GOING BEYOND CPU UTILIZATION WITH PROCESSOR COUNTER MONITOR

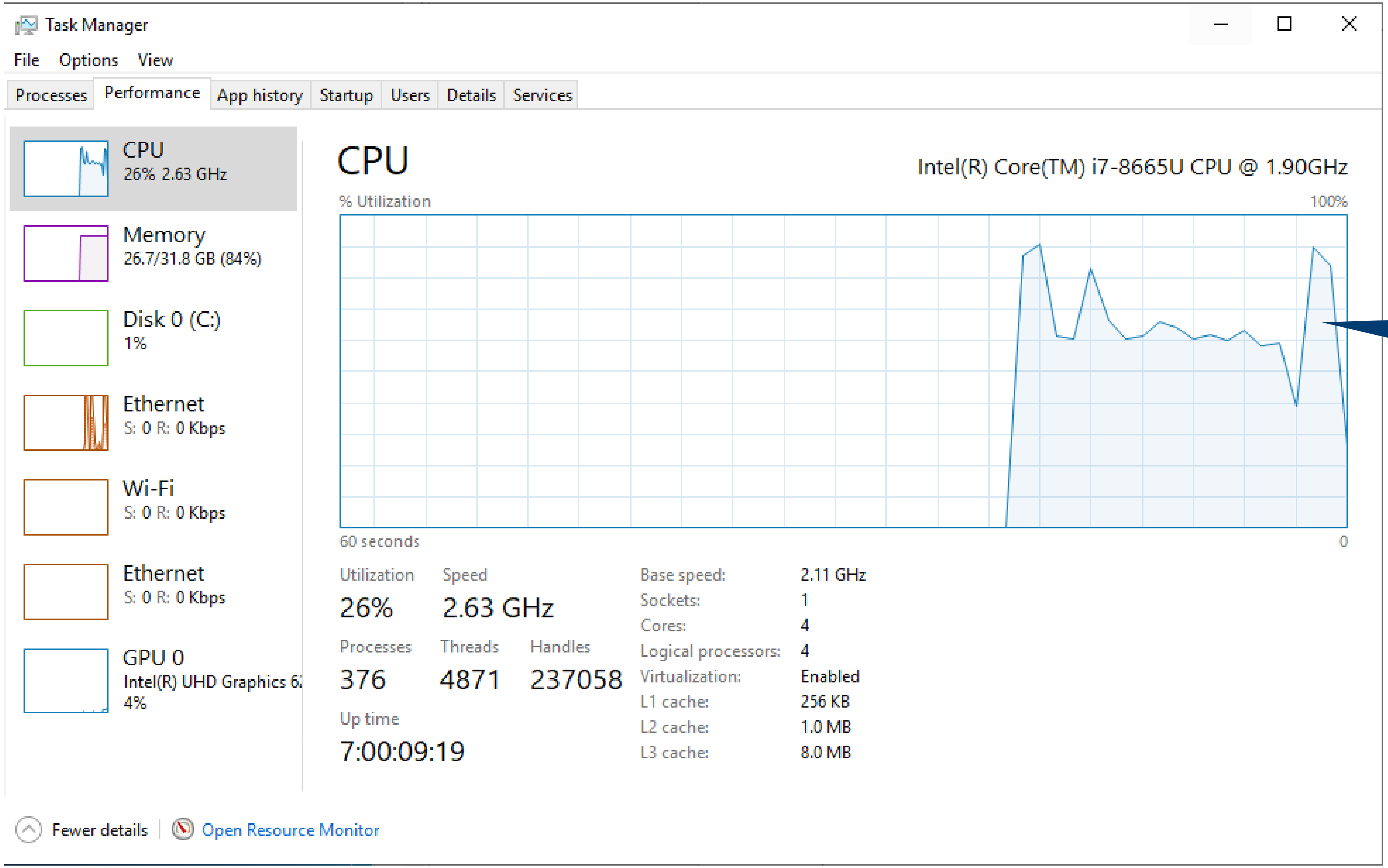
16-July-2020

Roman Dementiev on behalf PCM team

Outline

- What is PCM
- PCM Architecture
- PCM Tools

CPU Utilization



CPU
26% 2.63 GHz

Memory
26.7/31.8 GB (84%)

Disk 0 (C:)
1%

Ethernet
S: 0 R: 0 Kbps

Wi-Fi
S: 0 R: 0 Kbps

Ethernet
S: 0 R: 0 Kbps

GPU 0
Intel(R) UHD Graphics 62
4%

CPU

Intel(R) Core(TM) i7-8665U CPU @ 1.90GHz

Utilization 26%

Speed 2.63 GHz

Processes 376

Threads 4871

Handles 237058

Up time 7:00:09:19

Base speed:	2.11 GHz
Sockets:	1
Cores:	4
Logical processors:	4
Virtualization:	Enabled
L1 cache:	256 KB
L2 cache:	1.0 MB
L3 cache:	8.0 MB

what parts of processor how much utilized?

What is Processor Counter Monitor (PCM)

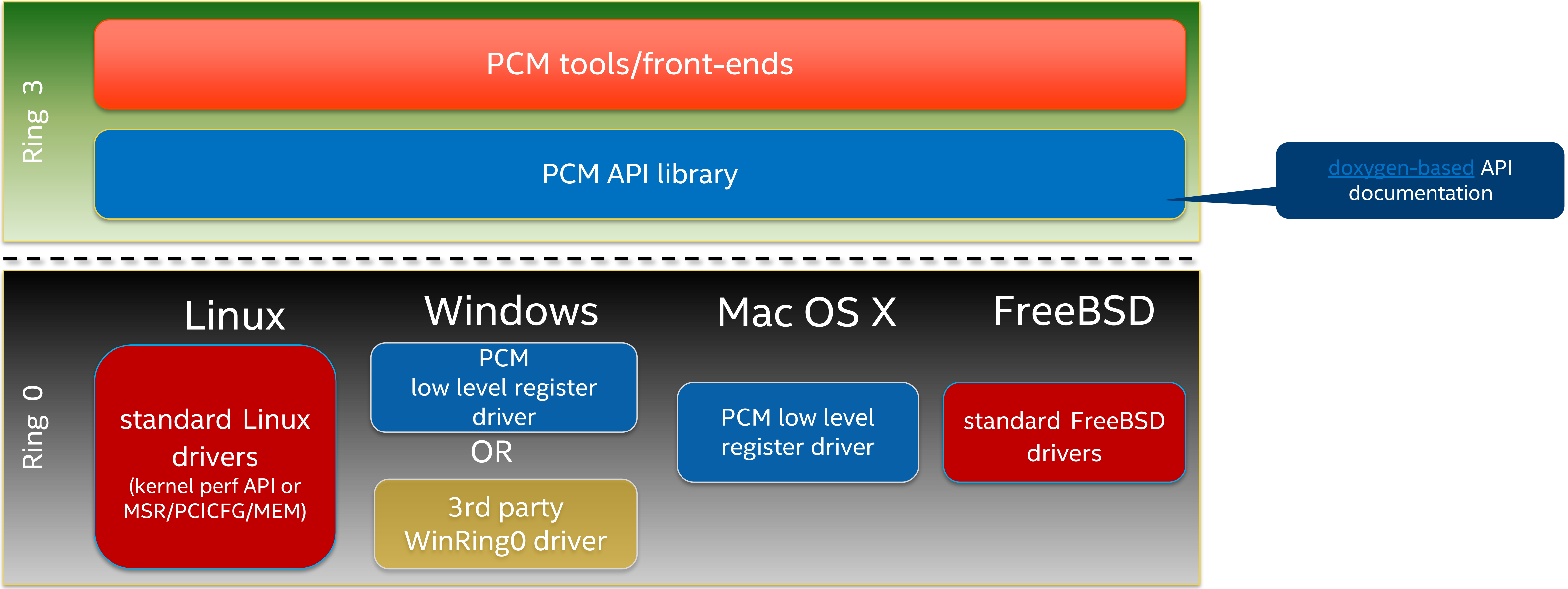
- Real-time tools and API exposing CPU statistics
- Open-source
- Xeon + Xeon Phi + Core + Atom
- Linux, FreeBSD, Windows, Apple OS X

