Debugging and Optimization Tools

Richard Gerber
NERSC
User Services Group Lead

Thanks to Woo-Sun Yang and Helen He
Outline

• Take-Aways
• Debugging
• Performance / Optimization
• NERSC “automatic” tools

Videos, presentations, and references:

http://www.nersc.gov/users/training/courses/CS267/

Help at NERSC
consult@nersc.gov
https://help.nersc.gov
Take-Aways

• Tools can help you find errors in your program and locate performance bottlenecks
• In the world of HPC parallel computing, there are few widely adopted standard tools – try some and see what works for you
  – Totalview and DDT debuggers
  – PAPI, Tau, various research tools, & vendor-specific performance tools
• Common code problems
• How tools work in general
• Be suspicious of outliers among parallel tasks
• Where to get more information
Debugging
What is a Bug?

- A bug is when your code
  - crashes
  - hangs (doesn’t finish)
  - gets inconsistent answers
  - produces wrong answers
  - behaves in any way you didn’t want it to

The term “bug” was popularized by Grace Hopper (motivated by the removal of an actual moth from a computer relay in 1947)
Common Causes of Bugs

• “Serial” (Sequential might be a better word)
  – Invalid memory references
  – Array reference out of bounds
  – Divide by zero
  – Use of uninitialized variables

• Parallel  Let’s concentrate on these
  – Unmatched sends/receives
  – Blocking receive before corresponding send
  – Out of order collectives
  – Race conditions
  – Unintentionally modifying shared memory structures
What to Do if You Have a Bug?

• Find It
  – You want to locate the part of your code that isn’t doing what it’s designed to do

• Fix It
  – Figure out how to solve it and implement a solution

• Check It
  – Run it to check for proper behavior
A is a bug that manifests only after someone reading source code or using the program in an unusual way now that it never should have worked in the first place, at which point the program promptly stops working for everybody until fixed.
Find It: Tools

• **printf, write**
  – Versatile, sometimes useful
  – Doesn’t scale well
  – Not interactive
  – Fishing expedition

• **Compiler / Runtime**
  – Bounds checking, exception handling
  – Dereferencing of NULL pointers
  – Function and subroutine interface checking

• **Serial gdb + friends**
  – GNU debugger, serial, command-line interface
  – See “man gdb”

• **Parallel debuggers**
  – DDT
  – Totalview

• **Memory debuggers**
  – MAP
  – Valgrind

See NERSC web site
https://www.nersc.gov/users/software/debugging-and-profiling/
This code hangs because both Task 0 and Task N-1 are blocking on MPI_Recv
NERSC NX SERVICE - X-WINDOWS ACCELERATION AT NERSC

Introduction

NX is a computer program that handles remote X Window System connections and it provides three benefits for NERSC users:

- **SPEED**: NX can greatly improve the performance of X Windows, allowing users with slow, high latency connections (e.g., on cell phone network, traveling in Africa) to use complex X Windows programs (such as rotating a plot in Matlab).

- **SESSION**: NX provides sessions that allow a user to disconnect from the session and reconnect to it at a later time while keeping the state of all running applications inside the session.

- **DESKTOP**: NX gives users a virtual desktop that's running at NERSC. You can customize the desktop according to your work requirement.
Compile & Start DDT

Compile for debugging

```
edison% make
cc -c -g hello.c
cc -o hello -g hello.o
```

Set up the parallel run environment

```
edison% qsub -I -V -lmppwidth=24
edison% cd $PBS_O_WORKDIR
```

Start the DDT debugger

```
edison% module load ddt
edison% ddt ./hello
```
Press Go and then Pause when code appears hung.

Task 0 is at line 44

At hang, tasks are in 3 different places.
What About Massive Parallelism?

• With 10K+ tasks/threads/streams it’s impossible to examine every parallel instance in detail
• Make use of statistics and summaries
• Look for tasks that are doing something different
  – Amount of memory used
  – Number of calculations performed (from counters)
  – Number of MPI calls
  – Wall time used
  – Time spent in I/O
  – One or a few tasks paused at a different line of code
• We (NERSC) have been advocating for this statistical view for some time
Vendors are starting to listen (DDT)
Debuggers on NERSC machines

• **Parallel debuggers with a graphical user interface**
  – DDT (Distributed Debugging Tool)
  – TotalView

• **Specialized debuggers on Hopper and Edison**
  – STAT (Stack Trace Analysis Tool)
    • Collect stack backtraces from all (MPI) tasks
  – ATP (Abnormal Termination Processing)
    • Collect stack backtraces from all (MPI) tasks when an application fails
  – CCDB (Cray Comparative Debugger)
    • Comparative debugging

• **Valgrind**
  – Suite of debugging and profiler tools

https://www.nersc.gov/users/training/courses/CS267/ for links to recent training presentations

SlideCourtesyofWoo-SunYang
STAT (Stack Trace Analysis Tool)

• Gathers stack backtraces (showing the function calling sequences leading up to the ones in the current stack frames) from all (MPI) processes and merges them into a single file (*.dot)
  – Results displayed graphically as a call tree showing the location in the code that each process is executing and how it got there
  – Can be useful for debugging a hung application
  – With the info learned from STAT, can investigate further with DDT or TotalView

