[Date]

Contents

- Value Prop – benefits of Intel DSA
- Workload targets for Intel DSA
- Metrics/results to help understand benefits of Intel DSA
- Software requirements, and how to integrate Software with Intel DSA

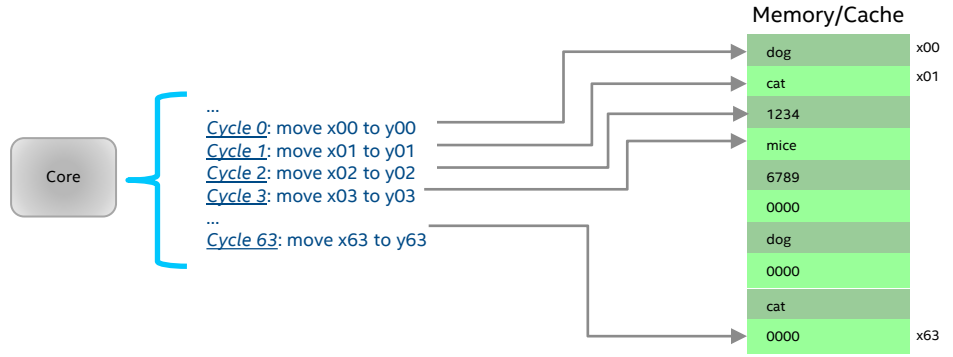
Intel DSA Overview & Benefits

Intel DSA – a data mover IP integrated on 4th Gen Xeon®

- Intel DSA Offloads data copy and data transformation operations (Move, DIF, CRC, Fill, Compare, Flush & Dual cast).
- ➢ **Freeing up CPU cycles (Increasing compute capacity).**
- ➢ **Accelerate data movement throughput**
- ~30GB/s throughput in each direction per Intel DSA instance
- Multiple Intel DSA instances per socket (E.g., up-to 4 Intel DSA in 4th Gen Intel® Xeon® XCC) providing up-to ~120GB/s throughput in each direction
- Leverages AIA and IOMMU features for efficient offload and scalable sharing
- ➢ 1 Intel DSA device available on all 4th Gen Intel® Xeon® SKUs.
- ➢ 1 to 4 Intel DSA Accelerator devices supported on select 4th Gen Intel® Xeon® SKUs

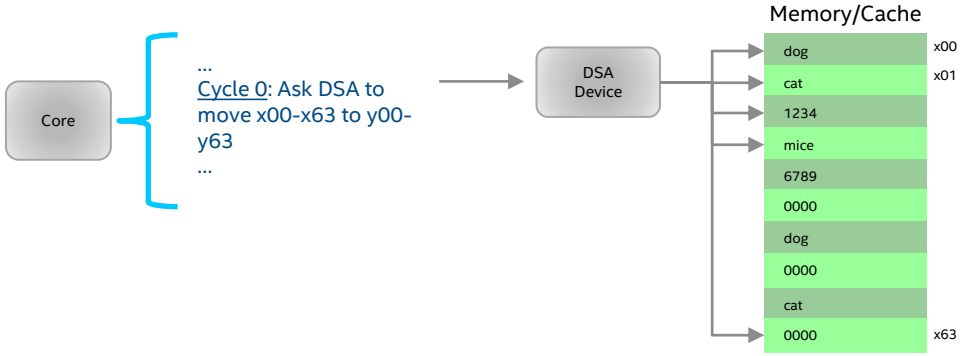
Memory Copy **without** Intel DSA

Cores move data; core cycles consumed



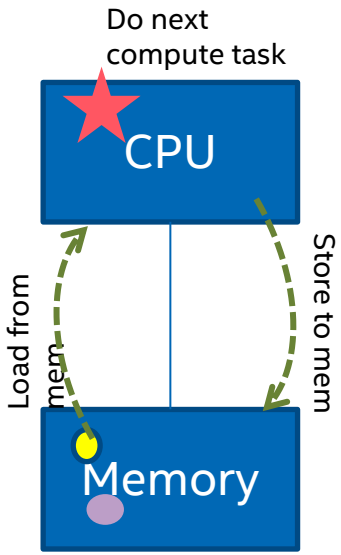
Memory Copy **with** Intel DSA

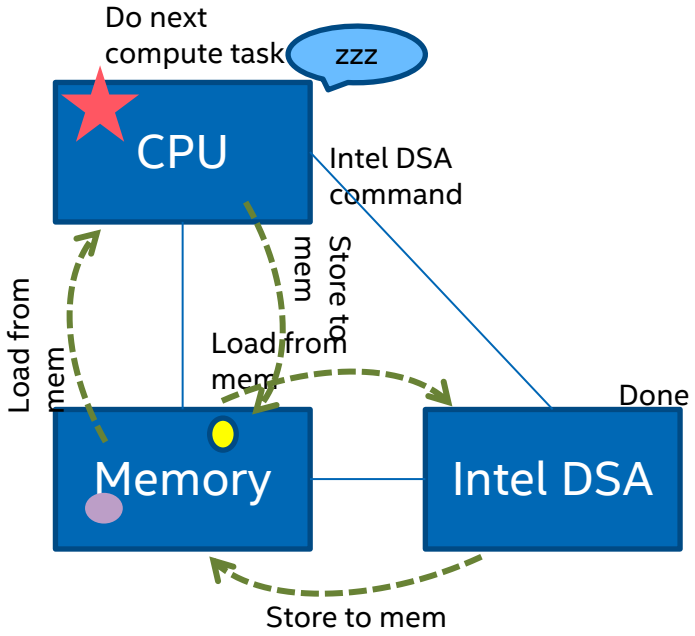
DSA moves data; core cycles freed

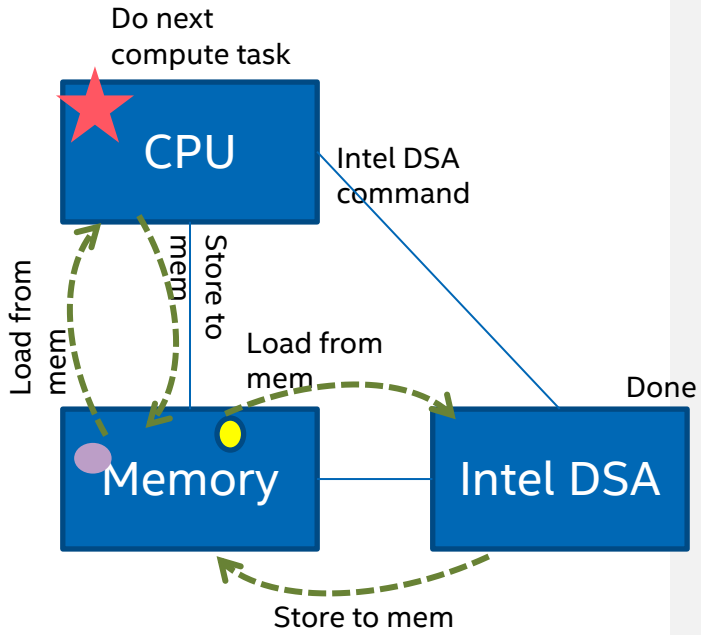


Animated Build

CPU->Intel DSA Offload Storyline



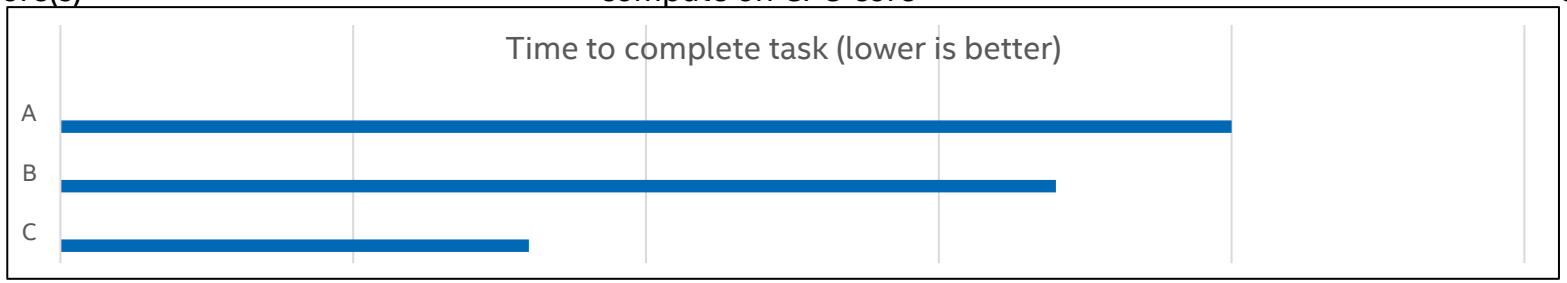




A Data movement and compute operations execute serially on CPU core(s)