- Ease of use – like the task manager/UNIX top: just run the binary and it will directly report the most common platform metrics (**cycles per instruction, cache misses, UPI and memory bandwidth**, etc)
- Real-time: No post-processing needed. You can watch what the system is doing while the test is running. For documenting and deeper analysis, the CSV output is used, but for a “quick test” post-processing is an additional burden.

PCM distribution

- Public open-source version: https://github.com/opcm/pcm
- Binaries:
  - Windows: https://ci.appveyor.com/project/opcm/pcm/history
  - Linux RPM: https://download.opensuse.org/repositories/home:/opcm/
  - PCM Docker server image:
    - https://github.com/opcm/pcm/blob/master/DOCKER_README.md
- Current PCM stats (July 2020):
  - >15000 downloads/month (github + dockerhub, does not include RPM and Windows)
  - 220 forks on github

PCM architecture



Ring 3

PCM tools/front-ends

PCM API library

doxygen-based API documentation

Ring 0

Linux

standard Linux drivers (kernel perf API or MSR/PCICFG/MEM)

Windows

PCM low level register driver

OR

3rd party WinRing0 driver

Mac OS X

PCM low level register driver

FreeBSD

standard FreeBSD drivers

PCM command line real-time utilities

pcm : basic processor monitoring utility (instructions per cycle, core frequency (including Intel(r) Turbo Boost Technology), mem ory and Intel(r) Quick Path Interconnect/Ultra Path Interconnect bandwidth, local and remote memory bandwidth, cache misses, core and CPU package sleep C-state residency, core and CPU package thermal headroom, cache utilization, CPU and memory energy consumption)

pcm-memory : monitor memory bandwidth (per-channel and per-DRAM DIMM rank)

pcm-latency : monitor L1 cache miss and DDR/PMM memory latency

pcm-pcie : monitor PCIe bandwidth per-socket

pcm-iio : monitor PCIe bandwidth per PCIe device

pcm-numa : monitor local and remote memory accesses

pcm-power : monitor sleep and energy states of processor, Intel(r) Quick Path Interconnect, DRAM memory, reasons of CPU frequency throttling and other energy-related metrics

pcm-tsx: monitor performance metrics for Intel(r) Transactional Synchronization Extensions

pcm-sensor-server : pcm collector exposing metrics over http in JSON or Prometheus (text based) format

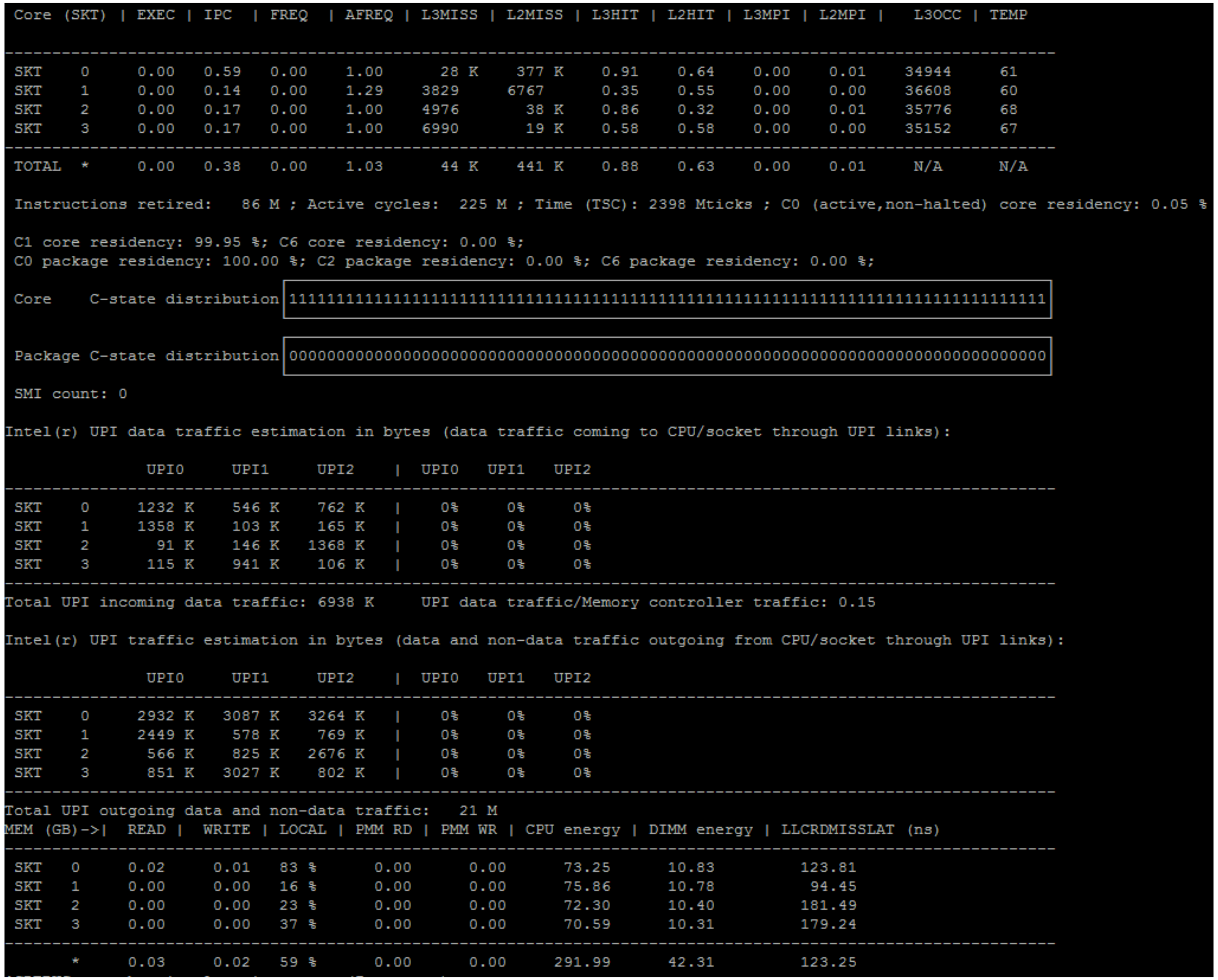
pcm daemon: pcm collector exposing metrics over shared memory (inter-process communication)

