Intel® Open Network Platform Release 2.1 Application Note on Resource Director Technology

SDN/NFV Solutions with Intel® Open Network Platform

Document Revision 1.0
March 2016

Revision History

Date	Revision	Comments
March 31, 2016	1.0	Initial document for release of Intel® Open Network Platform Release 2.1

Contents

1.0 Audience and Purpose 4
1.1 Summary 4
1.2 Intel® RDT Technology Overview 4
1.2.1 CMT and CAT 4
1.2.2 Intel® Xeon® Processor SKUs that Support CAT and CMT 5
1.2.3 The intel-cmt-cat Software Package 5
2.0 System Configuration 6
2.1 Hardware Components 6
2.2 Software Components 7
3.0 Service Assurance and Predictability 8
3.1 Test Setup 8
3.2 Results 9
3.2.1 Average Latency 10
3.2.2 More predictable Latency 10
3.3 Conclusions 11
Legal Information 14

1.0 Audience and Purpose

Intel® Open Network Platform (Intel® ONP) is a Reference Architecture that provides engineering guidance and ecosystem-enablement support to encourage widespread adoption of Software-Defined Networking (SDN) and Network Functions Virtualization (NFV) solutions in Telco, Enterprise, and Cloud. Intel® ONP is released in the form of a software stack and a set of documents available on 01.org (e.g. Intel® Open Network Platform Reference Architecture Guides, Performance Test Reports).

The primary audience for this application note are architects and engineers interested in previewing Intel® Resource Director Technology (RDT) and how they can use it to achieve service assurance, security, predictable latency and throughput in their NFV deployments with key software ingredients of Intel® ONP, such as Open vSwitch with DPDK and KVM on a Fedora platform.

1.1 Summary

Service assurance, security and predictable latency and throughput performance are system qualities that are important for Telco, Cloud and Enterprise deployments. This document provides a brief description on how they can be greatly improved by using Intel® Resource Director Technology (RDT). Full software enablement in open source projects such as the Linux Kernel, DPDK, OpenStack are still in progress. However, we are taking the opportunity to provide a preview of the Intel RDT technology using `pqos` utility in deployments with DPDK, OvS and KVM.

1.2 Intel® RDT Technology Overview

Intel Resource Director Technology (RDT) is a set of Intel technologies, namely Cache Monitoring Technology (CMT), Memory Bandwidth Monitoring (MBM), Cache Allocation Technology (CAT) and Code and Data Prioritization (CDP) Technology that provide the hardware framework to monitor and control the utilization of shared resources, like Last Level Cache (LLC) and main (DRAM) memory bandwidth. As multithreaded and multicore platform architectures continue to evolve, running workloads in single-threaded, multithreaded, or complex virtual machine environment such as in Network Functions Virtualization (NFV), the last level cache and memory bandwidth are key resources to manage and utilize based on the nature of workloads. Intel introduces CMT, MBM, CAT and CDP to manage these various workloads across shared resources.

Although this document strictly focuses on CAT and CMT, the reader can find more details on all of the aforementioned technologies and RDT in general in Appendix B: References.

1.2.1 CMT and CAT

Cache Monitoring Technology is a feature that allows an operating system (OS) or hypervisor or virtual machine monitor (VMM) to determine the cache usage by applications running on the platform. CMT can be used to do the following:

- Detect if the platform supports CMT monitoring capabilities via CPUID.
- Have the OS or VMM assign an ID for each application or VM that are scheduled to run on a core. This ID is called the Resource Monitoring ID (RMID).
- To monitor cache occupancy and memory bandwidth on a per-RMID basis.
- For an OS or VMM to read LLC occupancy and memory bandwidth for a given RMID at any time.

Cache Allocation Technology is a feature that allows an OS, hypervisor, or VMM to control allocation of a CPU's shared Last-Level Cache (LLC). Once CAT is configured, the processor allows access to portions of the cache according to the established Class Of Service (COS). The processor obeys the COS rules when it runs an application thread or application process. This can be accomplished by performing these steps:

- Determine if the CPU supports the CAT and CDP feature.
- Configure the COS to define the amount of resources (cache space) available. This configuration is at the processor level and is common to all logical processors.
- Associate each logical processor with an available COS.
- Run the application on the logical processor that uses the desired COS.