B Data movement offload to Intel DSA followed serially by compute on CPU core

C Concurrent execution of data movement using Intel DSA and compute on CPU core



Intel DSA Value Proposition in Summary

Intel DSA offloads data movement operations from CPU – freeing up Cores.
*Increasing effective 4th Gen Intel® Xeon® Processor Performance and Perf/Watt

Intel DSA on 4th Gen Intel Xeon processor offers performance gains compared to prior gen (CBDMA)
* Establishing Intel DSA leadership in Data movement acceleration.

Intel DSA provides improved CPU latency @ as-low-as 6 KB Transfer Size
*making Intel DSA available for synchronous copy operations

Intel DSA delivers better throughput @ as-low-as 2KB Transfer Size
* making Intel DSA full capability available for smaller Transfer size

Intel DSA is integrated into industry leading framework SPDK, DPDK, Libfabric and more..
* making its deployment easy for Storage, Networking & more Applications

Intel DSA Performance

Function
- Optimizing streaming data movement and transformation operations

Business Value
- Accelerated data protection for NVMe/TCP improving efficiency for data storage applications via CPU offload

Software Support
- Intel® Data Mover Library

Use Cases
- Virtualization, fast replication across non-transparent bridge, ERP, In-Memory Databases

Performance gains vs not using these accelerators

Data Integrity (Throughput)

Up to

1.7x

higher IOPs for large packet sequential reads with built-in Intel® DSA vs. ISA-L software

Performance gains vs prior generation products

Data Integrity (Throughput & Latency)

Up to

1.6x

higher IOPs and

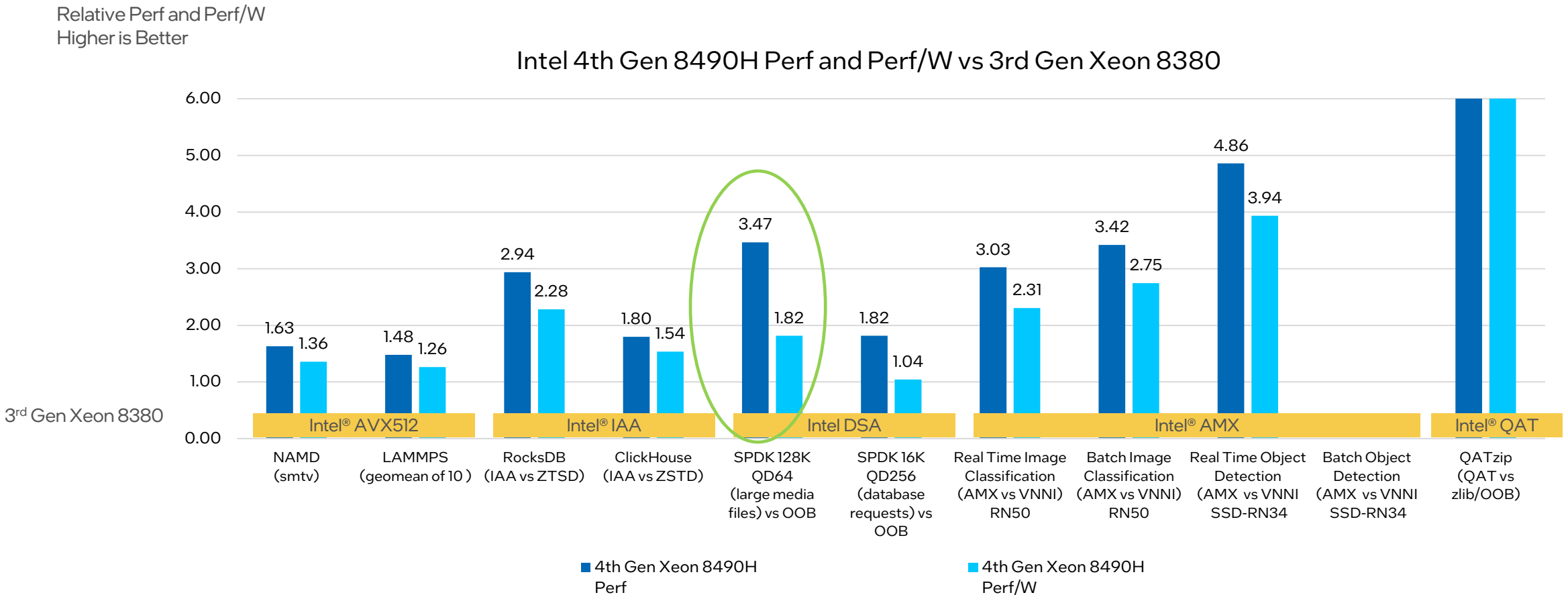
37%

Latency reduction for large packet sequential reads with built-in Intel® DSA vs. prior generation

See [N18] at https://edc.intel.com/content/www/us/en/products/performance/benchmarks/4th-generation-intel-xeon-scalable-processors/

Your results may vary.

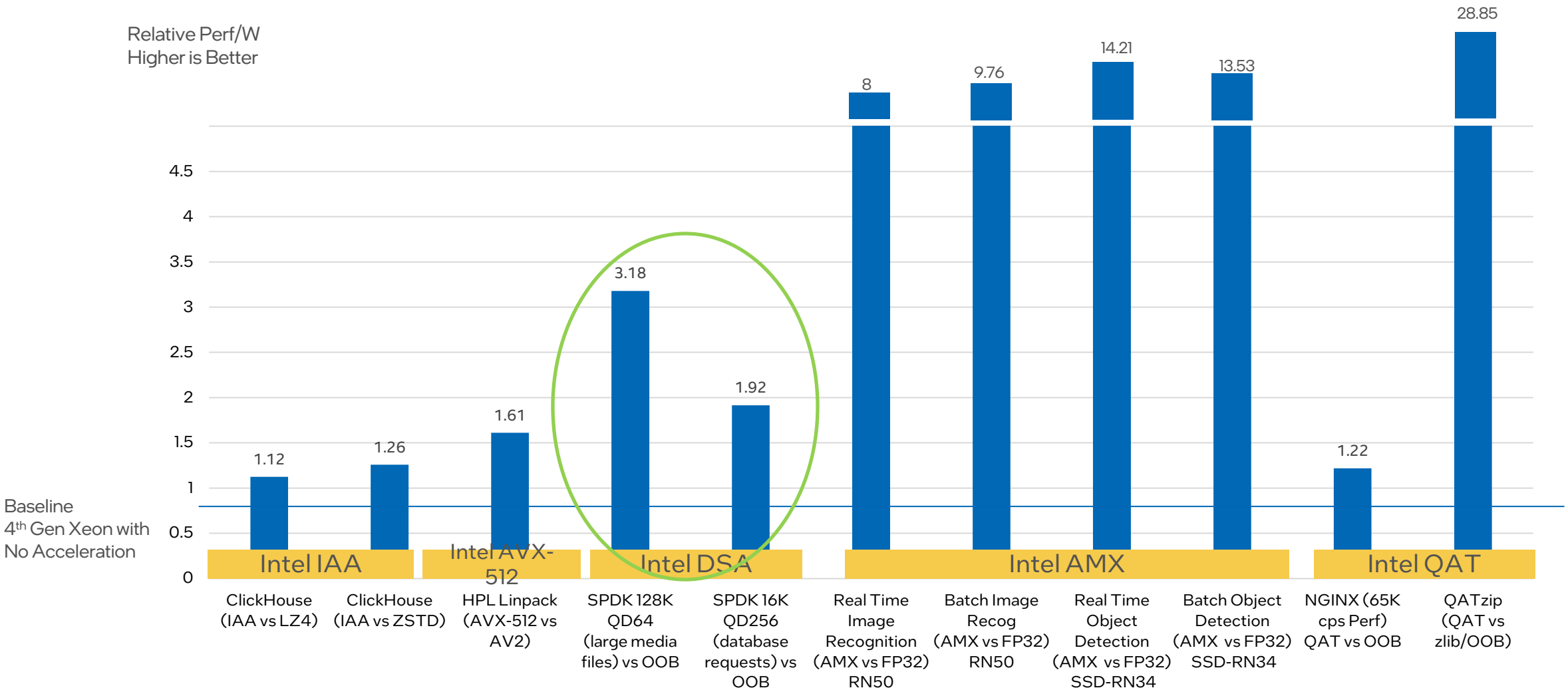
4th Gen Generational Accelerator Performance and Efficiency