pcm-core and pmu-query: query and monitor arbitrary processor core events

pcm-bw-histogram: collect memory bandwidth utilization histogram

pcm-msr/pcm-pcicfg: cross-platform register access utilities

pcm



```
 Core (SKT) | EXEC | IPC  | FREQ  | AFREQ | L3MISS | L2MISS | L3HIT | L2HIT | L3MPI | L2MPI |   L3OCC | TEMP

---------------------------------------------------------------------------------------------------------------
 SKT    0     0.00   0.59   0.00    1.00      28 K    377 K    0.91    0.64    0.00    0.01    34944     61
 SKT    1     0.00   0.14   0.00    1.29    3829     6767      0.35    0.55    0.00    0.00    36608     60
 SKT    2     0.00   0.17   0.00    1.00    4976       38 K    0.86    0.32    0.00    0.01    35776     68
 SKT    3     0.00   0.17   0.00    1.00    6990       19 K    0.58    0.58    0.00    0.00    35152     67
---------------------------------------------------------------------------------------------------------------
 TOTAL  *     0.00   0.38   0.00    1.03      44 K    441 K    0.88    0.63    0.00    0.01     N/A      N/A

 Instructions retired:   86 M ; Active cycles:  225 M ; Time (TSC): 2398 Mticks ; C0 (active,non-halted) core residency: 0.05 %

 C1 core residency: 99.95 %; C6 core residency: 0.00 %;
 C0 package residency: 100.00 %; C2 package residency: 0.00 %; C6 package residency: 0.00 %;

 Core    C-state distribution 11111111111111111111111111111111111111111111111111111111111111111111111111111111

 Package C-state distribution 00000000000000000000000000000000000000000000000000000000000000000000000000000000

 SMI count: 0

Intel(r) UPI data traffic estimation in bytes (data traffic coming to CPU/socket through UPI links):

               UPI0     UPI1     UPI2     | UPI0   UPI1   UPI2
---------------------------------------------------------------------------------------------------------------
 SKT    0     1232 K    546 K    762 K    |   0%     0%     0%
 SKT    1     1358 K    103 K    165 K    |   0%     0%     0%
 SKT    2       91 K    146 K   1368 K    |   0%     0%     0%
 SKT    3      115 K    941 K    106 K    |   0%     0%     0%
---------------------------------------------------------------------------------------------------------------
Total UPI incoming data traffic: 6938 K     UPI data traffic/Memory controller traffic: 0.15

Intel(r) UPI traffic estimation in bytes (data and non-data traffic outgoing from CPU/socket through UPI links):

               UPI0     UPI1     UPI2     | UPI0   UPI1   UPI2
---------------------------------------------------------------------------------------------------------------
 SKT    0     2932 K   3087 K   3264 K    |   0%     0%     0%
 SKT    1     2449 K    578 K    769 K    |   0%     0%     0%
 SKT    2      566 K    825 K   2676 K    |   0%     0%     0%
 SKT    3      851 K   3027 K    802 K    |   0%     0%     0%
---------------------------------------------------------------------------------------------------------------
Total UPI outgoing data and non-data traffic:   21 M
MEM (GB)->|  READ |  WRITE | LOCAL | PMM RD | PMM WR | CPU energy | DIMM energy | LLCRDMISSLAT (ns)
---------------------------------------------------------------------------------------------------------------
 SKT   0     0.02     0.01   83 %      0.00      0.00      73.25      10.83         123.81
 SKT   1     0.00     0.00   16 %      0.00      0.00      75.86      10.78          94.45
 SKT   2     0.00     0.00   23 %      0.00      0.00      72.30      10.40         181.49
 SKT   3     0.00     0.00   37 %      0.00      0.00      70.59      10.31         179.24
---------------------------------------------------------------------------------------------------------------
       *     0.03     0.02   59 %      0.00      0.00     291.99      42.31         123.25
```

instruction per cycle, cache hits/misses, cache usage temp headroom

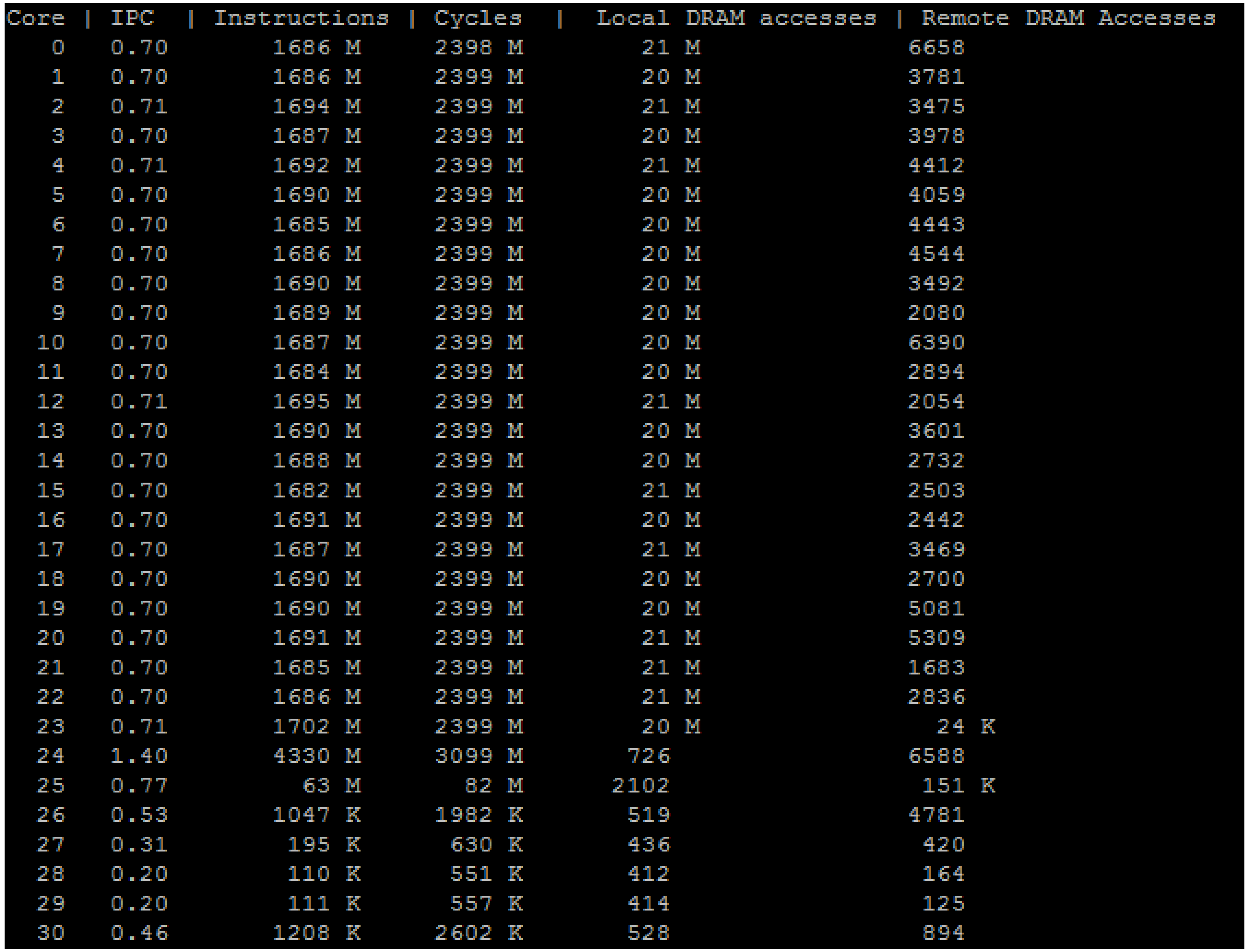
C-state core and package (sleep states)

