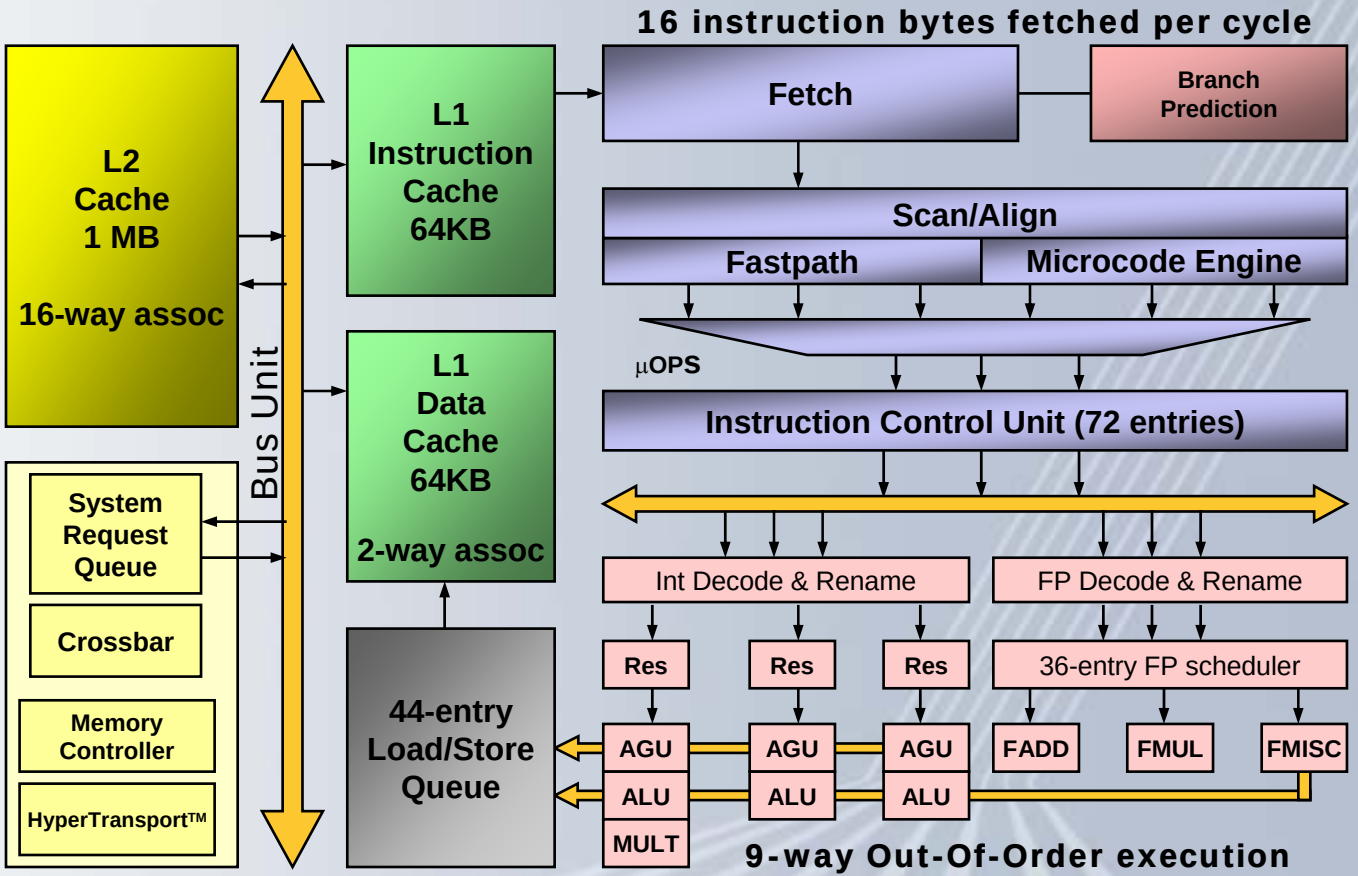
Using Hardware Performance Counters on the Cray XT

