DPDK

DATA PLANE DEVELOPMENT KIT

Lightning Talk #1

Optimizing Performance of Dataplane Software

HARRY VAN HAAREN

DPDK
DATA PLANE DEVELOPMENT KIT

DATA

INSIGHT

SOLUTION

- Record Data
- Understanding and Insight *
- Prototype Performant Code

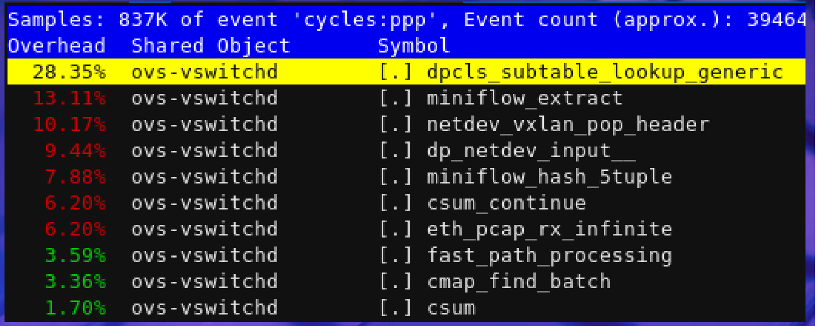
* May Require Experience

DPDK
DATA PLANE DEVELOPMENT KIT

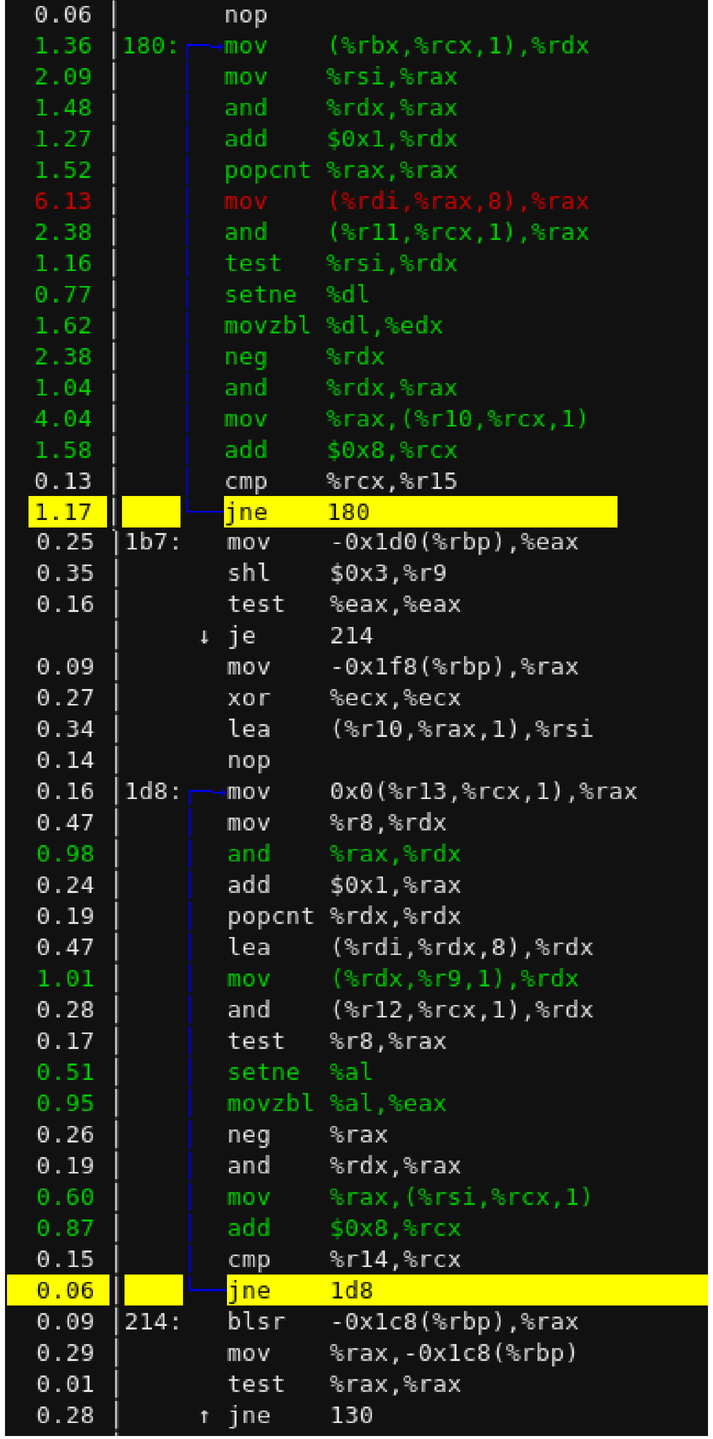
RECORD DATA

Tool:
Linux Perf

$ perf top –C1



```
Samples: 837K of event 'cycles:ppp', Event count (approx.): 39464
Overhead  Shared Object     Symbol
  28.35%  ovs-vswitchd      [.] dpcls_subtable_lookup_generic
  13.11%  ovs-vswitchd      [.] miniflow_extract
  10.17%  ovs-vswitchd      [.] netdev_vxlan_pop_header
   9.44%  ovs-vswitchd      [.] dp_netdev_input__
   7.88%  ovs-vswitchd      [.] miniflow_hash_5tuple
   6.20%  ovs-vswitchd      [.] csum_continue
   6.20%  ovs-vswitchd      [.] eth_pcap_rx_infinite
   3.59%  ovs-vswitchd      [.] fast_path_processing
   3.36%  ovs-vswitchd      [.] cmap_find_batch
   1.70%  ovs-vswitchd      [.] csum
```



```
 0.06 │        nop
 1.36 │180:    mov    (%rbx,%rcx,1),%rdx
 2.09 │        mov    %rsi,%rax
 1.48 │        and    %rdx,%rax
 1.27 │        add    $0x1,%rdx
 1.52 │        popcnt %rax,%rax
 6.13 │        mov    (%rdi,%rax,8),%rax
 2.38 │        and    (%r11,%rcx,1),%rax
 1.16 │        test   %rsi,%rdx
 0.77 │        setne  %dl
 1.62 │        movzbl %dl,%edx
 2.38 │        neg    %rdx
 1.04 │        and    %rdx,%rax
 4.04 │        mov    %rax,(%r10,%rcx,1)
 1.58 │        add    $0x8,%rcx
 0.13 │        cmp    %rcx,%r15
 1.17 │        jne    180
 0.25 │1b7:    mov    -0x1d0(%rbp),%eax
 0.35 │        shl    $0x3,%r9
 0.16 │        test   %eax,%eax
      │      ↓ je     214
 0.09 │        mov    -0x1f8(%rbp),%rax
 0.27 │        xor    %ecx,%ecx
 0.34 │        lea    (%r10,%rax,1),%rsi
 0.14 │        nop
 0.16 │1d8:    mov    0x0(%r13,%rcx,1),%rax
 0.47 │        mov    %r8,%rdx
 0.98 │        and    %rax,%rdx
 0.24 │        add    $0x1,%rax
 0.19 │        popcnt %rdx,%rdx
 0.47 │        lea    (%rdi,%rdx,8),%rdx
 1.01 │        mov    (%rdx,%r9,1),%rdx
 0.28 │        and    (%r12,%rcx,1),%rdx
 0.17 │        test   %r8,%rax
 0.51 │        setne  %al
 0.95 │        movzbl %al,%eax
 0.26 │        neg    %rax
 0.19 │        and    %rdx,%rax
 0.60 │        mov    %rax,(%rsi,%rcx,1)
 0.87 │        add    $0x8,%rcx
 0.15 │        cmp    %r14,%rcx
 0.06 │        jne    1d8
 0.09 │214:    blsr   -0x1c8(%rbp),%rax
 0.29 │        mov    %rax,-0x1c8(%rbp)
 0.01 │        test   %rax,%rax
 0.28 │      ↑ jne    130
```

Loop : Miniflow Unit 0
Iterations : **?**

Loop : Miniflow Unit 1
Iterations : **?**

“DATA – **INSIGHT** – SOLUTION”

1) Build a Mental Model of High Performance code	e.g. DPDK :)
2) Reason about Code Performance	Branches, Cache, SIMD …
3) Adjust Mental Model with Measured Data	Continuously Updating …
4) Prototype and Measure	Benchmarks + Real World …