UPI (cross-socket) traffic and link utilization

data and non-data (snoops, protocol overhead)

Consumed memory bandwidth (DRAM and PMEM), locality of access, CPU/memory energy, cache miss latency

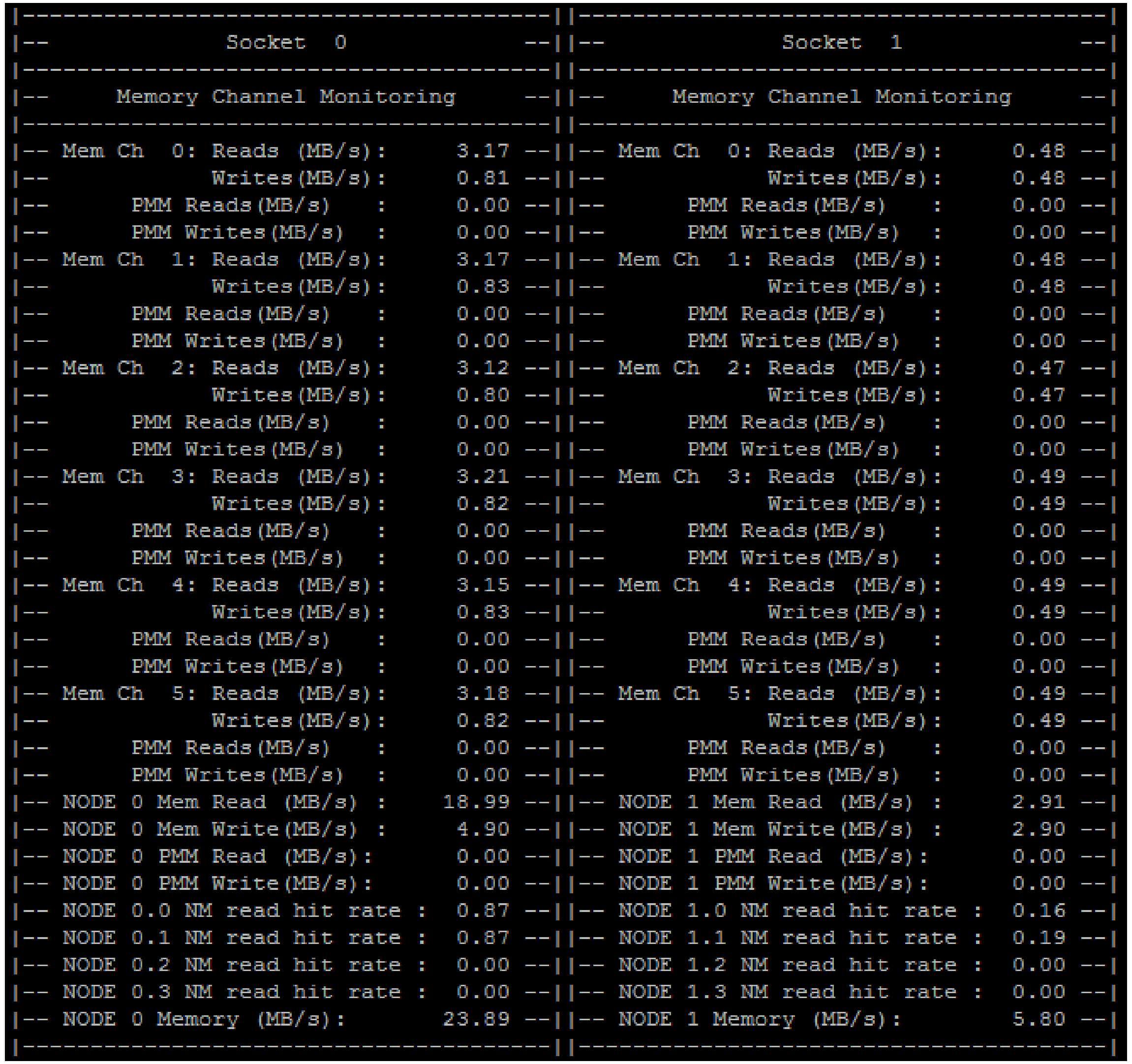
pcm-numa



```
Core | IPC   | Instructions | Cycles  |   Local DRAM accesses | Remote DRAM Accesses
  0   0.70        1686 M      2398 M         21 M                 6658
  1   0.70        1686 M      2399 M         20 M                 3781
  2   0.71        1694 M      2399 M         21 M                 3475
  3   0.70        1687 M      2399 M         20 M                 3978
  4   0.71        1692 M      2399 M         21 M                 4412
  5   0.70        1690 M      2399 M         20 M                 4059
  6   0.70        1685 M      2399 M         20 M                 4443
  7   0.70        1686 M      2399 M         20 M                 4544
  8   0.70        1690 M      2399 M         20 M                 3492
  9   0.70        1689 M      2399 M         20 M                 2080
 10   0.70        1687 M      2399 M         20 M                 6390
 11   0.70        1684 M      2399 M         20 M                 2894
 12   0.71        1695 M      2399 M         21 M                 2054
 13   0.70        1690 M      2399 M         20 M                 3601
 14   0.70        1688 M      2399 M         20 M                 2732
 15   0.70        1682 M      2399 M         21 M                 2503
 16   0.70        1691 M      2399 M         20 M                 2442
 17   0.70        1687 M      2399 M         21 M                 3469
 18   0.70        1690 M      2399 M         20 M                 2700
 19   0.70        1690 M      2399 M         20 M                 5081
 20   0.70        1691 M      2399 M         21 M                 5309
 21   0.70        1685 M      2399 M         21 M                 1683
 22   0.70        1686 M      2399 M         21 M                 2836
 23   0.71        1702 M      2399 M         20 M                   24 K
 24   1.40        4330 M      3099 M       726                    6588
 25   0.77          63 M        82 M      2102                     151 K
 26   0.53        1047 K      1982 K       519                    4781
 27   0.31         195 K       630 K       436                     420
 28   0.20         110 K       551 K       412                     164
 29   0.20         111 K       557 K       414                     125
 30   0.46        1208 K      2602 K       528                     894
```