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

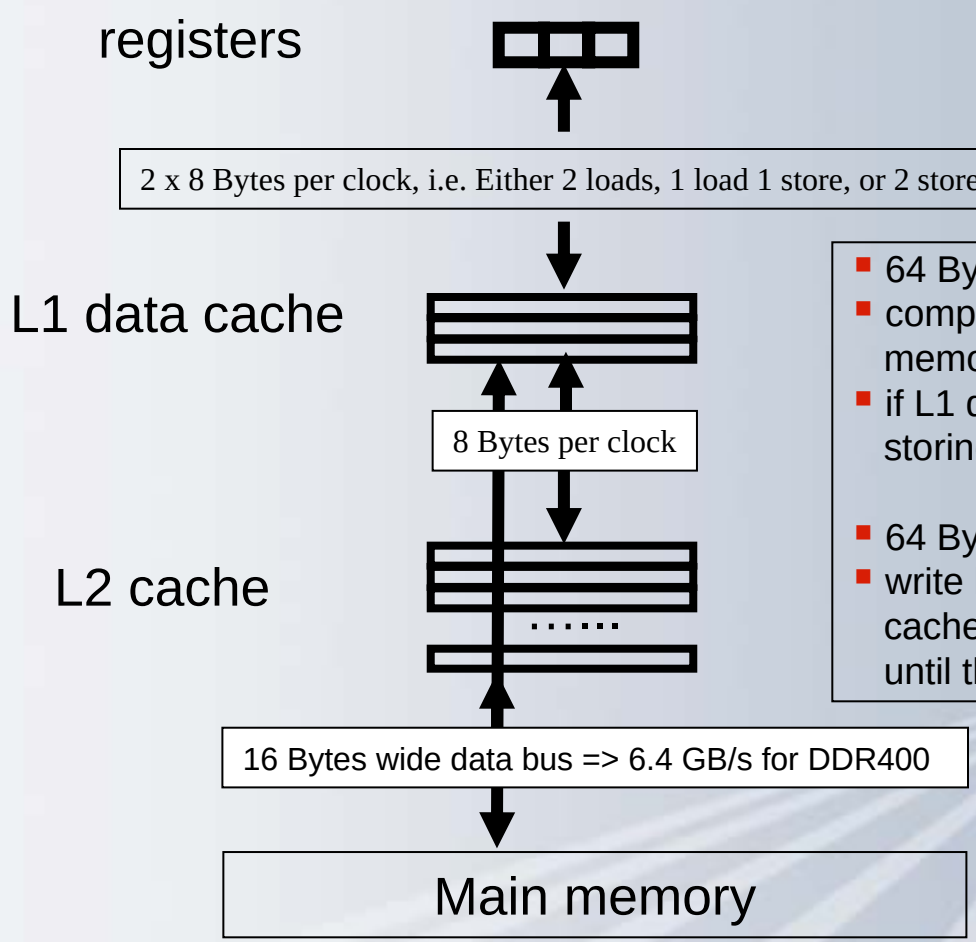
University of Bergen
Bergen, Norway
March 11-14, 2008

AMD Opteron Processor



- 36 entry FPU instruction scheduler
- 64-bit/80-bit FP Realized throughput (1 Mul + 1 Add)/cycle: 1.9 FLOPs/cycle
- 32-bit FP Realized throughput (2 Mul + 2 Add)/cycle: 3.4+ FLOPs/cycle

Simplified memory hierarchy on the AMD Opteron



- 64 Byte cache line
- complete data cache lines are loaded from main memory, if not in L2 cache
- if L1 data cache needs to be refilled, then storing back to L2 cache

- 64 Byte cache line
- write back cache: data offloaded from L1 data cache are stored here first until they are flushed out to main memory

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

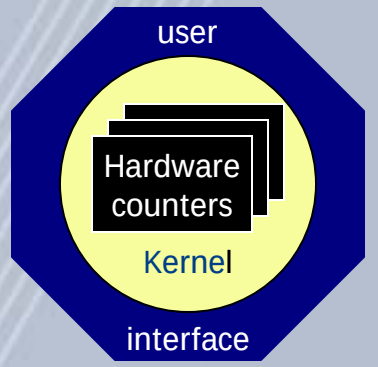
- Common set of events deemed relevant and useful for application performance tuning
  - Accesses to the memory hierarchy, cycle and instruction counts, functional units, pipeline status, etc.
  - The “papi_avail” utility shows which predefined events are available on the system – execute on compute 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 10 group on the Cray XT systems
        - » with CrayPat 3.2.1
    - 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_TOT_INS        Instructions completed
  PAPI_L1_DCA         Level 1 data cache accesses
  PAPI_FP_OPS         Floating point operations
  DATA_CACHE_MISSES   Data Cache Misses
  CYCLES_USER         User Cycles (approx, from clock ticks)

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_TOT_INS          2021.905  instr
    PAPI_L1_DCA           1275.739  refs
    PAPI_FP_OPS              0.000  ops
    DATA_CACHE_MISSES        7.702  misses
    CYCLES_USER              0.000  cycles
    Time                     2.054  microseconds
```

Hardware Performance Counters

