# HARDWARE-LEVEL PERFORMANCE ANALYSIS OF PLATFORM I/O

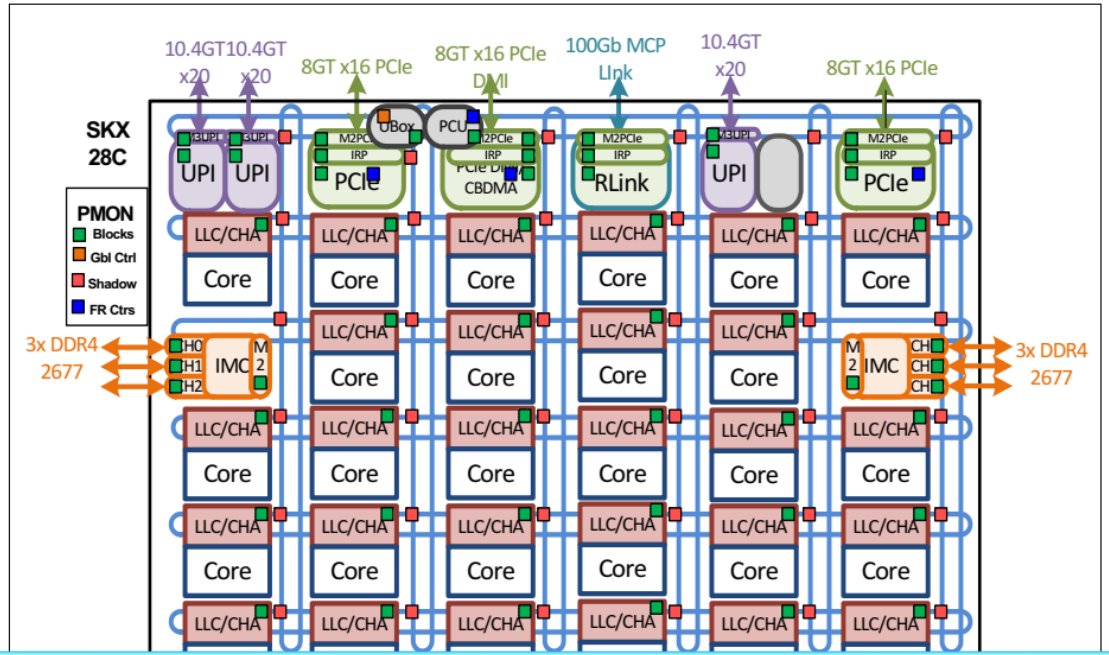
## or: How I Learned to Stop Worrying and Love Uncore

Roman Sudarikov, Perry Taylor, Jeremy Williamson, Neil Achtman

# Performance monitoring

- ***Core***
  - Execution units (ALU)
  - L1 caches
  - L2 caches
- ***Uncore***
  - Rest of the processor besides the core

**Figure 1-1. Skylake Server-28C EX Server Block Diagram**




**Hardware feedback through publically available tools**

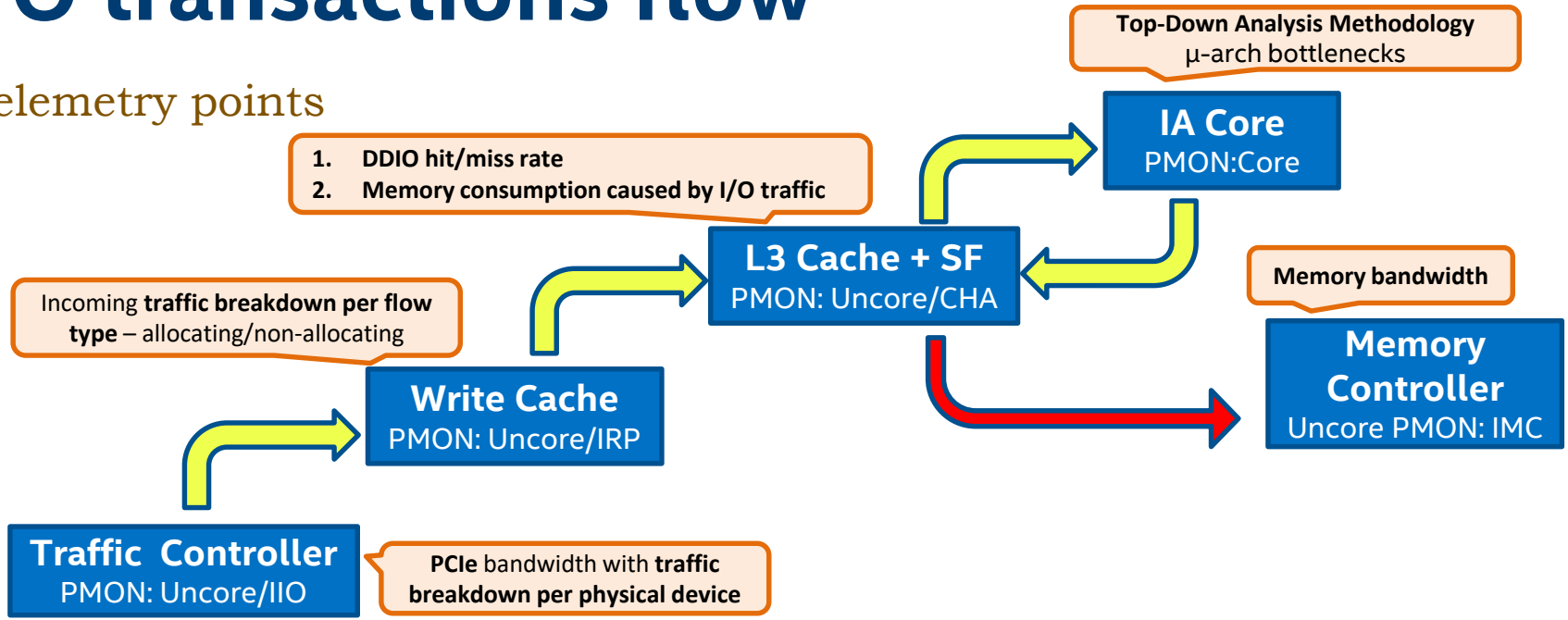
# Performance monitoring

- ***Core monitoring*** *can tell SW how it is performing on an iA core - e.g. was the code scheduled to execute efficiently? Were the tasks well balanced?*

- ***Uncore monitoring*** *can give SW a better sense where all that traffic was routed and what was involved in processing it*

# I/O transactions flow

## Telemetry points

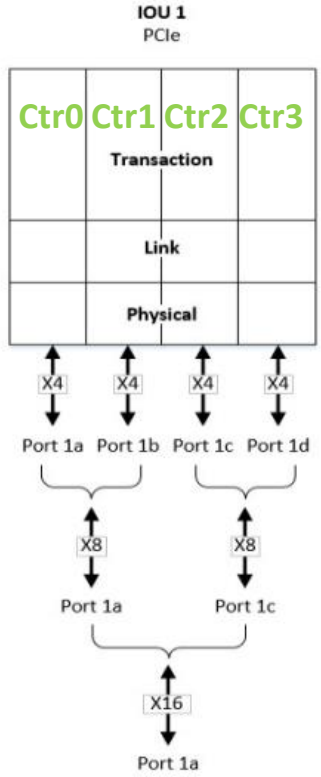



**There is so much happens with IO request on the way to the core which is beyond the scope of the core's PMU**

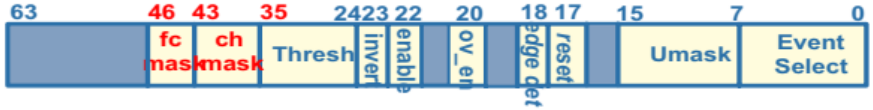
# Traffic controller

## Terminology

IIO – integrated IO
Outbound direction is from the root port towards the endpoint;
Inbound direction is from the endpoint towards the root port;
OTC – outbound traffic controller;
ITC – inbound traffic controller;




IIO counter control register for Skylake Server




fc_mask controls whether it looks at P, NP, or C;
ch_mask controls what PCIe port it looks at.