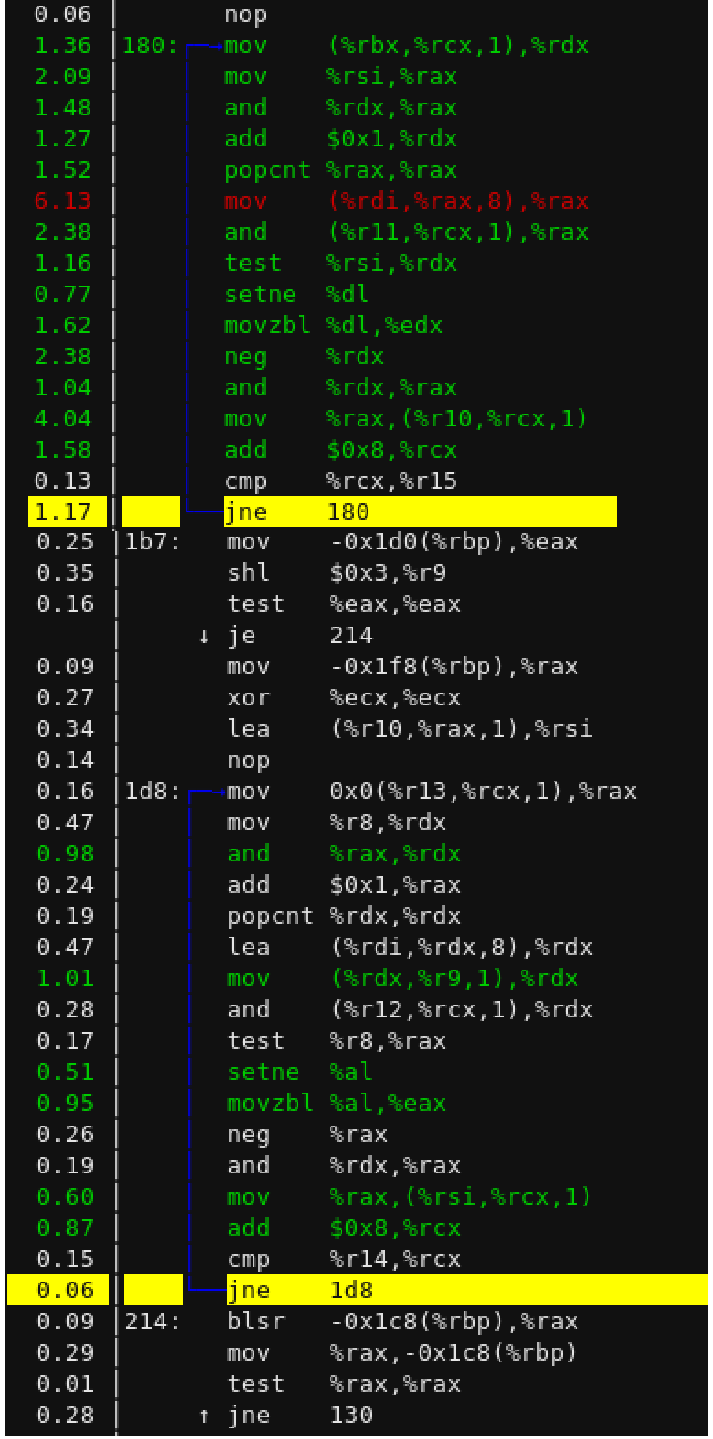
DPDK
DATA PLANE DEVELOPMENT KIT

INSIGHT

Investigate Code

Loops have Fixed Trip Counts

Use new Insight to Optimize!



```
0.06 │         nop
1.36 │180:┌─→mov    (%rbx,%rcx,1),%rdx
2.09 │    │  mov    %rsi,%rax
1.48 │    │  and    %rdx,%rax
1.27 │    │  add    $0x1,%rdx
1.52 │    │  popcnt %rax,%rax
6.13 │    │  mov    (%rdi,%rax,8),%rax
2.38 │    │  and    (%r11,%rcx,1),%rax
1.16 │    │  test   %rsi,%rdx
0.77 │    │  setne  %dl
1.62 │    │  movzbl %dl,%edx
2.38 │    │  neg    %rdx
1.04 │    │  and    %rdx,%rax
4.04 │    │  mov    %rax,(%r10,%rcx,1)
1.58 │    │  add    $0x8,%rcx
0.13 │    │  cmp    %rcx,%r15
1.17 │    └──jne    180
0.25 │1b7:   mov    -0x1d0(%rbp),%eax
0.35 │       shl    $0x3,%r9
0.16 │       test   %eax,%eax
     │     ↓ je     214
0.09 │       mov    -0x1f8(%rbp),%rax
0.27 │       xor    %ecx,%ecx
0.34 │       lea    (%r10,%rax,1),%rsi
0.14 │       nop
0.16 │1d8:┌─→mov    0x0(%r13,%rcx,1),%rax
0.47 │    │  mov    %r8,%rdx
0.98 │    │  and    %rax,%rdx
0.24 │    │  add    $0x1,%rax
0.19 │    │  popcnt %rdx,%rdx
0.47 │    │  lea    (%rdi,%rdx,8),%rdx
1.01 │    │  mov    (%rdx,%r9,1),%rdx
0.28 │    │  and    (%r12,%rcx,1),%rdx
0.17 │    │  test   %r8,%rax
0.51 │    │  setne  %al
0.95 │    │  movzbl %al,%eax
0.26 │    │  neg    %rax
0.19 │    │  and    %rdx,%rax
0.60 │    │  mov    %rax,(%rsi,%rcx,1)
0.87 │    │  add    $0x8,%rcx
0.15 │    │  cmp    %r14,%rcx
0.06 │    └──jne    1d8
0.09 │214:   blsr   -0x1c8(%rbp),%rax
0.29 │       mov    %rax,-0x1c8(%rbp)
0.01 │       test   %rax,%rax
0.28 │     ↑ jne    130
```

Loop : Miniflow Unit 0
Iterations : 3 X

Loop : Miniflow Unit 1
Iterations : 1 X

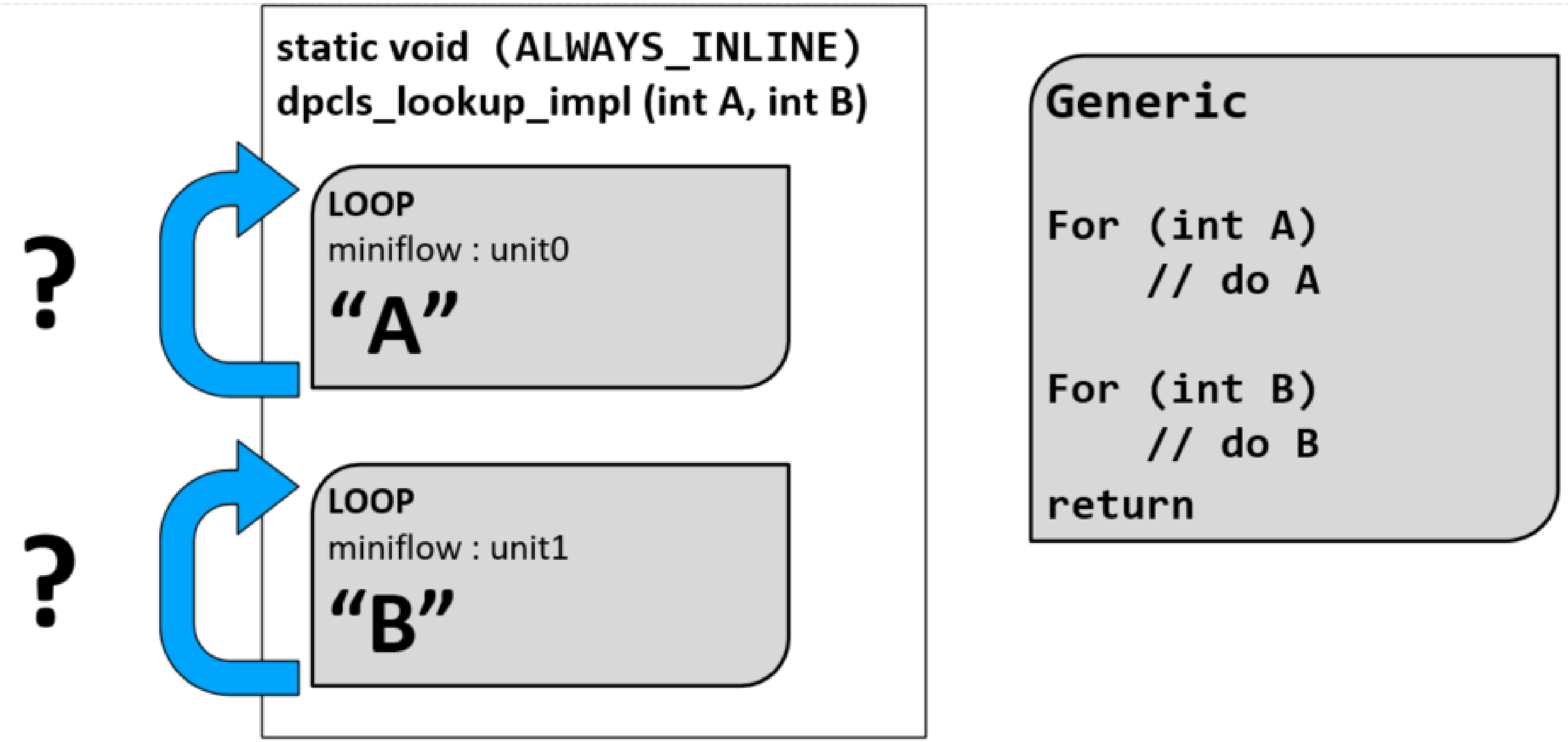
DPDK
DATA PLANE DEVELOPMENT KIT

SOLUTION

Constants

Constant Propagation

Inline Functions



static void (ALWAYS_INLINE)
dpcls_lookup_impl (int A, int B)

LOOP
miniflow : unit0
“A”

LOOP
miniflow : unit1
“B”

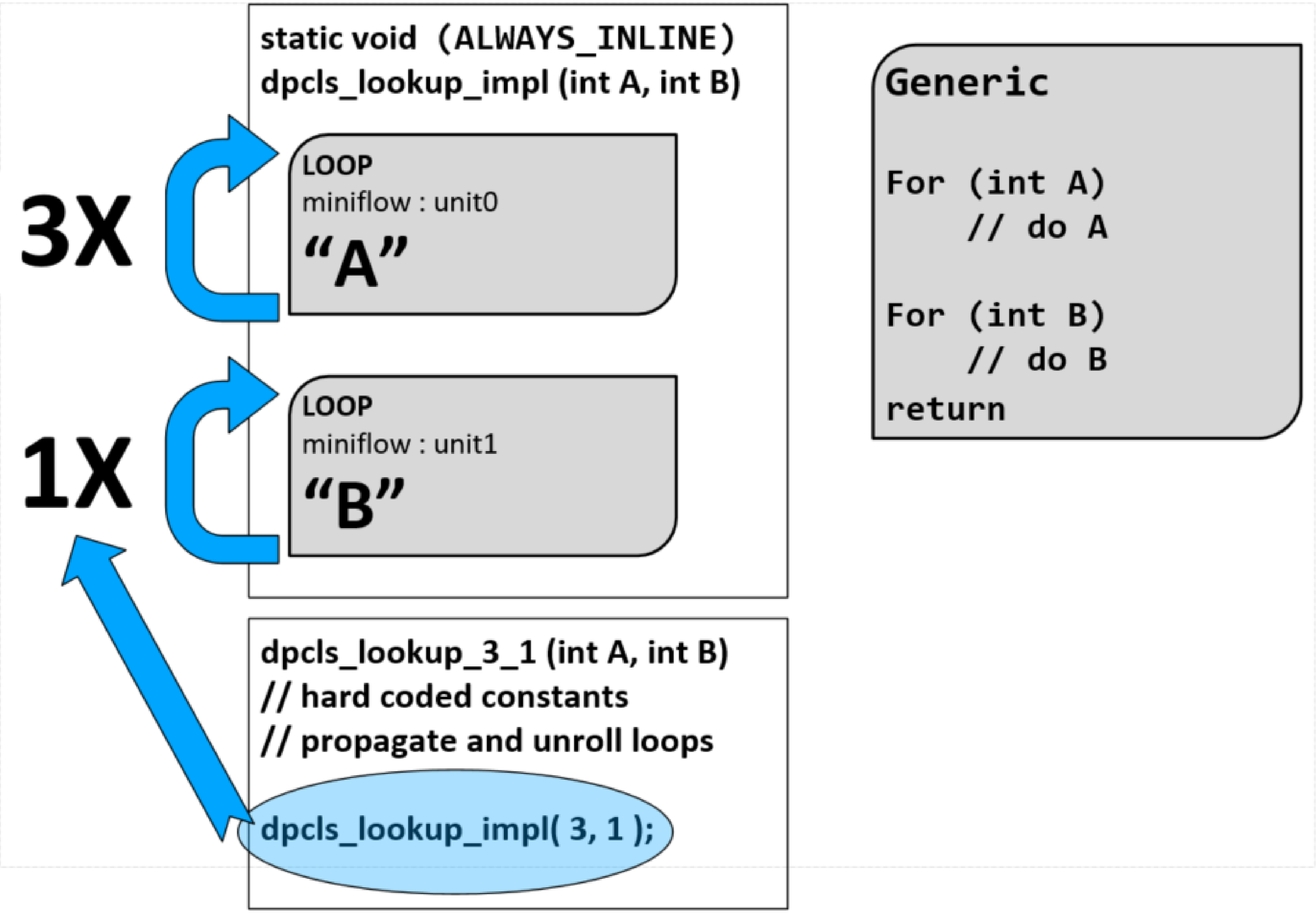
```
Generic

For (int A)
    // do A

For (int B)
    // do B
return
```

SOLUTION

Constants

Constant Propagation

Inline Functions



```
static void (ALWAYS_INLINE)
dpcls_lookup_impl (int A, int B)
```

3X

LOOP
miniflow : unit0
"A"

1X

LOOP
miniflow : unit1
"B"

```
dpcls_lookup_3_1 (int A, int B)
// hard coded constants
// propagate and unroll loops

dpcls_lookup_impl( 3, 1 );
```

