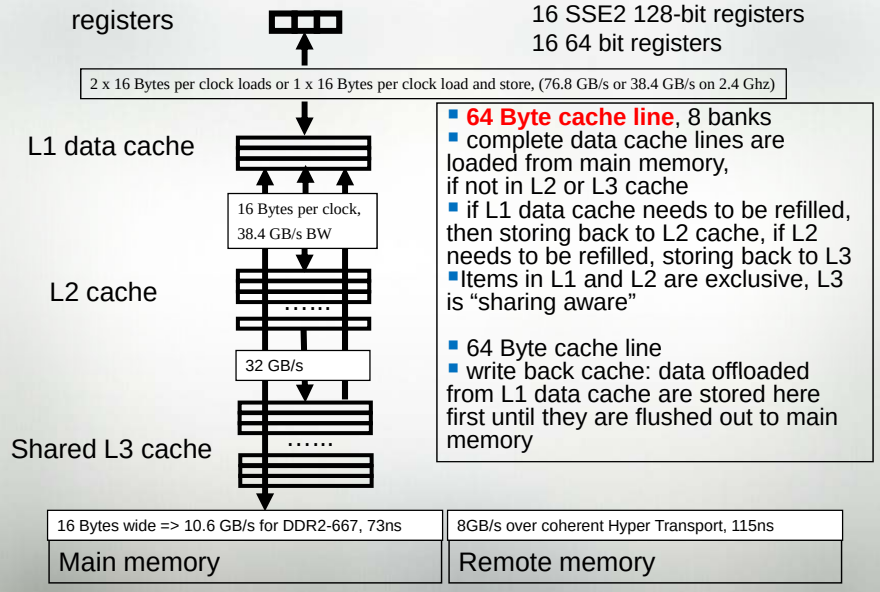
Using Hardware Performance Counters

Luiz DeRose
Programming Environments Director
Cray Inc.
ldr@cray.com

Simplified memory hierachy on the Quad Core AMD Opteron



Hardware Performance Counters

- AMD Opteron Hardware Performance Counters
  - **Four** 48-bit performance counters.
    - Each counter can monitor a single event
      - Count specific processor events
        - the processor increments the counter when it detects an occurrence of the event
        - (e.g., cache misses)
      - Duration of events
        - the processor counts the number of processor clocks it takes to complete an event
        - (e.g., the number of clocks it takes to return data from memory after a cache miss)
  - Time Stamp Counters (TSC)
    - Cycles (user time)

PAPI Predefined Events

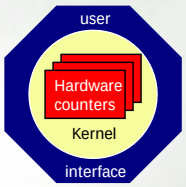
- Common set of events deemed relevant and useful for application performance tuning
  - Accesses to the memory hierarchy, cycle and instruction counts, functional units, pipeline status, etc.
  - The “papi_avail” utility shows which predefined events are available on the system – execute on compute node
- PAPI also provides access to native events
  - The “papi_native_avail” utility lists all AMD native events available on the system – execute on compute node
- Information on PAPI and AMD native events
  - pat_help counters
  - man papi_counters
  - For more information on AMD counters:
    - http://www.amd.com/us-en/assets/content_type/white_papers_and_tech_docs/26049.PDF

Hardware Counters Selection

- PAT_RT_HWPC <set number> | <event list>
  - Specifies hardware counter events to be monitored
    - A set number can be used to select a group of predefined hardware counters events (recommended)
      - CrayPat provides 19 groups on the Cray XT systems
    - Alternatively a list of hardware performance counter event names can be used
      - Maximum of 4 events
    - Both formats can be specified at the same time, with later definitions overriding previous definitions
    - Hardware counter events are not collected by default
    - Hardware counters collection is not supported with sampling on systems running Catamount on the compute nodes

Accuracy Issues

- Granularity of the measured code
  - If not sufficiently large enough, overhead of the counter interfaces may dominate



- Pay attention to what is not measured:
  - Out-of-order processors
  - Speculation
  - Lack of standard on what is counted
    - Microbenchmarks can help determine accuracy of the hardware counters
- For more information on AMD counters:
  - architecture manuals:
    - http://www.amd.com/us-en/assets/content_type/white_papers_and_tech_docs/26049.PDF