Your results may vary.

See slide #29 for configuration details

4th Gen Xeon Accelerators Efficiency



Your results may vary.

See slide #28 for configuration details

How to know if Intel DSA will help my workload

See “Related Specifications, Application Notes, and White Papers” under:

https://www.intel.com/content/www/us/en/developer/articles/technical/intel-sdm.html

Intel DSA Software stack and requirements

Applications

mTCP

Media Transport Lib

NVMe- oF

Frameworks

SPDK

VPP

Libraries

OFI Libfabric

Intel ® MPI

Intel® DML

Linux Page Copy/Clear***

Virtualization (Guest)

QEMU / KVM

OS & Kernel (Host)

Linux

RedHat

Ubuntu

SLES15 SP4

Azure Host ***

★ up-streamed

*** Out-of-Tree Code / To-be released

Resources Link

Software Support

Linux Driver:

- Intel® Data Accelerator Driver -> https://github.com/intel/idxd
- IDXD GitHub* repository -> https://github.com/intel/idxd-driver
- Accel-config (configuring DSA)-> https://github.com/intel/idxd-config

- Opensource blog at 01.org -> INTRODUCING THE INTEL® DATA STREAMING ACCELERATOR (INTEL® DSA)
- Intel DSA preliminary external specification -> Intel DSA Specification

Library:

- DPDK vSwitch -> http://doc.dpdk.org/guides/dmadevs/idxd.html
- Storage SPDK -> https://spdk.io/doc/idxd.html
- Intel® MPI Library -> https://www.intel.com/content/www/us/en/developer/tools/oneapi/mpi-library.html
- Media Transport Library -> https://github.com/OpenVisualCloud/Media-Transport-Library
- OFI Libfabric -> https://github.com/ofiwg/libfabric/releases (v1.17.x)
-
- Intel® Data Mover library (Intel® DML) v0.1.9-beta -> https://intel.github.io/DML/
- mTCP: user-level TCP stack for multicore systems. -> https://github.com/mtcp-stack/mtcp
- FD.io's Vector Packet Processor (VPP) -> https://s3-docs.fd.io/vpp/23.02/

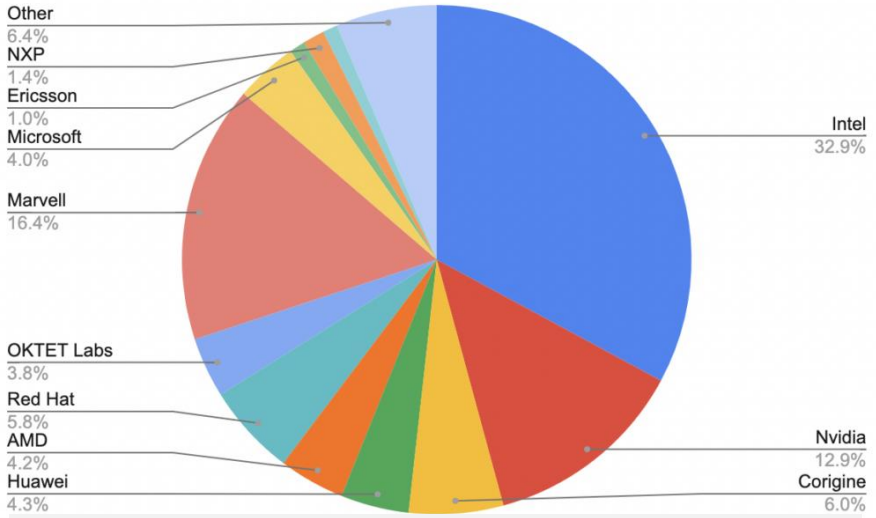
OVS-DPDK Perf Acceleration with Intel DSA

Intel Developer Zone Link: https://cdrdv2.intel.com/v1/dl/getContent/758697

What is DPDK?

A Data Plane Development Kit (DPDK) that consists of libraries to accelerate packet processing workloads running on a wide variety of CPU architectures.

DPDK's robust community of member organizations and cross-industry partners spans: hardware vendors, physical and virtual network drivers, and other open-source organizations that consume DPDK.
See below for a full list of up and down-stream open-source projects that consume DPDK.



Number of commits per employer

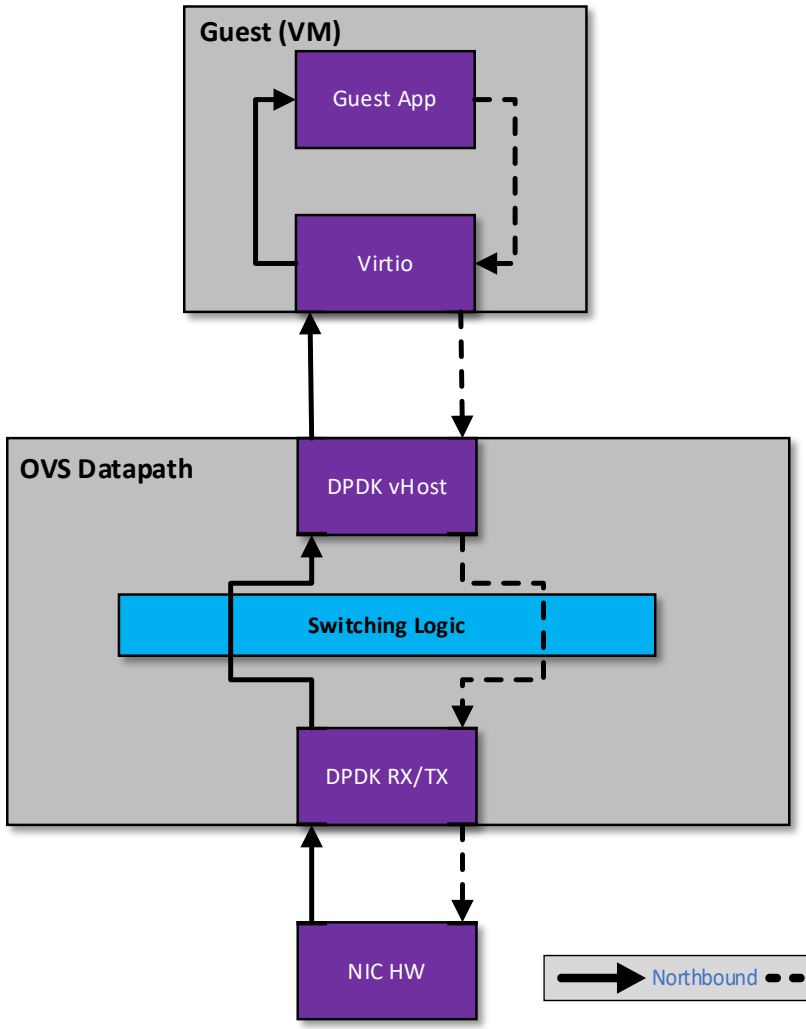
ANS– Accelerated Network Stack
BESS– Berkeley Extensible Software Switch
Butterfly– Connects Virtual Machines
Catnip– TCP Stack in Rust
DPVS– Layer-4 load balancer
dperf – Network load tester
FastClick– Highspeed dataplane
F-Stack– TCP Stack
IMTL – Real time and low latency media transport library
Lagopus– software OpenFlow 1.3 switch
Metronome – adaptive packet retrieval in DPDK
MoonGen– Packet generator
OpenDataPlane – Open DataPlane DPDK platform implementation
YANFF– NFF-Go -Network Function Framework for GO(former YANFF)