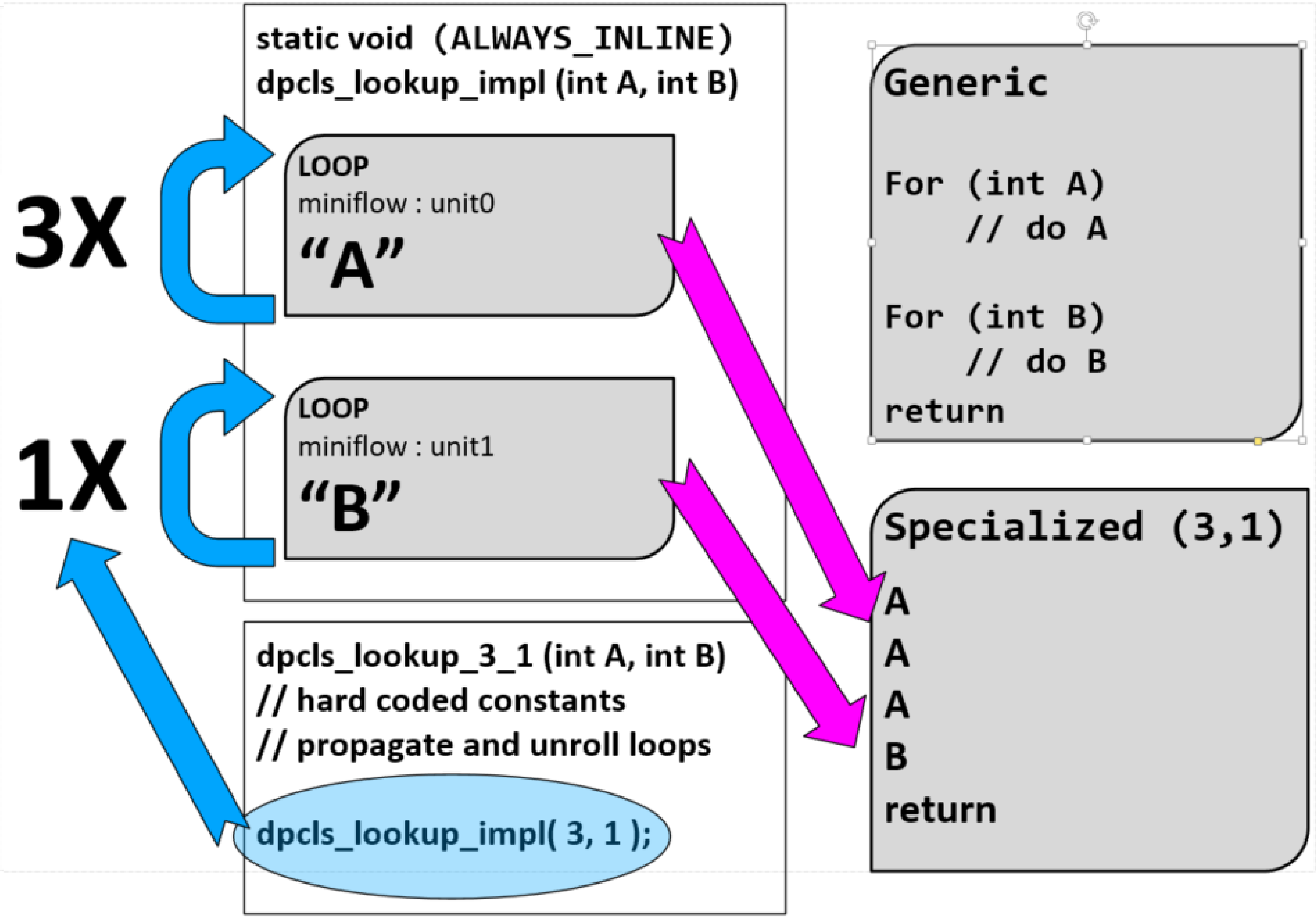
```
Generic

For (int A)
    // do A

For (int B)
    // do B
return
```

SOLUTION

Constants

Constant Propagation

Inline Functions



static void (ALWAYS_INLINE)
dpcls_lookup_impl (int A, int B)

3X

LOOP
miniflow : unit0
"A"

1X

LOOP
miniflow : unit1
"B"

dpcls_lookup_3_1 (int A, int B)
// hard coded constants
// propagate and unroll loops
dpcls_lookup_impl(3, 1);

```
Generic

For (int A)
    // do A

For (int B)
    // do B
return
```

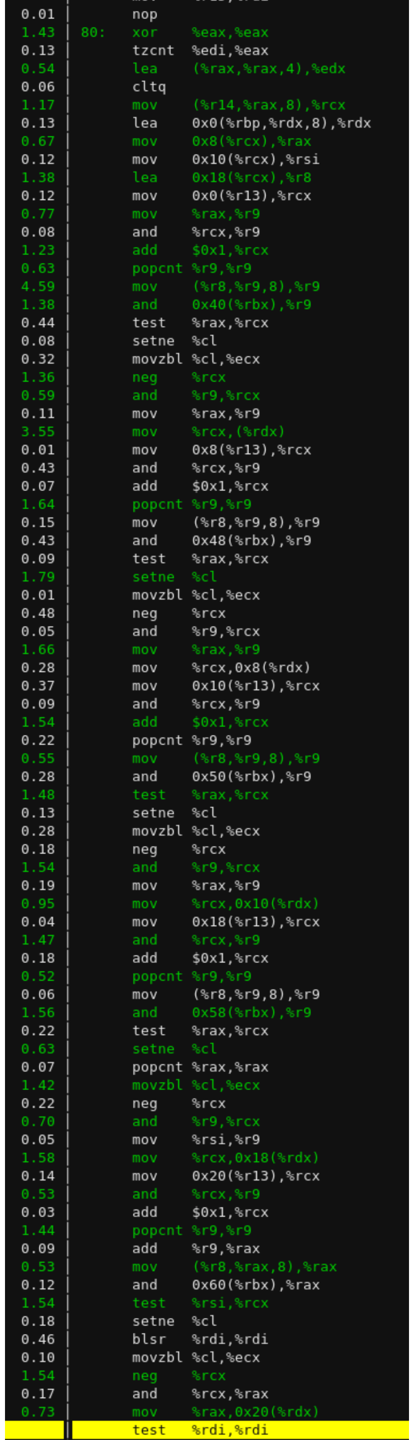
```
Specialized (3,1)
A
A
A
B
return
```

DPDK
DATA PLANE DEVELOPMENT KIT

VERIFY OPTIMIZATION

Flat code - No Loops!

- Start at the start
- Execute it all
- 1x per packet



```
 0.01 │        nop
 1.43 │  80:   xor    %eax,%eax
 0.13 │        tzcnt  %edi,%eax
 0.54 │        lea    (%rax,%rax,4),%edx
 0.06 │        cltq
 1.17 │        mov    (%r14,%rax,8),%rcx
 0.13 │        lea    0x0(%rbp,%rdx,8),%rdx
 0.67 │        mov    0x8(%rcx),%rax
 0.12 │        mov    0x10(%rcx),%rsi
 1.38 │        lea    0x18(%rcx),%r8
 0.12 │        mov    0x0(%r13),%rcx
 0.77 │        mov    %rax,%r9
 0.08 │        and    %rcx,%r9
 1.23 │        add    $0x1,%rcx
 0.63 │        popcnt %r9,%r9
 4.59 │        mov    (%r8,%r9,8),%r9
 1.38 │        and    0x40(%rbx),%r9
 0.44 │        test   %rax,%rcx
 0.08 │        setne  %cl
 0.32 │        movzbl %cl,%ecx
 1.36 │        neg    %rcx
 0.59 │        and    %r9,%rcx
 0.11 │        mov    %rax,%r9
 3.55 │        mov    %rcx,(%rdx)
 0.01 │        mov    0x8(%r13),%rcx
 0.43 │        and    %rcx,%r9
 0.07 │        add    $0x1,%rcx
 1.64 │        popcnt %r9,%r9
 0.15 │        mov    (%r8,%r9,8),%r9
 0.43 │        and    0x48(%rbx),%r9
 0.09 │        test   %rax,%rcx
 1.79 │        setne  %cl
 0.01 │        movzbl %cl,%ecx
 0.48 │        neg    %rcx
 0.05 │        and    %r9,%rcx
 1.66 │        mov    %rax,%r9
 0.28 │        mov    %rcx,0x8(%rdx)
 0.37 │        mov    0x10(%r13),%rcx
 0.09 │        and    %rcx,%r9
 1.54 │        add    $0x1,%rcx
 0.22 │        popcnt %r9,%r9
 0.55 │        mov    (%r8,%r9,8),%r9
 0.28 │        and    0x50(%rbx),%r9
 1.48 │        test   %rax,%rcx
 0.13 │        setne  %cl
 0.28 │        movzbl %cl,%ecx
 0.18 │        neg    %rcx
 1.54 │        and    %r9,%rcx
 0.19 │        mov    %rax,%r9
 0.95 │        mov    %rcx,0x10(%rdx)
 0.04 │        mov    0x18(%r13),%rcx
 1.47 │        and    %rcx,%r9
 0.18 │        add    $0x1,%rcx
 0.52 │        popcnt %r9,%r9
 0.06 │        mov    (%r8,%r9,8),%r9
 1.56 │        and    0x58(%rbx),%r9
 0.22 │        test   %rax,%rcx
 0.63 │        setne  %cl
 0.07 │        popcnt %rax,%rax
 1.42 │        movzbl %cl,%ecx
 0.22 │        neg    %rcx
 0.70 │        and    %r9,%rcx
 0.05 │        mov    %rsi,%r9
 1.58 │        mov    %rcx,0x18(%rdx)
 0.14 │        mov    0x20(%r13),%rcx
 0.53 │        and    %rcx,%r9
 0.03 │        add    $0x1,%rcx
 1.44 │        popcnt %r9,%r9
 0.09 │        add    %r9,%rax
 0.53 │        mov    (%r8,%rax,8),%rax
 0.12 │        and    0x60(%rbx),%rax
 1.54 │        test   %rsi,%rcx
 0.18 │        setne  %cl
 0.46 │        blsr   %rdi,%rdi
 0.10 │        movzbl %cl,%ecx
 1.54 │        neg    %rcx
 0.17 │        and    %rcx,%rax
 0.73 │        mov    %rax,0x20(%rdx)
      │        test   %rdi,%rdi
```

DPDK

DATA PLANE DEVELOPMENT KIT

Thanks / Questions

harry.van.haaren@intel.com

Using Perf and Hardware Counters

- **Optimized build with Debug Symbols**
- **Linux Perf**
  - cycles, cycle_activity.* , resource_stalls.* ,
    - Older kernel? Use https://github.com/andikleen/pmu-tools
  - Be curious
    - Test counters, do they highlight known bad code?
    - If so, that's a good counter to keep using in future
- **Perf Usage**
  - Stat first – overview of counts, build mental model, normalize to per packet cost
  - Report stats per physical core or per thread
  - use “minus capital i” flag to report per second : -I1000,
  - Top/Record – sample stack to see into SW