The number of local and remote DRAM memory accesses per-core

pcm-memory



```
|---------------------------------------||---------------------------------------|
|--             Socket  0             --||--             Socket  1             --|
|---------------------------------------||---------------------------------------|
|--     Memory Channel Monitoring     --||--     Memory Channel Monitoring     --|
|---------------------------------------||---------------------------------------|
|-- Mem Ch  0: Reads (MB/s):     3.17 --||-- Mem Ch  0: Reads (MB/s):     0.48 --|
|--            Writes(MB/s):     0.81 --||--            Writes(MB/s):     0.48 --|
|--      PMM Reads(MB/s)   :     0.00 --||--      PMM Reads(MB/s)   :     0.00 --|
|--      PMM Writes(MB/s)  :     0.00 --||--      PMM Writes(MB/s)  :     0.00 --|
|-- Mem Ch  1: Reads (MB/s):     3.17 --||-- Mem Ch  1: Reads (MB/s):     0.48 --|
|--            Writes(MB/s):     0.83 --||--            Writes(MB/s):     0.48 --|
|--      PMM Reads(MB/s)   :     0.00 --||--      PMM Reads(MB/s)   :     0.00 --|
|--      PMM Writes(MB/s)  :     0.00 --||--      PMM Writes(MB/s)  :     0.00 --|
|-- Mem Ch  2: Reads (MB/s):     3.12 --||-- Mem Ch  2: Reads (MB/s):     0.47 --|
|--            Writes(MB/s):     0.80 --||--            Writes(MB/s):     0.47 --|
|--      PMM Reads(MB/s)   :     0.00 --||--      PMM Reads(MB/s)   :     0.00 --|
|--      PMM Writes(MB/s)  :     0.00 --||--      PMM Writes(MB/s)  :     0.00 --|
|-- Mem Ch  3: Reads (MB/s):     3.21 --||-- Mem Ch  3: Reads (MB/s):     0.49 --|
|--            Writes(MB/s):     0.82 --||--            Writes(MB/s):     0.49 --|
|--      PMM Reads(MB/s)   :     0.00 --||--      PMM Reads(MB/s)   :     0.00 --|
|--      PMM Writes(MB/s)  :     0.00 --||--      PMM Writes(MB/s)  :     0.00 --|
|-- Mem Ch  4: Reads (MB/s):     3.15 --||-- Mem Ch  4: Reads (MB/s):     0.49 --|
|--            Writes(MB/s):     0.83 --||--            Writes(MB/s):     0.49 --|
|--      PMM Reads(MB/s)   :     0.00 --||--      PMM Reads(MB/s)   :     0.00 --|
|--      PMM Writes(MB/s)  :     0.00 --||--      PMM Writes(MB/s)  :     0.00 --|
|-- Mem Ch  5: Reads (MB/s):     3.18 --||-- Mem Ch  5: Reads (MB/s):     0.49 --|
|--            Writes(MB/s):     0.82 --||--            Writes(MB/s):     0.49 --|
|--      PMM Reads(MB/s)   :     0.00 --||--      PMM Reads(MB/s)   :     0.00 --|
|--      PMM Writes(MB/s)  :     0.00 --||--      PMM Writes(MB/s)  :     0.00 --|
|-- NODE 0 Mem Read (MB/s) :    18.99 --||-- NODE 1 Mem Read (MB/s) :     2.91 --|
|-- NODE 0 Mem Write(MB/s) :     4.90 --||-- NODE 1 Mem Write(MB/s) :     2.90 --|
|-- NODE 0 PMM Read (MB/s):      0.00 --||-- NODE 1 PMM Read (MB/s):      0.00 --|
|-- NODE 0 PMM Write(MB/s):      0.00 --||-- NODE 1 PMM Write(MB/s):      0.00 --|
|-- NODE 0.0 NM read hit rate :  0.87 --||-- NODE 1.0 NM read hit rate :  0.16 --|
|-- NODE 0.1 NM read hit rate :  0.87 --||-- NODE 1.1 NM read hit rate :  0.19 --|
|-- NODE 0.2 NM read hit rate :  0.00 --||-- NODE 1.2 NM read hit rate :  0.00 --|
|-- NODE 0.3 NM read hit rate :  0.00 --||-- NODE 1.3 NM read hit rate :  0.00 --|
|-- NODE 0 Memory (MB/s):       23.89 --||-- NODE 1 Memory (MB/s):        5.80 --|
|---------------------------------------||---------------------------------------|
```