# Traffic controller

Events

### 2.5.4 IIO Box Events Ordered By Code

The following table summarizes the directly measured IIO Box events.

| Symbol Name | Event Code | Ctrs | Max Inc/ Cyc | Description |
|---|---|---|---|---|
| NOTHING | 0x00 | | 0 | |
| CLOCKTICKS | 0x01 | 0-3 | 0 | Traffic Controller Clocks |
| MASK_MATCH_AND | 0x02 | 0-3 | 0 | AND Mask/match for debug bus |
| MASK_MATCH_OR | 0x03 | 0-3 | 0 | OR Mask/match for debug bus |
| LINK_NUM_RETRIES | 0x0e | | 0 | Num Link Retries |
| LINK_NUM_CORR_ERR | 0x0f | | 0 | Num Link Correctable Errors |
| MASK_MATCH | 0x21 | | 0 | Number packets that passed the Mask/Match Filter |
| VTD_OCCUPANCY | 0x40 | 0-3 | 0 | Intel® Virtualization Technology (Intel® VT) for Directed I/O (Intel® VT-d) Occupancy |
| VTD_ACCESS | 0x41 | 0-3 | 0 | Intel VT-d Access |
| SYMBOL_TIMES | 0x82 | | 0 | Symbol Times on Link |
| DATA_REQ_OF_CPU | 0x83 | 0-1 | 0 | Data requested of the CPU |
| TXN_REQ_OF_CPU | 0x84 | 0-3 | 0 | Number Transactions requested of the CPU |
| DATA_REQ_BY_CPU | 0xc0 | 2-3 | 0 | Data requested by the CPU |
| TXN_REQ_BY_CPU | 0xc1 | 0-3 | 0 | Number Transactions requested by the CPU |

https://software.intel.com/en-us/blogs/2014/07/11/documentation-for-uncore-performance-monitoring-units

# Traffic controller

## Application writes to device/device reads from system memory

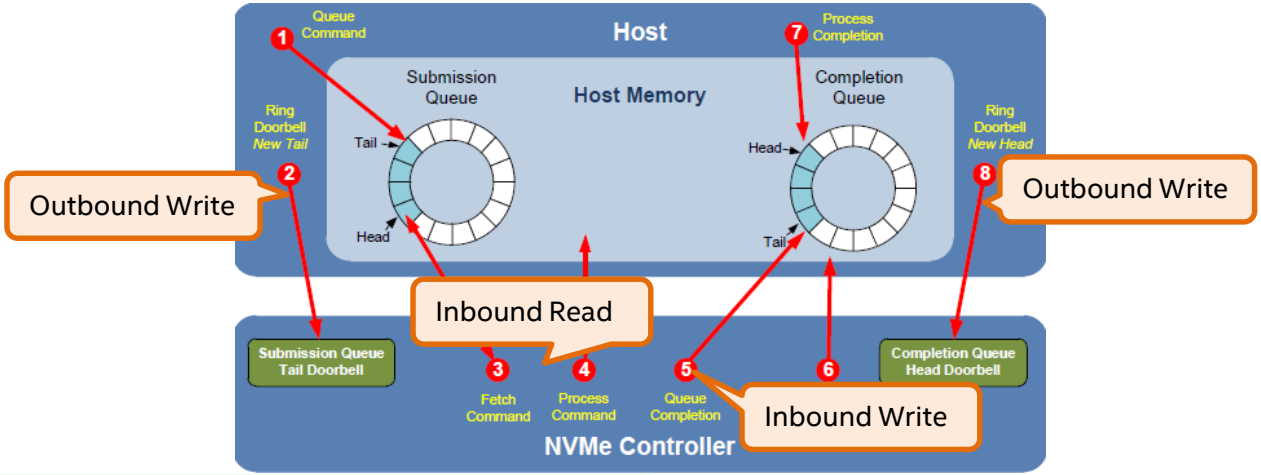
```
[root@nntvtune206 perf]# ./perf -r 'trtype:PCIe traddr:0000:da:00.0' -q 8 -o 4096 -w write -t 10 -c 0x1000000
Starting SPDK v19.04-pre / DPDK 18.11.0 initialization...
[ DPDK EAL parameters: perf --no-shconf -c 0x1000000 --base-virtaddr=0x200000000000 --file-prefix=spdk_pid10329 ]
EAL: Detected 96 lcore(s)
EAL: Detected 2 NUMA nodes
EAL: No free hugepages reported in hugepages-1048576kB
EAL: Probing VFIO support...
Initializing NVMe Controllers
Attaching to NVMe Controller at 0000:da:00.0
Attached to NVMe Controller at 0000:da:00.0 [8086:2701]
Associating INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) with lcore 24
Initialization complete. Launching workers.
Starting thread on core 24
========================================================
                                                                    Latency(us)
Device Information                                  :       IOPS      MB/s    Average        min        max
INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) from core 24  555166.80   2168.62      14.40       6.70     360.12
========================================================
Total                                               :  555166.80   2168.62      14.40       6.70     360.12

[root@nntvtune206 perf]#
```

| Uncore TC Events | Counts/sec | Comment |
|---|---|---|
| Inbound Write | 2,242,660 | Device updates CQ tail (x16B) |
| Inbound Read | 579,660,272 | Device fetches data (x4096B) + commands (x64B) |
| Outbound Write | 1,123,970 | CPU rings Doorbell (x4B) twice per each request |

# Traffic controller – explanation

## Application writes to device/device reads from system memory

Application reports:
**555166** IOPs




| Uncore TC Events | Counts/sec | Explanation | Comment |
|---|---|---|---|
| Inbound Write | 2,242,660 | = **555166** * 16B / 4B | Device updates CQ tail (x16B) |
| Inbound Read | 579,660,272 | = **555166** * (4096B + 64B) / 4B | Device fetches data (x4096B) + commands (x64B) |
| Outbound Write | 1,123,970 | = **555166** *2 | CPU rings Doorbell (x4B) twice |

