# 4th Gen Intel® Xeon® Processors On-Die Accelerators: Intel® Dynamic Load Balancer (Intel® DLB)

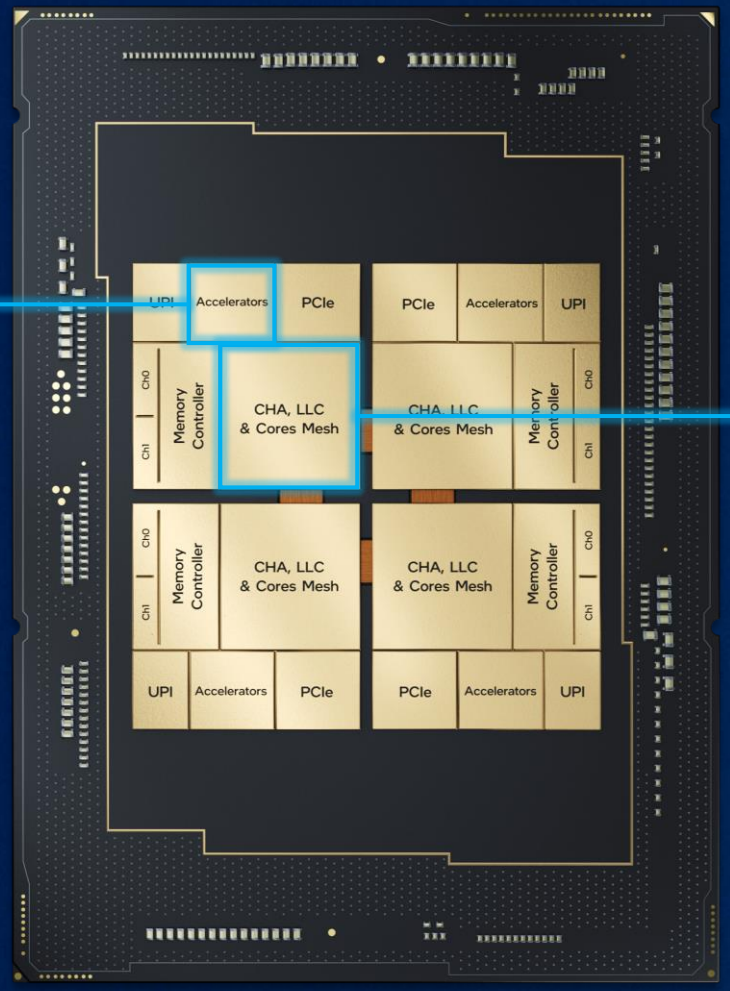
# Contents

# Accelerators on 4th Gen Intel Xeon

# 4th Gen Intel® Xeon® Processor Accelerator Architecture

- Intel DLB, Intel® Data Streaming Accelerator (Intel® DSA), Intel® QuickAssist Technology (Intel®QAT), and Intel® In-Memory Analytics Accelerator (Intel® IAA) sit on a "Data Accelerator Complex" (DAC) outside the CPUs CHA, LLC, and core mesh.

- Intel® Advanced Matrix Extensions (Intel® AMX) physically sits on each core.



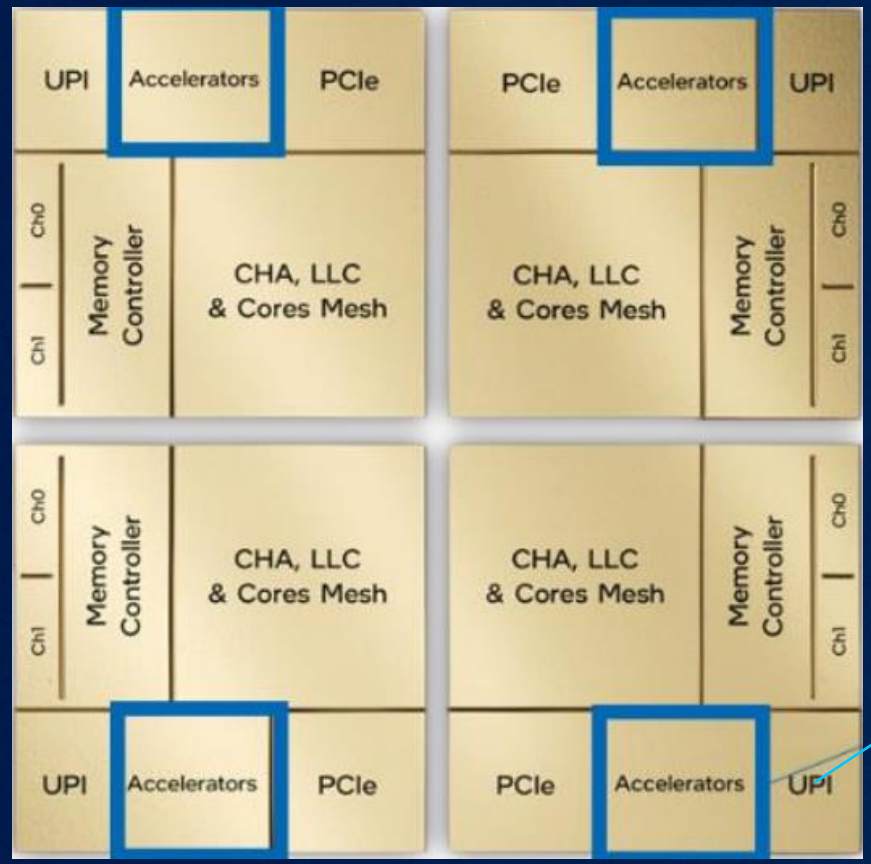

*INTEL DLB*
*INTEL DSA*
*INTEL IAA*
*INTEL QAT*

*INTEL AMX*

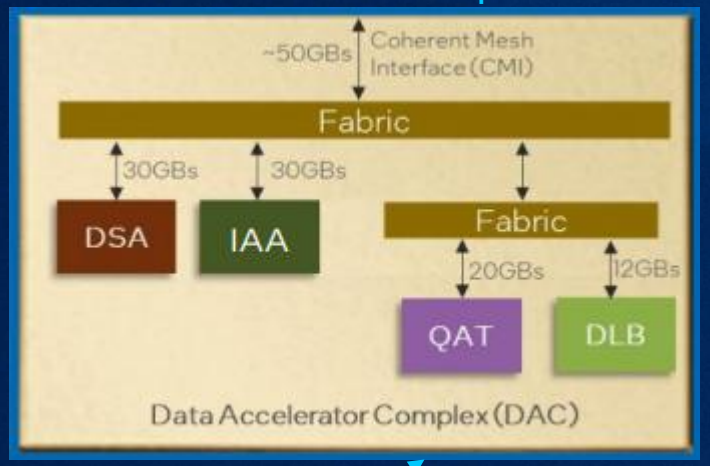
# 4th Gen Intel® Xeon® Processor Data Accelerator Complex (XCC)

## Data Accelerator Complex: Houses Intel DSA, Intel IAA, Intel QAT, and Intel DLB

- Accelerators are PCIe root-complex integrated end point devices.
- XCC clusters 1x Intel DSA, 1x Intel IAA, 1x Intel QAT and 1x Intel DLB device into a single DAC, and instantiates that DAC 4 times.
- Each accelerator is designed to operate independent of one another.
- CMI is the common path used by accelerators to access upstream CPU cache and memory, and generate PCIe messages.
- Accesses from downstream SW to the accelerators will also use CMI.
- Total CMI BW in each DAC is approximately equivalent to PCIe5 BW of 50GBs
- Each DAC has a separate and dedicated CMI.




Data Accelerator Complex



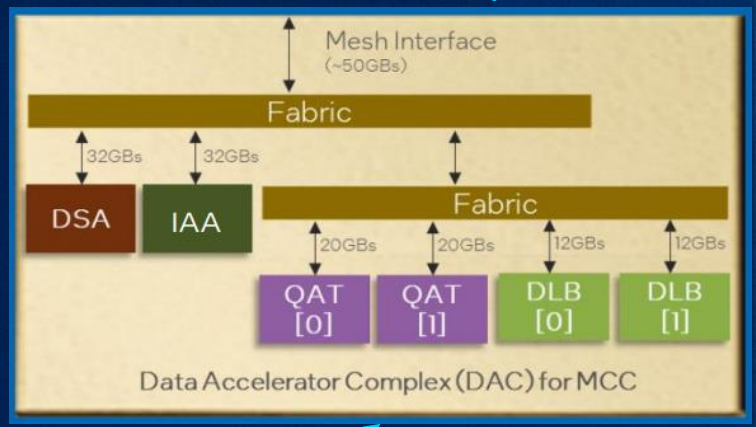

Data Accelerator Complex

# 4th Gen Intel® Xeon® Processor Data Accelerator Complex (MCC)

## Data Accelerator Complex: Houses Intel DSA, Intel IAA, Intel QAT, and Intel DLB

- Accelerators are PCIe root-complex integrated end point devices.
- MCC clusters 1x Intel DSA, 1x Intel IAA, 2x Intel QAT and 2x Intel DLB device into a single DAC, and instantiates that DAC once.
- Each accelerator is designed to operate independent of one another.
- CMI is the common path used by accelerators to access upstream CPU cache and memory, and generate PCIe messages.
- Accesses from downstream SW to the accelerators will also use CMI.
- Total CMI BW in each DAC is approximately equivalent to PCIe5 BW of 50GBs
- Each DAC has a separate and dedicated CMI.