DPDK

DATA PLANE DEVELOPMENT KIT

Lightning Talk #2

Data Plane Performance Monitor

KEITH WILES

Data Plane Performance Monitor

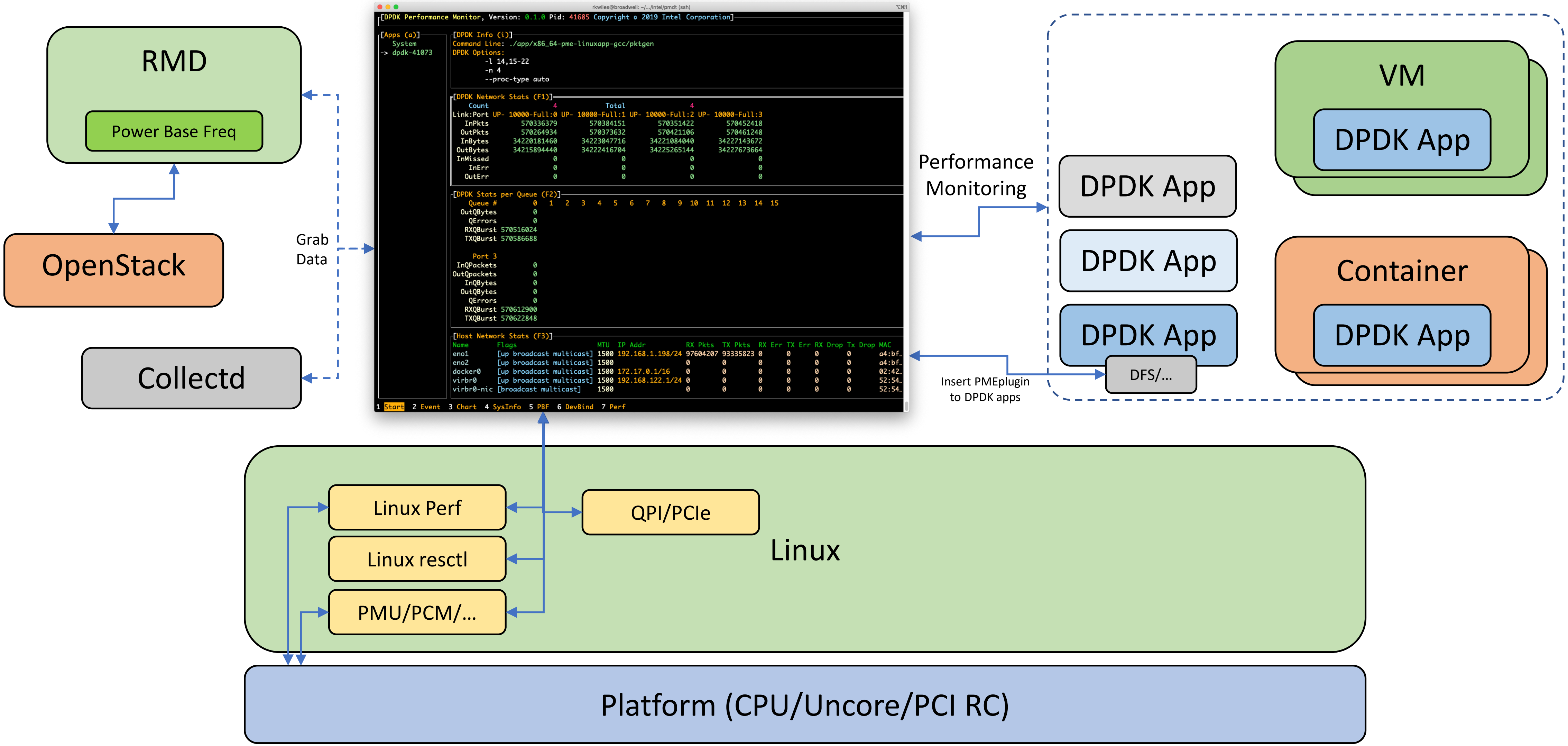
Performance Monitor Environment

2019/09/19

What is the Performance Monitor Tool

- Performance monitor tool is a standalone GoLang application
- Gathers metrics from DPDK applications and the system
- Displaying the data in a easy to read/understandable format
    - At this pointer in development the raw data is gathered and displayed using only an xTerm (VT100 cursor commands and ANSI color format)
- The tool gathers PCM, PMU, PBF, QPI, PCI and DPDK information allowing the user to see the data as simple set of charts or tables
- The goal of the tool is to analysis the data and give suggestions on how to improve performance or remove bottlenecks in the application and/or system

Performance Monitor Environment tool (PME)



```
[DPDK Performance Monitor, Version: 0.1.0 Pid: 41685 Copyright © 2019 Intel Corporation]

[Apps (a)]        [DPDK Info (i)]
   System         Command Line: ./app/x86_64-pme-linuxapp-gcc/pktgen
-> dpdk-41073     DPDK Options:
                          -l 14,15-22
                          -n 4
                          --proc-type auto

                  [DPDK Network Stats (F1)]
                      Count             4            Total             4
                  Link:Port UP- 10000-Full:0 UP- 10000-Full:1 UP- 10000-Full:2 UP- 10000-Full:3
                     InPkts       570336379       570384151       570351422       570452418
                    OutPkts       570264934       570373632       570421106       570461248
                    InBytes     34220181460     34223047716     34221084040     34227143672
                   OutBytes     34215894440     34222416704     34225265144     34227673664
                   InMissed               0               0               0               0
                      InErr               0               0               0               0
                     OutErr               0               0               0               0

                  [DPDK Stats per Queue (F2)]
                      Queue #         0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
                    OutQBytes         0
                      QErrors         0
                     RXQBurst 570516024
                     TXQBurst 570586688

                       Port 3
                   InQPackets         0
                  OutQpackets         0
                     InQBytes         0
                    OutQBytes         0
                      QErrors         0
                     RXQBurst 570612900
                     TXQBurst 570622848

                  [Host Network Stats (F3)]
                  Name        Flags                        MTU  IP Addr          RX Pkts  TX Pkts  RX Err TX Err RX Drop Tx Drop MAC
                  eno1        [up broadcast multicast] 1500 192.168.1.198/24 97604207 93335823 0      0      0       0       a4:bf…
                  eno2        [up broadcast multicast] 1500                  0        0        0      0      0       0       a4:bf…
                  docker0     [up broadcast multicast] 1500 172.17.0.1/16    0        0        0      0      0       0       02:42…
                  virbr0      [up broadcast multicast] 1500 192.168.122.1/24 0        0        0      0      0       0       52:54…
                  virbr0-nic  [broadcast multicast]    1500                  0        0        0      0      0       0       52:54…
1 Start  2 Event  3 Chart  4 SysInfo  5 PBF  6 DevBind  7 Perf
```

Gather data

- Use Prometheus to expose the metrics via the client web page
- Gather PMU register counters for each DPDK instance
  - Displayed in the PME tool along with analyzed information
  - Expose the PMU information via Prometheus for long term analysis
- Analyze the data and suggest solutions for known problems
  - Use AI to help determine hot spots or bottlenecks or …

Data Plane Performance Monitor Toolkit

- The tool will be open source code for developers and non-developers to utilize
  - BSD-3-Clause License
- The tool is not a DPDK application, but a tool and libraries for developers
- The toolkit contains the DFS library and other features TBD
- The tool is standalone and written in Go
- PCM/QPI/PCI data is gathered using a modified PCM-tool daemon, which the tool collects the information from a shared memory region

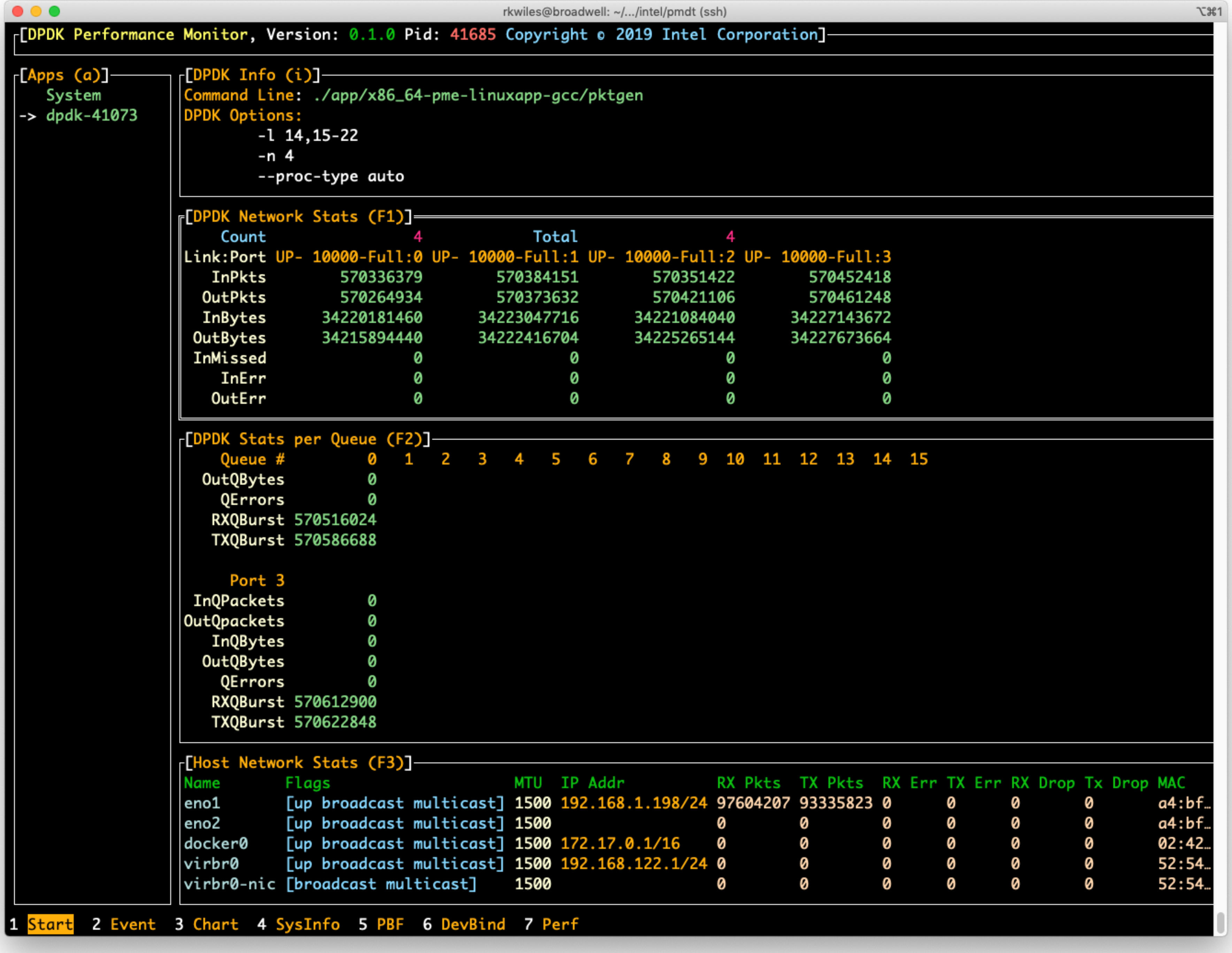
Panels in the performance monitor tool

The following slides show the current panels displayed by the tool and can/will change over time

Look at the screen as examples as we continue to define the metrics and how we display the data