# Traffic controller

## Application writes to device/device reads from system memory

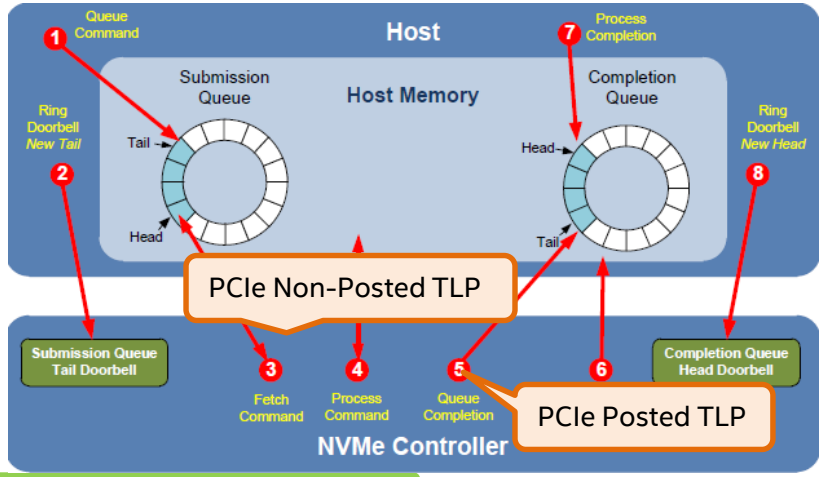
```
[root@nntvtune206 perf]# ./perf -r 'trtype:PCIe traddr:0000:da:00.0' -q 8 -o 4096 -w write -t 10 -c 0x1000000
Starting SPDK v19.04-pre / DPDK 18.11.0 initialization...
[ DPDK EAL parameters: perf --no-shconf -c 0x1000000 --base-virtaddr=0x200000000000 --file-prefix=spdk_pid10329 ]
EAL: Detected 96 lcore(s)
EAL: Detected 2 NUMA nodes
EAL: No free hugepages reported in hugepages-1048576kB
EAL: Probing VFIO support...
Initializing NVMe Controllers
Attaching to NVMe Controller at 0000:da:00.0
Attached to NVMe Controller at 0000:da:00.0 [8086:2701]
Associating INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) with lcore 24
Initialization complete. Launching workers.
Starting thread on core 24
========================================================
                                                                    Latency(us)
Device Information                                  :       IOPS      MB/s    Average        min        max
INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) from core 24  555166.80   2168.62      14.40       6.70     360.12
========================================================
Total                                               :  555166.80   2168.62      14.40       6.70     360.12

[root@nntvtune206 perf]#
```

| Uncore TC Events | Counts/sec | Comment |
|---|---|---|
| Inbound Posted TLPs | 559,893 | Device updates CQ tail (x16B) |
| Inbound Non-Posted TLPs | 5,070,096 | Device fetches data (x4096B) + commands (x64B)<br>MRRS=512B so 8 TLPs for Data + 1 for command |

# Traffic controller – explanation

## Application writes to device/device reads from system memory

Application reports:
**555166** IOPs




| Uncore TC Events | Counts/sec | Explanation | Comment |
|---|---|---|---|
| Inbound Posted TLPs | 559,893 | = **555166** | Device updates CQ tail (x16B) |
| Inbound Non-Posted TLPs | 5,070,096 | = **555166** * (8TLP+ 1TLP) | Device fetches data (x4096B) + commands (x64B)<br>MRRS=512B so 8 TLPs for Data + 1 for command |

# Traffic controller

## Application reads from device/device writes to system memory

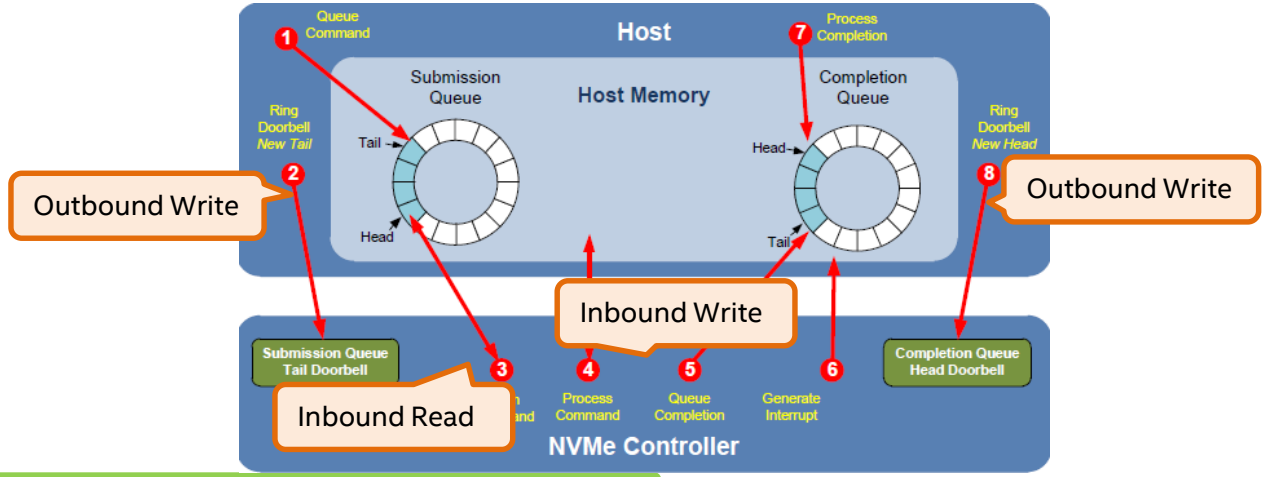
```
[root@nntvtune206 perf]# ./perf -r 'trtype:PCIe traddr:0000:da:00.0' -q 8 -o 4096 -w read -t 10 -c 0x1000000
Starting SPDK v19.04-pre / DPDK 18.11.0 initialization...
[ DPDK EAL parameters: perf --no-shconf -c 0x1000000 --base-virtaddr=0x200000000000 --file-prefix=spdk_pid10034 ]
EAL: Detected 96 lcore(s)
EAL: Detected 2 NUMA nodes
EAL: No free hugepages reported in hugepages-1048576kB
EAL: Probing VFIO support...
Initializing NVMe Controllers
Attaching to NVMe Controller at 0000:da:00.0
Attached to NVMe Controller at 0000:da:00.0 [8086:2701]
Associating INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) with lcore 24
Initialization complete. Launching workers.
Starting thread on core 24
========================================================
                                                                    Latency(us)
Device Information                                  :       IOPS      MB/s    Average        min        max
INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) from core 24   666966.10   2605.34      11.99       6.30     335.15
========================================================
Total                                               :   666966.10   2605.34      11.99       6.30     335.15

[root@nntvtune206 perf]#
```

| Uncore TC Events | Counts/sec | Comment |
|---|---|---|
| Inbound Write | 684,557,476 | Device delivers data (x4096B) + updates CQ tail (x16B) |
| Inbound Read | 10,644,464 | Device fetches commands (x64B) |
| Outbound Write | 1,330,200 | CPU rings Doorbell (x4B) twice |

# Traffic controller – explanation

## Application reads from device/device writes to system memory

Application reports:
**666966** IOPs




| Uncore TC Events | Counts/sec | Explanation | Comment |
|---|---|---|---|
| Inbound Write | 684,557,476 | = **666966** * (4096B + 64B) / 4B | Device delivers data (x4096B) + updates CQ tail (x16B) |
| Inbound Read | 10,644,464 | = **666966** * 64B / 4B | Device fetches commands (x64B) |
| Outbound Write | 1,330,200 | = **666966** *2 | CPU rings Doorbell (x4B) twice |