Data Accelerator Complex

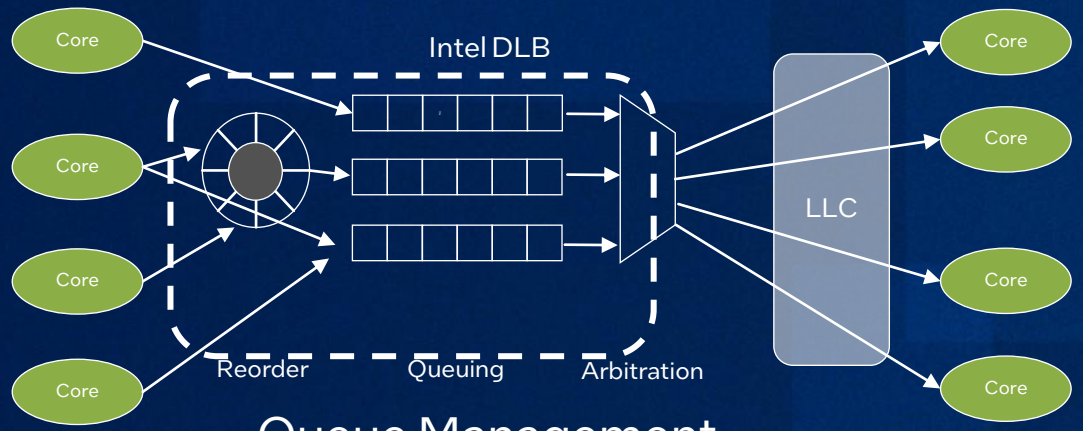



Data Accelerator Complex

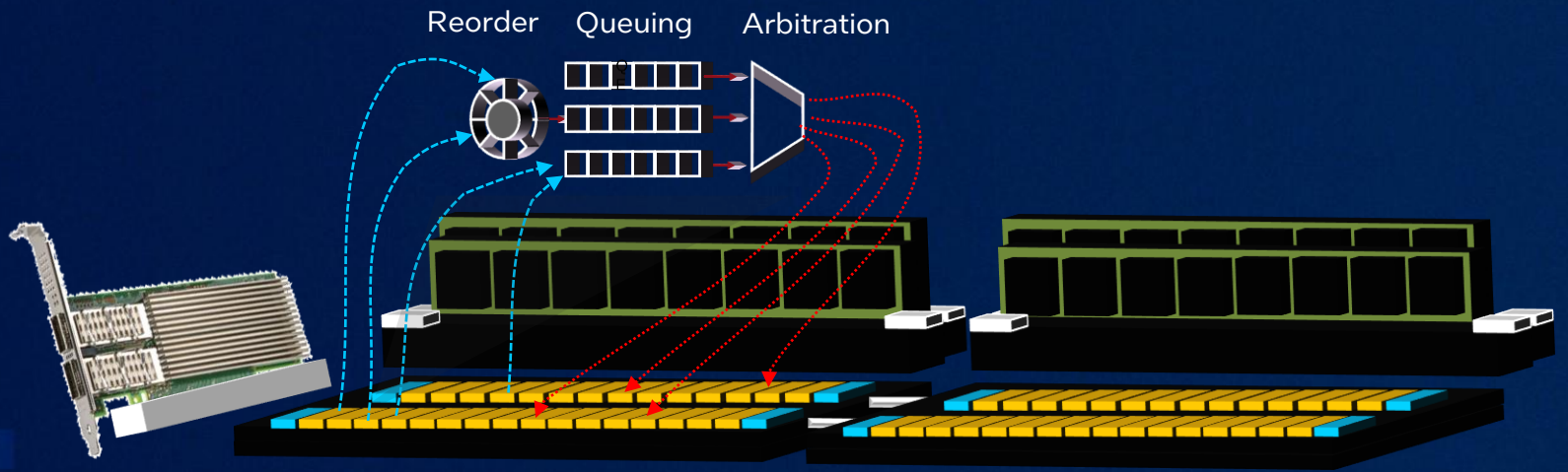
# Intel Dynamic Load Balancer

# Intel Dynamic Load Balancer

- HW-managed system of queues & arbiters linking producers and consumers
- Enables pipelined packet processing models for load balancing & packet queueing
- Enqueue & Dequeue capability from software




Queue Management,
Load Balancing,
Packet Prioritization



# Intel Dynamic Load Balancer Overview

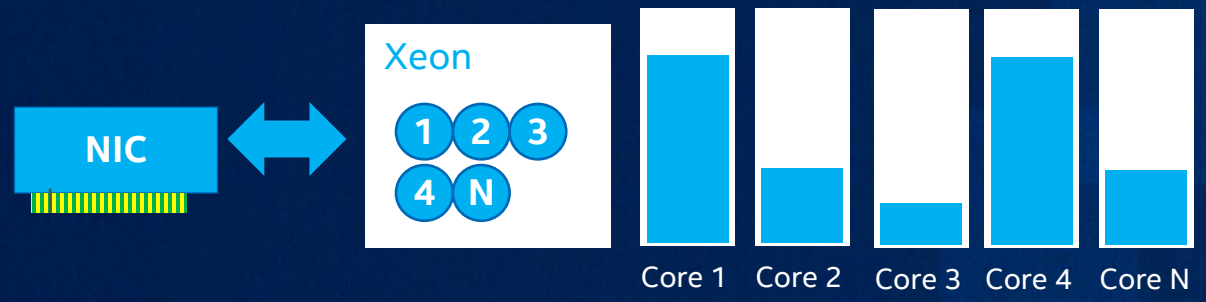
## What is it?

Dedicated HW (on-die PCIe root-complex endpoint devices) that provides intelligent dynamic, balanced distribution of network traffic across CPU cores.

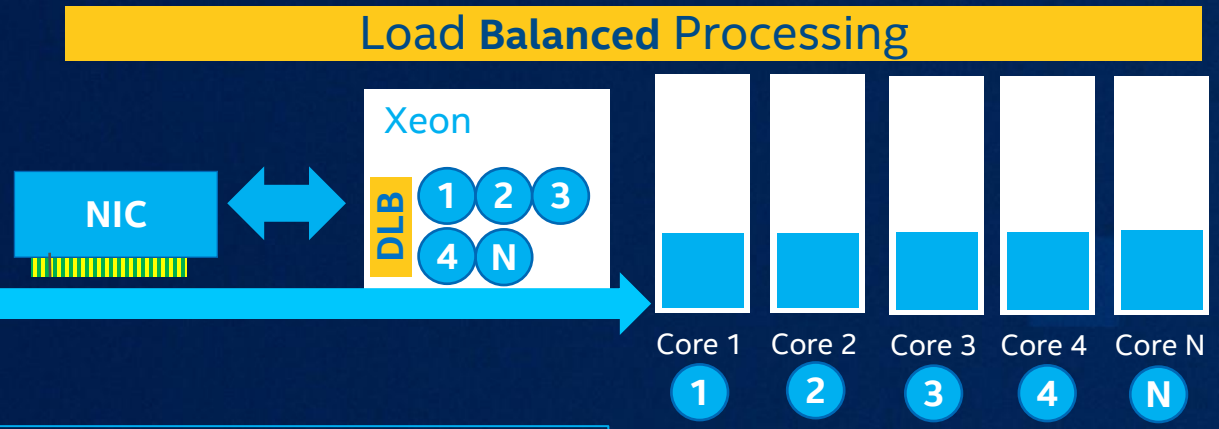
## Why it matters to customers

Packet distribution load balancing for efficient core utilization. Deterministic packet pipelines.

**Without Intel DLB**: CPU Utilization Per Core



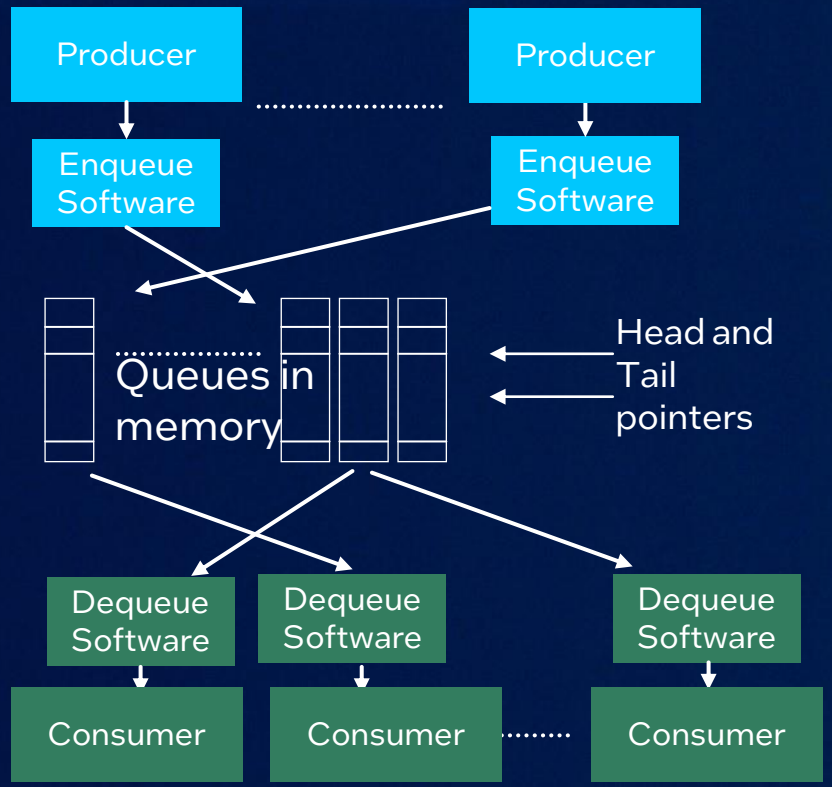

**With Intel DLB**: CPU Utilization Per Core



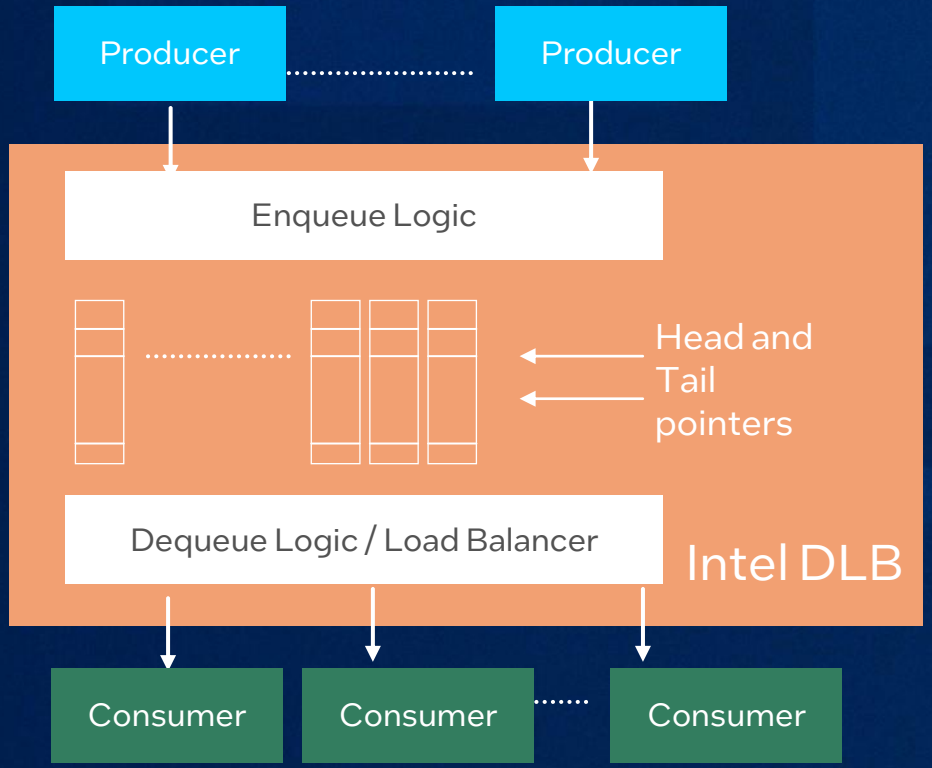

**TARGET WORKLOADS/USAGES**

- Streaming Data Processing
- IPSec Gateway Handling
- VPP Router
- Elephant Flow Handling
- Restore order of network data packets processed simultaneously by CPU cores