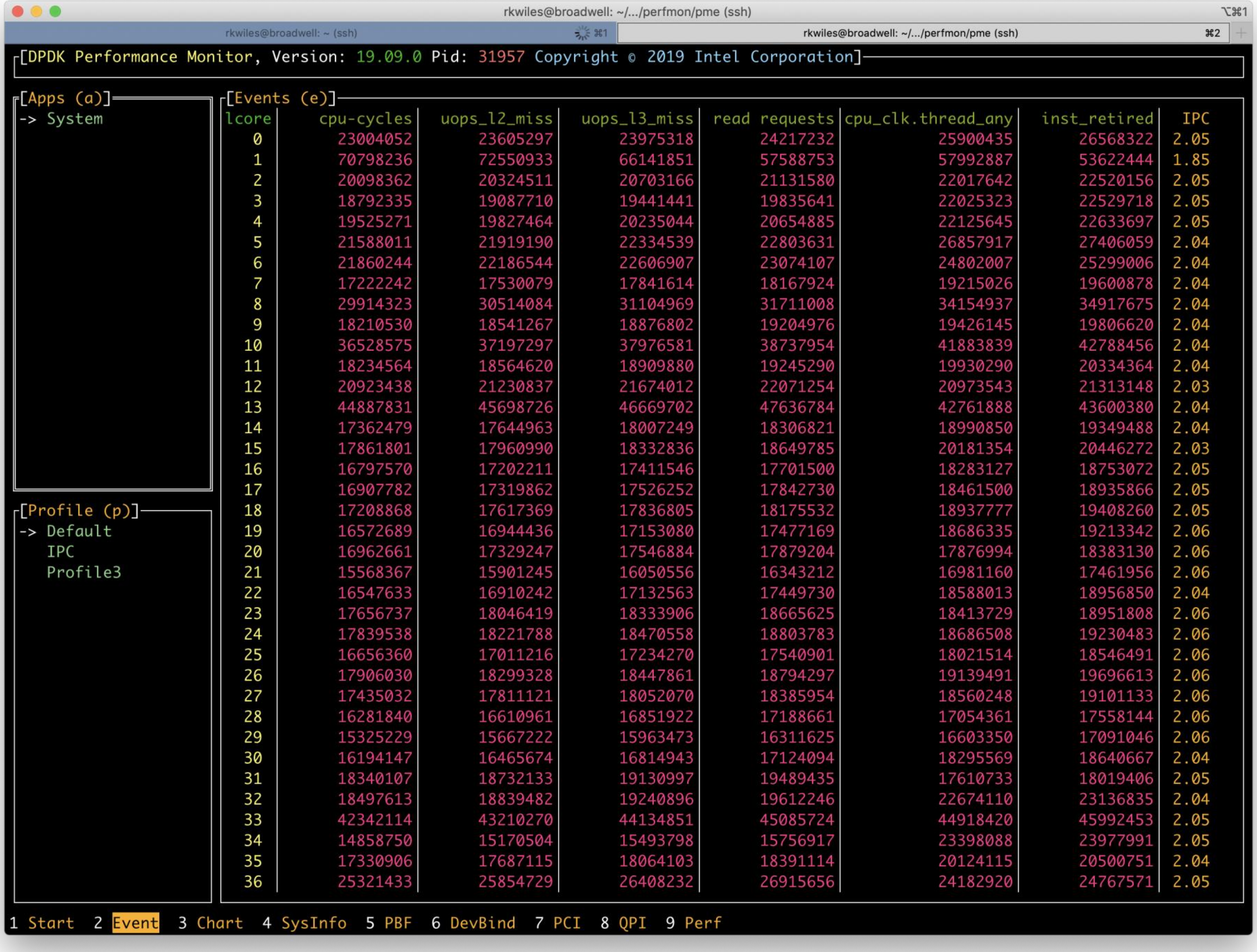
The goal is to analysis the data, then present solutions or suggestions to improve performance

PME Start Screen



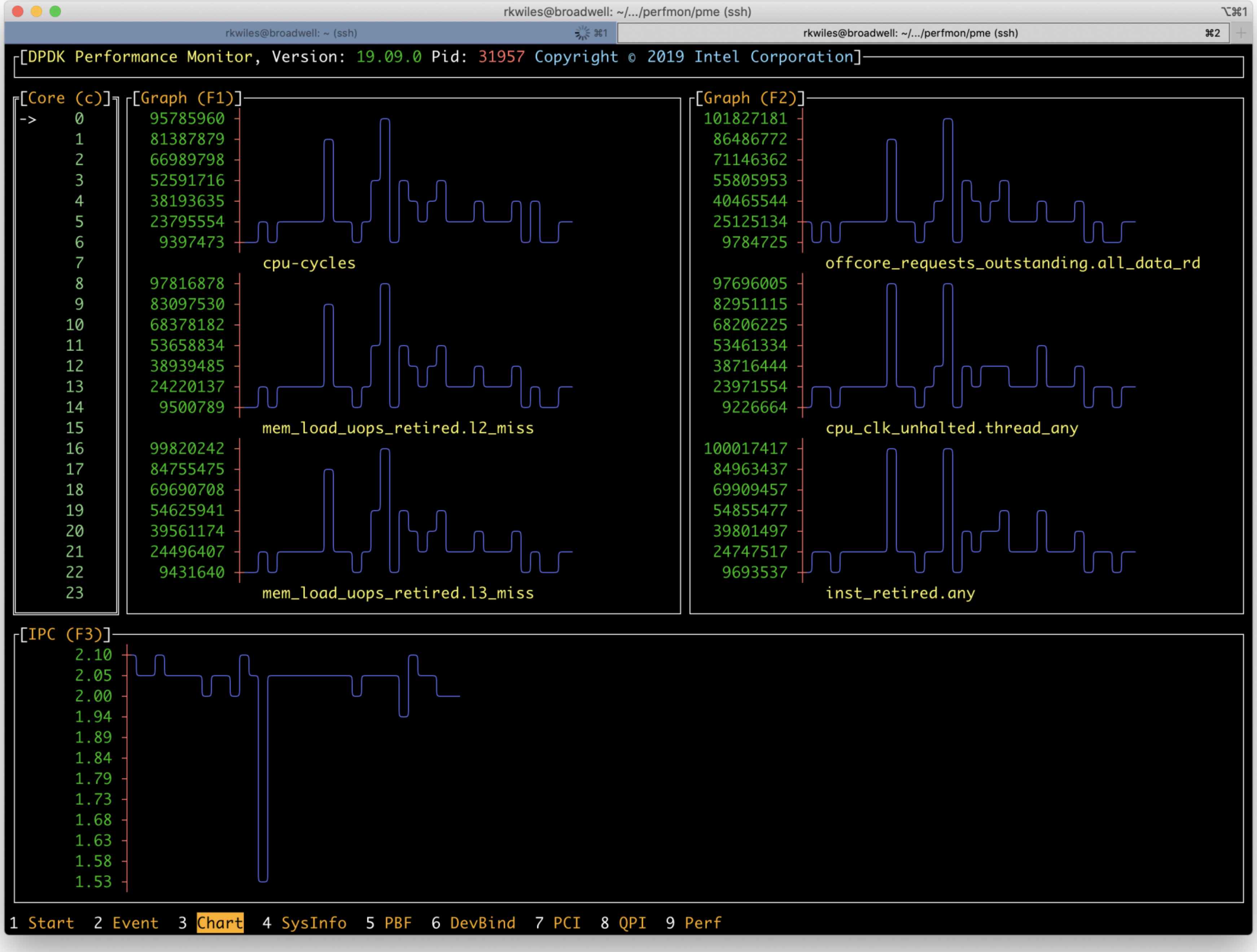
```
rkwiles@broadwell: ~/.../intel/pmdt (ssh)                                        ⌥⌘1
[DPDK Performance Monitor, Version: 0.1.0 Pid: 41685 Copyright © 2019 Intel Corporation]

[Apps (a)]        [DPDK Info (i)]
   System         Command Line: ./app/x86_64-pme-linuxapp-gcc/pktgen
-> dpdk-41073     DPDK Options:
                          -l 14,15-22
                          -n 4
                          --proc-type auto

                  [DPDK Network Stats (F1)]
                      Count          4             Total           4
                  Link:Port UP- 10000-Full:0 UP- 10000-Full:1 UP- 10000-Full:2 UP- 10000-Full:3
                     InPkts        570336379        570384151        570351422        570452418
                    OutPkts        570264934        570373632        570421106        570461248
                    InBytes      34220181460      34223047716      34221084040      34227143672
                   OutBytes      34215894440      34222416704      34225265144      34227673664
                   InMissed                0                0                0                0
                      InErr                0                0                0                0
                     OutErr                0                0                0                0

                  [DPDK Stats per Queue (F2)]
                      Queue #         0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
                    OutQBytes         0
                      QErrors         0
                     RXQBurst 570516024
                     TXQBurst 570586688

                       Port 3
                   InQPackets         0
                  OutQpackets         0
                     InQBytes         0
                    OutQBytes         0
                      QErrors         0
                     RXQBurst 570612900
                     TXQBurst 570622848

                  [Host Network Stats (F3)]
                  Name       Flags                       MTU  IP Addr          RX Pkts  TX Pkts  RX Err TX Err RX Drop Tx Drop MAC
                  eno1       [up broadcast multicast] 1500 192.168.1.198/24 97604207 93335823 0      0      0       0       a4:bf…
                  eno2       [up broadcast multicast] 1500                  0        0        0      0      0       0       a4:bf…
                  docker0    [up broadcast multicast] 1500 172.17.0.1/16    0        0        0      0      0       0       02:42…
                  virbr0     [up broadcast multicast] 1500 192.168.122.1/24 0        0        0      0      0       0       52:54…
                  virbr0-nic [broadcast multicast]    1500                  0        0        0      0      0       0       52:54…

1 Start  2 Event  3 Chart  4 SysInfo  5 PBF  6 DevBind  7 Perf
```

Events


rkwiles@broadwell: ~/.../perfmon/pme (ssh)
⌥⌘1
rkwiles@broadwell: ~ (ssh)
⌘1
rkwiles@broadwell: ~/.../perfmon/pme (ssh)
⌘2