Monitor memory per-channel and per DIMM rank

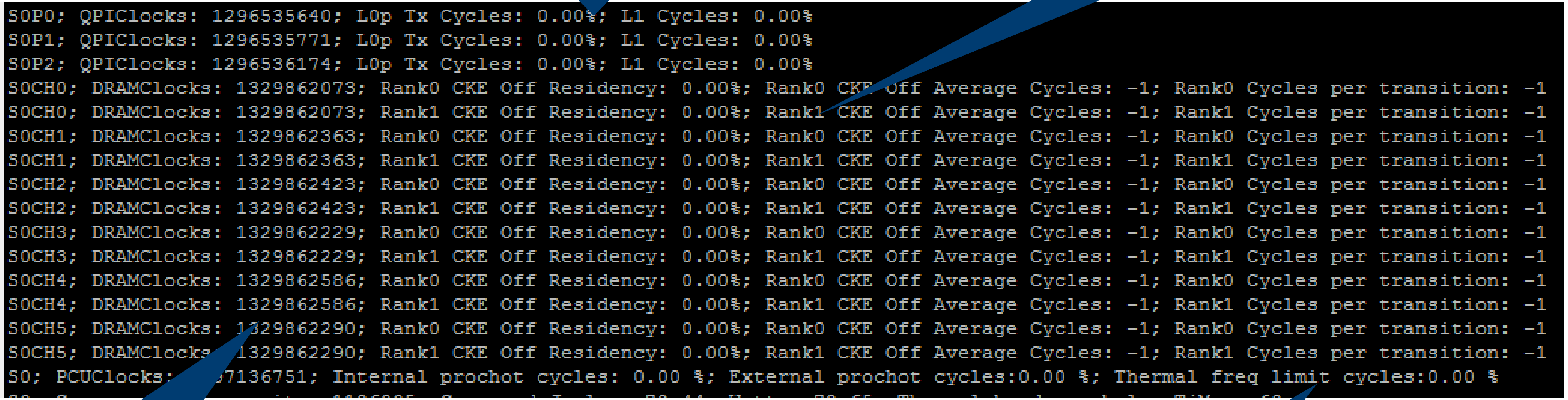
PMem channel bandwidth

PMem „Memory Mode“ DRAM cache hit rate

pcm-power

QPI/UPI L0p/L1 power saving states

DRAM power-saving residency and transition penalty



```
S0P0; QPIClocks: 1296535640; L0p Tx Cycles: 0.00%; L1 Cycles: 0.00%
S0P1; QPIClocks: 1296535771; L0p Tx Cycles: 0.00%; L1 Cycles: 0.00%
S0P2; QPIClocks: 1296536174; L0p Tx Cycles: 0.00%; L1 Cycles: 0.00%
S0CH0; DRAMClocks: 1329862073; Rank0 CKE Off Residency: 0.00%; Rank0 CKE Off Average Cycles: -1; Rank0 Cycles per transition: -1
S0CH0; DRAMClocks: 1329862073; Rank1 CKE Off Residency: 0.00%; Rank1 CKE Off Average Cycles: -1; Rank1 Cycles per transition: -1
S0CH1; DRAMClocks: 1329862363; Rank0 CKE Off Residency: 0.00%; Rank0 CKE Off Average Cycles: -1; Rank0 Cycles per transition: -1
S0CH1; DRAMClocks: 1329862363; Rank1 CKE Off Residency: 0.00%; Rank1 CKE Off Average Cycles: -1; Rank1 Cycles per transition: -1
S0CH2; DRAMClocks: 1329862423; Rank0 CKE Off Residency: 0.00%; Rank0 CKE Off Average Cycles: -1; Rank0 Cycles per transition: -1
S0CH2; DRAMClocks: 1329862423; Rank1 CKE Off Residency: 0.00%; Rank1 CKE Off Average Cycles: -1; Rank1 Cycles per transition: -1
S0CH3; DRAMClocks: 1329862229; Rank0 CKE Off Residency: 0.00%; Rank0 CKE Off Average Cycles: -1; Rank0 Cycles per transition: -1
S0CH3; DRAMClocks: 1329862229; Rank1 CKE Off Residency: 0.00%; Rank1 CKE Off Average Cycles: -1; Rank1 Cycles per transition: -1
S0CH4; DRAMClocks: 1329862586; Rank0 CKE Off Residency: 0.00%; Rank0 CKE Off Average Cycles: -1; Rank0 Cycles per transition: -1
S0CH4; DRAMClocks: 1329862586; Rank1 CKE Off Residency: 0.00%; Rank1 CKE Off Average Cycles: -1; Rank1 Cycles per transition: -1
S0CH5; DRAMClocks: 1329862290; Rank0 CKE Off Residency: 0.00%; Rank0 CKE Off Average Cycles: -1; Rank0 Cycles per transition: -1
S0CH5; DRAMClocks: 1329862290; Rank1 CKE Off Residency: 0.00%; Rank1 CKE Off Average Cycles: -1; Rank1 Cycles per transition: -1
S0; PCUClocks: [illegible]7136751; Internal prochot cycles: 0.00 %; External prochot cycles:0.00 %; Thermal freq limit cycles:0.00 %
```

DRAM speed/2

frequency throttling stats (thermal, current, power, etc)

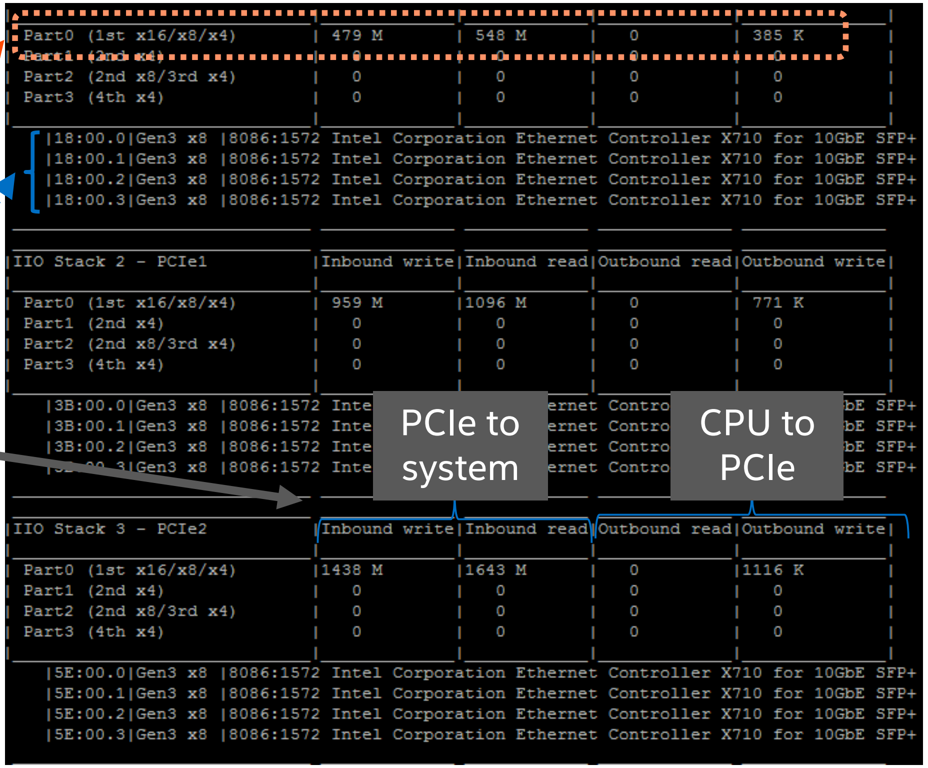
pcm-iio

Measure **individual** PCIe device **bandwidth at x4 granularity**

Enumerate downstream devices behind each IIO Stack

Monitor both inbound/outbound bandwidth

Can monitor **VT-d IOTLB miss rate** (opCode.txt)