Open vSwitch– Multilayer Open Virtual Switch
Packet-journey– Userland router
Pktgen-dpdk– Packet generator
PcapPlusPlus– C++ packet parsing framework
Ruru– Real-time TCP latency monitoring
Seastar– open-source C++ framework
SPDK– Storage Performance Development Kit
TLDK– TCP Stack
TRex– Stateful Traffic Generator
VPP– Fast Data Project
WARP17– Stateful Traffic Generator
mTCP– User-level TCP Stack
OPNFV– Open Platform for NFV

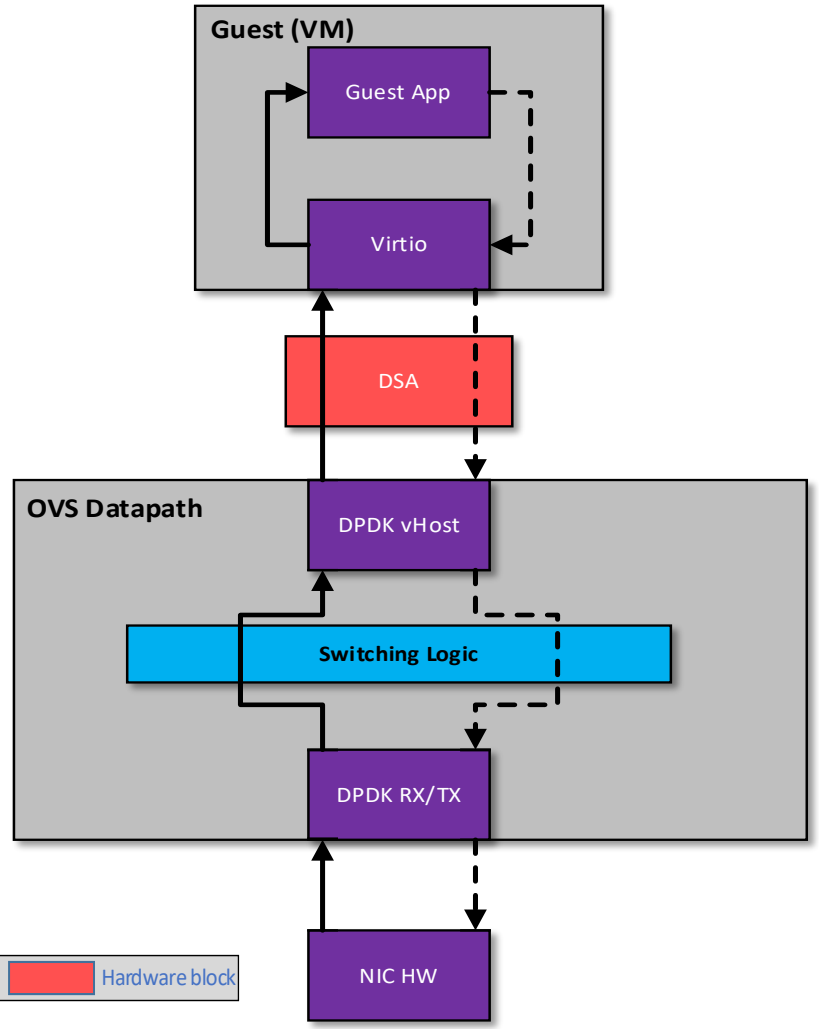
DPDK support for Intel DSA is described at http://doc.dpdk.org/guides/dmadevs/idxd.html.

Intel DSA Integration in DPDK/OVS

Phy-VM-Phy Datapath



Phy-VM-Phy Datapath with Intel DSA



Intel DSA : Protecting Data in the NVMe/TCP storage use case

https://ci.spdk.io/download/2022-virtual-forum-us/SPDK22-John-Kariuki-Protecting-NVMe-TCP-Data-with-Intel-DSA.pdf

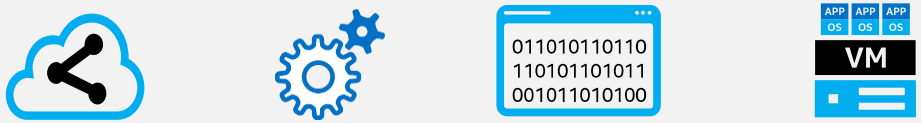
What is SPDK?

Storage Performance Development Kit (SPDK) is a set of tools and libraries for writing high-performance, scalable, user-mode storage applications.

- Intel Established & Lead
- Vibrant, Multi-vendor, Global Community
- Storage Tools, Libraries, Drivers, & Applications
- User-space, poll mode, lockless, asynchronous
- Significant performance and efficiency!
  - 14 Million NVMe IO/core/sec
- Rich Feature Set
- **Broad adoption in Cloud & Enterprise**
- Open Source & BSD Licensed

Participate/Learn More https://SPDK.io

FEATURE CATEGORIES



Network Protocols | Services | Drivers | Virtualization

POPULAR USE CASES

Database | SDS | IaaS | Enterprise | Infrastructure Offload

OPTIMIZED FOR

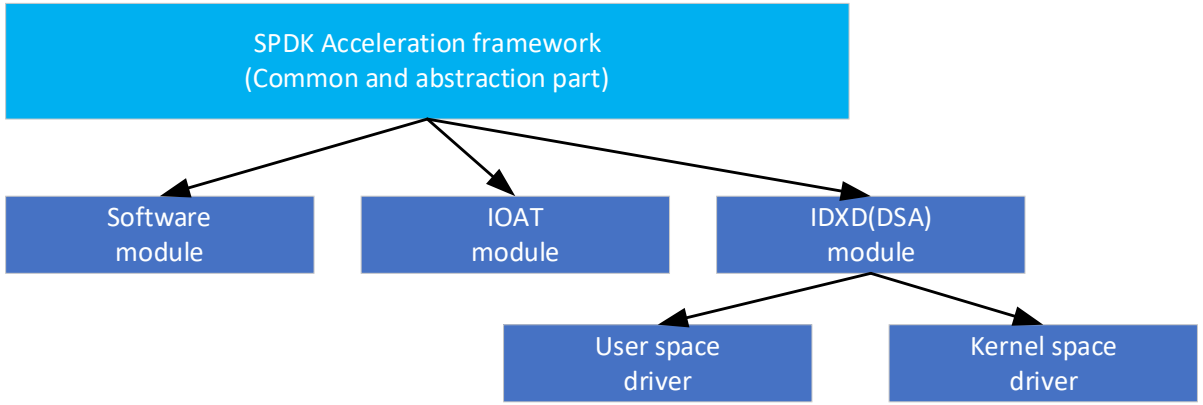
CPUs & Chipsets | NICs IPUs | FPGAs & Accelerators | SSDs