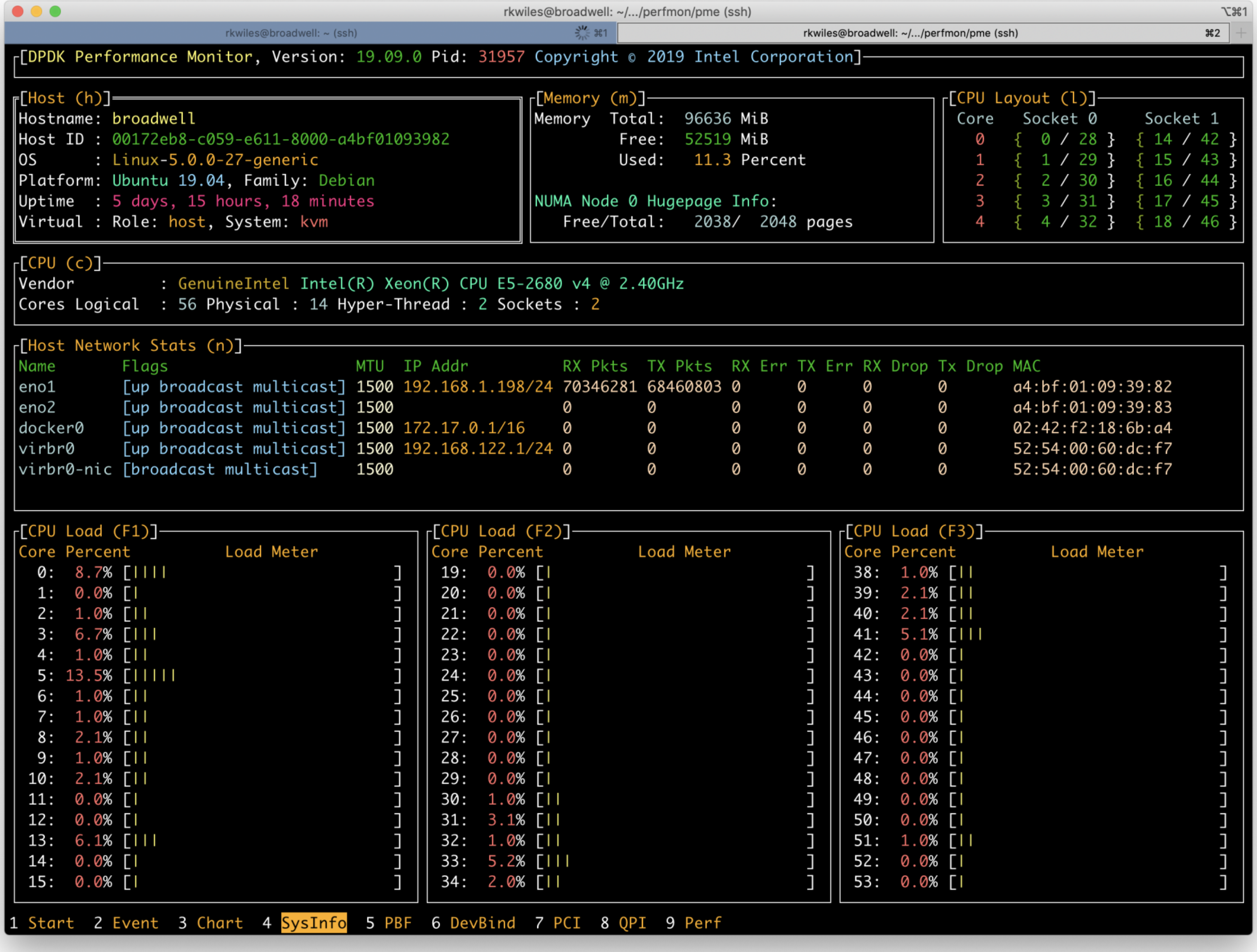
```
┌[DPDK Performance Monitor, Version: 19.09.0 Pid: 31957 Copyright © 2019 Intel Corporation]

┌[Apps (a)]           ┌[Events (e)]
-> System             lcore   cpu-cycles  uops_l2_miss  uops_l3_miss  read requests cpu_clk.thread_any  inst_retired  IPC
                          0     23004052      23605297      23975318       24217232            25900435      26568322 2.05
                          1     70798236      72550933      66141851       57588753            57992887      53622444 1.85
                          2     20098362      20324511      20703166       21131580            22017642      22520156 2.05
                          3     18792335      19087710      19441441       19835641            22025323      22529718 2.05
                          4     19525271      19827464      20235044       20654885            22125645      22633697 2.05
                          5     21588011      21919190      22334539       22803631            26857917      27406059 2.04
                          6     21860244      22186544      22606907       23074107            24802007      25299006 2.04
                          7     17222242      17530079      17841614       18167924            19215026      19600878 2.04
                          8     29914323      30514084      31104969       31711008            34154937      34917675 2.04
                          9     18210530      18541267      18876802       19204976            19426145      19806620 2.04
                         10     36528575      37197297      37976581       38737954            41883839      42788456 2.04
                         11     18234564      18564620      18909880       19245290            19930290      20334364 2.04
                         12     20923438      21230837      21674012       22071254            20973543      21313148 2.03
                         13     44887831      45698726      46669702       47636784            42761888      43600380 2.04
                         14     17362479      17644963      18007249       18306821            18990850      19349488 2.04
                         15     17861801      17960990      18332836       18649785            20181354      20446272 2.03
                         16     16797570      17202211      17411546       17701500            18283127      18753072 2.05
                         17     16907782      17319862      17526252       17842730            18461500      18935866 2.05
┌[Profile (p)]           18     17208868      17617369      17836805       18175532            18937777      19408260 2.05
-> Default               19     16572689      16944436      17153080       17477169            18686335      19213342 2.06
   IPC                   20     16962661      17329247      17546884       17879204            17876994      18383130 2.06
   Profile3              21     15568367      15901245      16050556       16343212            16981160      17461956 2.06
                         22     16547633      16910242      17132563       17449730            18588013      18956850 2.04
                         23     17656737      18046419      18333906       18665625            18413729      18951808 2.06
                         24     17839538      18221788      18470558       18803783            18686508      19230483 2.06
                         25     16656360      17011216      17234270       17540901            18021514      18546491 2.06
                         26     17906030      18299328      18447861       18794297            19139491      19696613 2.06
                         27     17435032      17811121      18052070       18385954            18560248      19101133 2.06
                         28     16281840      16610961      16851922       17188661            17054361      17558144 2.06
                         29     15325229      15667222      15963473       16311625            16603350      17091046 2.06
                         30     16194147      16465674      16814943       17124094            18295569      18640667 2.04
                         31     18340107      18732133      19130997       19489435            17610733      18019406 2.05
                         32     18497613      18839482      19240896       19612246            22674110      23136835 2.04
                         33     42342114      43210270      44134851       45085724            44918420      45992453 2.05
                         34     14858750      15170504      15493798       15756917            23398088      23977991 2.05
                         35     17330906      17687115      18064103       18391114            20124115      20500751 2.04
                         36     25321433      25854729      26408232       26915656            24182920      24767571 2.05

1 Start  2 Event  3 Chart  4 SysInfo  5 PBF  6 DevBind  7 PCI  8 QPI  9 Perf
```

Event Charts



```
[DPDK Performance Monitor, Version: 19.09.0 Pid: 31957 Copyright © 2019 Intel Corporation]

[Core (c)]
->    0
      1
      2
      3
      4
      5
      6
      7
      8
      9
     10
     11
     12
     13
     14
     15
     16
     17
     18
     19
     20
     21
     22
     23

1 Start  2 Event  3 Chart  4 SysInfo  5 PBF  6 DevBind  7 PCI  8 QPI  9 Perf
```

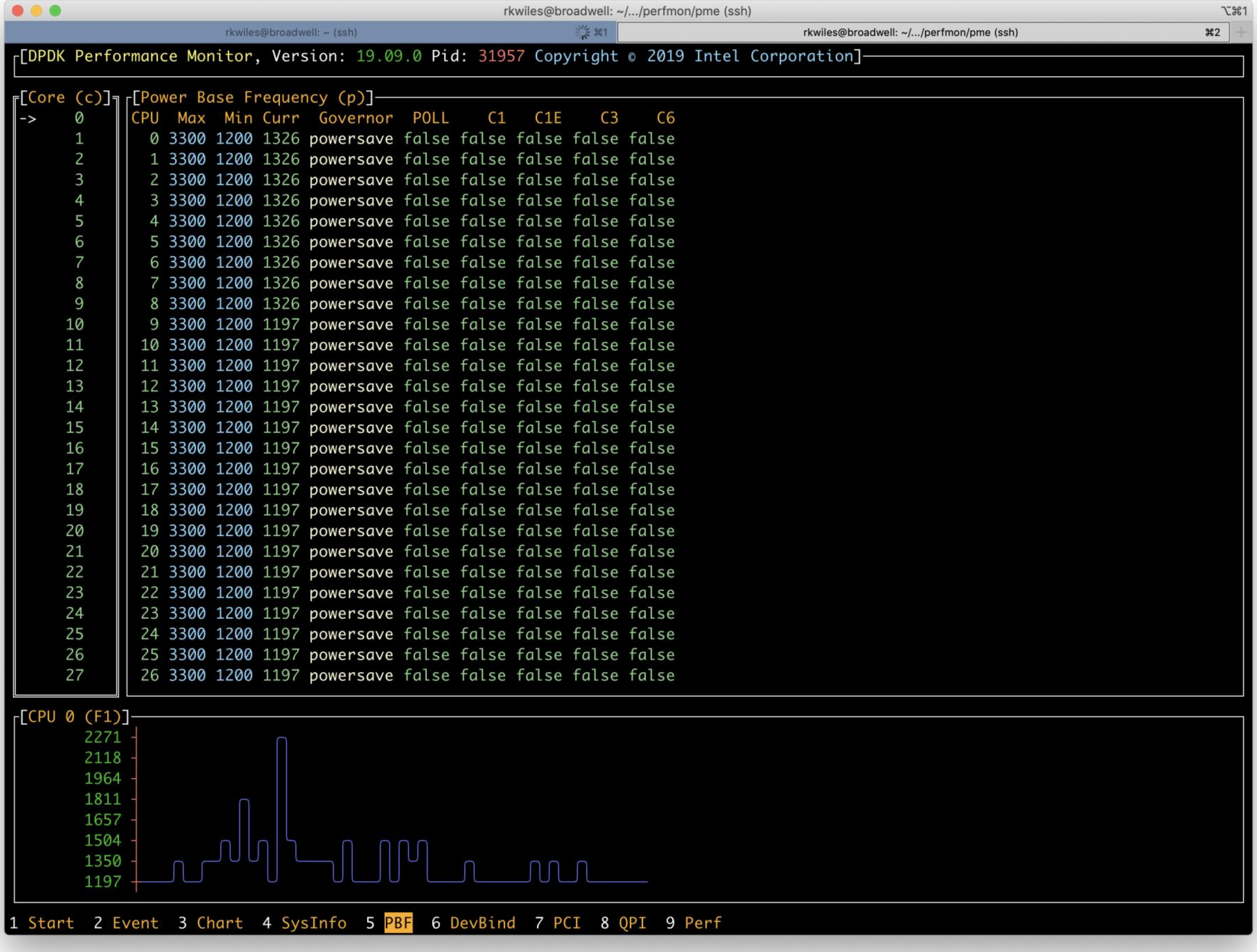
SysInfo



rkwiles@broadwell: ~/.../perfmon/pme (ssh)
⌥⌘1
rkwiles@broadwell: ~ (ssh)
⌘1
rkwiles@broadwell: ~/.../perfmon/pme (ssh)
⌘2