# Queue Management : Before & After Intel DLB




With Intel DLB



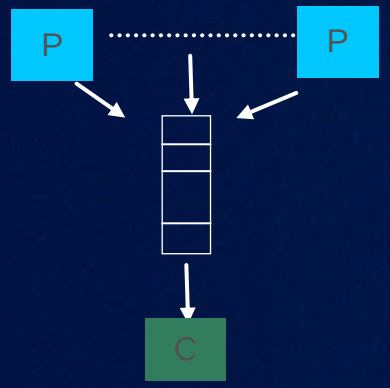

**SW Queue Management (Without Intel DLB)**

- Queues and pointers stored in system memory
- Queues managed by software
- Requires lock to enqueue/dequeue
- Impacted by lock latency, memory latency, cache behaviors, polling multiple queues

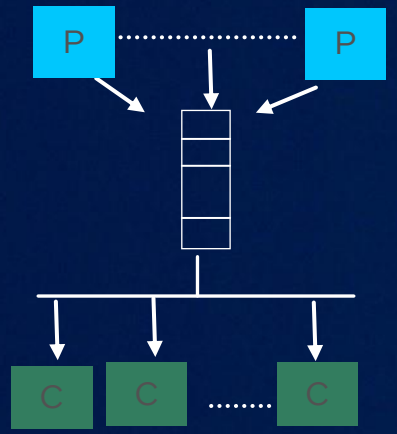
**With Intel DLB:**

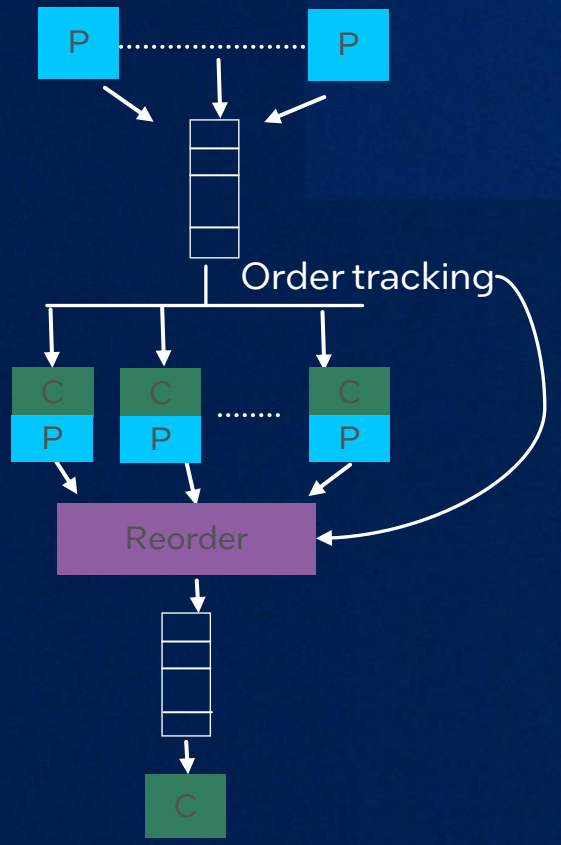
- Queues and pointers stored in DLB local memory
- Queues managed by hardware
- Improvement in number of operations per second
- More complex load balancing algorithms
- Better determinism, cycles freed on cores

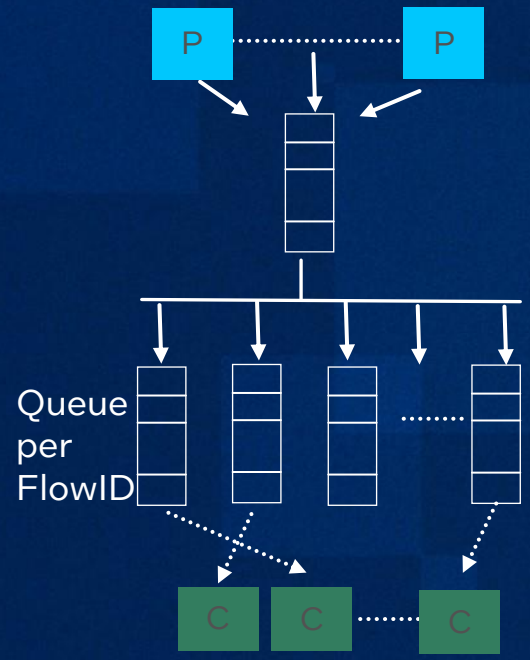
# Intel DLB Traffic Types




**Type: Direct**
m->1 communication




Dynamic affinity
FlowID ->Consumer

Non-Atomic Types

**Type: Unordered**

m->n load balanced across Cs

**Type: Ordered**

m->n load balanced across workers with order recovery

**Type: Atomic**

m->n load balancing with dynamic affinity. Allows Cons to operate on per-flow variables without atomics

Load Balanced Types

# Intel Dynamic Load Balancer Performance Snap Shot

## Function
- Dynamic redistribution of data load across cores when static NIC distribution causes a load-imbalance

## Business Value
- Improves system performance related to handling network data on multi-core Intel® Xeon® Scalable processors
- Improved performance for distributed processing, dynamic load balancing and dynamic network processing reordering

## Software Support
- Intel® Data Mover Library

## Use Cases
- IPSec security gateway, VPP router, UPF, vSwitch, Streaming data processing, Elephant flow handling

**Performance gains vs not using these accelerators**

### Microservices

Up to

**96%**

lower latency at the same throughput with built-in Intel® DLB vs. software for Istio ingress gateway

**Performance gains vs prior generation products**

### Microservices

Up to

**89%**

lower latency and

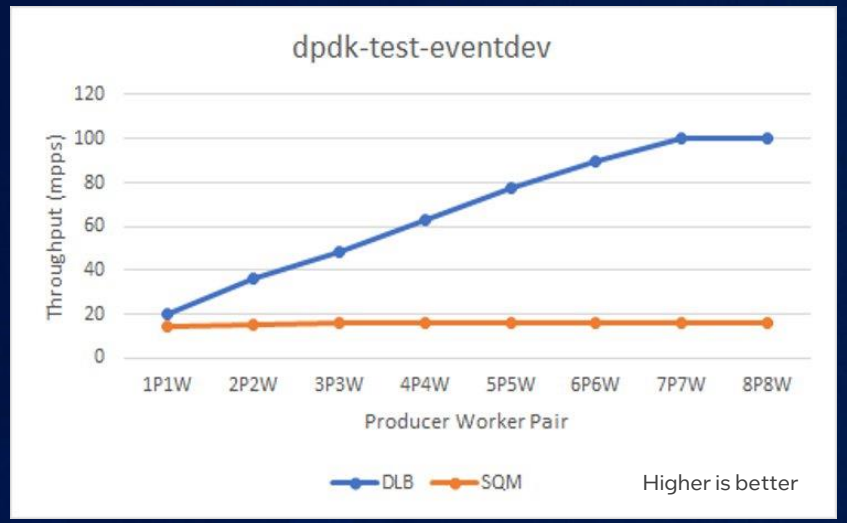
**57%**

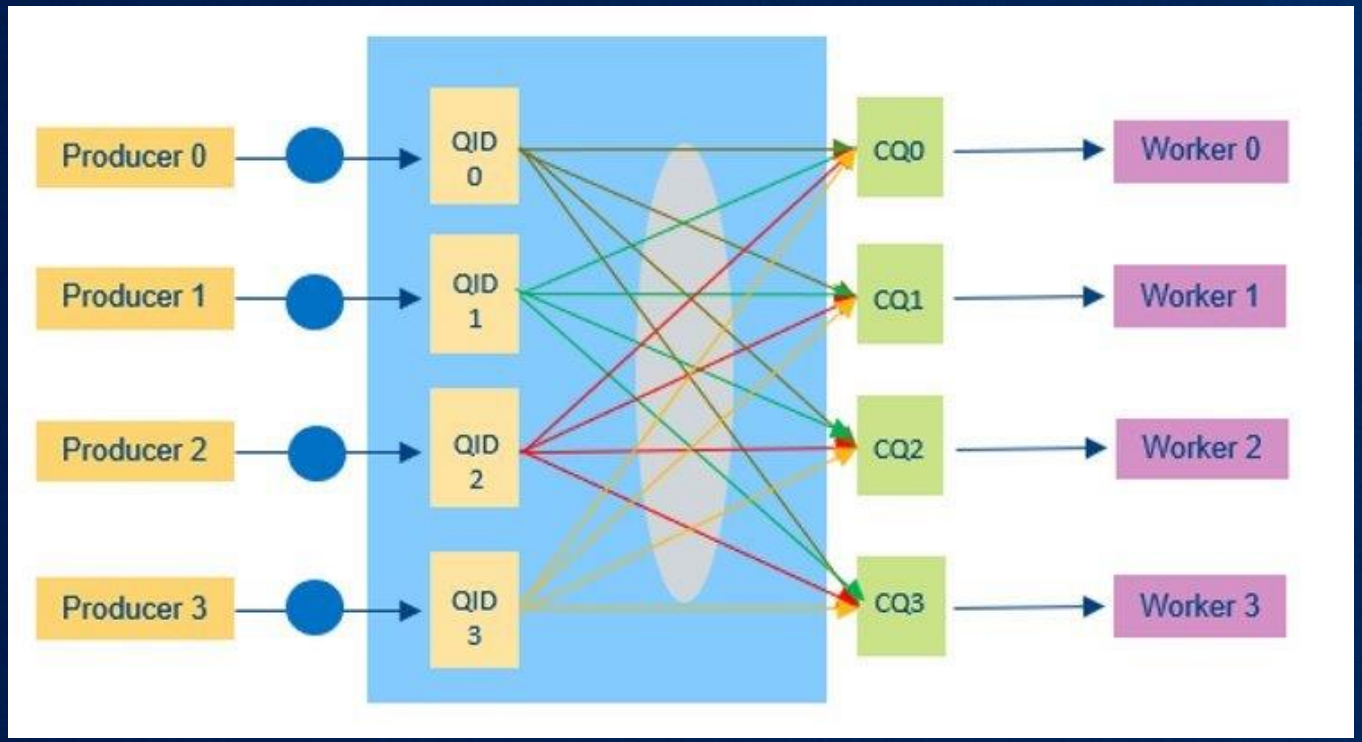
lower CPU utilization at same core count with built-in Intel® DLB vs. prior generation

See [W6] at https://edc.intel.com/content/www/us/en/products/performance/benchmarks/4th-generation-intel-xeon-scalable-processors/