Hardware Performance Counters

```
Hardware performance counter events:
  PAPI_L1_DCM  Level 1 data cache misses
  CYCLES_RTC   User Cycles (approx, from rtc)
  PAPI_L1_DCA  Level 1 data cache accesses
  PAPI_TLB_DM  Data translation lookaside buffer misses
  PAPI_FP_OPS  Floating point operations

Estimated minimum overhead per call of a traced function,
  which was subtracted from the data shown in this report
  (for raw data, use the option:  -s overhead=include):
    PAPI_L1_DCM      8.040  misses
    PAPI_TLB_DM      0.005  misses
    PAPI_L1_DCA    474.080  refs
    PAPI_FP_OPS      0.000  ops
    CYCLES_RTC    1863.680  cycles
    Time             0.693  microseconds
```


PAT_RT_HWPC=1 (Summary with TLB)

```
  PAPI_TLB_DM  Data translation lookaside buffer misses
  PAPI_L1_DCA  Level 1 data cache accesses
  PAPI_FP_OPS  Floating point operations
  DC_MISS      Data Cache Miss
  User_Cycles  Virtual Cycles
==========================================================================
USER
--------------------------------------------------------------------------
  Time%                                            98.3%
  Time                                         4.434402 secs
  Imb.Time                                           -- secs
  Imb.Time%                                          --
  Calls                        0.001M/sec         4500.0 calls
  PAPI_L1_DCM                 14.820M/sec       65712197 misses
  PAPI_TLB_DM                  0.902M/sec        3998928 misses
  PAPI_L1_DCA                333.331M/sec     1477996162 refs
  PAPI_FP_OPS                445.571M/sec     1975672594 ops
  User time (approx)           4.434 secs    11971868993 cycles  100.0%Time
  Average Time per Call                        0.000985 sec
  CrayPat Overhead : Time         0.1%
  HW FP Ops / User time      445.571M/sec     1975672594 ops   4.1%peak(DP)
  HW FP Ops / WCT            445.533M/sec
  Computational intensity       0.17 ops/cycle    1.34 ops/ref
  MFLOPS (aggregate)         1782.28M/sec
  TLB utilization             369.60 refs/miss   0.722 avg uses
  D1 cache hit,miss ratios     95.6% hits         4.4% misses
  D1 cache utilization (misses)  22.49 refs/miss   2.811 avg hits
==========================================================================
```

PAT_RT_HWPC=1
Flat profile data
Hard counts
Derived metrics

PAT_RT_HWPC=0 (Summary with Instructions)

```
  PAPI_TOT_INS Instructions count
  PAPI_L1_DCA  Level 1 data cache accesses
  PAPI_FP_OPS  Floating point operations
  PAPI_L1_DCM  Data Cache Miss
  User_Cycles  Virtual Cycles

USER
------------------------------------------------------------------------
  Time%                                          98.6%
  Time                                       4.442352 secs
  Imb.Time                                        -- secs
  Imb.Time%                                       --
  Calls                      0.001M/sec       4500.0 calls
  PAPI_L1_DCM               14.807M/sec     65771688 misses
  PAPI_TOT_INS             530.210M/sec   2355221562 instr
  PAPI_L1_DCA              332.718M/sec   1477953890 refs
  PAPI_FP_OPS              444.765M/sec   1975672594 ops
  User time (approx)         4.442 secs  11993557493 cycles  100.0%Time
  Average Time per Call                     0.000987 sec
  CrayPat Overhead : Time         0.1%
  HW FP Ops / User time    444.765M/sec   1975672594 ops   4.1%peak(DP)
  HW FP Ops / WCT          444.736M/sec
  HW FP Ops / Inst                              83.9%
  Computational intensity       0.16 ops/cycle    1.34 ops/ref
  Instr per cycle                                 0.20 inst/cycle
  MIPS                     2120.84M/sec
  MFLOPS (aggregate)       1779.06M/sec
  Instructions per LD & ST   62.8% refs          1.59 inst/ref
  D1 cache hit,miss ratios   95.5% hits          4.5% misses
  D1 cache utilization (misses)  22.47 refs/miss   2.809 avg hits
========================================================================
```