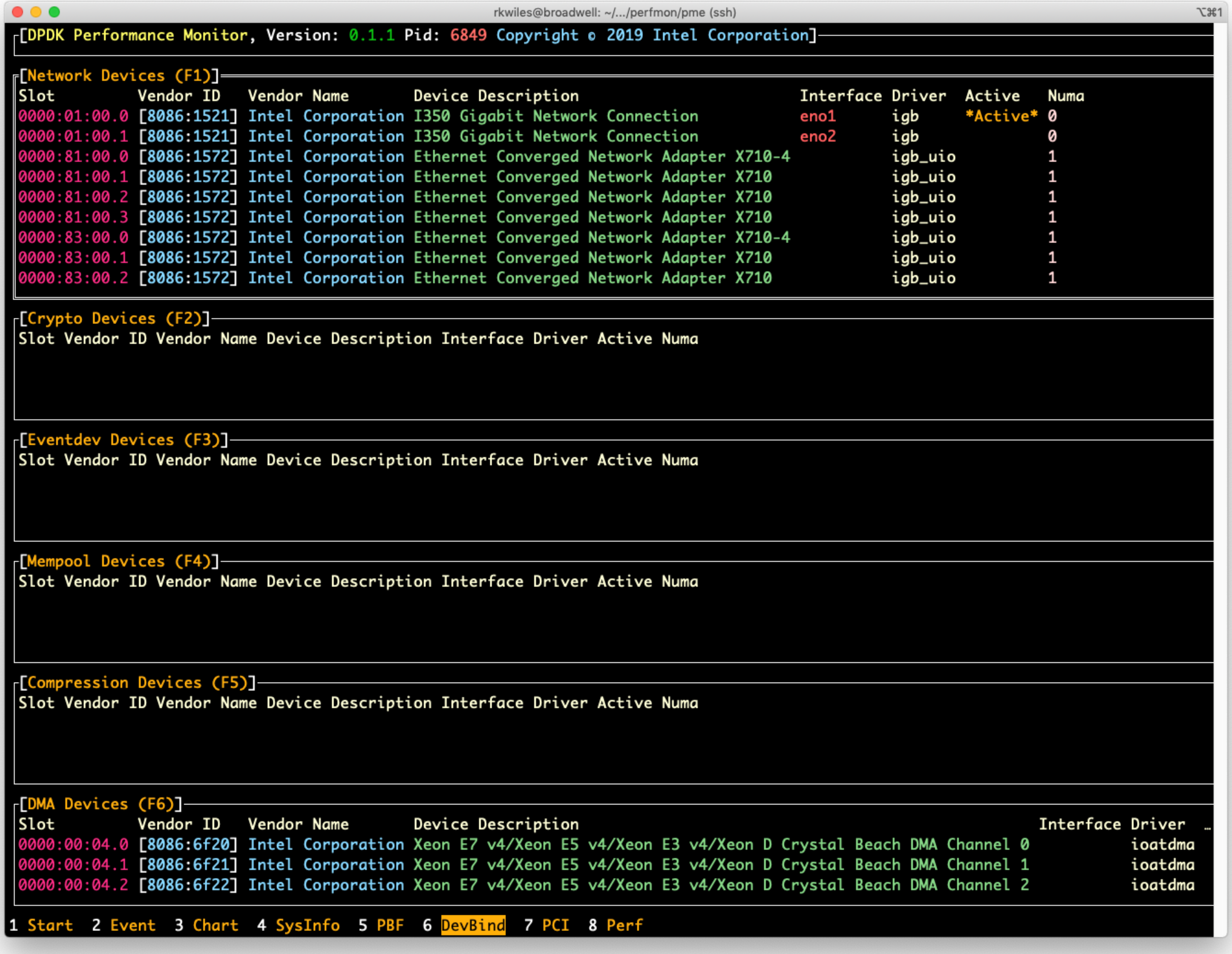
```
[DPDK Performance Monitor, Version: 19.09.0 Pid: 31957 Copyright © 2019 Intel Corporation]

[Host (h)]                                              [Memory (m)]                                  [CPU Layout (l)]
Hostname: broadwell                                     Memory  Total:   96636 MiB                     Core    Socket 0        Socket 1
Host ID : 00172eb8-c059-e611-8000-a4bf01093982                   Free:   52519 MiB                      0    {  0 / 28 }   { 14 / 42 }
OS      : Linux-5.0.0-27-generic                                 Used:    11.3 Percent                  1    {  1 / 29 }   { 15 / 43 }
Platform: Ubuntu 19.04, Family: Debian                                                                  2    {  2 / 30 }   { 16 / 44 }
Uptime  : 5 days, 15 hours, 18 minutes                  NUMA Node 0 Hugepage Info:                      3    {  3 / 31 }   { 17 / 45 }
Virtual : Role: host, System: kvm                          Free/Total:    2038/  2048 pages             4    {  4 / 32 }   { 18 / 46 }

[CPU (c)]
Vendor          : GenuineIntel Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz
Cores Logical   : 56 Physical : 14 Hyper-Thread : 2 Sockets : 2

[Host Network Stats (n)]
Name        Flags                      MTU  IP Addr          RX Pkts  TX Pkts  RX Err TX Err RX Drop Tx Drop MAC
eno1        [up broadcast multicast] 1500 192.168.1.198/24 70346281 68460803 0      0      0       0       a4:bf:01:09:39:82
eno2        [up broadcast multicast] 1500                  0        0        0      0      0       0       a4:bf:01:09:39:83
docker0     [up broadcast multicast] 1500 172.17.0.1/16    0        0        0      0      0       0       02:42:f2:18:6b:a4
virbr0      [up broadcast multicast] 1500 192.168.122.1/24 0        0        0      0      0       0       52:54:00:60:dc:f7
virbr0-nic  [broadcast multicast]    1500                  0        0        0      0      0       0       52:54:00:60:dc:f7

[CPU Load (F1)]                          [CPU Load (F2)]                          [CPU Load (F3)]
Core Percent       Load Meter            Core Percent       Load Meter            Core Percent       Load Meter
  0:   8.7% [||||                    ]     19:   0.0% [|                     ]     38:   1.0% [||                    ]
  1:   0.0% [|                       ]     20:   0.0% [|                     ]     39:   2.1% [||                    ]
  2:   1.0% [||                      ]     21:   0.0% [|                     ]     40:   2.1% [||                    ]
  3:   6.7% [|||                     ]     22:   0.0% [|                     ]     41:   5.1% [|||                   ]
  4:   1.0% [||                      ]     23:   0.0% [|                     ]     42:   0.0% [|                     ]
  5:  13.5% [|||||                   ]     24:   0.0% [|                     ]     43:   0.0% [|                     ]
  6:   1.0% [||                      ]     25:   0.0% [|                     ]     44:   0.0% [|                     ]
  7:   1.0% [||                      ]     26:   0.0% [|                     ]     45:   0.0% [|                     ]
  8:   2.1% [||                      ]     27:   0.0% [|                     ]     46:   0.0% [|                     ]
  9:   1.0% [||                      ]     28:   0.0% [|                     ]     47:   0.0% [|                     ]
 10:   2.1% [||                      ]     29:   0.0% [|                     ]     48:   0.0% [|                     ]
 11:   0.0% [|                       ]     30:   1.0% [||                    ]     49:   0.0% [|                     ]
 12:   0.0% [|                       ]     31:   3.1% [||                    ]     50:   0.0% [|                     ]
 13:   6.1% [|||                     ]     32:   1.0% [||                    ]     51:   1.0% [||                    ]
 14:   0.0% [|                       ]     33:   5.2% [|||                   ]     52:   0.0% [|                     ]
 15:   0.0% [|                       ]     34:   2.0% [||                    ]     53:   0.0% [|                     ]
1 Start  2 Event  3 Chart  4 SysInfo  5 PBF  6 DevBind  7 PCI  8 QPI  9 Perf
```

Power Based Frequency


rkwiles@broadwell: ~/.../perfmon/pme (ssh)
⌥⌘1
rkwiles@broadwell: ~ (ssh)
⌘1
rkwiles@broadwell: ~/.../perfmon/pme (ssh)
⌘2

```
┌[DPDK Performance Monitor, Version: 19.09.0 Pid: 31957 Copyright © 2019 Intel Corporation]

┌[Core (c)]┐ ┌[Power Base Frequency (p)]
->      0    CPU  Max  Min Curr  Governor  POLL    C1   C1E    C3    C6
        1      0 3300 1200 1326 powersave false false false false false
        2      1 3300 1200 1326 powersave false false false false false
        3      2 3300 1200 1326 powersave false false false false false
        4      3 3300 1200 1326 powersave false false false false false
        5      4 3300 1200 1326 powersave false false false false false
        6      5 3300 1200 1326 powersave false false false false false
        7      6 3300 1200 1326 powersave false false false false false
        8      7 3300 1200 1326 powersave false false false false false
        9      8 3300 1200 1326 powersave false false false false false
       10      9 3300 1200 1197 powersave false false false false false
       11     10 3300 1200 1197 powersave false false false false false
       12     11 3300 1200 1197 powersave false false false false false
       13     12 3300 1200 1197 powersave false false false false false
       14     13 3300 1200 1197 powersave false false false false false
       15     14 3300 1200 1197 powersave false false false false false
       16     15 3300 1200 1197 powersave false false false false false
       17     16 3300 1200 1197 powersave false false false false false
       18     17 3300 1200 1197 powersave false false false false false
       19     18 3300 1200 1197 powersave false false false false false
       20     19 3300 1200 1197 powersave false false false false false
       21     20 3300 1200 1197 powersave false false false false false
       22     21 3300 1200 1197 powersave false false false false false
       23     22 3300 1200 1197 powersave false false false false false
       24     23 3300 1200 1197 powersave false false false false false
       25     24 3300 1200 1197 powersave false false false false false
       26     25 3300 1200 1197 powersave false false false false false
       27     26 3300 1200 1197 powersave false false false false false

┌[CPU 0 (F1)]
        2271 ┤
        2118 ┤
        1964 ┤
        1811 ┤
        1657 ┤
        1504 ┤
        1350 ┤
        1197 ┤

1 Start  2 Event  3 Chart  4 SysInfo  5 PBF  6 DevBind  7 PCI  8 QPI  9 Perf
```

DevBind

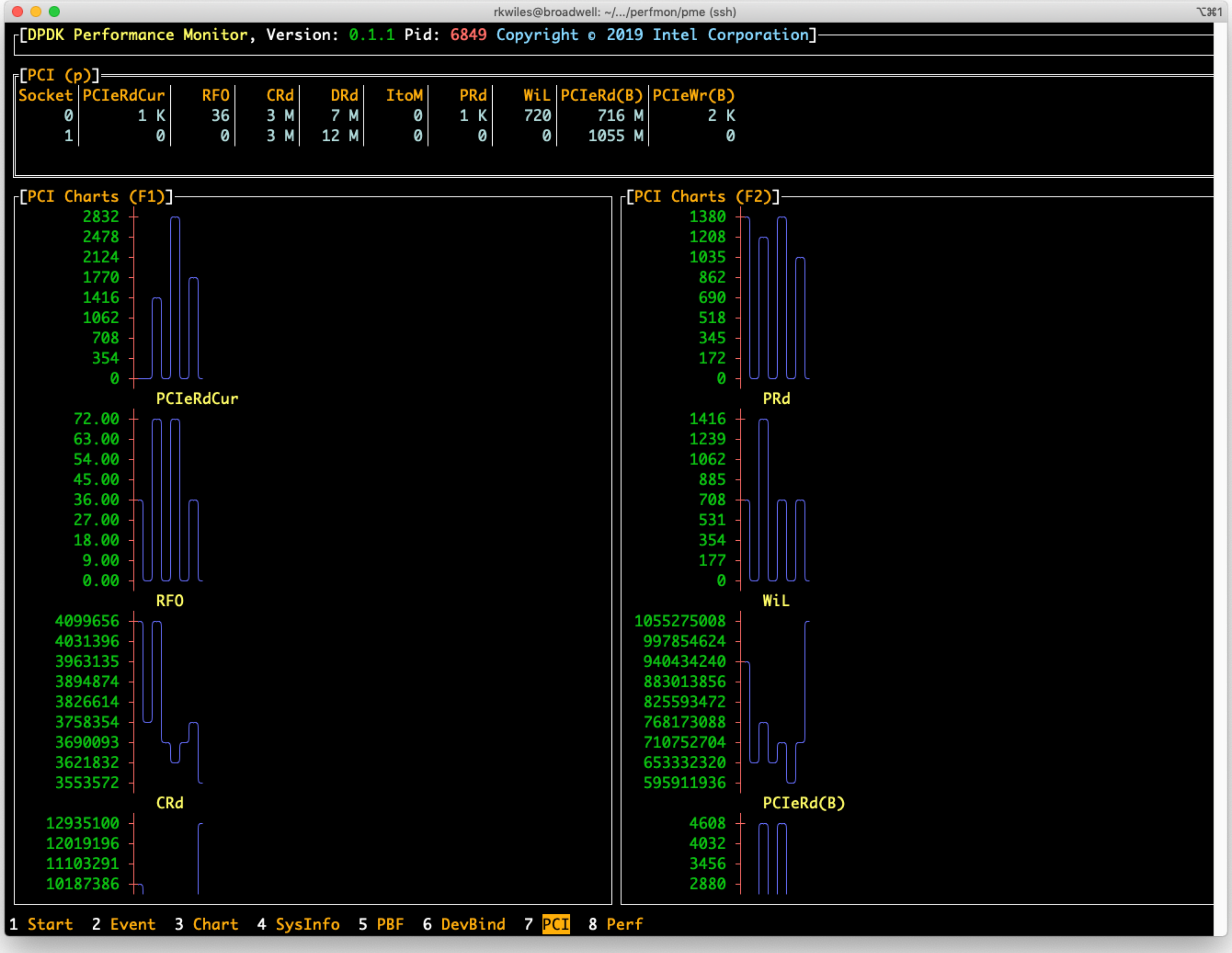


rkwiles@broadwell: ~/.../perfmon/pme (ssh)