1.2.2 Intel® Xeon® Processor SKUs that Support CAT and CMT

Following SKUs of Intel® Xeon® processors support both CAT and CMT:

- Intel® Xeon® processor E5-2658 v3
- Intel® Xeon® processor E5-2658A v3
- Intel® Xeon® processor E5-2648L v3
- Intel® Xeon® processor E5-2628L v3
- Intel® Xeon® processor E5-2618L v3
- Intel® Xeon® processor E5-2608L v3
- All SKUs of Intel® Xeon® processor D product family
- All SKUs of Intel® Xeon® processor E5-2600 v4 product family.

1.2.3 The intel-cmt-cat Software Package

The intel-cmt-cat is a software package that provides basic support for Cache Monitoring Technology (CMT), Memory Bandwidth Monitoring (MBM), Cache Allocation Technology (CAT) and Code and Data Prioritization (CDP) Technology, and includes the `pqos` utility. Refer to https://github.com/01org/intel-cmt-cat for specific information with regard to CMT, MBM, CAT and CDP software package details.

After compilation the `pqos` executable can be used to configure the last level cache allocation feature and monitor the last level cache occupancy as well as memory bandwidth.

The `pqos` utility can be used by typing the following commands:

- `./pqos -h`
  This option will display extensive help page. Please refer to "-h" option for usage details.
- `./pqos -s`
  Shows current CAT, CMT and MBM configuration.
- `./pqos -T`
  Provides top like monitoring output
- `./pqos -f FILE`
  Loads the commands from selected file

2.0 System Configuration

2.1 Hardware Components

Table 2-1 Intel® Xeon® processor E5-2658 v3 platform- hardware ingredients

Item	Description
Server Platform	Intel® Server Board S2600WTT, Formerly Wildcat Pass 2 x PCI-E 3.0 x16 slot, 1 x PCI-E 3.0 x8 slot
Processor	Intel® Xeon® processor E5-2658 v3 2.20 GHz 2 Sockets (NUMA Nodes), 12 cores/Socket, 12 threads, 2.2 GHz, 30 MB Last Level Cache
Memory	64 GB 1600MHZ DDR3L ECC CL11 SODIMM 1.35V
BIOS	SE5C610.86B.01.01.0011.081020151200 Hyper-Threading: Disabled Intel® Virtualization Technology (Intel® VT-x): Enabled Intel® Virtualization Technology for Directed I/O (Intel® VT-d): Disabled C-State: Disabled CPU P-State Control Enhanced Intel® SpeedStep® Technology: Disabled Fan PWM Offset: 100 CPU Power and Performance Policy: Performance QPI/DMI: Auto
Network Interfaces	1 x Intel® Ethernet X710-DA4 Adapter (Total: 4 Ports) http://ark.intel.com/products/83965/Intel-Ethernet-Converged-Network-Adapter-X710-DA4 Tested with Intel® FTLX8571D3BCV-IT transceivers
Local Storage	Intel® SSD DC S3500 Series Formerly Wolfsville SSDSC2BB120G4 120 GB SSD 2.5in SATA 6 Gb/s
Security options	UEFI Secure Boot

2.2 Software Components

The Table 2-2 describes functions of the software ingredients along with their version or configuration. For open-source components, a specific commit ID set is used for this integration. Note that the commit IDs detailed in the table are used as they are the latest working set at the time of this release.