# Traffic controller

## Application reads from device/device writes to system memory

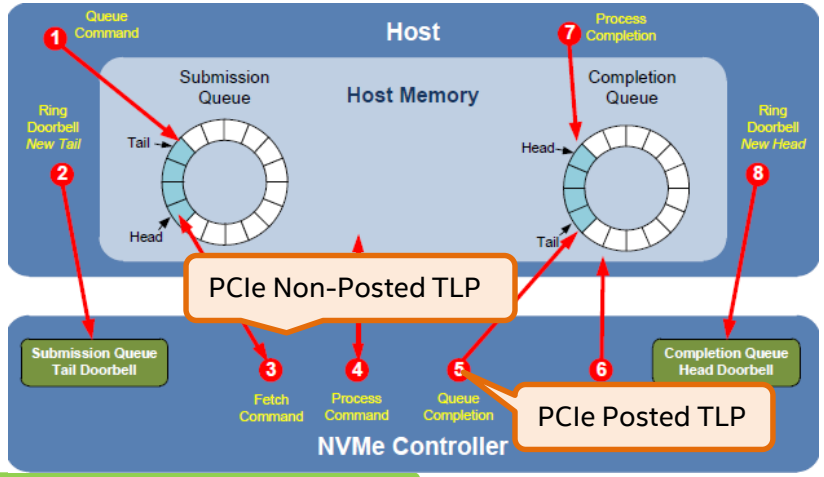
```
[root@nntvtune206 perf]# ./perf -r 'trtype:PCIe traddr:0000:da:00.0' -q 8 -o 4096 -w read -t 10 -c 0x1000000
Starting SPDK v19.04-pre / DPDK 18.11.0 initialization...
[ DPDK EAL parameters: perf --no-shconf -c 0x1000000 --base-virtaddr=0x200000000000 --file-prefix=spdk_pid10034 ]
EAL: Detected 96 lcore(s)
EAL: Detected 2 NUMA nodes
EAL: No free hugepages reported in hugepages-1048576kB
EAL: Probing VFIO support...
Initializing NVMe Controllers
Attaching to NVMe Controller at 0000:da:00.0
Attached to NVMe Controller at 0000:da:00.0 [8086:2701]
Associating INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) with lcore 24
Initialization complete. Launching workers.
Starting thread on core 24
========================================================
                                                                    Latency(us)
Device Information                                  :       IOPS      MB/s    Average        min        max
INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) from core 24  666966.10   2605.34      11.99       6.30     335.15
========================================================
Total                                               :  666966.10   2605.34      11.99       6.30     335.15

[root@nntvtune206 perf]#
```

| Uncore TC Events | Counts/sec | Comment |
|---|---|---|
| Inbound Posted TLPs | 11,316,994 | Device delivers data (x4096B) + updates CQ tail (x16B)<br>MPS=256B so 16 TLPs for Data + 1 for CQ tail |
| Inbound Non-Posted TLPs | 666,375 | Device fetches commands (x64B) |

# Traffic controller – explanation

## Application writes to device/device reads from system memory

Application reports:
**666966** IOPs




| Uncore TC Events | Counts/sec | Explanation | Comment |
|---|---|---|---|
| Inbound Posted TLPs | 11,316,994 | = **666966** * (16TLP+ 1TLP) | Device delivers data (x4096B) + updates CQ tail (x16B)<br>MPS=256B so 16 TLPs for Data + 1 for CQ tail |
| Inbound Non-Posted TLPs | 666,375 | = **666966** | Device fetches commands (x64B) |

# Write Cache

## Events

| Symbol Name | Event Code | Ctrs | Max Inc/ Cyc | Description |
|---|---|---|---|---|
| CLOCKTICKS | 0x01 | | 0 | IRP Clocks |
| TxC_BL_DRS_INSERTS | 0x02 | 0-1 | 0 | BL DRS Egress Inserts |
| TxC_BL_NCB_INSERTS | 0x03 | 0-1 | 0 | BL NCB Egress Inserts |
| TxC_BL_NCS_INSERTS | 0x04 | 0-1 | 0 | BL NCS Egress Inserts |
| TxC_BL_DRS_CYCLES_FULL | 0x05 | 0-1 | 0 | BL DRS Egress Cycles Full |
| TxC_BL_NCB_CYCLES_FULL | 0x06 | 0-1 | 0 | BL NCB Egress Cycles Full |
| TxC_BL_NCS_CYCLES_FULL | 0x07 | 0-1 | 0 | BL NCS Egress Cycles Full |
| TxC_BL_DRS_OCCUPANCY | 0x08 | 0-1 | 0 | BL DRS Egress Occupancy |
| TxC_BL_NCB_OCCUPANCY | 0x09 | 0-1 | 0 | BL NCB Egress Occupancy |
| TxC_BL_NCS_OCCUPANCY | 0x0a | 0-1 | 0 | BL NCS Egress Occupancy |
| TxC_AK_INSERTS | 0x0b | 0-1 | 0 | AK Egress Allocations |
| TxS_REQUEST_OCCUPANCY | 0x0c | 0-1 | 0 | Outbound Request Queue Occupancy |
| TxS_DATA_INSERTS_NCB | 0x0d | 0-1 | 0 | Outbound Read Requests |
| TxS_DATA_INSERTS_NCS | 0x0e | 0-1 | 0 | Outbound Read Requests |
| CACHE_TOTAL_OCCUPANCY | 0x0f | 0-1 | 0 | Total Write Cache Occupancy |
| COHERENT_OPS | 0x10 | 0-1 | 0 | Coherent Ops |
| TRANSACTIONS | 0x11 | 0-1 | 0 | Inbound Transaction Count |

https://software.intel.com/en-us/blogs/2014/07/11/documentation-for-uncore-performance-monitoring-units