```
==========================================================
USER / sweep_
----------------------------------------------------------
  Time%                                            97.1%
  Time                                         3.205429
  Imb.Time                                     0.055034
  Imb.Time%                                        1.7%
  Calls                                              12
  DATA_CACHE_MISSES          25.967M/sec       82965477 misses
  PAPI_TOT_INS             2018.073M/sec     6447742786 instr
  PAPI_L1_DCA               783.213M/sec     2502366829 refs
  PAPI_FP_OPS               590.020M/sec     1885113173 ops
  User time (approx)          3.195 secs     7029000000 cycles
  Cycles                      3.195 secs     7029000000 cycles
  User time (approx)          3.195 secs     7029000000 cycles
  Utilization rate                                99.7%
  Instr per cycle                                  0.92 inst/cycle
  HW FP Ops / Cycles                               0.27 ops/cycle
  HW FP Ops / User time     590.020M/sec     1885113173 ops          13.4%peak
  HW FP Ops / WCT           588.100M/sec
  HW FP Ops / Inst                                29.2%
  Computation intensity                            0.75 ops/ref
  MIPS                    96867.50M/sec
  MFLOPS                  28320.95M/sec
  Instructions per LD ST                           2.58 inst/ref
  LD & ST per D1 miss                             30.16 refs/miss
  D1 cache hit ratio                              96.7%
  LD ST per Instructions                          38.8%
```

PAT_RT_HWPC=0
Flat profile data
Hard counts
Derived metrics

Hardware Performance Counters

```
  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 / sweep_
------------------------------------------------------------------------
  Time%                                          97.5%
  Time                                       3.230138
  Imb.Time                                   0.102759
  Imb.Time%                                      3.1%
  Calls                                           576
  PAPI_TLB_DM                 23.637M/sec      76402282 misses
  PAPI_L1_DCA              42084.031M/sec  136028363961 ops
  PAPI_FP_OPS              31627.272M/sec  102228944600 ops
  DC_MISS                   1196.211M/sec    3866518161 ops
  User time                    3.232 secs    7757528615 cycles
  Utilization rate                              100.0%
  HW FP Ops / Cycles                             13.18 ops/cycle
  HW FP Ops / User time    31627.272M/sec  102228944600 ops        13.7%peak
  HW FP Ops / WCT          31627.272M/sec
  Computation intensity                          0.75 ops/ref
  LD & ST per TLB miss                        1780.42 refs/miss
  LD & ST per D1 miss                           35.18 refs/miss
  D1 cache hit ratio                            97.2%
  % TLB misses / cycle                           0.0%
```

PAT_RT_HWPC=2 (Cache Info)

```
  PAPI_L1_DCA          Level 1 data cache accesses
  DC_L2_REFILL_MOESI   Refill from L2. Cache bits: Modified Owner Exclusive Shared Invalid
  DC_SYS_REFILL_MOESI  Refill from system. Cache bits: Modified Owner Exclusive Shared Invalid
  BU_L2_REQ_DC         Internal L2 request - DC fill
  User_Cycles          Virtual Cycles
=============================================================================
USER / sweep_
-----------------------------------------------------------------------------
  Time%                                            97.6%
  Time                                           3.229766
  Imb.Time                                       0.102401
  Imb.Time%                                         3.1%
  Calls                                              576
  PAPI_L1_DCA             42089.216M/sec   136029595956 ops
  DC_L2_REFILL_MOESI       1194.502M/sec     3860550881 ops
  DC_SYS_REFILL_MOESI       275.938M/sec      891814098 ops
  BU_L2_REQ_DC             1228.504M/sec     3970443642 req
  User time                   3.232 secs     7756643255 cycles
  Utilization rate                              100.0%
  L1 Data cache misses     1470.440M/sec     4752364979 misses
  LD & ST per D1 miss                            28.62 refs/miss
  D1 cache hit ratio                             96.5%
  LD & ST per D2 miss                           152.53 refs/miss
  D2 cache hit ratio                             77.5%
  L2 cache hit ratio                             81.2%
  Total cache hit ratio                          99.3%
  Effective Reuse                                 2.38 refs/byte
  Memory to D1 refill       275.938M/sec      891814098 lines
  Memory to D1 bandwidth  16841.928MB/sec   57076102272 bytes
  L2 to Dcache bandwidth  72906.587MB/sec  247075256384 bytes
=============================================================================
```

PAT_RT_HWPC=3 (Bandwidth)