Table 2-2 Software Versions

Software Component	Function	Version/Configuration
Fedora 23	Host Operating System	Fedora 23 Server x86_64
KVM4NFV Kernel	Host Operating System Kernel	4.1.10-rt10, KVM4NFV kernel, Bramhaputra.1.0 tag
Fedora 20	Guest Operating System Kernel	Real Time kernel: 3.14.16-rt34-M_L_E_X2-VM
QEMU-KVM	Virtualization technology	QEMU-KVM version: 2.3.1-7.fc22.x86_64 libvirt version: 1.2.13.1-3.fc22.x86_64
DPDK	Network stack bypass and libraries for packet processing; includes user space vhost drivers	DPDK Release 2.2.0 frozen at a38e5ec15e3fe615b94f3cc5edca5974dab325ab
Open vSwitch	vSwitch	OvS Release 2.5.0, frozen at 61c4e39460a7db3be7262a3b2af767a84167a9d8 used for Open vSwitch 2.5 with DPDK
Cache Allocation Technology (CAT) / Cache Monitoring Technology (CMT)	Resource Director Technology components	intel-cmt-cat commit id: 1c473f93a2639f9d564ed7869bd1ef7a725bd513
Intel® Ethernet Drivers	Ethernet drivers	Driver Version: i40e 1.4.25 Firmware Version: 5.02 0x80002284 0.0.0

3.0 Service Assurance and Predictability

This section describes how the technology can be used to achieve reduced and more predictable latency in a VM with a (simulated) noisy neighbor. It might be worth mentioning that the noisy neighbor can be either another VM/VNF that just happens to generate a lot of traffic, or it can be a malicious neighbor trying to starve the forwarding VM from resources such as cache and affect its performance. In this case it is shown below that CAT can protect and isolate VMs/VNFs from each other avoiding the negative effects of an intentionally malicious (or not) noisy neighbor.

3.1 Test Setup

The test setup for measuring the performance (latency and throughput) of the VM0 is shown in Figure 3-1. This setup shows an OvS with DPDK-based deployment with the VM under test (VM0) that does Layer 2 packet forwarding while the second VM (VM1) is running memtester continuously, simulating a noisy neighbor which would create cache thrashing, affecting the performance of the packet forwarding VM0 because of the forced cache evictions. The Figure 3-1 shows which core is running what process, and the Class of Service (COS) that has been assigned to it/them.