SPDK Community insights

- 49 active contributors created 534 commits just last quarter
  - 14+ companies
- 200+ unique visitors to GitHub each day
- 2021 'SPDK, PMDK, & Intel® Platform Performance Analyzer Virtual Forum':
  - US event over 500 attendees, spanning 130 companies and 35 universities
  - PRC event over 400 attendees, from around 200 companies and organizations
- Multiple third party maintainers
- 'SPDK 2022 Hackathon' completed – April 28th

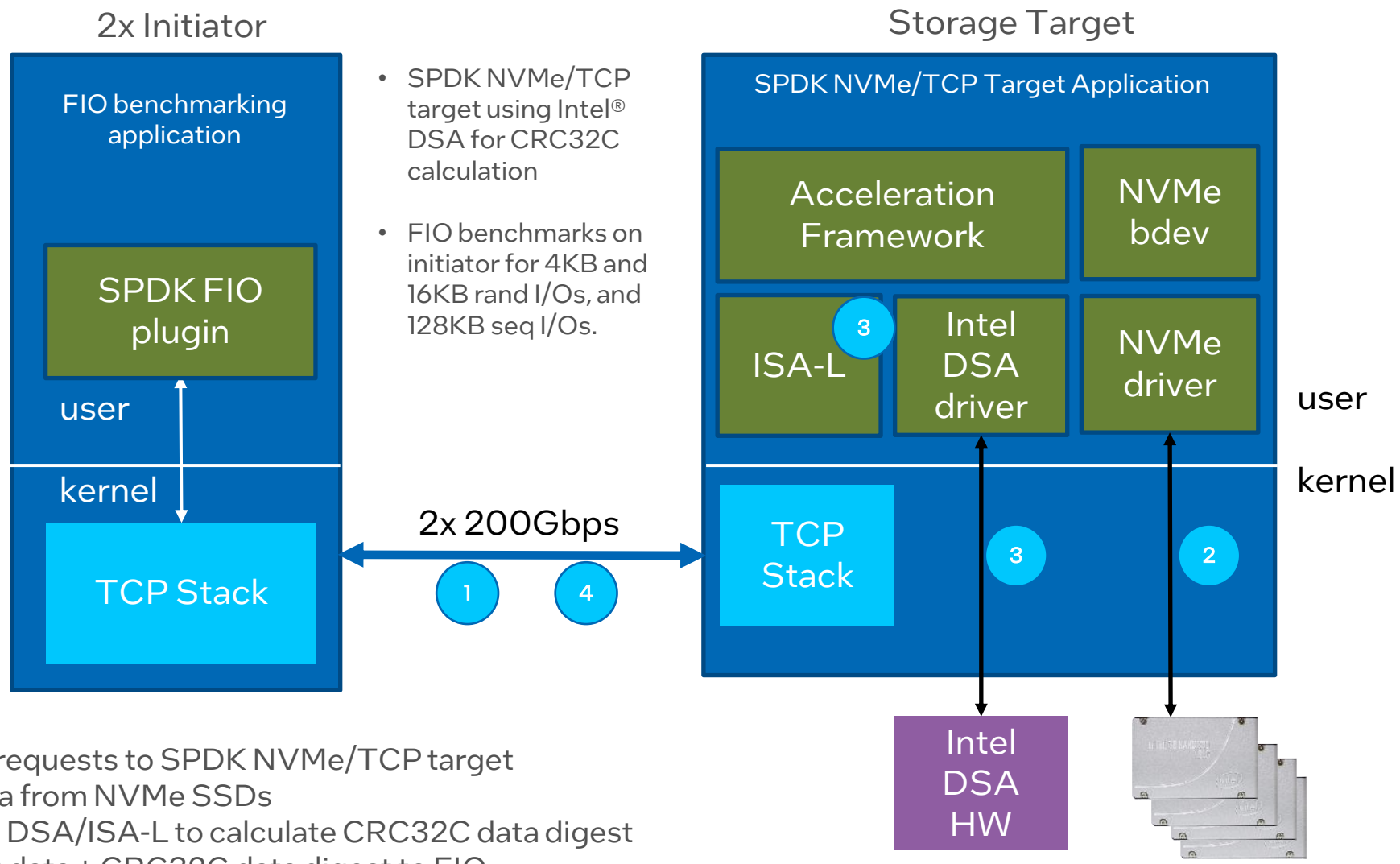
SPDK Acceleration Framework

- A framework for abstracting general acceleration capabilities:
  - With HW engines like Intel DSA, IOAT, QAT, etc.
  - Designed for SW defined infra/storage: SW plug-in modules for environments without HW accelerators.
  - Asynchronous workflow: application uses CPU for other work while HW accelerator is moving/transforming data.
- Accelerated Functions: CRC32CC, copy, fill, compare, dual cast, copy_crc32c, de/compress
- SPDK provides libs/apps for organization building enterprise, SDS, object store solutions.
- NVMe-oF target: User-space storage target, presenting block devices over ethernet fabrics uses the accel FW to offload CRC32C digest calculation to DSA.
- Link to Documentation



Other SPDK-based app of Accel FW: App replicating data over NTB use SPDK Accel FW for dualcast operation

NVMe-oF/TCP Setup



- SPDK NVMe/TCP target using Intel® DSA for CRC32C calculation
- FIO benchmarks on initiator for 4KB and 16KB rand I/Os, and 128KB seq I/Os.

Data Flow:

1. FIO submits I/O requests to SPDK NVMe/TCP target
2. Target reads data from NVMe SSDs
3. Target uses Intel DSA/ISA-L to calculate CRC32C data digest
4. Target sends I/O data + CRC32C data digest to FIO

How Do You Know if Intel DSA is Right for You?

- **Look at where Intel DSA can be utilized for your workloads:**
  - Performance acceleration - for Networking, Storage, Memory-management & AI/HPC/Applications.
  - Scalability - Build product with performance scaling need of 1 to 4 Intel DSA.
  - Efficiency – Significant core saving translating to Significant Performance/Watt improvements
- **Specific Questions**
  1. Where do you need efficient "data movement" for Storage, Networking and other applications?
  2. Where do you need consistent & efficient data movement for Perf/core & Perf/Watt Improvement?
  3. Connect with Intel for engagement and design-in and potential applications enablement support