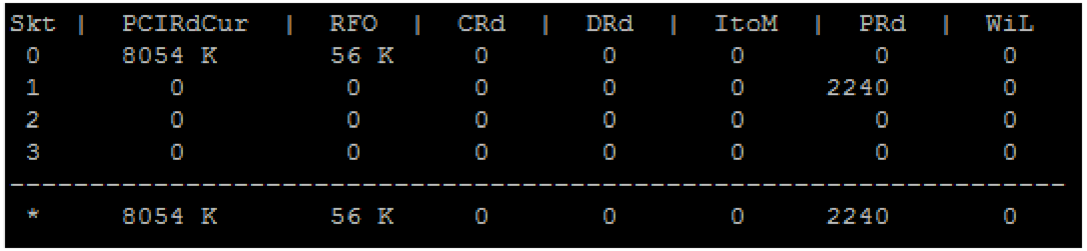


```
| Part0  (1st x16/x8/x4)       | 479 M        | 548 M       |    0         | 385 K        |
| Part1  (2nd x4)              |   0          |   0         |    0         |   0          |
| Part2  (2nd x8/3rd x4)       |   0          |   0         |    0         |   0          |
| Part3  (4th x4)              |   0          |   0         |    0         |   0          |
|                              |              |             |              |              |
   |18:00.0|Gen3 x8 |8086:1572 Intel Corporation Ethernet Controller X710 for 10GbE SFP+
   |18:00.1|Gen3 x8 |8086:1572 Intel Corporation Ethernet Controller X710 for 10GbE SFP+
   |18:00.2|Gen3 x8 |8086:1572 Intel Corporation Ethernet Controller X710 for 10GbE SFP+
   |18:00.3|Gen3 x8 |8086:1572 Intel Corporation Ethernet Controller X710 for 10GbE SFP+

|IIO Stack 2 - PCIe1           |Inbound write|Inbound read|Outbound read|Outbound write|
|                              |              |             |              |              |
| Part0  (1st x16/x8/x4)       | 959 M        |1096 M       |    0         | 771 K        |
| Part1  (2nd x4)              |   0          |   0         |    0         |   0          |
| Part2  (2nd x8/3rd x4)       |   0          |   0         |    0         |   0          |
| Part3  (4th x4)              |   0          |   0         |    0         |   0          |
|                              |              |             |              |              |
   |3B:00.0|Gen3 x8 |8086:1572 Inte              ernet Contro              bE SFP+
   |3B:00.1|Gen3 x8 |8086:1572 Inte              ernet Contro              bE SFP+
   |3B:00.2|Gen3 x8 |8086:1572 Inte              ernet Contro              bE SFP+
   |3B:00.3|Gen3 x8 |8086:1572 Inte              ernet Contro              bE SFP+

|IIO Stack 3 - PCIe2           |Inbound write|Inbound read|Outbound read|Outbound write|
|                              |              |             |              |              |
| Part0  (1st x16/x8/x4)       |1438 M        |1643 M       |    0         |1116 K        |
| Part1  (2nd x4)              |   0          |   0         |    0         |   0          |
| Part2  (2nd x8/3rd x4)       |   0          |   0         |    0         |   0          |
| Part3  (4th x4)              |   0          |   0         |    0         |   0          |
|                              |              |             |              |              |
   |5E:00.0|Gen3 x8 |8086:1572 Intel Corporation Ethernet Controller X710 for 10GbE SFP+
   |5E:00.1|Gen3 x8 |8086:1572 Intel Corporation Ethernet Controller X710 for 10GbE SFP+
   |5E:00.2|Gen3 x8 |8086:1572 Intel Corporation Ethernet Controller X710 for 10GbE SFP+
   |5E:00.3|Gen3 x8 |8086:1572 Intel Corporation Ethernet Controller X710 for 10GbE SFP+
```

Read/write bandwidth for PCIe-connected devices:

- SSD/disk
- Network
- Graphics
- FPGA
- etc

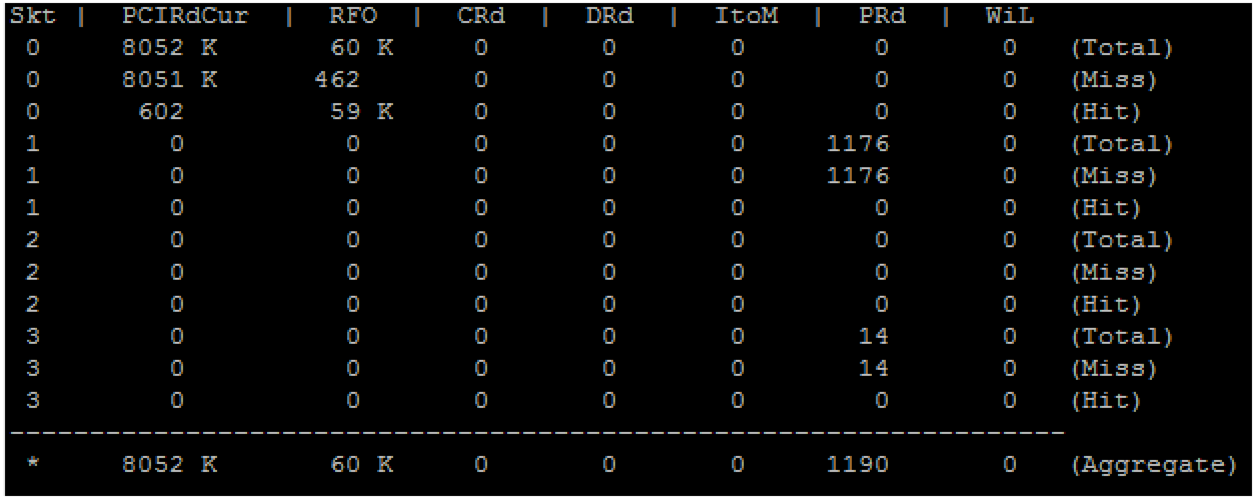
pcm-pcie (socket-level PCIe stats)



```
Skt |  PCIRdCur   |  RFO   |  CRd   |  DRd   |  ItoM   |  PRd   |  WiL
 0     8054 K       56 K      0        0        0        0        0
 1        0          0        0        0        0      2240       0
 2        0          0        0        0        0        0        0
 3        0          0        0        0        0        0        0
-------------------------------------------------------------------
 *     8054 K       56 K      0        0        0      2240       0
```