# Intel DLB vs SQM (Software Queue Manager)

| P-W Pairs | SQM (mpps) | DLB (mpps) |
|---|---|---|
| 1P1W | 14.68 | 20.38 |
| 2P2W | 15.07 | 36.22 |
| 3P3W | 15.97 | 48.43 |
| 4P4W | 16.01 | 63.10 |
| 5P5W | 16.02 | 77.17 |
| 6P6W | 15.91 | 89.49 |
| 7P7W | 15.89 | 99.99 |

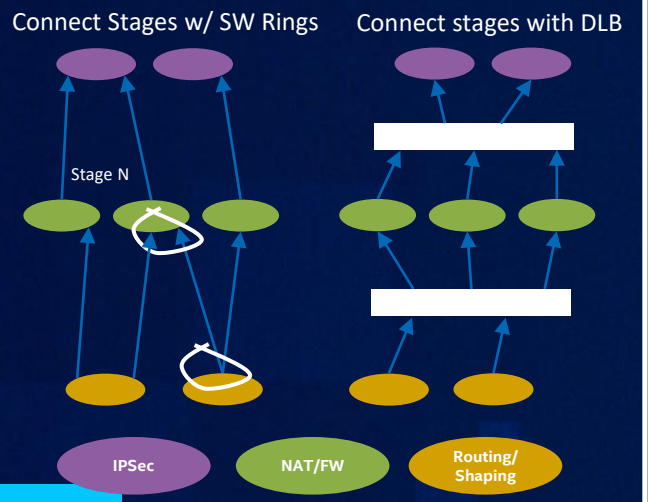


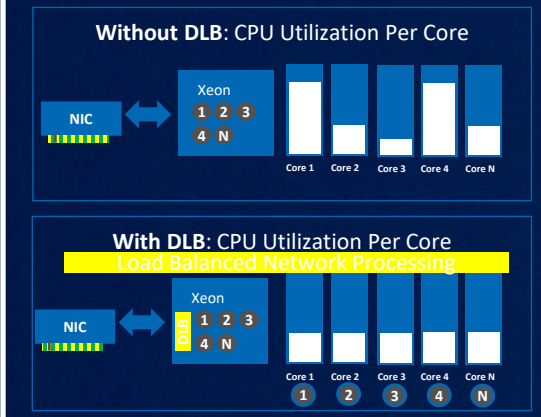



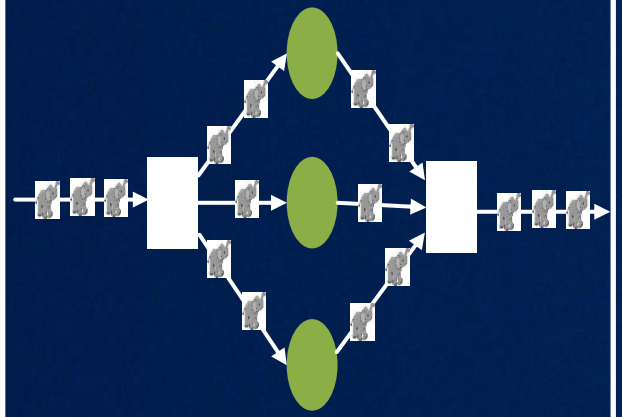

Results may vary. See Config 1

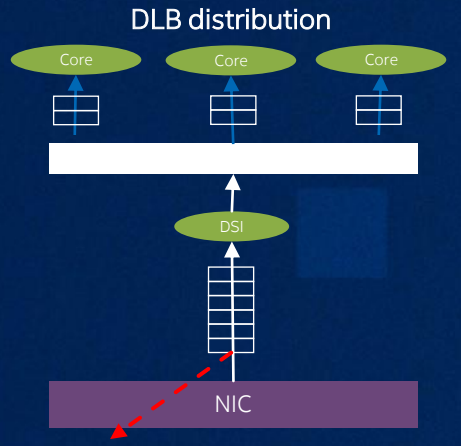
# Intel DLB – Packet Application Value Drivers & Markets

| Multi-function Pipeline (SD-WAN, Edge) | Atomic Load Balancing (Routers, NGFW, LB, Data Movement & Storage) | Elephant Flow Management (VPN/Security GW) | Packet Latency Bounding (Web Servers, Storage, Cloud) |
|---|---|---|---|



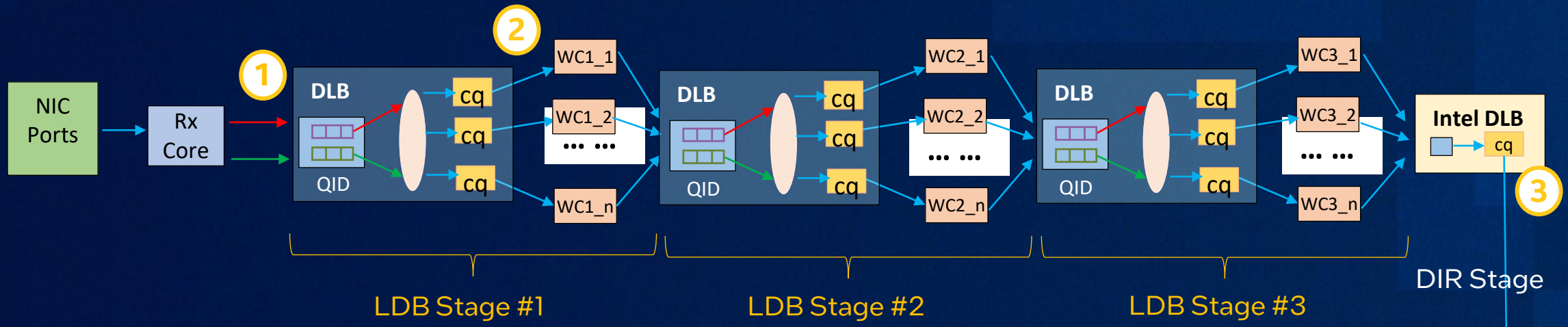






All workers serve to minimize wait time

# Intel DLB Processing Flow Summary



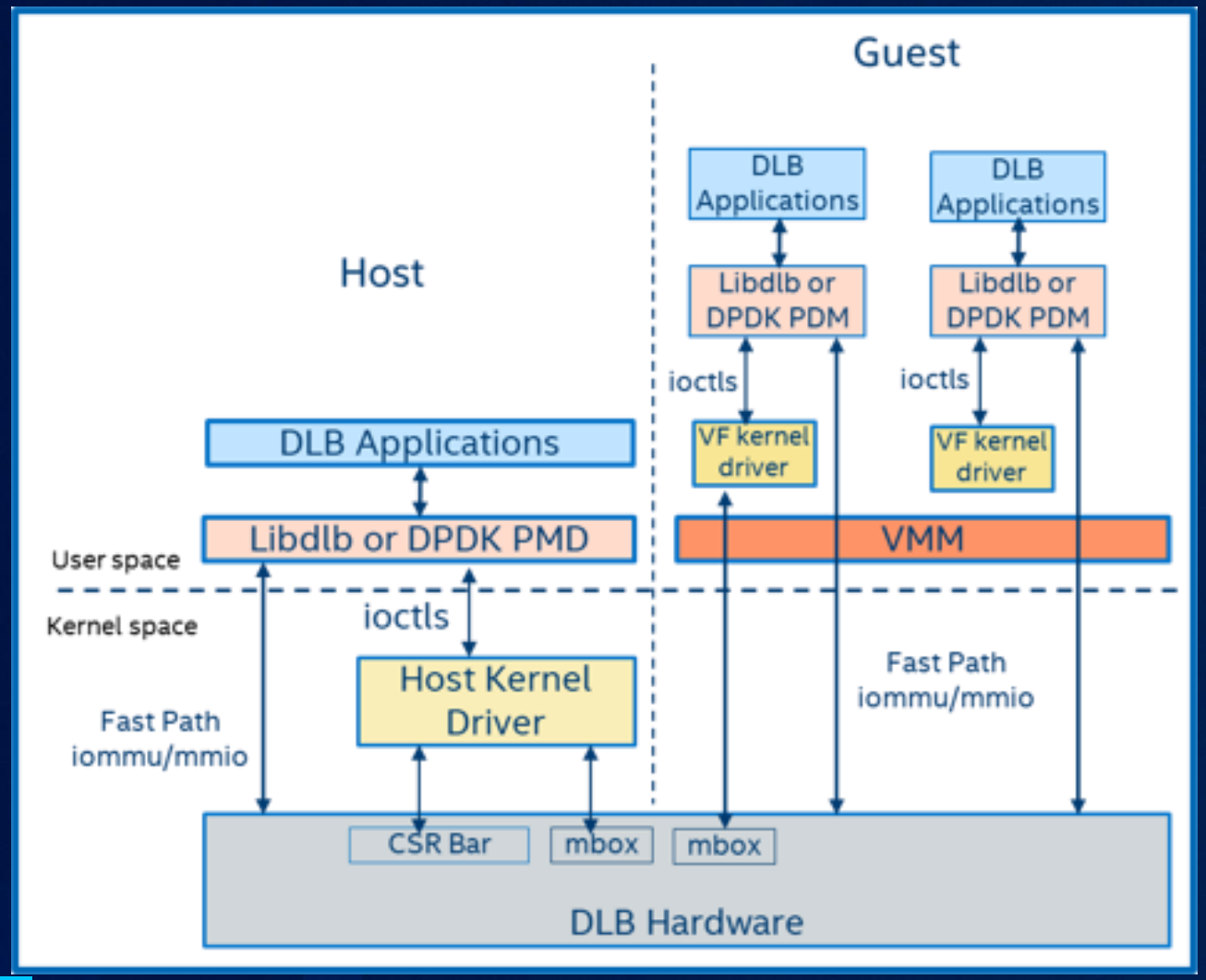

1. Intel DLB receives and stores a list of 8x 8byte pointers (QIDs) to packets in memory, via Rx core
2. Intel DLB uses QIDs to direct associated packets to cores for execution, only targeting cores that, with these packets, will still operate at max performance and efficiency.
3. Tx cores transmit data back out once all packet instructions are complete

**LDB** = Load balanced
**DIR** = Direct linked
**DC** = Distributor core

**WC** = Worker Core
**CQ** = Consumer Queue
**WC2_1** = worker core #1 of stage #2

High priority traffic
Normal traffic
Mixed traffic

# Intel DLB: How Does It Interface With Your WL?




Linux kernel driver:

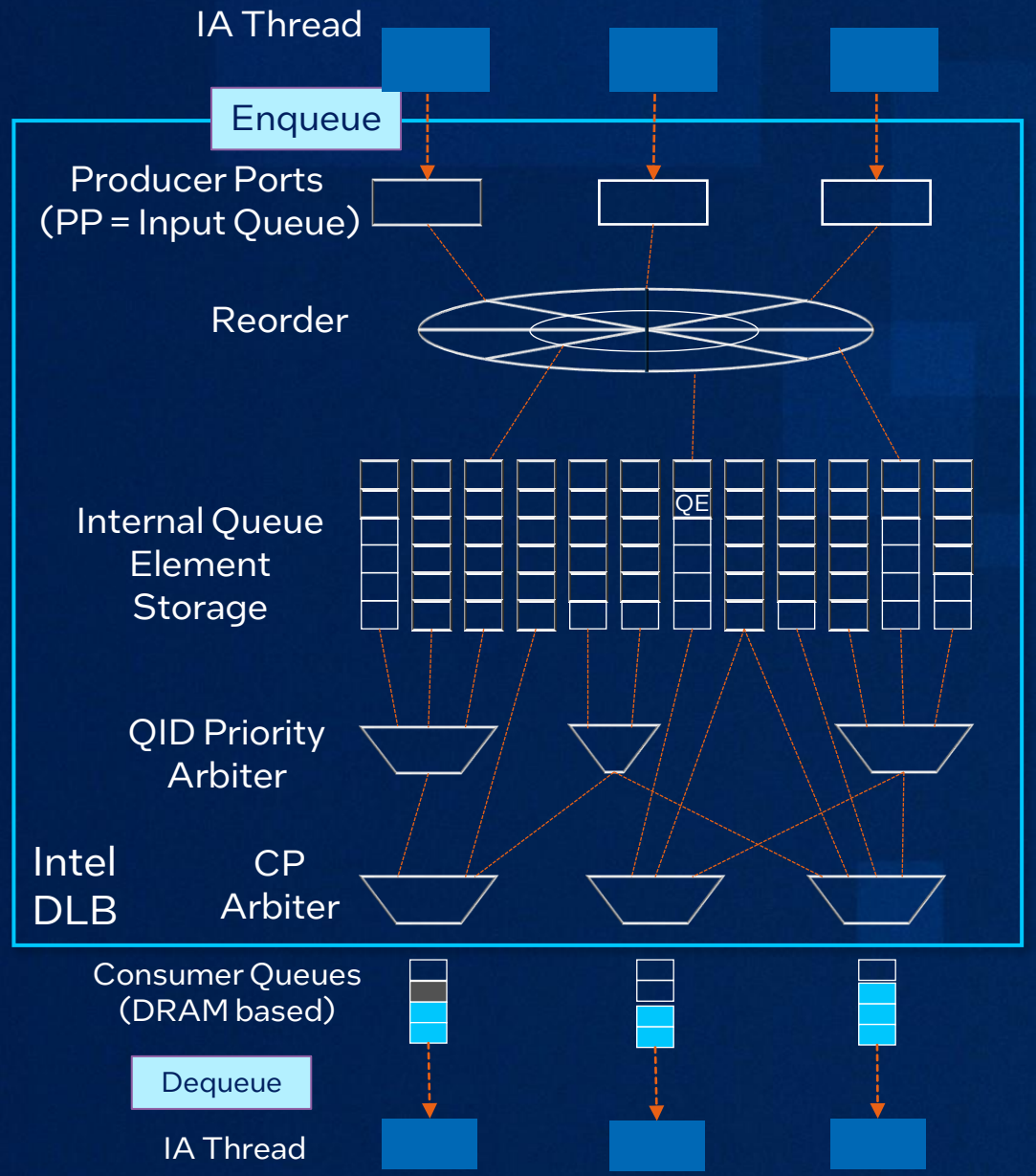
- Provides Intel DLB device configuration, resource management, and Physical function (PF) – to- Virtual Function (VF) communication via ioctls, mmap, and sysfs interfaces.

User-space libraries:

- DPDK Poll Mode Driver (PMD)
  - Bifurcated PMD (shown): configuration requests via the kernel driver
  - PF PMD (not shown): configuration done by PMD
- Libdlb – user space library that runs on the top of dlb kernel driver.

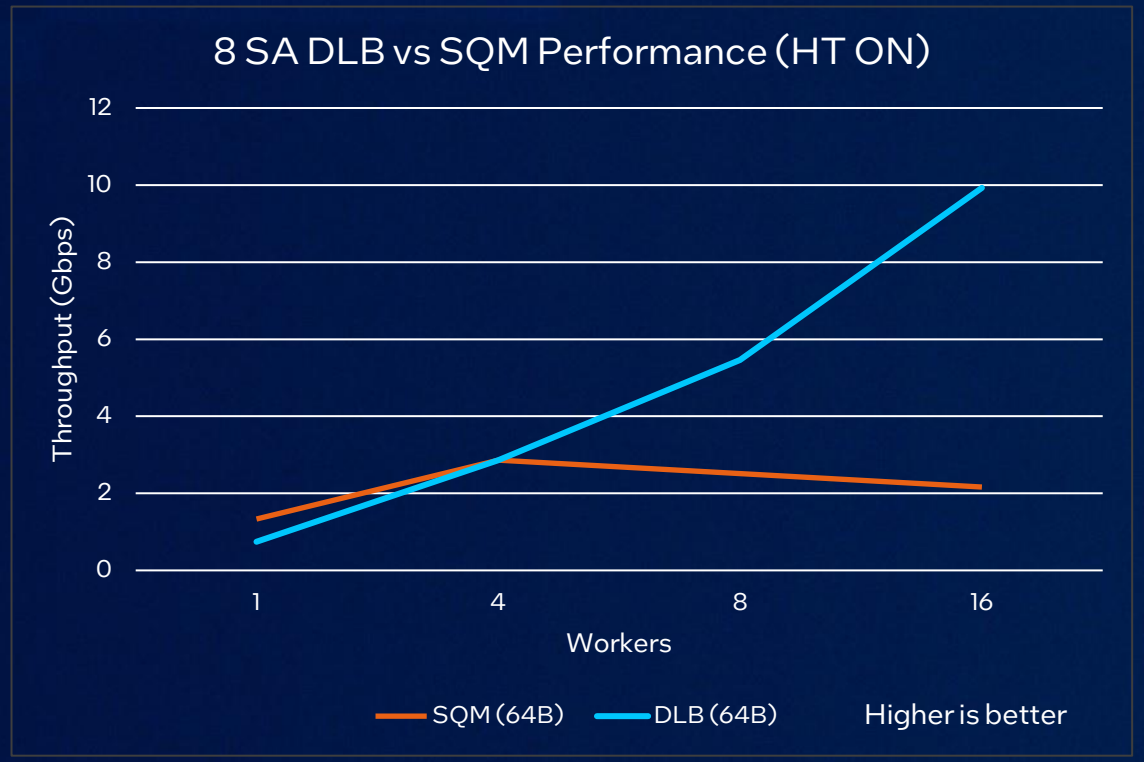
# Intel DLB: Under the Hood

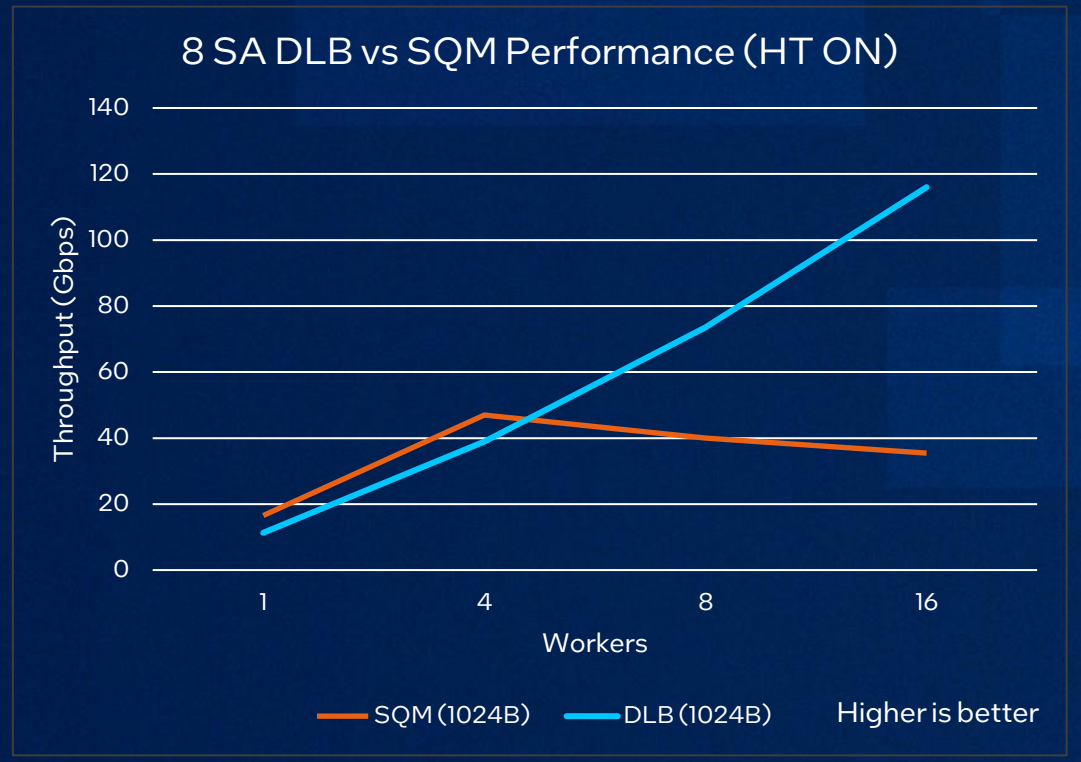
- Thread writes HCW to PP with any transmitted packets and noting completion if required.
- QE, unless reordered, gets inserted into specified QID. Reordered QEs are held until order recovered.
- Eventually bubbles to top of QID, becomes available for scheduling.
- Scheduling done across CQs which have available space.
- For each such CQ, select only from up to 8 mapped QIDs. Both QID and QE priorities apply.
- QE entry written to CQ in memory.



# Applications

# Performance IPsec (Intel DLB Vs SW) - Gbps




## Why it matters to customers

Higher Performance - More efficient load balancing translates into higher performance for packet throughput.