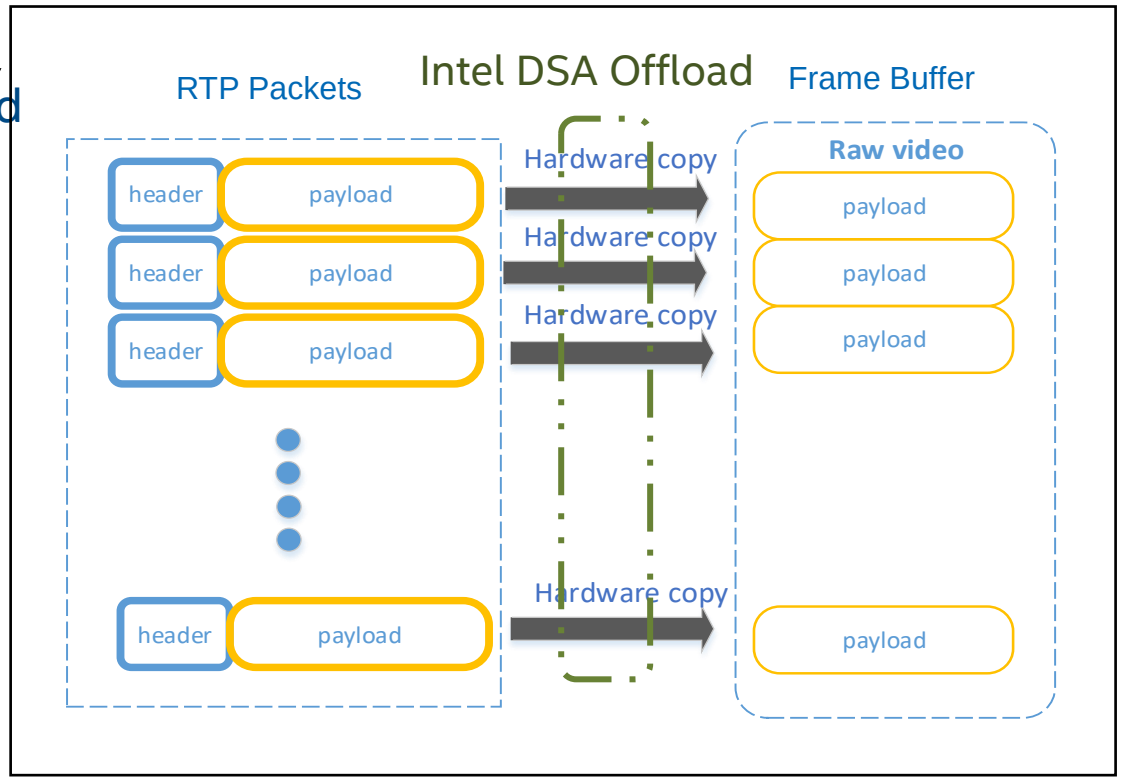
More workload examples

Intel DSA Acceleration for Media Transportation

Workload Description : Media Transport

Use case -Smart Stadium

Receiver End

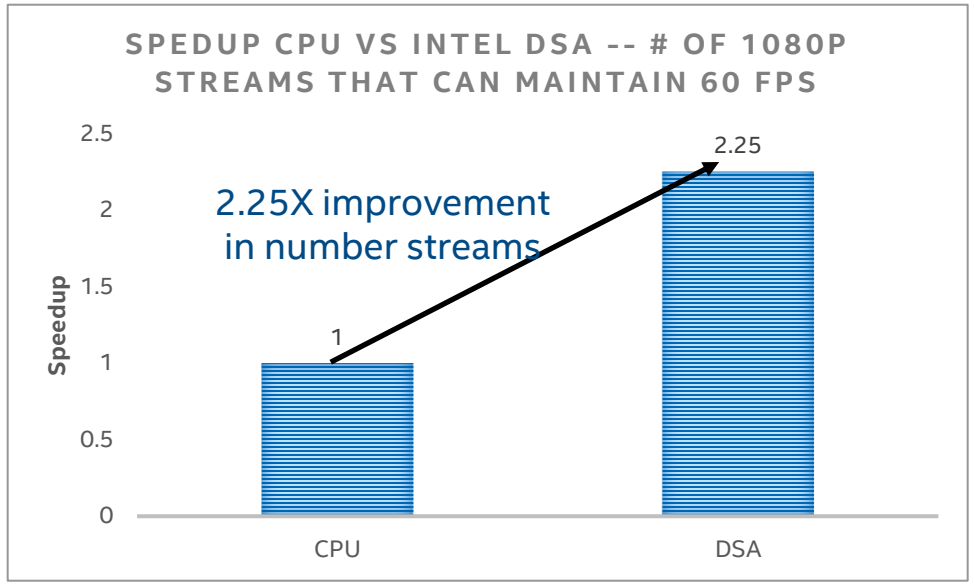


- SMPTE ST 2110 family of standards from SMPTE (Society of Motion Picture and Television Engineers) is a digital video transmission standards over IP networks.
- Intel® Media Transport library provides functions for video transmission & receiver for ST 2110
- After a video packet is received from NIC, RTP packet is processed, and the payload is copied to the raw video frame buffer
- Intel DSA engine accelerates this memory copy

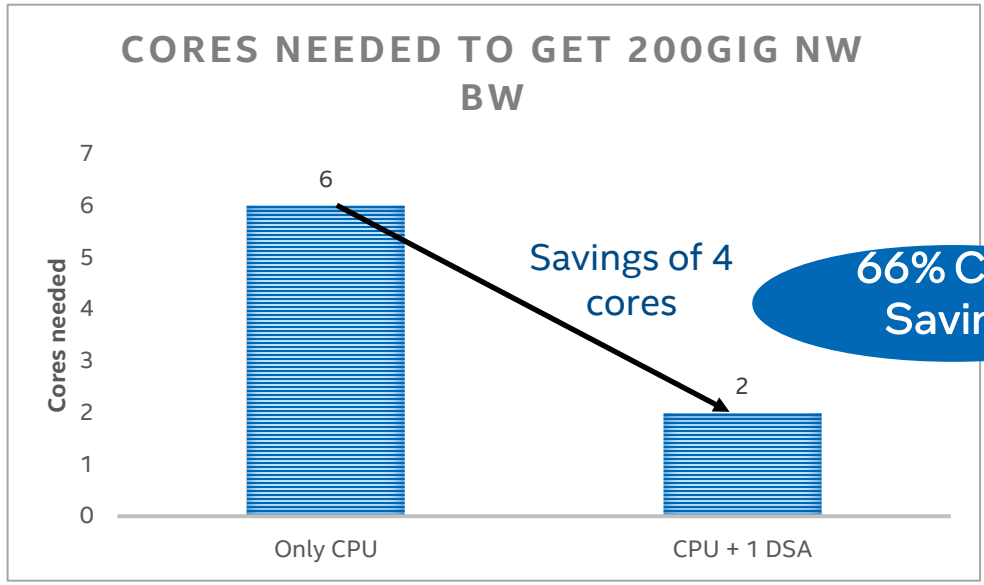
Intel DSA Acceleration for Media Transport Library

Context

- Significant amount of CPU cycles are spent to copy the payload from RTP packets to video frame buffer.
- Offloading this to Intel DSA, increases the number of video streams @1080p,60FPS by 2.25X at iso core.
- At the same time, the maximum network bandwidth can be achieved with fewer cores when offloaded to Intel DSA.
- Use cases: media streaming from a stadium or a large gathering.



ISO Core(2 cores)



ISO Network(200 Gbps -54 streams 1080p@60 fps)

See [N204] at https://edc.intel.com/content/www/us/en/products/performance/benchmarks/4th-generation-intel-xeon-scalable-processors/
Intel DSA config: 1 instance, 4 engine per instance, 8 work queue per instance and ATS disabled per work queue
TX_Flush interval : 50 usec , Batch size: none , Size of copy 5MB. Your results may vary.
Per disclosure reference: https://github.com/OpenVisualCloud/Media-Transport-Library

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more on the Performance Index site.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details.

No product or component can be absolutely secure.

Your costs and results may vary.

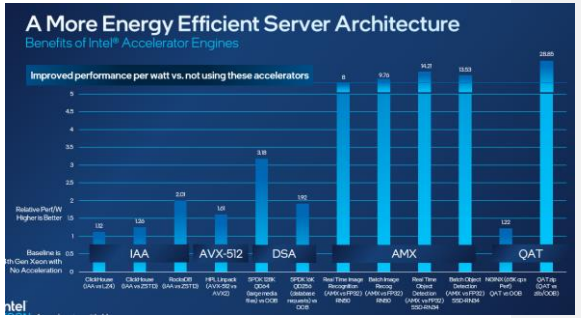
Intel technologies may require enabled hardware, software or service activation.