# Write Cache

## Application reads from device/device writes to system memory

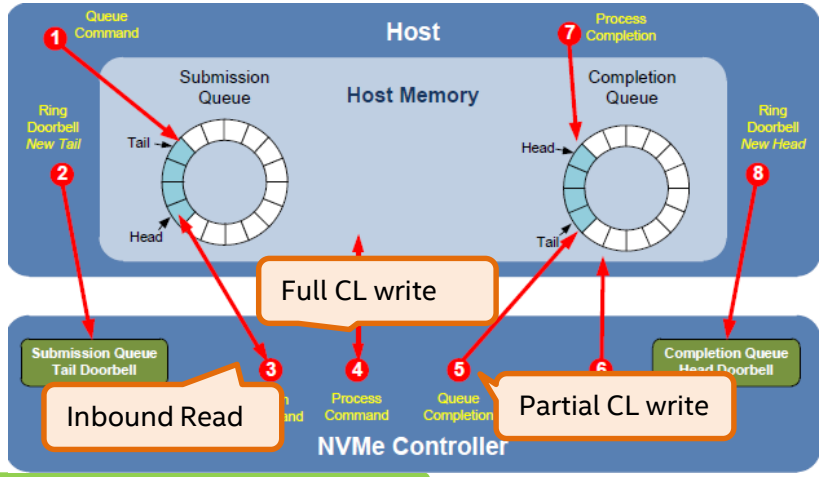
```
[root@nntvtune206 perf]# ./perf -r 'trtype:PCIe traddr:0000:da:00.0' -q 8 -o 4096 -w read -t 10 -c 0x1000000
Starting SPDK v19.04-pre / DPDK 18.11.0 initialization...
[ DPDK EAL parameters: perf --no-shconf -c 0x1000000 --base-virtaddr=0x200000000000 --file-prefix=spdk_pid10034 ]
EAL: Detected 96 lcore(s)
EAL: Detected 2 NUMA nodes
EAL: No free hugepages reported in hugepages-1048576kB
EAL: Probing VFIO support...
Initializing NVMe Controllers
Attaching to NVMe Controller at 0000:da:00.0
Attached to NVMe Controller at 0000:da:00.0 [8086:2701]
Associating INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) with lcore 24
Initialization complete. Launching workers.
Starting thread on core 24
========================================================
                                                                    Latency(us)
Device Information                                  :       IOPS      MB/s    Average        min        max
INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) from core 24   666966.10   2605.34      11.99       6.30     335.15
========================================================
Total                                               :       666966.10   2605.34      11.99       6.30     335.15

[root@nntvtune206 perf]#
```

| Uncore IRP Events | Counts/sec | Comment |
|---|---|---|
| Full cache line write | 42,268,836 | Device delivers data (x4096B) |
| Partial cache line Write | 666,934 | Device updates CQ tail (x16B) |
| Read | 666,940 | Device fetches commands (x64B) |

# Write Cache – explanation

## Application reads from device/device writes to system memory

Application reports:
**666966** IOPs




| Uncore TC Events | Counts/sec | Explanation | Comment |
|---|---|---|---|
| Full cache line write | 42,268,836 | = **666966** * 4096B / 64B | Device delivers data (x4096B) |
| Partial cache line Write | 666,934 | = **666966** | Device updates CQ tail (x16B) |
| Read | 666,940 | = **666966** | Device fetches commands (x64B) |

# Write Cache

## Application writes to device/device reads from system memory

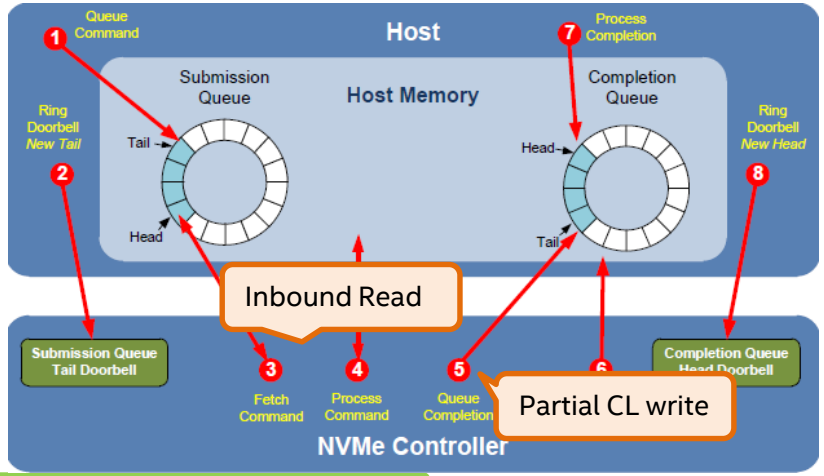
```
[root@nntvtune206 perf]# ./perf -r 'trtype:PCIe traddr:0000:da:00.0' -q 8 -o 4096 -w write -t 10 -c 0x1000000
Starting SPDK v19.04-pre / DPDK 18.11.0 initialization...
[ DPDK EAL parameters: perf --no-shconf -c 0x1000000 --base-virtaddr=0x200000000000 --file-prefix=spdk_pid10329 ]
EAL: Detected 96 lcore(s)
EAL: Detected 2 NUMA nodes
EAL: No free hugepages reported in hugepages-1048576kB
EAL: Probing VFIO support...
Initializing NVMe Controllers
Attaching to NVMe Controller at 0000:da:00.0
Attached to NVMe Controller at 0000:da:00.0 [8086:2701]
Associating INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) with lcore 24
Initialization complete. Launching workers.
Starting thread on core 24
========================================================
                                                                    Latency(us)
Device Information                                  :       IOPS      MB/s    Average        min        max
INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) from core 24  555166.80   2168.62      14.40       6.70     360.12
========================================================
Total                                               :  555166.80   2168.62      14.40       6.70     360.12

[root@nntvtune206 perf]#
```

| Uncore IRP Events | Counts/sec | Comment |
|---|---|---|
| Full cache line write | 0 | |
| Partial cache line Write | 565,389 | Device updates CQ tail (x16B) |
| Read | 36,403,315 | Device fetches data (x4096B) + commands (x64B) |

# Write Cache – explanation

## Application writes to device/device reads from system memory

Application reports:
**555166** IOPs




| Uncore TC Events | Counts/sec | Explanation | Comment |
|---|---|---|---|
| Full cache line write | 0 | | |
| Partial cache line Write | 565,389 | = **555166** | Device updates CQ tail (x16B) |
| Read | 36,403,315 | = **555166** * (4096B + 64B) / 64B | Device fetches data (x4096B) + commands (x64B) |