PAT_RT_HWPC=2 (L1 and L2 Metrics)

```
========================================================================
USER
------------------------------------------------------------------------
  Time%                                          98.3%
  Time                                       4.436808 secs
  Imb.Time                                        -- secs
  Imb.Time%                                       --
  Calls                      0.001M/sec       4500.0 calls
  DATA_CACHE_REFILLS:
    L2_MODIFIED:L2_OWNED:
    L2_EXCLUSIVE:L2_SHARED       9.821M/sec      43567825 fills
  DATA_CACHE_REFILLS_FROM_SYSTEM:
    ALL                        24.743M/sec     109771658 fills
  PAPI_L1_DCM                  14.824M/sec      65765949 misses
  PAPI_L1_DCA                 332.960M/sec    1477145402 refs
  User time (approx)            4.436 secs   11978286133 cycles  100.0%Time
  Average Time per Call                         0.000986 sec
  CrayPat Overhead : Time         0.1%
  D1 cache hit,miss ratios       95.5% hits          4.5% misses
  D1 cache utilization (misses)  22.46 refs/miss    2.808 avg hits
  D1 cache utilization (refills)  9.63 refs/refill  1.204 avg uses
  D2 cache hit,miss ratio        28.4% hits         71.6% misses
  D1+D2 cache hit,miss ratio     96.8% hits          3.2% misses
  D1+D2 cache utilization        31.38 refs/miss    3.922 avg hits
  System to D1 refill          24.743M/sec     109771658 lines
  System to D1 bandwidth      1510.217MB/sec   7025386144 bytes
  D2 to D1 bandwidth           599.398MB/sec   2788340816 bytes
========================================================================
```

PAT_RT_HWPC=3 (Bandwidth)

```
========================================================================
USER
------------------------------------------------------------------------
  Time%                                             98.1%
  Time                                           4.426578 secs
  Imb.Time                                             -- secs
  Imb.Time%                                            --
  Calls                          0.001M/sec       4500.0 calls
  QUADWORDS_WRITTEN_TO_SYSTEM:
    ALL                         99.459M/sec    440235508 ops
  DATA_CACHE_REFILLS:
    L2_MODIFIED:L2_OWNED:
    L2_EXCLUSIVE:L2_SHARED       9.850M/sec     43597156 fills
  DATA_CACHE_REFILLS_FROM_SYSTEM:
    ALL                         24.799M/sec    109769976 fills
  DATA_CACHE_LINES_EVICTED:ALL  59.484M/sec    263295896 ops
  User time (approx)             4.426 secs  11951052461 cycles  100.0%Time
  Average Time per Call                          0.000984 sec
  CrayPat Overhead : Time         0.1%
  System to D1 refill           24.799M/sec    109769976 lines
  System to D1 bandwidth      1513.635MB/sec  7025278496 bytes
  D2 to D1 bandwidth           601.168MB/sec  2790217984 bytes
  L2 to System BW per core    1517.619MB/sec  7043768128 bytes
========================================================================
```


PAT_RT_HWPC=5 (Floating point mix)

```
========================================================================
USER
------------------------------------------------------------------------
  Time%                                             98.4%
  Time                                           4.426552 secs
  Imb.Time                                             -- secs
  Imb.Time%                                            --
  Calls                          0.001M/sec       4500.0 calls
  RETIRED_MMX_AND_FP_INSTRUCTIONS:
    PACKED_SSE_AND_SSE2  454.860M/sec   2013339518 instr
  PAPI_FML_INS           156.443M/sec    692459506 ops
  PAPI_FAD_INS           289.908M/sec   1283213088 ops
  PAPI_FDV_INS             7.418M/sec     32834786 ops
  User time (approx)       4.426 secs  11950955381 cycles  100.0%Time
  Average Time per Call                          0.000984 sec
  CrayPat Overhead : Time   0.1%
  HW FP Ops / Cycles                              0.17 ops/cycle
  HW FP Ops / User time  446.351M/sec   1975672594 ops   4.1%peak(DP)
  HW FP Ops / WCT        446.323M/sec
  FP Multiply / FP Ops                           35.0%
  FP Add / FP Ops                                65.0%
  MFLOPS (aggregate)     1785.40M/sec
========================================================================
```