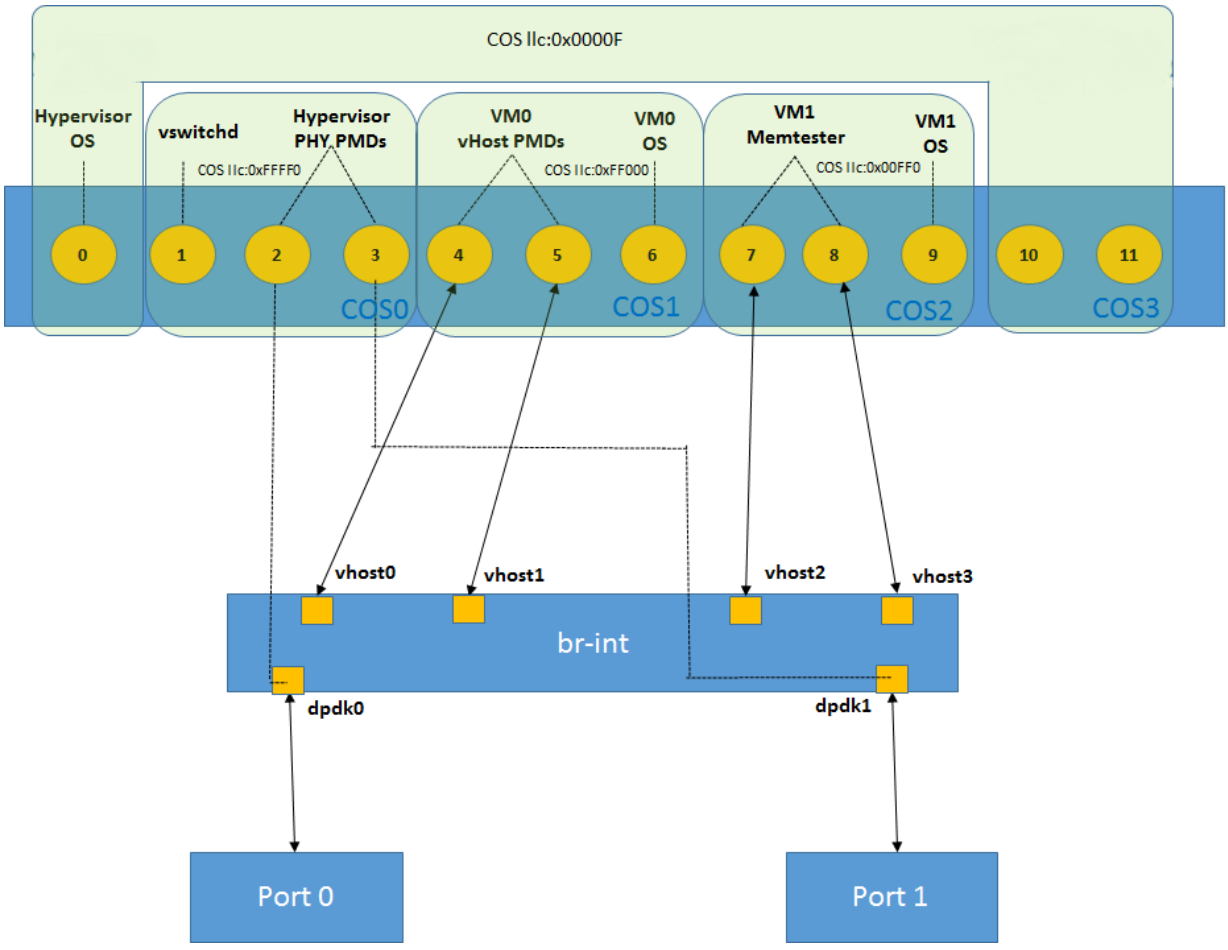

Figure 3-1 Demonstration of predictable latency with simulated noisy neighbor

The following classes of service have been defined and given to the corresponding cores, so that:

- Each VM has non-overlapping cache ways, COS llc:1 for VM0 and COS llc:2 for VM1
- Both VMs have overlapping cache ways with the larger COS that is shared between the VMs and the OvS cores (llc:0). OvS Cores includes cores dedicated for vSwitchd process and the two PMD threads.
- The rest of the cores on that socket that run OS related threads will use the remaining non-overlapping cache ways

Table 3-1 Definition of classes of service

```
alloc-class-set: llc:0=0xffff0;llc:1=0xff000;llc:2=0x00ff0; llc:3=0x0000f
alloc-assoc-set: llc:0=1-3
alloc-assoc-set: llc:1=4-6
alloc-assoc-set: llc:2=7-9
alloc-assoc-set: llc:3=0,10,11
```

The COS settings can be set using the below steps:

1. Save the COS definition from Table 3-1 to .cfg configuration file, for example onp_cat.cfg.
2. Modify the platform COS settings using the `pqos` utility and configuration file:
   ```
   # ./pqos -f onp_cat.cfg
   ```
3. Verify that the COS settings are updated using:
   ```
   # ./pqos -s -v
   ```
4. The LLC usage can be monitored per core basis using:
   ```
   # ./pqos -r -T
   ```
   It can also be monitored per group core / VM, as described in https://github.com/01org/intel-cmt-cat/wiki/Usage-Examples#example-cmtmbm-usage-scenario.

The VM under test, VM0 uses DPDK based application, `testpmd`, to do the Layer 2 forwarding with two PMD threads using following settings:

```
# ./testpmd -c 0x3 -n 4 -socket-mem 1024 -- --burst=64 -i --txd=2048 \
      --rxd=2048 -txqflags=0xf00 -disable-hw-vlan
```

The second VM, VM1 uses `memtester`, a Linux utility, to act as noisy neighbor using following settings:

```
# ./memtester 100M > /dev/null
```

Memtester utilizes 100% of two cores in the VM. The QEMU threads of VM1 that use 100% CPU are affinitized to two isolated cores out of the three cores available in the VM using `taskset`. This simulates a noisy neighbor VM that has two PMDs utilizing two cores 100%.