# Write Cache

- **Inbound Write Flows:**

  Two phases

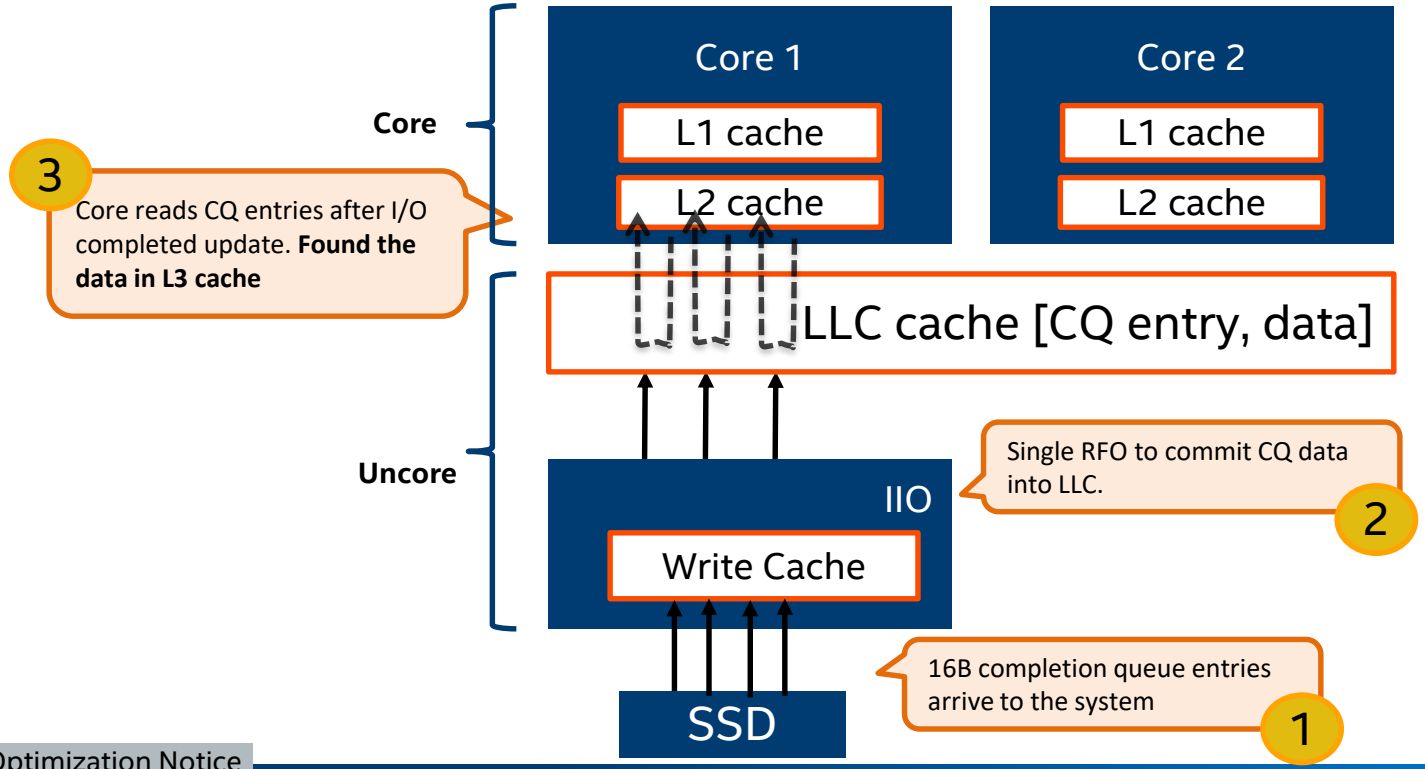
  - Get line ownership for IIO
    - full cache line write, ownership without data (I2M)
    - partial cache line write, ownership with data (RFO)
  - IIO delivers data to LLC, releases ownership (WbMtoI)

- **IIO Snoop assumption:**

  - IIO treats all core snoops as invalidating

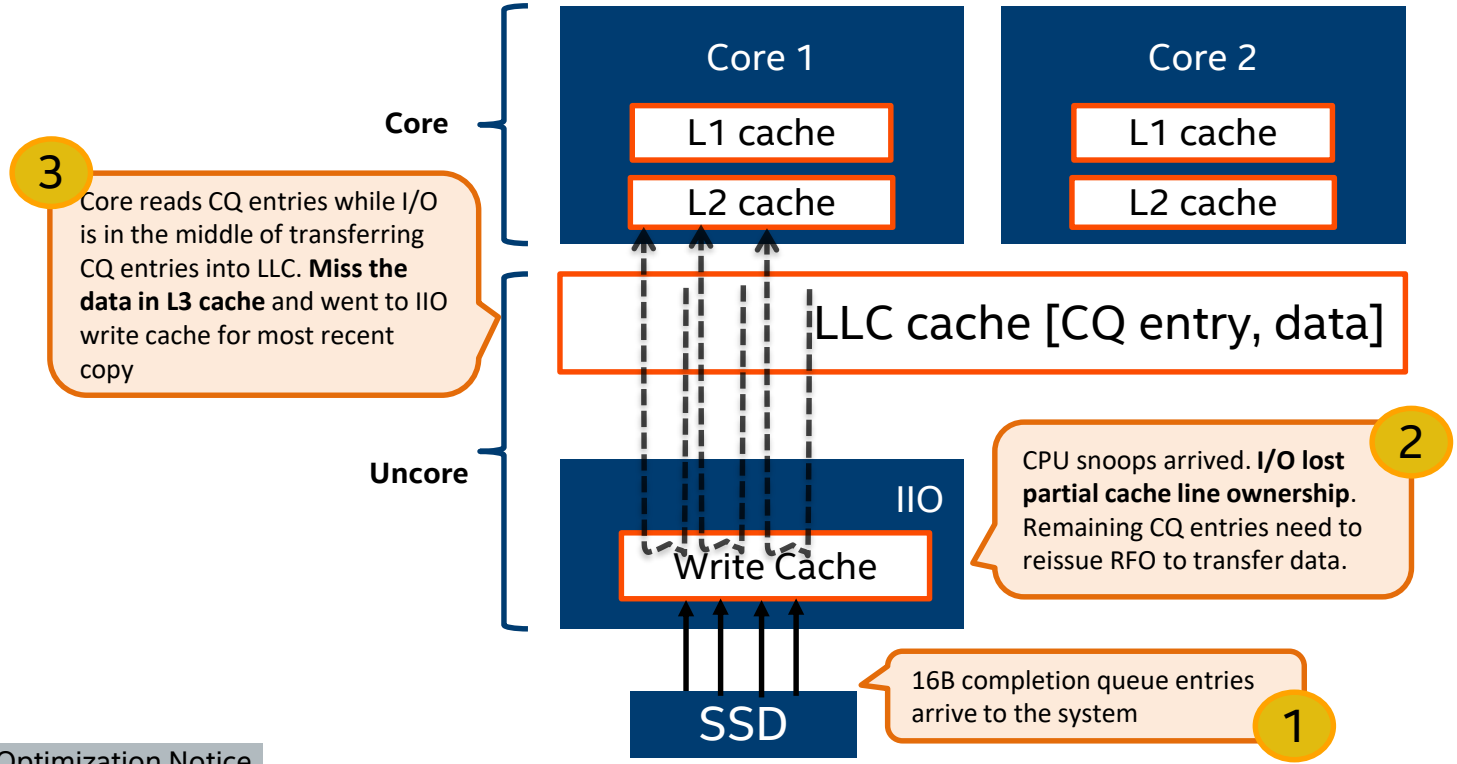
# Write Cache

## Basic flow without conflicts – core is polling just right



# Write Cache

## Core/IO conflict flow – core polling is ahead of IO



# Write Cache

## Core/IO interaction

```
[root@nntvtune206 perf]# ./perf -r 'trtype:PCIe traddr:0000:da:00.0' -q 8 -o 4096 -w write -t 10 -c 0x1000000
Starting SPDK v19.04-pre / DPDK 18.11.0 initialization...
[ DPDK EAL parameters: perf --no-shconf -c 0x1000000 --base-virtaddr=0x200000000000 --file-prefix=spdk_pid10329 ]
EAL: Detected 96 lcore(s)
EAL: Detected 2 NUMA nodes
EAL: No free hugepages reported in hugepages-1048576kB
EAL: Probing VFIO support...
Initializing NVMe Controllers
Attaching to NVMe Controller at 0000:da:00.0
Attached to NVMe Controller at 0000:da:00.0 [8086:2701]
Associating INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) with lcore 24
Initialization complete. Launching workers.
Starting thread on core 24
========================================================
                                                                    Latency(us)
Device Information                                  :       IOPS      MB/s    Average        min        max
INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) from core 24  555166.80   2168.62      14.40       6.70     360.12
========================================================
Total                                               :  555166.80   2168.62      14.40       6.70     360.12

[root@nntvtune206 perf]#
```

| Uncore IRP Events | Counts/sec | Comment |
|---|---|---|
| Lost Forward | 0 | Snoop pulled away ownership before a write was committed |
| SNOOP_RESP.HIT_M | 555,721 | HIT_M means cores missed LLC and snoop into IIO for completion queue entries |