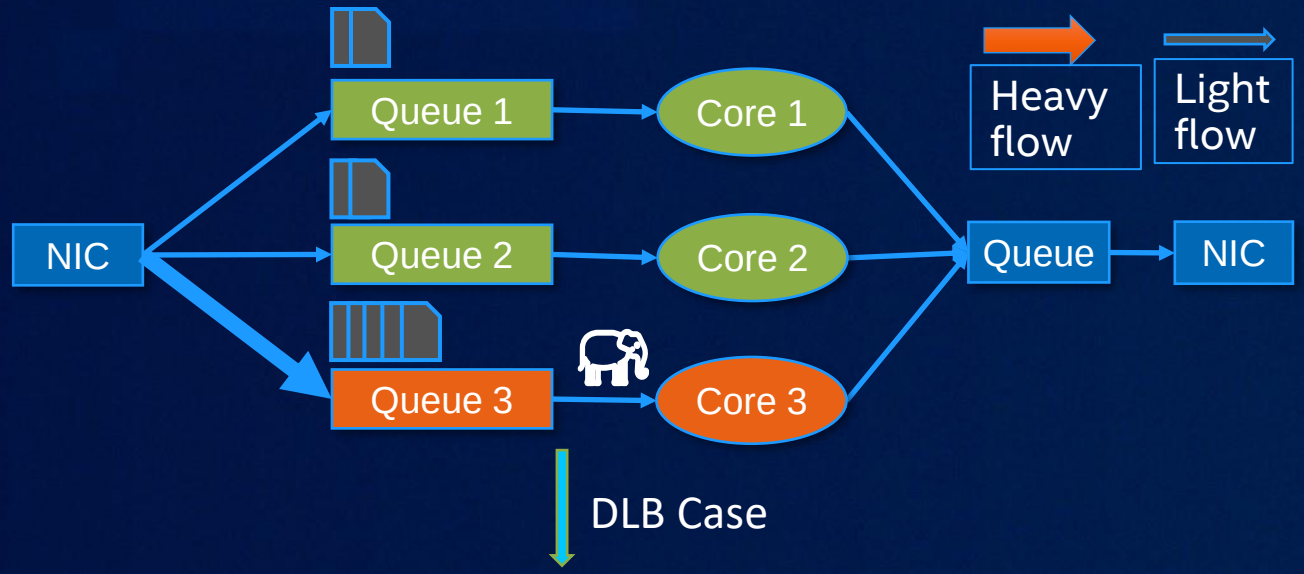
Packet flows across more than a core(An Elephant Flow) are effectively balanced across cores with the Intel DLB.

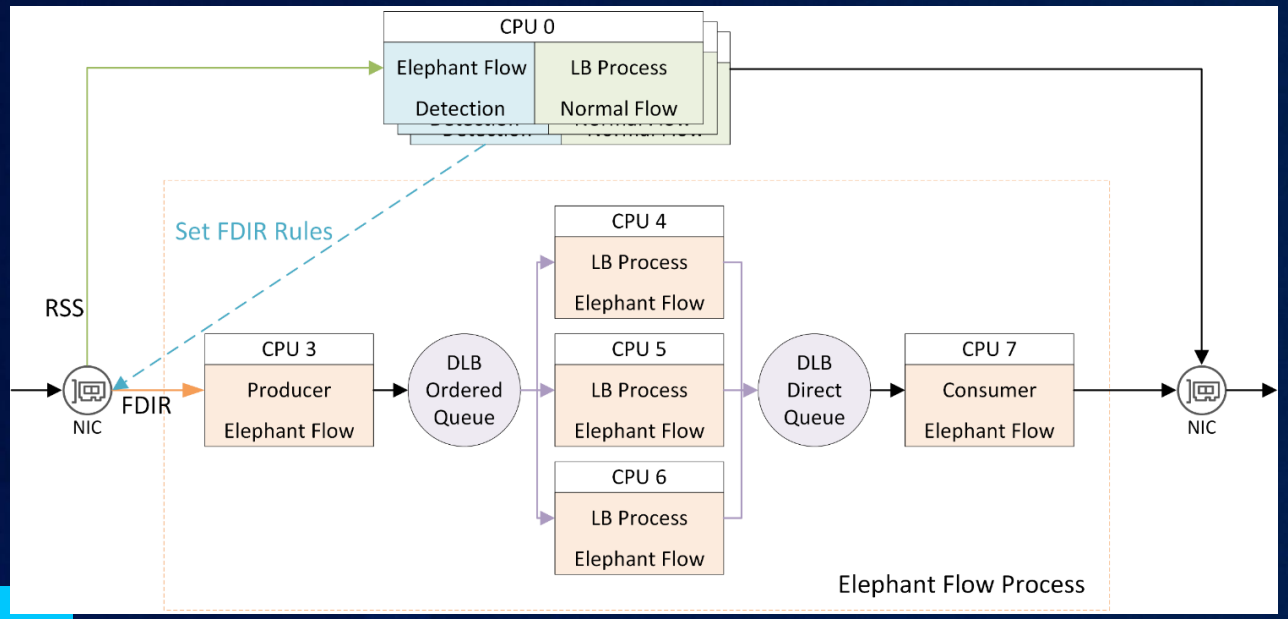
Note: The configuration consists of 1 HT core for producer, 1 HT core for consumer (both on a single physical core) and # worker cores (separate physical cores). SQM configuration consists of 1 EXTRA service core for scheduling.

Results using pre-production 4th Gen Intel® Xeon® Scalable Processor, systems, and software.
Performance varies by part, use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.
See backup for workloads and configurations. Results may vary .
All product plans, roadmaps, and performance are subject to change without notice.

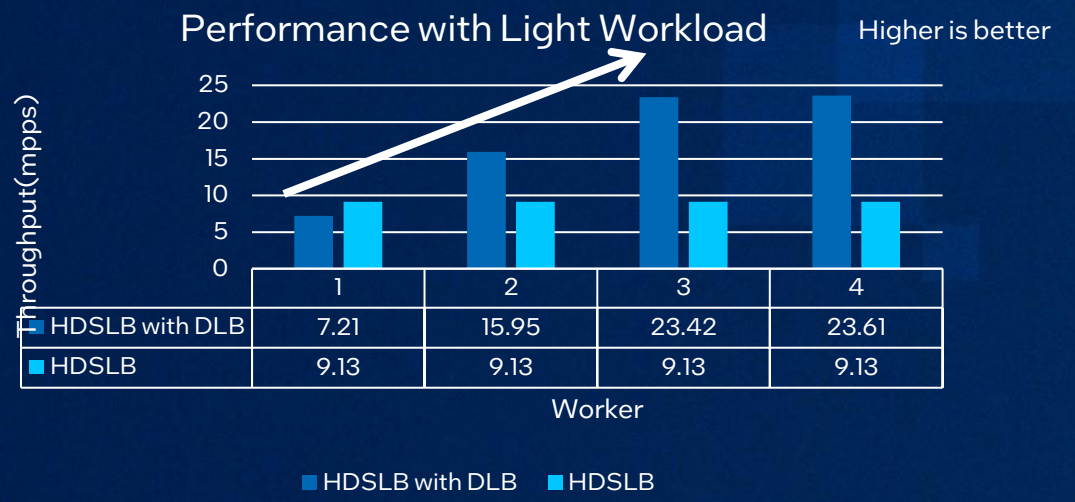
Results may vary. See Config 2

# Elephant Flow Management in Security Load Balancer




Intel DLB PoC with leading cloud vendors
Shows greater throughput with the addition of Intel DLB distributing packets across cores



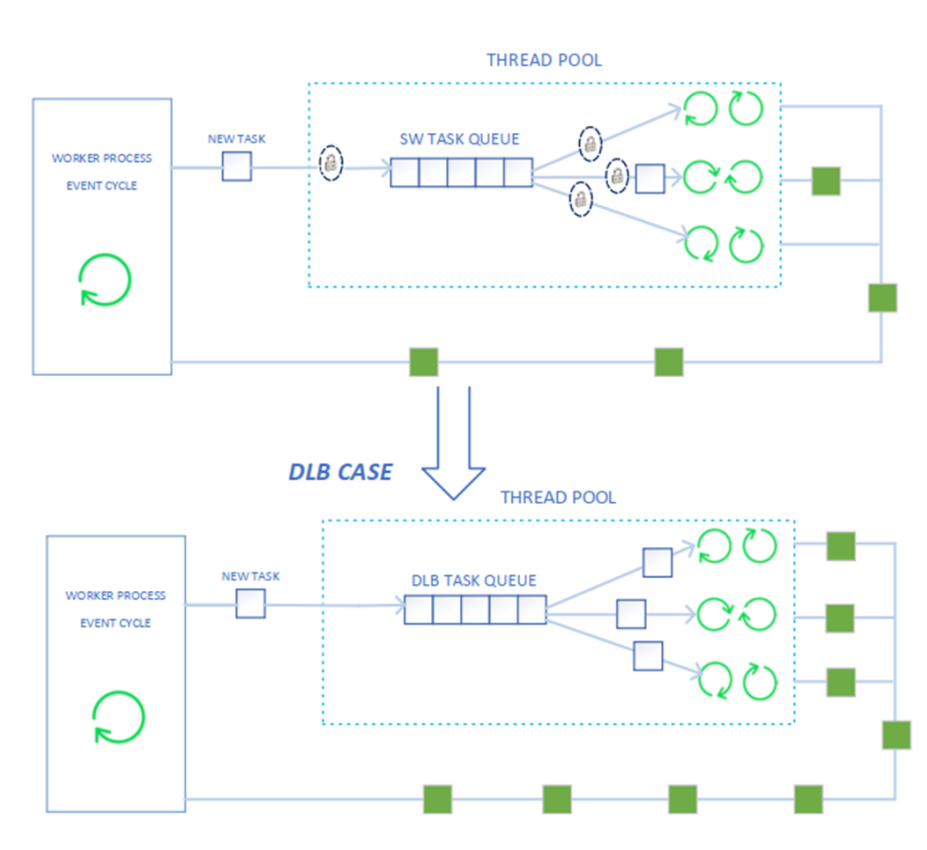

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| HDSLB with DLB | 7.21 | 15.95 | 23.42 | 23.61 |
| HDSLB | 9.13 | 9.13 | 9.13 | 9.13 |