Availability of accelerators varies by SKU. Visit https://ark.intel.com/content/www/us/en/ark/products/series/228622/4th-generation-intel-xeon-scalable-processors.html

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.


intel

Resources and Configurations



A More Energy Efficient Server Architecture

Up to 1.12x and 1.26x higher performance/W using 4th Gen Xeon Scalable w/Intel Analytics Accelerator vs LZ4 and Zstd on ClickHouse
1-node, 2x pre-production 4th Gen Intel Xeon Scalable processor (60 cores) with integrated Intel In-Memory Analytics Accelerator (Intel IAA), Number of IAA device utilized=8(2 sockets active), on pre-production Intel platform and software, HT On, Turbo On, SNC off, Total Memory 1024GB (16x64GB DDR5 4800), microcode 0x2b0000a1, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.18.12-051812-generic, QPL v0.1.21, accel-config-v3.4.6.4, gcc 11.2, Clickhouse 21.12, Star Schema Benchmark, tested by Intel November 2022.

Up to 2.01x higher performance/W using 4th Gen Xeon Scalable w/Intel Analytics Accelerator vs Zstd on RocksDB
1-node, 2x pre-production 4th Gen Intel Xeon Scalable Processor (60 cores) with integrated Intel In-Memory Analytics Accelerator (Intel IAA), on pre-production Intel platform and software, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 4800), microcode 0x2b0000a1, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.18.12-051812-generic, QPL v0.2.1,accel-config-v3.4.6.4, ZSTD v1.5.2, RocksDB v6.4.6 (db_bench), tested by Intel November 2022.

Up to 1.61 higher performance/W using 4th Gen Xeon Scalable w/AVX-512 vs AVX2 on Linpack
1-node, 2x pre-production 4th Gen Intel® Xeon® Scalable processor (60 core), on pre-production Supermicro SYS-221H-TNR and software with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000c0, HT On, Turbo On, SNC 4, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, One API BaseKit 2022.2.0.262, One API HPC 2022.2.0.191, Linpack ver 2.3, tested by Intel November 2022.

Up to 3.18x and 1.92x higher performance/W using 4th Gen Xeon Scalable w/Data Streaming Accelerator vs out-of-box OS software on SPDK NVMe TCP
1-node, 2x pre-production 4th Gen Intel Xeon Scalable processor (60 core) with integrated Intel Data Streaming Accelerator (Intel DSA), DSA device utilized=1(1 active socket), on pre-production Intel platform and software with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000a1, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x 1.92TB Intel® SSDSC2KG01, 4x 1.92TB Samsung PM1733, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 2x100GbE, FIO v3.30, SPDK 22.05, tested by Intel November 2022.

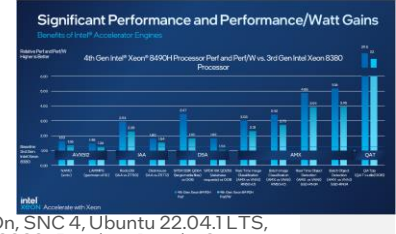
Up to 8x and 9.76x higher performance/W using 4th Gen Xeon Scalable w/Advanced Matrix Extensions using AMX vs VNNI instructions on ResNet50 Image Processing
1-node, 2x pre-production 4th Gen Intel® Xeon® Scalable processor (60 core) with Intel® Advanced Matrix Extensions (Intel AMX), on pre-production Supermicro SYS-221H-TNR with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000c0, HT On, Turbo On, SNC Off, CentOS Stream 8, 5.19.16-301.fc37.x86_64, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Intel TF 2.10, AI Model=Resnet 50 v1_5, best scores achieved: BS1 FP32 8 cores/instance (max. 15ms SLA), BS1 INT8 2 cores/instance (max. 15ms SLA), BS1 AMX 1 core/instance (max. 15ms SLA), BS16 FP32 5 cores/instance, BS16 INT8 5 cores/instance, BS16 AMX 5 cores/instance, using physical cores, tested by Intel November 2022.

Up to 14.21x and 13.53x higher performance/W using 4th Gen Intel Xeon Scalable w/Advanced Matrix Extensions using AMX vs VNNI instructions on SSD-ResNet34 on Object Detection
1-node, 2x pre-production 4th Gen Intel® Xeon® Scalable processor (60 core) with Intel® Advanced Matrix Extensions (Intel AMX), Intel platform with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000a1, HT On, Turbo On, SNC Off, CentOS Stream 8, 5.19.16-301.fc37.x86_64, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Intel TF 2.10, AI Model=SSD-ResNet34, best scores achieved: BS1 FP32 60 cores/instance (max. 100ms SLA), BS1 INT8 4 cores/instance (max. 100ms SLA), BS1 AMX 4 core/instance (max. 100ms SLA), BS8 FP32 8 cores/instance, BS2 INT8 1 cores/instance, BS2 AMX 1 cores/instance, using physical cores, tested by Intel November 2022.

Up to 1.22x higher performance/W using 4th Gen Intel Xeon Scalable w/QuickAssist Accelerator vs out-of-box software on NGINX TLS Handshake.
QAT Accelerator: 1-node, 2x pre-production 4th Gen Intel Xeon Scalable Processor (60 cores) with integrated Intel QuickAssist Accelerator (Intel QAT), Number of QAT device utilized=4(1 socket active), on pre-production Intel platform and software with DDR5 memory total 1024GB (16x64 GB), microcode 0x2b0000a1, HT On, Turbo Off, SNC Off, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 1x100GbE, QAT engine v0.6.14, QAT v20.l.0.9.1, NGINX 1.20.1, OpenSSL 1.1.1l, IPP crypto v2021_5, IPSec v1.1 , TLS 1.3 AES_128_GCM_SHA256, ECDHE-X25519-RSA2K, 65K CPS target SLA, tested by Intel November 2022. Out of box configuration: 1-node, 2x pre-production 4th Gen Intel Xeon Scalable Processor (60 cores) with integrated Intel QuickAssist Accelerator (Intel QAT), Number of QAT device utilized=0, on pre-production Intel platform and software with DDR5 memory total 1024GB (16x64 GB), microcode 0x2b0000a1, HT On, Turbo Off, SNC Off, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 1x100GbE, NGINX 1.20.1, OpenSSL 1.1.1l, TLS 1.3 AES_128_GCM_SHA256, ECDHE-X25519-RSA2K, 65K CPS target SLA, tested by Intel November 2022.

Up to 28.85x higher performance/W using 4th Gen Intel Xeon Scalable w/QuickAssist Accelerator vs out-of-box zlib on QATzip compression
1-node, 2x pre-production 4th Gen Intel® Xeon Scalable Processor (60 core) with integrated Intel QuickAssist Accelerator (Intel QAT), QAT device utilized=8(2 sockets active), on pre-production Intel platform and software with DDR5 memory Total 1024GB (16x64 GB), microcode 0x2b0000a1, HT On, Turbo Off, SNC Off, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, QAT v20.l.0.9.1 , QATzip v1.0.9, tested by Intel November 2022.

Resources and Configurations



Significant Performance and Performance /Watt Gains:

NAMD

New Configuration: 1-node, 2x pre-production 4th Gen Intel® Xeon® Scalable processor (60 core), on pre-production Supermicro SYS-221H-TNR and software with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000c0, HT On, Turbo On, SNC 4, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, One API BaseKit 2022.2.0.262, One API HPC 2022.2.0.191, NAMD release-2-15-alpha-1, charm v6.10.2, tcl core-8-5-branch, benchmark from NAMD v2.13, tested by Intel November 2022. Baseline: 1-node, 2x production 3rd Gen Intel Xeon Scalable Processors(40 cores) on Supermicro SYS-220U-TNR , DDR4 memory total 1024GB (16x64 GB), microcode 0xd000375, HT On, Turbo On, SNC On, HT On, Turbo On, SNC On, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, One API BaseKit 2022.2.0.262, One API HPC 2022.2.0.191, NAMD release-2-15-alpha-1, charm v6.10.2, tcl core-8-5-branch, benchmark from NAMD v2.13, tested by Intel November 2022.