3.2 Results

The results below show the average latencies of the forwarding VM0 for different packet sizes. In both test runs shown below `memtester` was ran in a second VM1 (simulating a noisy neighbor):

- NoCAT - is a test run without using CAT.
- withCAT - is a test run using CAT.

All the tests were run using RFC 2544 with 0.1% acceptable packet loss, and then acquiring the latency numbers at the highest acceptable throughput rate using IxNetwork from Ixia. The methodology uses 2000 bidirectional flows across the two Physical ports of the hypervisor. The COS settings defined in Table 3-1 were used.

3.2.1 Average Latency

Latency in forwarding VM0 is negatively impacted when `memtester` is run in the noisy neighbor VM1 and CAT is not invoked. When the `pqos` utility is run and CAT gets utilized (see Table 3-1), the latency of the forwarding VM0 stays low. The X axis represents frame sizes while the Y axis represents average latency in µs.

Figure 3-2 Demonstration of reduced average latency with noisy neighbor

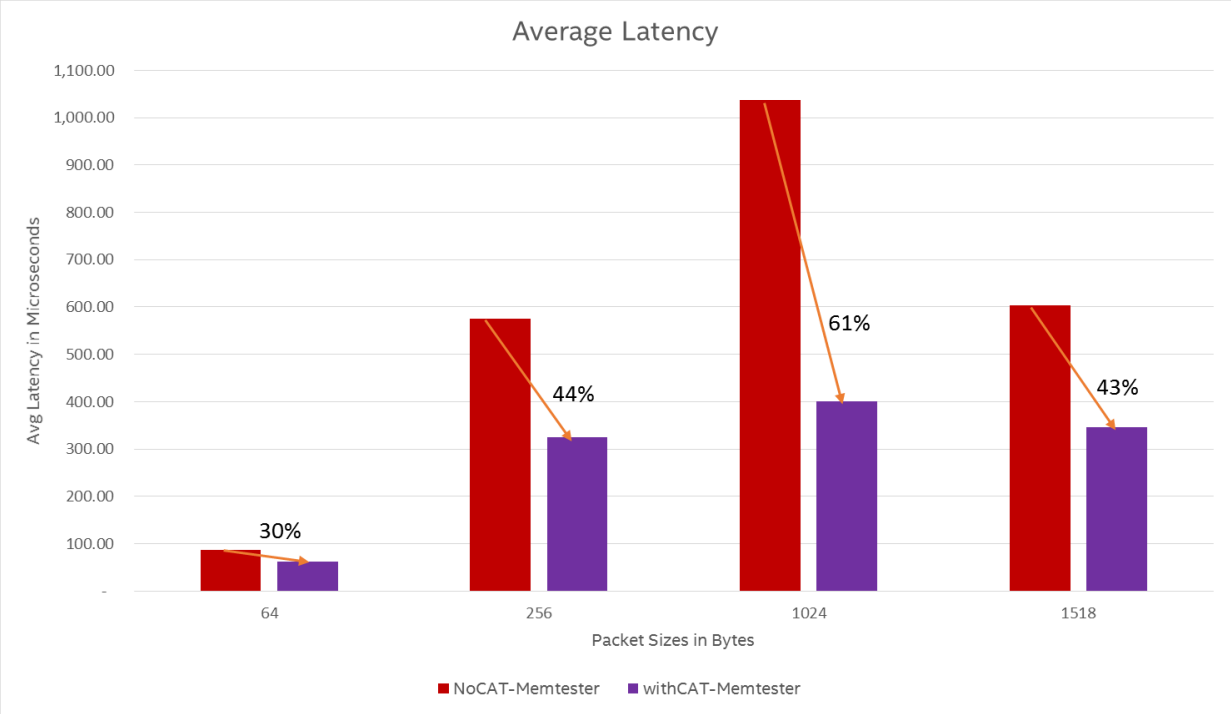

Table 3-2 Results for average latency with and without CAT

Packet Size	Avg Latency without CAT [µs]	Avg Latency with CAT [µs]	Delta
64	87.06	61.36	30%
256	576.11	324.62	44%
1024	1,036.62	400.09	61%
1512	603.15	346.13	43%