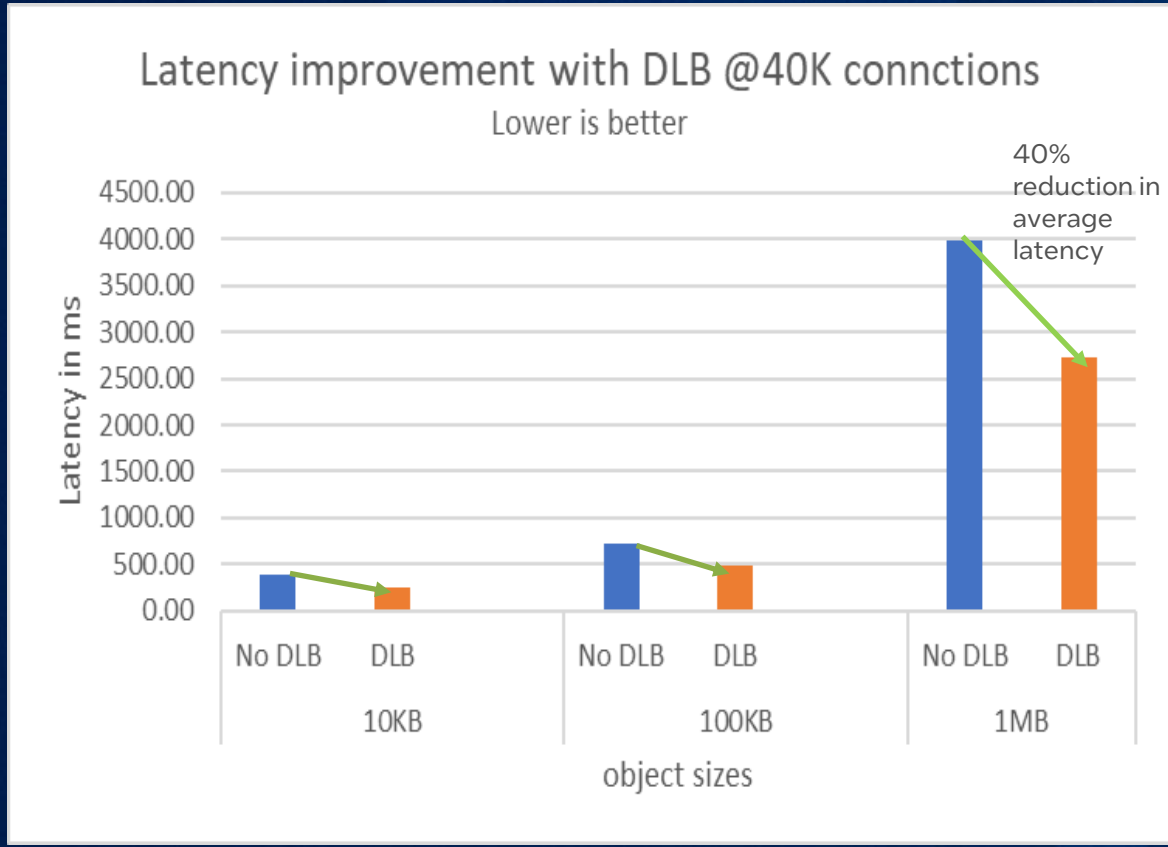
## Why it matters to customers

Higher Performance – More Efficient distribution of flows to workers results in higher performance vs SW solutions.

Results may vary. See Config 3

# Intel DLB in Nginx Webserver – Primary Usage : Load Balancing




## Why it matters to customers

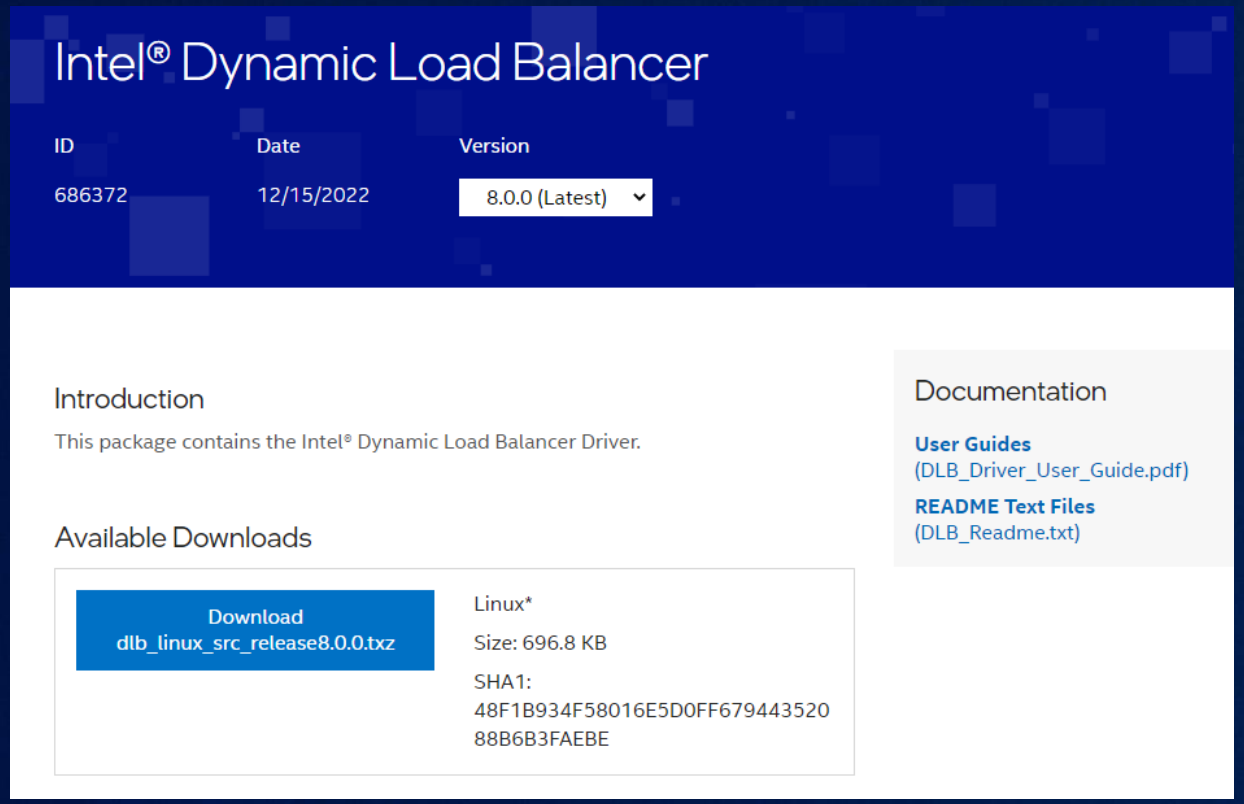
- Reduced Latency - Faster response times for TLS applications – Faster Content Delivery,
- More Connections per Sec = More Client traffic.
- Up to 1.4X improvement in average latency for DLB vs NO DLB on nginx for higher number of connections with operations >1 million across all object types.

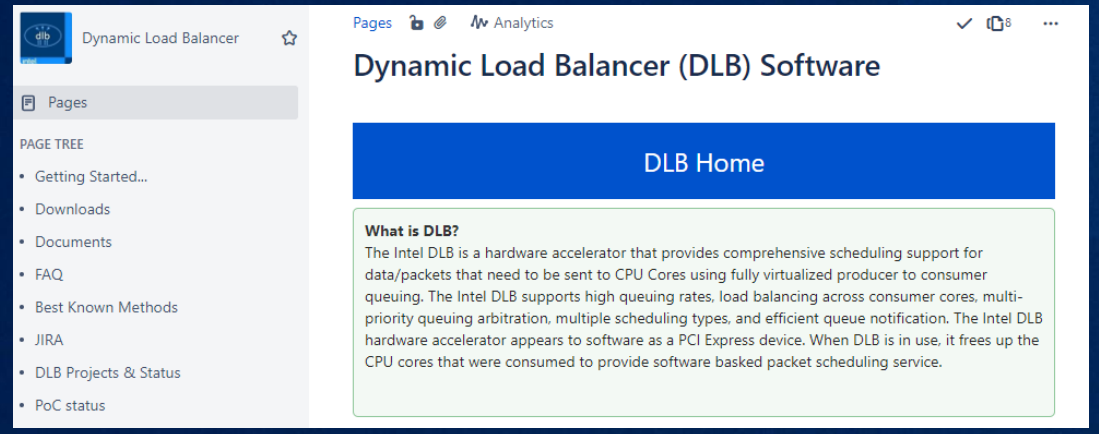
Results may vary. See Config 4

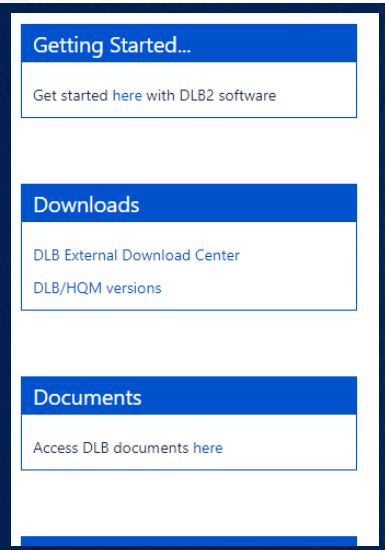
# Enabling

# Where to get the SW & What is enabled?

- Releases are posted externally and internally.




https://www.intel.com/content/www/us/en/download/686372/intel-dynamic-load-balancer.html?
https://wiki.ith.intel.com/pages/viewpage.action?spaceKey=dlb&title=Dynamic+Load+Balancer+%28DLB%29+Software

# Intel DLB Value Proposition

## Core Recovery

Elimination of Software Scheduler yields CPU cores back for application usage.

## Performance

Finer work distribution allows for improved packet processing/data movement performance.

## Determinism

Work can be fed to cores evenly and with an efficiency that provides fine granularity and controlled latencies for applications like data plane policing or packet processing pipelines.

## Lower Latency

Finer tuned performance & work distribution provide for lower latencies in IPSec, PDCP, TLS protocols.

**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

# Thank You

# Performance Configurations

# System Config 1 - Eventdev

| | | |
|---|---|---|
| Platform Hardware | | ac06pdk-ArcherCity |
| | Manufacturer | Intel Corporation |
| | Product | ArcherCity |
| | Sockets | 1 |
| | CPU Model | Genuine Intel(R) CPU $0000%@ |
| | Base Frequency | 1000 MHz |
| | Maximum Frequency | 1.001GHz |
| | All-core Maximum Frequency | 1.9GHz |
| | Cores per CPU | 56 |
| | Total CPU Threads | 112 |
| | TDP | 300 watts |
| | Family | 6 |
| | Model | 143 |
| | Stepping | 2 |
| | Architecture | Unknown Intel |
| | DIMM Config | 1 x 32 GB <OUT OF SPEC> 4000 MT/s |
| | PMEM Firmware Version | |
| | MemTotal | 32636812 kB |
| Platform Config | | |
| | Microcode | 0x8c0002d0 |
| | BIOS Version | EGSDCRB1.86B.0058.D16.2104291438 |
| | BIOS Settings | |
| | NUMA Nodes | 1 |
| | SNC | Disabled |
| | Thread(s) per core | 2 |
| | Turbo | Enabled |
| | Power & Perf Policy | Performance |
| | Prefetchers | DCU HW, DCU IP, L2 HW, L2 Adj. |
| System Software | | |
| | Name | ac06pdk-ArcherCity |
| | Time | Thu Jul 8 10:17:15 PM UTC 2021 |
| | OS | Ubuntu 21.04 |
| | Kernel | 5.11.0-18-generic |
| | GCC Version | gcc (Ubuntu 10.3.0-1ubuntu1) 10.3.0 |
| | GLIBC Version | ldd (Ubuntu GLIBC 2.33-0ubuntu5) 2.33 |
| | Binutils Version | GNU ld (GNU Binutils for Ubuntu) 2.36.1 |
| | Frequency Driver | acpi-cpufreq |
| | Frequency Governer Setting | performance |
| | PMEM Mode | |
| | Huge Pages Total | 8 |
| | Huge Pages Size | 1048576 kB |
| | Transparent Huge Pages | madvise |
| | Automatic NUMA Balancing | Disabled |
| | IRQ Balance | Enabled |
| CVEs | | |
| | CVE-2017-5753 | OK |
| | CVE-2017-5715 | OK |
| | CVE-2017-5754 | OK |
| | CVE-2018-3640 | OK |
| | CVE-2018-3639 | OK |
| | CVE-2018-3615 | OK |
| | CVE-2018-3620 | OK |
| | CVE-2018-3646 | OK |