```
  DC_L2_REFILL_MOESI   Refill from L2. Cache bits: Modified Owner Exclusive Shared Invalid
  DC_SYS_REFILL_MOESI  Refill from system. Cache bits: Modified Owner Exclusive Shared Invalid
  DC_COPYBACK_MOESI    Copyback. Cache bits: Modified Owner Exclusive Shared Invalid
  SI_QUAD_WRITE        Quadwords Written to System: Quadword write transfer
  User_Cycles          Virtual Cycles
=====================================================================
USER / sweep_
---------------------------------------------------------------------
  Time%                                          97.5%
  Time                                       3.229972
  Imb.Time                                   0.102654
  Imb.Time%                                      3.1%
  Calls                                           576
  DC_L2_REFILL_MOESI       1196.069M/sec     3865961067 ops
  DC_SYS_REFILL_MOESI       275.951M/sec      891933730 ops
  DC_COPYBACK_MOESI        1472.059M/sec     4758021002 ops
  SI_QUAD_WRITE            2130.392M/sec     6885899519 ops
  User time                   3.232 secs     7757332017 cycles
  Utilization rate                              100.0%
  L1 Data cache misses     1472.020M/sec     4757894797 misses
  L2 cache hit ratio                             81.3%
  Memory to D1 refill       275.951M/sec      891933730 lines
  Memory to D1 bandwidth  16842.692MB/sec   57083758720 bytes
  L2 to Dcache bandwidth  73002.276MB/sec  247421508288 bytes
  Dcache to L2 bandwidth  89847.351MB/sec  304513344128 bytes
  L2 to Memory bandwidth  16253.602MB/sec   55087196152 bytes
=====================================================================
```

PAT_RT_HWPC=4 (HT Data Transfers)

```
  SI_QUAD_WRITE     Quadwords Written to System: Quadword write transfer
  HT_LL_MEM_XFR     HyperTransport data transfer from local memory to local memory
  HT_LL_IO_XFR      HyperTransport data transfer from local memory to local IO
  HT_LL_IO_MEM_XFR  HyperTransport data transfer from local IO to local memory
  User_Cycles       Virtual Cycles
=============================================================================
USER / sweep_
-----------------------------------------------------------------------------
  Time%                                              97.6%
  Time                                             3.229909
  Imb.Time                                         0.102472
  Imb.Time%                                            3.1%
  Calls                                                576
  SI_QUAD_WRITE             2130.809M/sec    6886944175 ops
  HT_LL_MEM_XFR             2229.315M/sec    7205322525 ops
  HT_LL_IO_XFR                 0.316M/sec       1022602 ops
  HT_LL_IO_MEM_XFR             4.333M/sec      14003499 ops
  User time                    3.232 secs    7756989744 cycles
  Utilization rate                              100.0%
  L2 to Memory bandwidth  16256.785MB/sec  55095553400 bytes
```

PAT_RT_HWPC=5 (FP & Vectorization)

```
  PAPI_FML_INS            Floating point multiply instructions
  PAPI_FAD_INS            Floating point add instructions
  FR_FPU_SSE_SSE2_PACKED  Retired FPU instructions - Combined packed SSE and SSE2 instructions
  FR_FPU_SSE_SSE2_SCALAR  Retired FPU instructions - Combined scalar SSE and SSE2 instructions
  User_Cycles             Virtual Cycles
=============================================================================
USER / sweep_
-----------------------------------------------------------------------------
  Time%                                          97.5%
  Time                                       3.230243
  Imb.Time                                   0.102413
  Imb.Time%                                      3.1%
  Calls                                           576
  PAPI_FML_INS            14751.121M/sec   47682418308 instr
  PAPI_FAD_INS            16874.595M/sec   54546466150 instr
  FR_FPU_SSE_SSE2_PACKED                           0 instr
  FR_FPU_SSE_SSE2_SCALAR  43000.358M/sec  138996966424 instr
  User time                   3.232 secs    7757905716 cycles
  Utilization rate                             100.0%
  HW FP Ops / Cycles                            13.18 ops/cycle
  HW FP Ops / User time   31625.716M/sec  102228884458 ops        13.7%peak
  HW FP Ops / WCT         31625.716M/sec
  FP Multiply / FP Ops                          46.6%
  FP Add / FP Ops                               53.4%
When compiled with fastsse:
=============================================================================
USER / sweep_
-----------------------------------------------------------------------------
  Time%                                          97.0%
  Time                                       2.577571
  Imb.Time                                   0.101843
  Imb.Time%                                      3.9%
  Calls                                           576
  PAPI_FML_INS            16061.952M/sec   41438628312 instr
  PAPI_FAD_INS            18681.139M/sec   48195934483 instr
  FR_FPU_SSE_SSE2_PACKED   9315.154M/sec   24032397312 instr
  FR_FPU_SSE_SSE2_SCALAR  39220.314M/sec  101185461233 instr
  User time                   2.580 secs    6191819596 cycles
  Utilization rate                             100.0%
  HW FP Ops / Cycles                            14.48 ops/cycle
  HW FP Ops / User time   34743.091M/sec   89634562795 ops        15.1%peak
  HW FP Ops / WCT         34743.091M/sec
  FP Multiply / FP Ops                          46.2%
  FP Add / FP Ops                               53.8%
```

PAT_RT_HWPC=6 (Stalls / Resources Idle)