3.2.2 More Predictable Latency

The results show that using CAT leads not only to reduced average latency, but also more predictable latency as depicted in Figure 3-3. In this figure the total number of packets equal to about 700 Million packets (aggregate among all packet sizes, from 64B up to 1512B). When running memtester without CAT (simulating a noisy neighbor), the main concentration of the packets are shown in the 75-100µs, 500-750µs and >1000µs buckets, but when CAT is introduced there is a clear shift to lower latency buckets, with main concentration of the packets in the 50µs-75µs and 250-500µs ones. This proves that using Intel RDT will help latency sensitive applications executing in a Virtual Network Functions environment provide a more predictable latency.

Figure 3-3 Demonstration of better predictable latency with noisy neighbor

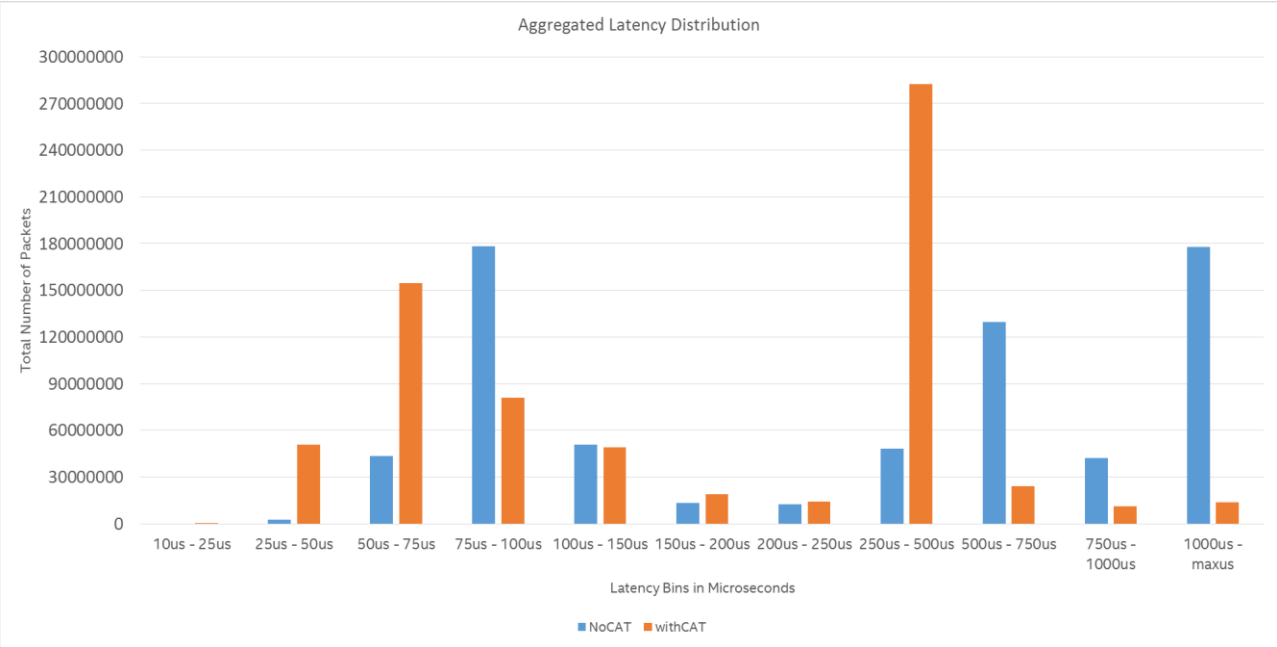

Table 3-3 Latency distribution of approximately 700 Million injected packets aggregated in ranges

Latency Bins	NoCAT	withCAT
min - 25µs	1	116
25µs - 50µs	2,579,899	50,758,603
50µs - 75µs	43,603,792	154,478,784
75µs - 100µs	178,374,563	80,978,896
100µs - 150µs	51,028,651	48,901,761
150µs - 200µs	13,462,738	19,074,613
200µs - 250µs	12,688,404	14,032,668
250µs - 500µs	48,264,741	282,393,165
500µs - 750µs	129,556,913	24,143,727
750µs - 1000µs	42,281,092	11,116,133
1000µs - max	177,818,136	13,955,846

3.3 Conclusions

Intel Resource Director Technology, and in this example CAT, can provide a better platform service assurance. By using CAT data access latency is reduced and also becomes more predictable. Intel® Resource Director Technology also has a strong security aspect as VMs can be safeguarded from malicious noisy neighbors which deploy denial of service (DoS) attack techniques and starve legitimate VMs from their cache resources.