2021.27.07
test date

| | | |
|---|---|---|
| | CVE-2018-12126 | OK |
| | CVE-2018-12130 | OK |
| | CVE-2018-12127 | OK |
| | CVE-2019-11091 | OK |
| | CVE-2019-11135 | OK |
| | CVE-2018-12207 | OK |
| | CVE-2020-0543 | OK |
| Drive(s) | | |
| | ** NAME ** | sda |
| | MODEL | WDC_WDS500G2B0A-00SM50 |
| | SIZE | 465.8G |
| | FSTYPE | |
| | RQ-SIZE | 64 |
| | MIN-IO | 512 |
| | NUMA Node | |
| | | |
| | ** NAME ** | sda2 |
| | MODEL | |
| | SIZE | 465.3G |
| | FSTYPE | ext4 |
| | RQ-SIZE | 64 |
| | MIN-IO | 512 |
| | NUMA Node | |
| | | |
| | ** NAME ** | sda1 |
| | MODEL | |
| | SIZE | 512M |
| | FSTYPE | vfat |
| | RQ-SIZE | 64 |
| | MIN-IO | 512 |
| | NUMA Node | |
| NIC(s) | | |
| | ** Name ** | enp1s0 |
| | Model | Ethernet Controller I225-LM |
| | Speed | 1000Mb/s |
| | Link | yes |
| | Driver Version | 5.11.0-18-generic |
| | Firmware Version | |
| | NUMA Node | 0 |
| | | |
| | ** Name ** | ens2f1 |
| | Model | Ethernet Controller XXV710 for 25GbE SFP28 |
| | Speed | Unknown! |
| | Link | no |
| | Driver Version | 5.11.0-18-generic |
| | Firmware Version | 6.01 0x80003554 1.1747.0 |
| | NUMA Node | 0 |
| GPU(s) | | |
| | Model | |
| | Version | |
| | Driver | |
| svr_info | | |
| | version | 1.2.4 2021-02-10 816ff182 internal |

# System Config 2 - IPSec

| | |
|---|---|
| Name | spr-dlb-ipsec-sys |
| Time | Thu 11 Jan 2022 |
| Manufacturer | Intel Corporation |
| Product Name | ArcherCity |
| BIOS Version | EGSDCRB1.86B.0062.D04.2107281526, ITP Tuning applied |
| OS | Ubuntu 20.04.3 LTS |
| Kernel | 5.4.0-40-generic |
| Microcode | 0x8d000260 |
| IRQ Balance | Disabled |
| CPU Model | **SPR D0, QDF: QYE3** |
| Base Frequency | 1.8GHz |
| Maximum Frequency | 4.0GHz |
| All-core Maximum Frequency | 2.7GHz |
| CPU(s) | 120 |
| Thread(s) per Core | 2 |
| Core(s) per Socket | 60 |
| Socket(s) | 1 |
| NUMA Node(s) | 1 |
| Prefetchers | DCU HW, DCU IP, L2 HW, L2 Adj. |
| Turbo | Disabled |
| PPIN(s) | |
| Power & Perf Policy | Performance |
| TDP | 350 watts |
| Frequency Driver | |
| Frequency Governor | |
| Frequency (MHz) | 1800 |
| Max C-State | 1 |
| | |
| Installed | 256GB (8x32GB <OUT OF SPEC> 4800MT/s [4800MT/s]) - Dual Rank |
| Huge Pages Size | 1048576 kB |
| Transparent Huge Pages | Never |
| Automatic NUMA Balancing | Disabled |
| NIC Summary | Intel Corporation Ethernet Controller E810-C for QSFP (rev 02)<br>2x100 GbE link |
| Drive Summary | 1x 240G INTEL_SSD |

| | |
|---|---|
| Workload | DPDK* IPsec with Intel® Dynamic Load Balancer (DLB) and Software queue management (SQM) |
| OS | Ubuntu* 20.04.2 LTS |
| Kernel | 5.4.0-40-generic |
| Workload Version | DPDK* IPsec -20.08 |
| Compiler | gcc 9.3.0 |
| CPU Utilization (active cores) | 100% |

# System Config 3 - HDSLB

| | |
|---|---|
| Name | gtk-12-3-2 |
| Time | Thu Nov 3 01:38:06 PM UTC 2022 |
| Manufacturer | Intel Corporation |
| Product Name | ArcherCity |
| BIOS Version | EGSDCRB1.SYS.8901.P01.2209200243 |
| OS | Ubuntu 22.04 LTS |
| Kernel | 5.15.0-27-generic |
| Microcode | 0x2b0000a1 |
| IRQ Balance | Disabled |
| QDF/Stepping | E5 Stepping |
| Base Frequency | 1.8GHz |
| Maximum Frequency | 3.6GHz |
| All-core Maximum Frequency | 2.8GHz |
| CPU(s) | 104 |
| Thread(s) per Core | 2 |
| Core(s) per Socket | 52 |
| Socket(s) | 1 |
| NUMA Node(s) | 1 |
| Prefetchers | L2 HW, L2 Adj., DCU HW, DCU IP |
| Turbo | Disabled |
| PPIN(s) | 28d1f8c9350f4a0f |
| Power & Perf Policy | Performance |
| TDP | 300 watts |
| Frequency Driver | |
| Frequency Governer | |
| Frequency (MHz) | 3600 |
| Max C-State | 1 |
| Installed Memory | 256GB (8x32GB DDR5 4800 MT/s [4800 MT/s]) |
| Huge Pages Size | 1048576 kB |
| Transparent Huge Pages | madvise |
| Automatic NUMA Balancing | Disabled |
| NIC Summary | 2x Ethernet Controller E810-C for QSFP |
| Drive Summary | INTEL SSDSC2KB240G8 |

| | Config1 (baseline) |
|---|---|
| Workload & version | HDSLB |
| Compiler | Ninja 1.10.1<br>gcc (Ubuntu 11.2.0-19ubuntu1) 11.2.0 |
| DPDK | 21.11.0 |
| CPU Utilization | 100% per core |

# System Config 4 - NGINX

| | | |
|---|---|---|
| Name | spr03--qyk8--bck47 | spr03--qyk8--bck47 |
| Time | Tue Mar 29 08:00:25 UTC 2022 | Tue Mar 29 07:19:53 UTC 2022 |
| Manufacturer | Intel Corporation | Intel Corporation |
| Product Name | EAGLESTREAM | EAGLESTREAM |
| BIOS Version | EGSDCRB1.86B.0071.D03.2112251345 | EGSDCRB1.86B.0071.D03.2112251345 |
| OS | CentOS Linux 8 | CentOS Linux 8 |
| Kernel | 5.12.0-0507.intel_next.10_26_po.49.x86_64+server | 5.12.0-0507.intel_next.10_26_po.49.x86_64+server |
| Microcode | 0x8d0003f0 | 0x8d0003f0 |
| IRQ Balance | Disabled | Disabled |
| QDF/Stepping | QYK8 | QYK8 |
| Base Frequency | 1.9GHz | 1.9GHz |
| Maximum Frequency | 4.0GHz | 4.0GHz |
| All-core Maximum Frequency | 3.0GHz | 3.0GHz |
| CPU(s) | 112 | 112 |
| Thread(s) per Core | 2 | 2 |
| Core(s) per Socket | 56 | 56 |
| Socket(s) | 1 | 1 |
| NUMA Node(s) | 1 | 1 |
| Prefetchers | DCU HW, DCU IP, L2 HW, L2 Adj. | DCU HW, DCU IP, L2 HW, L2 Adj. |

| | | |
|---|---|---|
| Turbo | Enabled | Enabled |
| PPIN(s) | D9d80af20beaa4a9 | d9d80af20beaa4a9 |
| Power & Perf Policy | Performance | Performance |
| TDP | 350 watts | 350 watts |
| Frequency Driver | intel_pstate | intel_pstate |
| Frequency Governer | Performance | Performance |
| Frequency (MHz) | 3596 | 1900 |
| Max C-State | 9 | 9 |
| Installed Memory | 128GB (4x32GB 4800MT/s [4800MT/s]) | 128GB (4x32GB 4800MT/s [4800MT/s]) |
| Huge Pages Size | 2048 kB | 2048 kB |
| Transparent Huge Pages | always | always |
| Automatic NUMA Balancing | *Not Found* | *Not Found* |
| NIC Summary | Ethernet Controller I225-LM, Ethernet Controller XXV710 for 25GbE SFP28, Ethernet Controller XXV710 for 25GbE SFP28, Ethernet Controller E810-C for QSFP, Ethernet Controller E810-C for QSFP | Ethernet Controller I225-LM, Ethernet Controller XXV710 for 25GbE SFP28, Ethernet Controller XXV710 for 25GbE SFP28, Ethernet Controller E810-C for QSFP, Ethernet Controller E810-C for QSFP |
| Drive Summary | SSDPF2KX038TZ, SSDPE2KX040T8 | SSDPF2KX038TZ, SSDPE2KX040T8 |

| | Config1 (baseline)- CPU only | Config2 (new) – with DLB |
|---|---|---|
| Workload1 & version | nginx/1.16.1 | nginx/1.16.1 |
| Compiler | gcc (GCC) 8.5.0 20210514 (Red Hat 8.5.0-4) | gcc (GCC) 8.5.0 20210514 (Red Hat 8.5.0-4) |
| DPDK | NA | NA |
| wrk ( client side) | wrk 4.2.0 [epoll] Copyright (C) 2012 Will Glozer | wrk 4.2.0 [epoll] Copyright (C) 2012 Will Glozer |
| BKC# | 47 | 47 |
| DLB Driver | NA | RELEASE_VER_7.4.0-V1 |
| Iterations and result choice (median, average, min, max) | 180sec, average of 3 runs | 180sec |
| Raw Results (latency in ms) | 3990 | 2720 |
| Operating Frequency | 2.72Ghz | 2.72Ghz |
| CPU Utilization | 100% | 98% |


intel