# The LLC coherence engine (CHA)

Events

| Symbol Name | Event Code | Ctrs | Max Inc/ Cyc | Description |
|---|---|---|---|---|
| CLOCKTICKS | 0x00 | 0-3 | 0 | Uncore Clocks |
| RxC_OCCUPANCY | 0x11 | 0 | 0 | Ingress (from CMS) Occupancy |
| RxC_INSERTS | 0x13 | 0-3 | 0 | Ingress (from CMS) Allocations |
| CORE_PMA | 0x17 | 0-3 | 0 | Core PMA Events |
| RxC_IRQ0_REJECT | 0x18 | | 0 | Ingress (from CMS) Request Queue Rejects |
| RxC_IRQ1_REJECT | 0x19 | | 0 | Ingress (from CMS) Request Queue Rejects |
| COUNTER0_OCCUPANCY | 0x1f | 0-3 | 0 | Counter 0 Occupancy |
| RxC_PRQ0_REJECT | 0x20 | | 0 | Ingress (from CMS) Request Queue Rejects |
| RxC_PRQ1_REJECT | 0x21 | | 0 | Ingress (from CMS) Request Queue Rejects |
| RxC_IPQ0_REJECT | 0x22 | | 0 | Ingress Probe Queue Rejects |
| RxC_IPQ1_REJECT | 0x23 | | 0 | Ingress Probe Queue Rejects |
| RxC_ISMQ0_REJECT | 0x24 | | 0 | ISMQ Rejects |

https://software.intel.com/en-us/blogs/2014/07/11/documentation-for-uncore-performance-monitoring-units

# The LLC coherence engine (CHA)

## Data Direct I/O

```
[root@nntvtune206 perf]# ./perf -r 'trtype:PCIe traddr:0000:da:00.0' -q 8 -o 4096 -w write -t 10 -c 0x1000000
Starting SPDK v19.04-pre / DPDK 18.11.0 initialization...
[ DPDK EAL parameters: perf --no-shconf -c 0x1000000 --base-virtaddr=0x200000000000 --file-prefix=spdk_pid10329 ]
EAL: Detected 96 lcore(s)
EAL: Detected 2 NUMA nodes
EAL: No free hugepages reported in hugepages-1048576kB
EAL: Probing VFIO support...
Initializing NVMe Controllers
Attaching to NVMe Controller at 0000:da:00.0
Attached to NVMe Controller at 0000:da:00.0 [8086:2701]
Associating INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) with lcore 24
Initialization complete. Launching workers.
Starting thread on core 24
========================================================
                                                                    Latency(us)
Device Information                                  :       IOPS      MB/s    Average        min        max
INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) from core 24   555166.80   2168.62      14.40       6.70     360.12
========================================================
Total                                               :   555166.80   2168.62      14.40       6.70     360.12

[root@nntvtune206 perf]#
```

| Uncore IRP Events | Counts/sec | Comment |
|---|---|---|
| TOR_INSERTS.PRQ_MISS.RFO | 0 | The LLC/SF/MLC complex were able to service the request without involving Intel UPI and/or any RD/WR to a local memory controller |
| TOR_INSERTS.PRQ_HIT.RFO | 555,721 | |

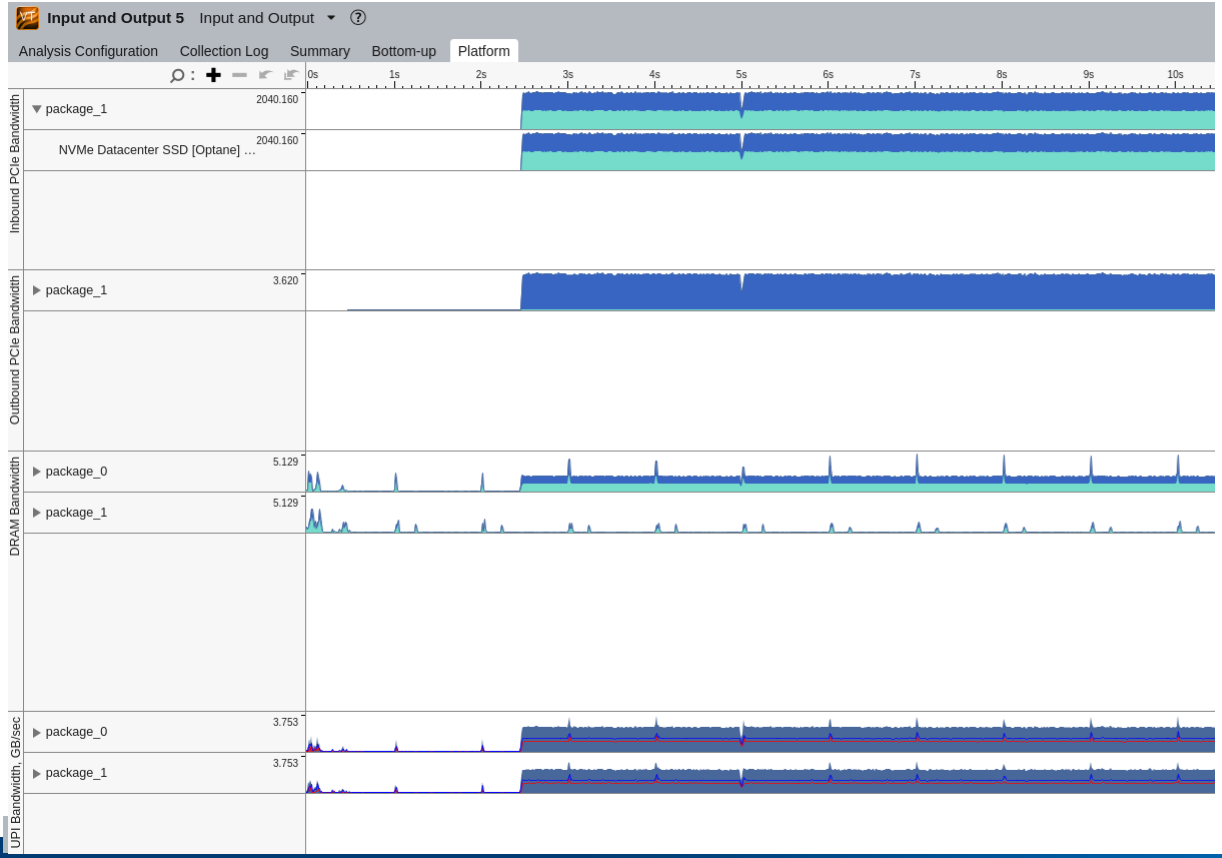
# The LLC coherence engine (CHA)