Appendix A: Acronyms and Abbreviations

Abbreviation	Description
CAT	Cache Allocation Technology
CDP	Code and Data Prioritization Technology
CMT	Cache Monitoring Technology
COS	Class of Service
DoS	Denial of Service
DPDK	Data Plane Development Kit
Intel® ONP	Intel® Open Network Platform
KVM	Kernel-based Virtual Machine
LLC	Last Level Cache
MBM	Memory Bandwidth Monitoring
NFV	Network Functions Virtualization
OS	Operating System
OvS	Open vSwitch
QoS	Quality of Service
RDT	Resource Director Technology
RMID	Resource Monitoring ID
VM	Virtual Machine
VMM	Virtual Machine Monitor

Appendix B: References

Reference	Location
Intel® 64 and IA-32 Architectures Software Developer's Manuals	http://www.intel.com/content/www/us/en/processors/architectures-software-developer-manuals.html v055, Vol 3b. Chapter 17.15 and 17.16, covers CMT, CAT, MBM and CDP
Intel, Cache Monitoring and Cache Allocation Technologies landing page	http://www.intel.com/content/www/us/en/communications/cache-monitoring-cache-allocation-technologies.html
CMT, MBM, CAT and CDP public software library/utility	https://01.org/packet-processing/cache-monitoring-technology-memory-bandwidth-monitoring-cache-allocation-technology-code-and-data
CMT, MBM, CAT and CDP public software library/utility GitHub Project	https://github.com/01org/intel-cmt-cat
Intel, "Enabling NFV to Deliver on its Promise"	http://www.intel.com/content/www/us/en/communications/nfv-packet-processing-brief.html
CAT cgroup kernel patches	http://marc.info/?l=linux-kernel&m=142620227328406&w=2
Christos Kozyrakis et al, "Heracles: Improving Resource Efficiency at Scale"	http://csl.stanford.edu/~christos/publications/2015.heracles.isca.pdf, 2015
Introduction to CMT Blog	https://software.intel.com/en-us/blogs/2014/06/18/benefit-of-cache-monitoring
Discussion of RMIDs and CMT Software Interfaces Blog	https://software.intel.com/en-us/blogs/2014/12/11/intel-s-cache-monitoring-technology-software-visible-interfaces
Use Models and Example Data using CMT Blog	https://software.intel.com/en-us/blogs/2014/12/11/intels-cache-monitoring-technology-use-models-and-data
Software Supports and Tools: Intel's CMT: Software Support and Tools	https://software.intel.com/en-us/blogs/2014/12/11/intels-cache-monitoring-technology-software-support-and-tools
Intel Platform Shared Resource Monitoring and CAT	http://smackerelofopinion.blogspot.com/2015/11/intel-platform-shared-resource.html
Intel, "Increasing Platform Determinism with Platform Quality of Service for the Data Plane Development Kit"	http://www.intel.com/content/www/us/en/communications/increasing-platform-determinism-pqos-dpdk-white-paper.html

Legal Information

By using this document, in addition to any agreements you have with Intel, you accept the terms set forth below.

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.

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

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.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request. Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Intel technologies may require enabled hardware, specific software, or services activation. Check with your system manufacturer or retailer. 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.

All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. 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.

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.

Intel does not control or audit third-party web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Intel Corporation may have patents or pending patent applications, trademarks, copyrights, or other intellectual property rights that relate to the presented subject matter. The furnishing of documents and other materials and information does not provide any license, express or implied, by estoppel or otherwise, to any such patents, trademarks, copyrights, or other intellectual property rights.