⌥⌘1

```
[DPDK Performance Monitor, Version: 0.1.1 Pid: 6849 Copyright © 2019 Intel Corporation]

[Network Devices (F1)]
Slot         Vendor ID   Vendor Name       Device Description                       Interface Driver  Active   Numa
0000:01:00.0 [8086:1521] Intel Corporation I350 Gigabit Network Connection           eno1      igb     *Active* 0
0000:01:00.1 [8086:1521] Intel Corporation I350 Gigabit Network Connection           eno2      igb              0
0000:81:00.0 [8086:1572] Intel Corporation Ethernet Converged Network Adapter X710-4           igb_uio          1
0000:81:00.1 [8086:1572] Intel Corporation Ethernet Converged Network Adapter X710             igb_uio          1
0000:81:00.2 [8086:1572] Intel Corporation Ethernet Converged Network Adapter X710             igb_uio          1
0000:81:00.3 [8086:1572] Intel Corporation Ethernet Converged Network Adapter X710             igb_uio          1
0000:83:00.0 [8086:1572] Intel Corporation Ethernet Converged Network Adapter X710-4           igb_uio          1
0000:83:00.1 [8086:1572] Intel Corporation Ethernet Converged Network Adapter X710             igb_uio          1
0000:83:00.2 [8086:1572] Intel Corporation Ethernet Converged Network Adapter X710             igb_uio          1

[Crypto Devices (F2)]
Slot Vendor ID Vendor Name Device Description Interface Driver Active Numa

[Eventdev Devices (F3)]
Slot Vendor ID Vendor Name Device Description Interface Driver Active Numa

[Mempool Devices (F4)]
Slot Vendor ID Vendor Name Device Description Interface Driver Active Numa

[Compression Devices (F5)]
Slot Vendor ID Vendor Name Device Description Interface Driver Active Numa

[DMA Devices (F6)]
Slot         Vendor ID   Vendor Name       Device Description                                                  Interface Driver  …
0000:00:04.0 [8086:6f20] Intel Corporation Xeon E7 v4/Xeon E5 v4/Xeon E3 v4/Xeon D Crystal Beach DMA Channel 0           ioatdma
0000:00:04.1 [8086:6f21] Intel Corporation Xeon E7 v4/Xeon E5 v4/Xeon E3 v4/Xeon D Crystal Beach DMA Channel 1           ioatdma
0000:00:04.2 [8086:6f22] Intel Corporation Xeon E7 v4/Xeon E5 v4/Xeon E3 v4/Xeon D Crystal Beach DMA Channel 2           ioatdma

1 Start  2 Event  3 Chart  4 SysInfo  5 PBF  6 DevBind  7 PCI  8 Perf
```

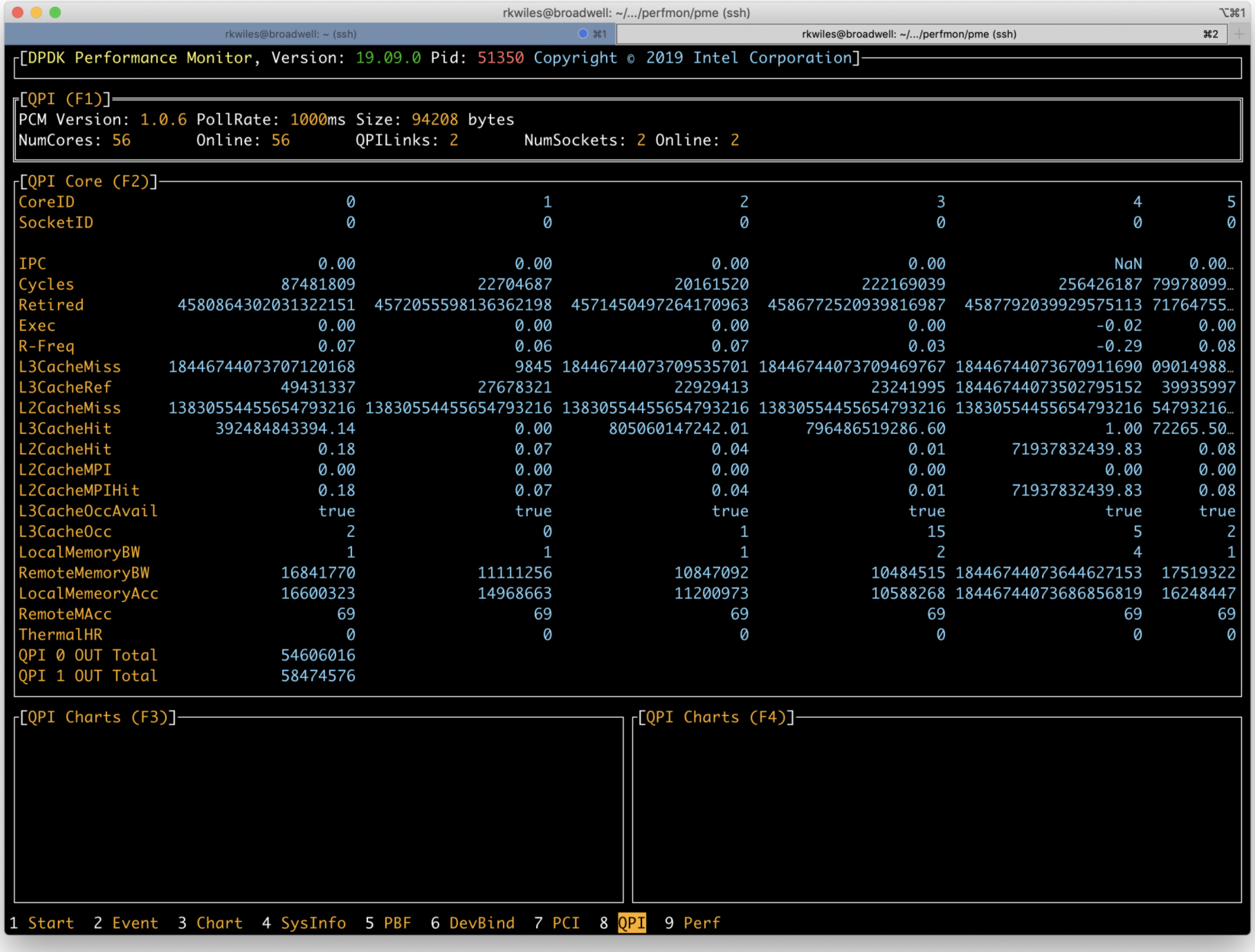
PCI PCM



```
rkwiles@broadwell: ~/.../perfmon/pme (ssh)                                   ⌥⌘1
┌[DPDK Performance Monitor, Version: 0.1.1 Pid: 6849 Copyright © 2019 Intel Corporation]

┌[PCI (p)]
Socket|PCIeRdCur|   RFO|   CRd|   DRd|  ItoM|   PRd|   WiL|PCIeRd(B)|PCIeWr(B)
     0|      1 K|    36|   3 M|   7 M|     0|   1 K|   720|    716 M|      2 K
     1|        0|     0|   3 M|  12 M|     0|     0|     0|   1055 M|        0

┌[PCI Charts (F1)]                                ┌[PCI Charts (F2)]

1 Start  2 Event  3 Chart  4 SysInfo  5 PBF  6 DevBind  7 PCI  8 Perf
```

QPI

(ignore the data it is just a place holder)


rkwiles@broadwell: ~/.../perfmon/pme (ssh)
⌥⌘1
rkwiles@broadwell: ~ (ssh)
⌘1
rkwiles@broadwell: ~/.../perfmon/pme (ssh)
⌘2

```
[DPDK Performance Monitor, Version: 19.09.0 Pid: 51350 Copyright © 2019 Intel Corporation]

[QPI (F1)]
PCM Version: 1.0.6 PollRate: 1000ms Size: 94208 bytes
NumCores: 56       Online: 56       QPILinks: 2       NumSockets: 2 Online: 2

[QPI Core (F2)]
CoreID                             0                    1                    2                    3                    4         5
SocketID                           0                    0                    0                    0                    0         0

IPC                             0.00                 0.00                 0.00                 0.00                  NaN     0.00…
Cycles                      87481809             22704687             20161520            222169039            256426187 79978099…
Retired          4580864302031322151  4572055598136362198  4571450497264170963  4586772520939816987  4587792039929575113 71764755…
Exec                            0.00                 0.00                 0.00                 0.00                -0.02      0.00
R-Freq                          0.07                 0.06                 0.07                 0.03                -0.29      0.08
L3CacheMiss     18446744073707120168                 9845 18446744073709535701 18446744073709469767 18446744073670911690 09014988…
L3CacheRef                  49431337             27678321             22929413             23241995 18446744073502795152  39935997
L2CacheMiss     13830554455654793216 13830554455654793216 13830554455654793216 13830554455654793216 13830554455654793216 54793216…
L3CacheHit           392484843394.14                 0.00      805060147242.01      796486519286.60                 1.00 72265.50…
L2CacheHit                      0.18                 0.07                 0.04                 0.01       71937832439.83      0.08
L2CacheMPI                      0.00                 0.00                 0.00                 0.00                 0.00      0.00
L2CacheMPIHit                   0.18                 0.07                 0.04                 0.01       71937832439.83      0.08
L3CacheOccAvail                 true                 true                 true                 true                 true      true
L3CacheOcc                         2                    0                    1                   15                    5         2
LocalMemoryBW                      1                    1                    1                    2                    4         1
RemoteMemoryBW              16841770             11111256             10847092             10484515 18446744073644627153  17519322
LocalMemeoryAcc             16600323             14968663             11200973             10588268 18446744073686856819  16248447
RemoteMAcc                        69                   69                   69                   69                   69        69
ThermalHR                          0                    0                    0                    0                    0         0
QPI 0 OUT Total             54606016
QPI 1 OUT Total             58474576

[QPI Charts (F3)]                                                  [QPI Charts (F4)]

1 Start  2 Event  3 Chart  4 SysInfo  5 PBF  6 DevBind  7 PCI  8 QPI  9 Perf
```

Questions?

One ask is for help with development and features

Backup slides

How does the tool work?

- To use the tool simply login to a Linux based machine and start the tool running inside an xterm.
- The tool scans the system and finds the DPDK processes by looking at the DPDK fuse filesystem (DFS)
    - The DFS is a library linked or dynamically loaded to a DPDK application
    - The DPDK application does not need to know about the DFS unless the application wants to install files into the FUSE filesystem
    - The DFS creates a base directory @ /dpdk, then creates files and directories
    - Each DPDK instance gets its own directory e.g. /dpdk/dpdk-<PID>
- The tool reads the DFS files to gather information about DPDK

How does it work?

- After the tool has located DPDK applications it displays the data in a set of panels or screens
  - The DPDK applications can be detected dynamically by the tool
- The data in the DFS /dpdk filesystem is a collection of files and directories
  - The data in these files can be any format JSON, binary, ASCII, …
  - The data is created and presented when the files are open (on demand data)
- The application can add files and directories to the filesystem and these files or directories can be dynamically add/removed
- Other applications like collectd or Prometheus(client) can be taught to gather the information from the filesystem

Steps to Observability & Analysis

- Add a few more metrics to DPDK
  - Empty polls counters, histogram of packet RX/TX requests, …
- Enable the metrics in DPDK to be on by default (if possible)
- Expose the metrics from DPDK in the FUSE filesystem
- Gather the metrics and display the data via Prometheus if enabled
  - Perform analysis on the data to determine Hot Spots
  - Provide solutions to remove hot spots or increase performance

Add a few more metrics

- DPDK metrics (possible counters added)
    - Counter for data requests that fail
        - request mempool, pktmbuf, ring, … when no data available
    - Counters for empty RX polling rte_eth_dev_rx_burst() per port/queue
        - Helps to determine idleness of the application core
    - Counters in PMD to measure performance of efficiency of Rx/TX ring
    - Counters in mempool, rte_ring and rte_malloc for allocation/free count
    - Counters for mempool, pktmbuf, … on allocation/free memory or pktmbufs
- A few more data points will need to be added over time

Exposing metric data from DPDK

- DFS (DPDK File System)
- DFS is a FUSE based filesystem similar to /proc or /sys
    - Data is exported in any format needed or multiple formats JSON, ASCII text, binary, …
    - Extract the data via any method just by reading the filesystem files
    - Able to find DPDK applications and command line used to start the application
- Other possible solutions to gather the data are possible