## Data Direct I/O

```
[root@nntvtune206 perf]# ./perf -r 'trtype:PCIe traddr:0000:da:00.0' -q 8 -o 4096 -w write -t 10
Starting SPDK v19.04-pre / DPDK 18.11.0 initialization...
[ DPDK EAL parameters: perf --no-shconf -c 0x1 --base-virtaddr=0x200000000000 --file-prefix=spdk_pid26733 ]
EAL: Detected 96 lcore(s)
EAL: Detected 2 NUMA nodes
EAL: No free hugepages reported in hugepages-1048576kB
EAL: Probing VFIO support...
Initializing NVMe Controllers
Attaching to NVMe Controller at 0000:da:00.0
Attached to NVMe Controller at 0000:da:00.0 [8086:2701]
Associating INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) with lcore 0
Initialization complete. Launching workers.
Starting thread on core 0
========================================================
                                                                    Latency(us)
Device Information                                  :       IOPS      MB/s    Average        min        max
INTEL SSDPED1K375GA  (PHKS733500A1375AGN  ) from core 0:  556001.10    171.88      14.38       7.06   12670.69
========================================================
Total                                               :  556001.10    171.88      14.38       7.06   12670.69

[root@nntvtune206 perf]#
```

| Uncore IRP Events | Counts/sec | Comment |
|---|---|---|
| TOR_INSERTS.PRQ_MISS.RFO | 555,721 | The LLC/SF/MLC complex were **NOT** able to service the request without involving Intel UPI and/or any RD/WR to a local memory controller |
| TOR_INSERTS.PRQ_HIT.RFO | 0 | |

# Intel VTune





## Uncore Event Count

| Uncore Event Type | Uncore Event Count |
|---|---|
| UNC_IIO_DATA_REQ_BY_CPU.MEM_READ.PART2[UNIT0] | 0 |
| UNC_IIO_DATA_REQ_BY_CPU.MEM_READ.PART2[UNIT1] | 1,894 |
| UNC_IIO_DATA_REQ_BY_CPU.MEM_READ.PART2[UNIT2] | 0 |
| UNC_IIO_DATA_REQ_BY_CPU.MEM_READ.PART2[UNIT3] | 2,431 |
| UNC_IIO_DATA_REQ_BY_CPU.MEM_READ.PART2[UNIT4] | 0 |
| UNC_IIO_DATA_REQ_BY_CPU.MEM_WRITE.PART2[UNIT0] | 0 |
| UNC_IIO_DATA_REQ_BY_CPU.MEM_WRITE.PART2[UNIT1] | 66 |
| UNC_IIO_DATA_REQ_BY_CPU.MEM_WRITE.PART2[UNIT2] | 0 |
| UNC_IIO_DATA_REQ_BY_CPU.MEM_WRITE.PART2[UNIT3] | 8,647,859 |
| UNC_IIO_DATA_REQ_BY_CPU.MEM_WRITE.PART2[UNIT4] | 0 |
| UNC_IIO_DATA_REQ_OF_CPU.MEM_READ.PART2[UNIT0] | 0 |
| UNC_IIO_DATA_REQ_OF_CPU.MEM_READ.PART2[UNIT1] | 0 |
| UNC_IIO_DATA_REQ_OF_CPU.MEM_READ.PART2[UNIT2] | 0 |
| UNC_IIO_DATA_REQ_OF_CPU.MEM_READ.PART2[UNIT3] | 2,473,251,568 |
| UNC_IIO_DATA_REQ_OF_CPU.MEM_READ.PART2[UNIT4] | 0 |
| UNC_IIO_DATA_REQ_OF_CPU.MEM_WRITE.PART2[UNIT0] | 0 |
| UNC_IIO_DATA_REQ_OF_CPU.MEM_WRITE.PART2[UNIT1] | 0 |
| UNC_IIO_DATA_REQ_OF_CPU.MEM_WRITE.PART2[UNIT2] | 0 |
| UNC_IIO_DATA_REQ_OF_CPU.MEM_WRITE.PART2[UNIT3] | 2,421,334,788 |
| UNC_IIO_DATA_REQ_OF_CPU.MEM_WRITE.PART2[UNIT4] | 0 |
| UNC_I_COHERENT_OPS.PCITOM[UNIT0] | 284 |
| UNC_I_COHERENT_OPS.PCITOM[UNIT1] | 0 |
| UNC_I_COHERENT_OPS.PCITOM[UNIT2] | 160 |
| UNC_I_COHERENT_OPS.PCITOM[UNIT3] | 129,309,921 |
| UNC_I_COHERENT_OPS.PCITOM[UNIT4] | 0 |
| UNC_I_COHERENT_OPS.RFO[UNIT0] | 104 |
| UNC_I_COHERENT_OPS.RFO[UNIT1] | 0 |
| UNC_I_COHERENT_OPS.RFO[UNIT2] | 414 |
| UNC_I_COHERENT_OPS.RFO[UNIT3] | 4,200,571 |
| UNC_I_COHERENT_OPS.RFO[UNIT4] | 0 |
| UNC_I_FAF_INSERTS[UNIT0] | 10,097 |
| UNC_I_FAF_INSERTS[UNIT1] | 0 |
| UNC_I_FAF_INSERTS[UNIT2] | 9,163 |
| UNC_I_FAF_INSERTS[UNIT3] | 152,927,863 |
| UNC_I_FAF_INSERTS[UNIT4] | 0 |
| UNC_M_CAS_COUNT.RD[UNIT0] | 94,577,213 |
| UNC_M_CAS_COUNT.RD[UNIT1] | 94,609,076 |
| UNC_M_CAS_COUNT.RD[UNIT2] | 0 |
| UNC_M_CAS_COUNT.RD[UNIT3] | 0 |
| UNC_M_CAS_COUNT.RD[UNIT4] | 0 |

# Recap

## Hardware-Level Performance Analysis of Platform I/O

| Total Bytes Transferred | | | Data B/W | | TLP/sec | | | Average Transfer Size | | | DDIO Usage | | CPU/IO Conflicts |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Read | Write | | Read | Write | P | NP | C | P | NP | C | Hit | Miss | |
| | Full | Partial | | | | | | | | | | | |
| 6798660 | 7863 | 8666790 | 7986660 | 6798066 | 769 | 7968 | 0 | 679 | 7968 | 0 | 4358 | 86766 | Not Detected |

- TLP – Transaction Layer Packet
- P – Posted, NP – NonPosted, C – Completions
- DDIO – Intel Data Direct I/O
- Full/Partial - @64B aligned/not aligned

# Summary

- Building performance tuning methodology on top of server Uncore perfmon continues to be an area of ongoing research and development
- Uncore perfmon events are proved to be really useful for debugging performance of I/O intensive applications, especially in the field
- Intel VTune Amplifier forms all-in-one approach putting the all necessary information in one place for developers to analyze various CPU/IO interactions

## Contacts


– Roman Sudarikov roman.sudarikov@intel.com

# Reference Information

- Documents and links
  - SDM w/PMU chapter – https://www-ssl.intel.com/content/www/us/en/architecture-and-technology/64-ia-32-architectures-software-developer-vol-3a-part-1-manual.html
  - List of publically available Uncore PMON docs – JKT -> SKX - https://software.intel.com/en-us/blogs/2014/07/11/documentation-for-uncore-performance-monitoring-units
  - Intel VTune Amplifier - https://software.intel.com/en-us/vtune
  - “Intel® Xeon® Scalable Family (Purley) Platform Integrated I/O Performance Guide”, ref#598895