PAT_RT_HWPC=12 (QC Vectorization)

```
========================================================================
USER
------------------------------------------------------------------------
  Time%                                            98.3%
  Time                                         4.434163 secs
  Imb.Time                                           -- secs
  Imb.Time%                                          --
  Calls                        0.001M/sec       4500.0 calls
  RETIRED_SSE_OPERATIONS:
    SINGLE_ADD_SUB_OPS:
    SINGLE_MUL_OPS                                    0 ops
  RETIRED_SSE_OPERATIONS:
    DOUBLE_ADD_SUB_OPS:
    DOUBLE_MUL_OPS          225.224M/sec    998097162 ops
  RETIRED_SSE_OPERATIONS:
    SINGLE_ADD_SUB_OPS:
    SINGLE_MUL_OPS:OP_TYPE                            0 ops
  RETIRED_SSE_OPERATIONS:
    DOUBLE_ADD_SUB_OPS:
    DOUBLE_MUL_OPS:OP_TYPE  445.818M/sec   1975672594 ops
  User time (approx)           4.432 secs  11965243964 cycles  99.9%Time
  Average Time per Call                        0.000985 sec
  CrayPat Overhead : Time      0.1%
========================================================================
```


Vectorization Example

```
========================================================================
USER / calc2_
------------------------------------------------------------------------
  Time%                                            28.2%
  Time                                         0.600875 secs
  Imb.Time                                     0.069872 secs
  Imb.Time%                                       11.9%
  Calls                        864.9 /sec         500.0 calls
  RETIRED_SSE_OPERATIONS:
    SINGLE_ADD_SUB_OPS:
    SINGLE_MUL_OPS                                    0 ops
  RETIRED_SSE_OPERATIONS:
    DOUBLE_ADD_SUB_OPS:
    DOUBLE_MUL_OPS          369.139M/sec    213408500 ops
  RETIRED_SSE_OPERATIONS:
    SINGLE_ADD_SUB_OPS:
    SINGLE_MUL_OPS:OP_TYPE                            0 ops
  RETIRED_SSE_OPERATIONS:
    DOUBLE_ADD_SUB_OPS:
    DOUBLE_MUL_OPS:OP_TYPE  369.139M/sec    213408500 ops
  User time (approx)           0.578 secs  1271875000 cycles  96.2%Time

 When compiled with fast:
 ========================================================================
 USER / calc2_
 ------------------------------------------------------------------------
   Time%                                           24.3%
   Time                                        0.485654 secs
   Imb.Time                                    0.146551 secs
   Imb.Time%                                      26.4%
   Calls                        0.001M/sec         500.0 calls
   RETIRED_SSE_OPERATIONS:
     SINGLE_ADD_SUB_OPS:
     SINGLE_MUL_OPS                                   0 ops
   RETIRED_SSE_OPERATIONS:
     DOUBLE_ADD_SUB_OPS:
     DOUBLE_MUL_OPS          208.641M/sec    103016531 ops
   RETIRED_SSE_OPERATIONS:
     SINGLE_ADD_SUB_OPS:
     SINGLE_MUL_OPS:OP_TYPE                           0 ops
   RETIRED_SSE_OPERATIONS:
     DOUBLE_ADD_SUB_OPS:
     DOUBLE_MUL_OPS:OP_TYPE  415.628M/sec    205216531 ops
   User time (approx)           0.494 secs  1135625000 cycles  100.0%Time
```

How do I interpret these derived metrics?

- The following thresholds are guidelines to identify if optimization is needed:
  - **Computational Intensity: < 0.5 ops/ref**
    - This is the ratio of FLOPS by L&S
    - Measures how well the floating point unit is being used
  - **FP Multiply / FP Ops** or **FP Add / FP Ops**: **< 25%**
  - **Vectorization: < 1.5**

Memory Hierarchy Thresholds

- **TLB utilization: < 90.0%**
  - Measures how well the memory hierarchy is being utilized with regards to TLB
  - This metric depends on the computation being single precision or double precision
    - A page has 4 Kbytes. So, one page fits 512 double precision words or 1024 single precision words
  - TLB utilization < 1 indicates that not all entries on the page are being utilized between two TLB misses
- **D1 cache utilization: < 1 (D1+D2 cache utilization: < 1)**
  - A cache line has 64 bytes (8 double precision words or 16 single precision words)
  - D1 cache utilization < 1 indicates that not all entries on the cache line are being utilized between two cache misses
- **D1 cache hit (or miss) ratios: < 90% (> 10%)**
- **D2 (L2) cache hit (or miss) ratios: < 95% (> 5%)**
- **D1 + D2 cache hit (or miss) ratios: < 92% (> 8%)**
  - D1 and D2 caches on the Opteron are complementary
  - This metric provides a view of the Total Cache hit (miss) ratio

CRAY
THE SUPERCOMPUTER COMPANY

Using Hardware Performance Counters

Questions / Comments
Thank You!

Luiz DeRose (ldr@cray.com)