```
  PAPI_FPU_IDL    Cycles floating point units are idle
  PAPI_STL_ICY    Cycles with no instruction issue
  PAPI_RES_STL    Cycles stalled on any resource
  IC_FETCH_STALL  Instruction fetch stall
  User_Cycles     Virtual Cycles
==========================================================================
USER / sweep_
--------------------------------------------------------------------------
  Time%                                              96.9%
  Time                                           2.619334
  Imb.Time                                       0.141884
  Imb.Time%                                          5.2%
  Calls                                               576
  PAPI_FPU_IDL           0.209 secs      500570926.75 cycles
  PAPI_STL_ICY           0.029 secs  70217803.2916667 cycles
  PAPI_RES_STL           1.725 secs     4140098264.75 cycles
  IC_FETCH_STALL         1.930 secs  4631824703.22917 cycles
  User time              2.622 secs  6292743345.83333 cycles
  Utilization rate                                100.0%
  Total time stalled     1.725 secs     4140098264.75 cycles  65.8%
  Time I Fetch Stalled   1.930 secs  4631824703.22917 cycles  73.6%
  Avg Time FPUs idle     0.104 secs     250285463.375 cycles   4.0%
  Time Decoder empty     0.029 secs  70217803.2916667 cycles   1.1%
==========================================================================
```

PAT_RT_HWPC=7 (Stalls/ Resources Full)

```
  FR_DECODER_EMPTY             Nothing to dispatch - decoder empty
  FR_DISPATCH_STALLS           Dispatch stalls - D2h or DAh combined
  FR_DISPATCH_STALLS_FULL_FPU  Dispatch stall when FPU is full
  FR_DISPATCH_STALLS_FULL_LS   Dispatch stall when LS is full
  User_Cycles                  Virtual Cycles
==========================================================================
USER / sweep_
--------------------------------------------------------------------------
  Time%                                                  97.0%
  Time                                                2.618878
  Imb.Time                                            0.142062
  Imb.Time%                                               5.3%
  Calls                                                    576
  FR_DECODER_EMPTY             1281.985M/sec        3360773456 ops
  FR_DISPATCH_STALLS              1.725 secs    4139586865.875 cycles
  FR_DISPATCH_STALLS_FULL_FPU     1.118 secs  2683961106.04167 cycles
  FR_DISPATCH_STALLS_FULL_LS      0.438 secs  1050422214.02083 cycles
  User time                       2.622 secs   6291691461.4375 cycles
  Utilization rate                                    100.0%
  Total time stalled              1.725 secs    4139586865.875 cycles  65.8%
  Avg Time FPUs stalled           0.559 secs  1341980553.02083 cycles  21.3%
  Avg Time LSs stalled            0.219 secs  525211107.010417 cycles   8.3%
  Time Decoder empty              1.400 secs        3360773456 cycles  53.4%
==========================================================================
```

PAT_RT_HWPC Other Sets

```
-----------------------------------------------------------------
Set 8: Branches
  PAPI_BR_TKN   Conditional branch instructions taken
  PAPI_BR_MSP   Conditional branch instructions mispredicted
  PAPI_TOT_INS  Instructions completed
  IC_MISS       IC Miss
  User_Cycles   Virtual Cycles


-----------------------------------------------------------------

Set 9: Instructions
  PAPI_L2_ICM   Level 2 instruction cache misses
  PAPI_L1_ICA   Level 1 instruction cache accesses
  IC_MISS       IC Miss
  IC_L2_REFILL  Refill from L2
  User_Cycles   Virtual Cycles
```

Using Hardware Performance Counters on the Cray XT

Questions / Comments

Thank You!

University of Bergen
Bergen, Norway
March 11-14, 2008

Luiz DeRose (ldr@cray.com)

CRAY
THE SUPERCOMPUTER COMPANY