PCIe transfer events by type

„-e“ option



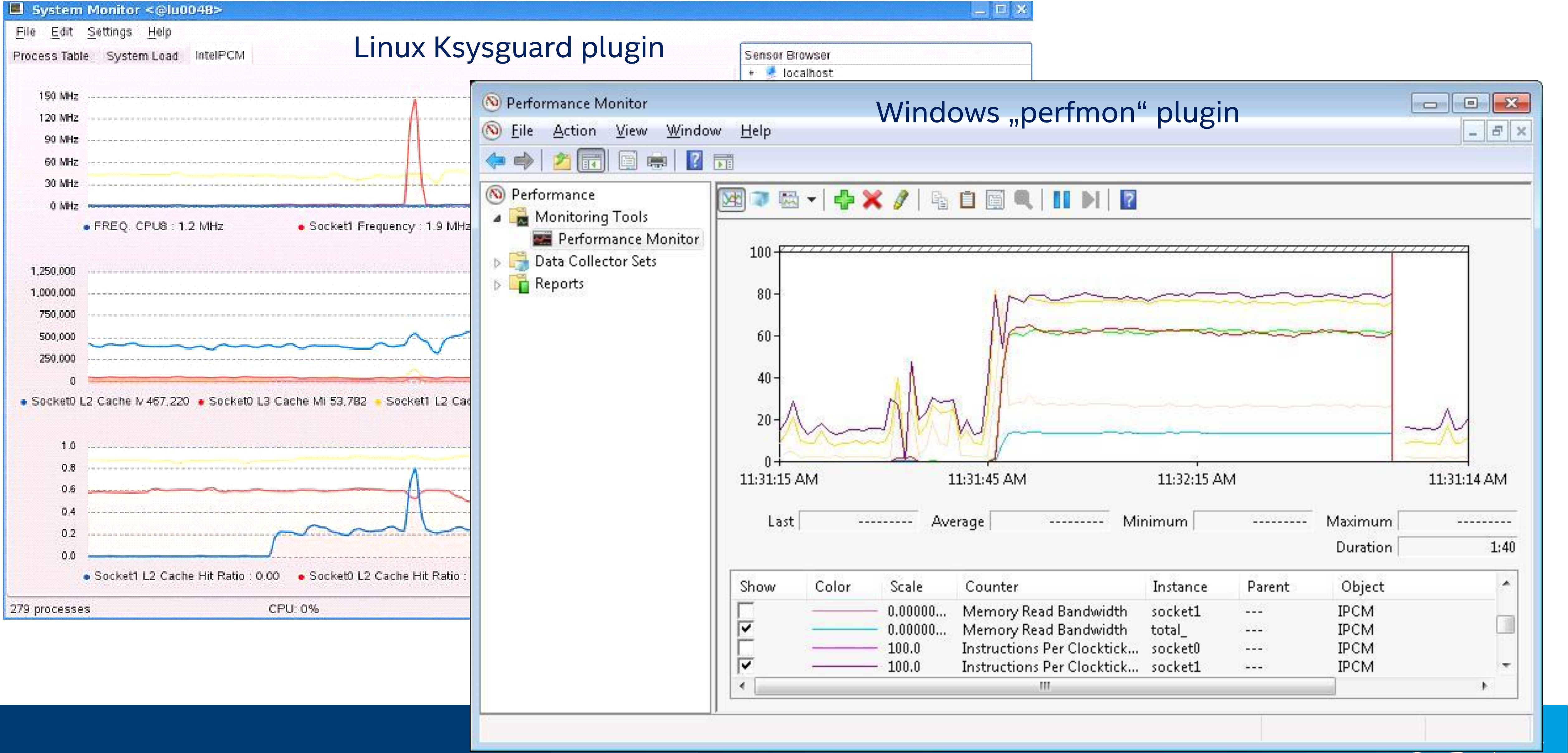
```
Skt |  PCIRdCur   |  RFO   |  CRd   |  DRd   |  ItoM   |  PRd   |  WiL
 0     8052 K       60 K      0        0        0        0        0     (Total)
 0     8051 K      462        0        0        0        0        0     (Miss)
 0      602         59 K      0        0        0        0        0     (Hit)
 1        0          0        0        0        0      1176       0     (Total)
 1        0          0        0        0        0      1176       0     (Miss)
 1        0          0        0        0        0        0        0     (Hit)
 2        0          0        0        0        0        0        0     (Total)
 2        0          0        0        0        0        0        0     (Miss)
 2        0          0        0        0        0        0        0     (Hit)
 3        0          0        0        0        0       14        0     (Total)
 3        0          0        0        0        0       14        0     (Miss)
 3        0          0        0        0        0        0        0     (Hit)
-------------------------------------------------------------------
 *     8052 K       60 K      0        0        0      1190       0     (Aggregate)
```

PCIe LLC cache optimization efficiciency (Intel DDIO tech)

see white-paper for detailed analysis method

white-paper

PCM real-time graphical front-ends



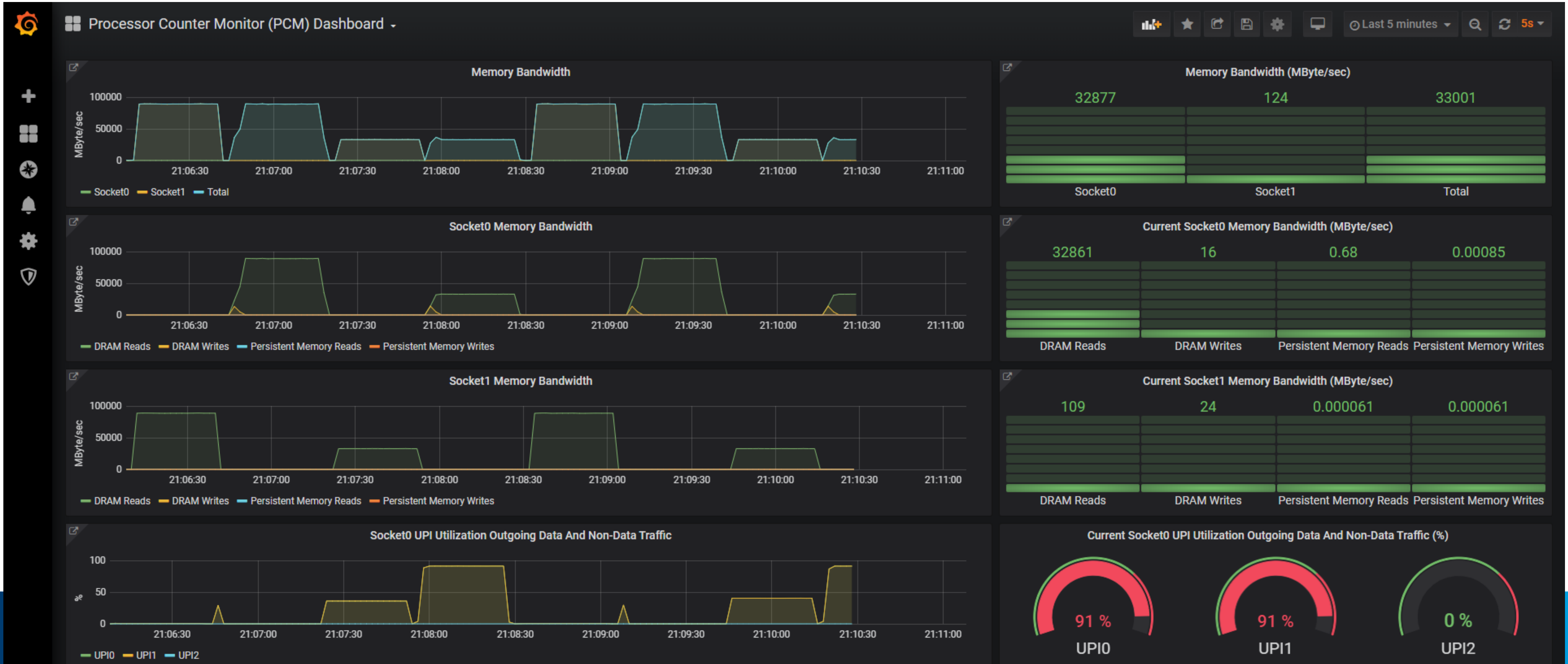
Linux Ksysguard plugin

Windows „perfmon“ plugin

Show	Color	Scale	Counter	Instance	Parent	Object
□		0.00000...	Memory Read Bandwidth	socket1	---	IPCM
☑		0.00000...	Memory Read Bandwidth	total_	---	IPCM
□		100.0	Instructions Per Clocktick...	socket0	---	IPCM
☑		100.0	Instructions Per Clocktick...	socket1	---	IPCM

pcm-sensor-server (JSON, prometheus over http)

Grafana real-time CPU dashboard (in browser):



Summary

- PCM „double-clicks“ on what exactly is busy inside the processor
- Real-time
- Easy to use for novice users
- Open-source
- Supports API and standard data export interfaces (csv, JSON, prometheus)

intel®

experience
what's inside™