LAMMPS

New Configuration: 1-node, 2x pre-production 4th Gen Intel® Xeon® Scalable processor (60 core), on pre-production Supermicro SYS-221H-TNR and software with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000c0, HT On, Turbo On, SNC 4, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, One API BaseKit 2022.2.0.262, One API HPC 2022.2.0.191, LAMMPS update 2 for Stable release 23 June 2022, tested by Intel November 2022. Baseline: 1-node, 2x production 3rd Gen Intel Xeon Scalable Processors(40 cores) on Supermicro SYS-220U-TNR , DDR4 memory total 1024GB (16x64 GB), microcode 0xd000375, HT On, Turbo On, SNC On, HT On, Turbo On, SNC On, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, One API BaseKit 2022.2.0.262, One API HPC 2022.2.0.191, LAMMPS update 2 for Stable release 23 June 2022, tested by Intel November 2022.

RocksDB

New Configuration: 1-node, 2x pre-production 4th Gen Intel Xeon Scalable Processor (60 cores) with integrated Intel In-Memory Analytics Accelerator (Intel IAA), on pre-production Intel platform and software, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 4800), microcode 0x2b0000a1, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.18.12-051812-generic, QPL v0.2.1,accel-config-v3.4.6.4, ZSTD v1.5.2, RocksDB v6.4.6 (db_bench), tested by Intel November 2022. Baseline: 1-node, 2x production 3rd Gen Intel Xeon Scalable Processors(40 cores) on Supermicro SYS-220U-TNR , HT On, Turbo On, SNC Off, Total Memory 1024GB (16x64GB DDR4 3200), microcode 0xd000375, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.18.12-051812-generic, ZSTD v1.5.2, RocksDB v6.4.6 (db_bench), tested by Intel November 2022.

ClickHouse

New Configuration: 1-node, 2x pre-production 4th Gen Intel Xeon Scalable processor (60 cores) with integrated Intel In-Memory Analytics Accelerator (Intel IAA), Number of IAA device utilized=8(2 sockets active), on pre-production Intel platform and software, HT On, Turbo On, SNC off, Total Memory 1024GB (16x64GB DDR5 4800), microcode 0x2b0000a1, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.18.12-051812-generic, QPL v0.1.21, accel-config-v3.4.6.4, gcc 11.2, Clickhouse 21.12, Star Schema Benchmark, tested by Intel November 2022. Baseline: 1-node, 2x production 3rd Gen Intel Xeon Scalable Processors(40 cores) on Supermicro SYS-220U-TNR , HT On, Turbo On, SNC Off, Total Memory 1024GB (16x64GB DDR4 3200), microcode 0xd000375, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.18.12-051812-generic, gcc 11.2, Clickhouse 21.12, Star Schema Benchmark, tested by Intel November 2022.

SPDK

New Configuration: 1-node, 2x pre-production 4th Gen Intel Xeon Scalable processor (60 core) with integrated Intel Data Streaming Accelerator (Intel DSA), DSA device utilized=1(1 active socket), on pre-production Intel platform and software with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000a1, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x 1.92TB Intel® SSDSC2KG01, 4x 1.92TB Samsung PM1733, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 2x100GbE, FIO v3.30, SPDK 22.05, tested by Intel November 2022. Baseline: 1-node, 2x production 3rd Gen Intel Xeon Scalable Processors(40 cores) on Supermicro SYS-220U-TNR , DDR4 memory total 1024GB (16x64 GB), HT On, Turbo On, SNC Off, microcode 0xd000375, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x 1.92TB Intel SSDSC2KG01, 4x 1.92TB Samsung PM1733, 1x Intel Ethernet Network Adapter E810-2CQDA2, 2x100GbE, FIO v3.30, SPDK 22.05, tested by Intel November 2022.

ResNet-50

New Configuration: 1-node, 2x pre-production 4th Gen Intel® Xeon® Scalable processor (60 core) with Intel® Advanced Matrix Extensions (Intel AMX), on pre-production Supermicro SYS-221H-TNR with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000c0, HT On, Turbo On, SNC Off, CentOS Stream 8, 5.19.16-301.fc37.x86_64, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Intel TF 2.10, AI Model=Resnet 50 v1_5, best scores achieved: BS1 FP32 8 cores/instance (max. 15ms SLA), BS1 INT8 2 cores/instance (max. 15ms SLA), BS1 AMX 1 core/instance (max. 15ms SLA), BS16 FP32 5 cores/instance, BS16 INT8 5 cores/instance, BS16 AMX 5 cores/instance, using physical cores, tested by Intel November 2022. Baseline: 1-node, 2x production 3rd Gen Intel Xeon Scalable Processors(40 cores) on Supermicro SYS-220U-TNR , DDR4 memory total 1024GB (16x64 GB), microcode 0xd000375, HT On, Turbo On, SNC Off, CentOS Stream 8, 5.19.16-301.fc37.x86_64, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Intel TF 2.10, AI Model=Resnet 50 v1_5, best scores achieved: BS1 FP32 8 cores/instance (max. 15ms SLA), BS1 INT8 2 cores/instance (max. 15ms SLA), BS16 FP32 5 cores/instance, BS16 INT8 5 cores/instance, using physical cores, tested by Intel November 2022.

SSD-ResNet-34

New Configuration: 1-node, 2x pre-production 4th Gen Intel® Xeon® Scalable processor (60 core) with Intel® Advanced Matrix Extensions (Intel AMX), Intel platform with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000a1, HT On, Turbo On, SNC Off, CentOS Stream 8, 5.19.16-301.fc37.x86_64, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Intel TF 2.10, AI Model=SSD-ResNet34, best scores achieved: BS1 FP32 60 cores/instance (max. 100ms SLA), BS1 INT8 4 cores/instance (max. 100ms SLA), BS1 AMX 4 core/instance (max. 100ms SLA), BS8 FP32 8 cores/instance, BS2 INT8 1 cores/instance, BS2 AMX 1 cores/instance, using physical cores, tested by Intel November 2022. Baseline: 1-node, 2x production 3rd Gen Intel Xeon Scalable Processors(40 cores) on Supermicro SYS-220U-TNR , DDR4 memory total 1024GB (16x64 GB), microcode 0xd000375, HT On, Turbo On, SNC Off, CentOS Stream 8, 5.19.16-301.fc37.x86_64, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Intel TF 2.10, AI Model= SSD-ResNet34, best scores achieved: BS1 FP32 40 cores/instance (max. 100ms SLA), BS1 INT8 10 cores/instance (max. 100ms SLA), BS16 FP32 4 cores/instance, BS16 INT8 1 cores/instance, using physical cores, tested by Intel November 2022.

QAT.zip

New Configuration: 1-node, 2x pre-production 4th Gen Intel® Xeon Scalable Processor (60 core) with integrated Intel QuickAssist Accelerator (Intel QAT), QAT device utilized=8(2 sockets active), on pre-production Intel platform and software with DDR5 memory Total 1024GB (16x64 GB), microcode 0x2b0000a1, HT On, Turbo Off, SNC Off, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, QAT v20.l.0.9.1 , QATzip v1.0.9, tested by Intel November 2022. Baseline: 1-node, 2x production 3rd Gen Intel Xeon Scalable Processors(40 cores) on Supermicro SYS-220U-TNR , DDR4 memory total 1024GB (16x64 GB), microcode 0xd000375, HT On, Turbo Off, SNC Off, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, QAT v20.l.0.9.1 , QATzip v1.0.9, tested by Intel November 2022.