• Works for MPI, CAF and UPC, but not OpenMP

• For more info:
  – ‘intro_stat’, ‘STAT’, ‘statview’ and ‘statgui’ man pages
Hung application with STAT (Cont’d)

Ranks 1 & 2 are here
Rank 0 is here
Rank 3 is here

Slide Courtesy of Woo-Sun Yang
ATP (Abnormal Termination Processing)

- ATP gathers stack backtraces from all processes of a failing application
  - Invokes STAT underneath
  - Output in atpMergedBT.dot and atpMergedBT_line.dot (which shows source code line numbers), which are to be viewed with statview

- By default, the atp module is loaded on Hopper and Edison, but ATP is *not* enabled; to enable:
  - `setenv ATP_ENABLED 1`  # csh/tcsh
  - `export ATP_ENABLED=1`  # sh/bash/ksh

- For more info
  - ‘intro_atp’ man page
CCDB (Cray Comparative Debugger)

• Find a bug introduced in a version, by running two versions side by side and comparing data between them

• GUI

• Supports MPI; doesn’t support threading

• For info:
  – ccdb man page and help pages
  – lgdb man page and help pages
Running CCDB

```bash
% qsub -IV -lmppwidth=48,walltime=30:00 -q debug
% cd $PBS_O_WORKDIR
% module load cray-ccdb
% ccdb
```

Request enough nodes to run two apps simultaneously.

Slide Courtesy of Woo-Sun Yang
Valgrind

• Suite of debugging and profiler tools
• Tools include
  – memcheck: memory error and memory leaks detection
  – cachegrind: a cache and branch-prediction profiler
  – callgrind: a call-graph generating cache and branch prediction profiler
  – massif, dhat (exp-dhat): heap profilers
  – helgrind, drd: pthreads error detectors
• For info:
Performance / Optimization
Performance Questions

• How can we tell if a program is performing well? Or isn’t? What is “good”?

• If performance is not “good,” can we identify the causes?

• What can we do about it?
Is Your Code Performing Well?

• No single answer, but
  – Does is scale well?
  – Is MPI time <20% of total run time?
  – Is I/O time <10% of total run time?
  – Is it load balanced?
  – If GPU code, does GPU+Processor perform better than 2 Processors?

• “Theoretical” CPU performance vs. “Real World” performance in a highly parallel environment
  – Cache-based x86 processors: >10% of theoretical is pretty good
  – GPUs, Xeon Phi: >few% in today’s real full HPC applications pretty good? This your challenge!
What can we do about it

- Minimize latency effects (aggregate messages)
- Maximize work vs. communication
- Minimize data movement (recalculate vs. send)
- Use the “most local” memory
- Take advantage of vector (SIMD) capabilities
- Use large-block I/O
- Use a balanced strategy for I/O
  - Avoid “too many” tasks accessing a single file, but “too many” files performs poorly (~1000s)
  - Use “enough” I/O tasks to maximum I/O bandwidth, but “too many” causes contention 1/node
Performance Landscape
Can We Identify the Causes? Use Tools

• Vendor Tools:
  – CrayPat on Crays
  – INTEL VTune

• Community Tools:
  – TAU (U. Oregon via ACTS)
  – PAPI (Performance API)
  – gprof
  – HPC Toolkit
  – Scalasca

• NERSC “automatic” and/or easy-to-use tools
  – e.g. IPM, Darshan

See NERSC web site
https://www.nersc.gov/users/software/debugging-and-profiling/
Example: CrayPat

• Suite of tools that provides a wide range of performance-related information

• Can be used for both sampling and tracing
  – with or without hardware or network performance counters
  – Built on PAPI

• Supports Fortran, C, C++, UPC, MPI, Coarray Fortran, OpenMP, Pthreads, SHMEM

• Man pages
  – intro_craypat(1), intro_app2(1), intro_papi(1)
Using CrayPat

1. **Access the tools**
   - module load perftools

2. **Build your application; keep .o files**
   - make clean
   - make

3. **Instrument application**
   - pat_build ... a.out
   - Result is a new file, a.out+pat

4. **Run instrumented application to get top time consuming routines**
   - aprun ... a.out+pat
   - Result is a new file XXXXX.xf (or a directory containing .xf files)

5. **Run pat_report on that new file; view results**
   - pat_report XXXXX.xf > my_profile
   - view my_profile
   - Also produces a new file: XXXXX.ap2 that can be viewed with apprentice GUI application
Cray perftools and perftools-lite

• **Reports:**
  – execution time
  – memory high water mark
  – aggregate FLOPS rate
  – top time consuming user function
  – MPI information
  – IO information
  – hardware performance counters
  – load balance ...

• **Start with perftools-lite**

• **Available on Hopper and Edison.**

• **Documentation:**

---

**Number of PEs (MPI ranks):** 240  
**Numbers of PEs per Node:** 24  
**Number of Cores per Socket:** 12  
**Execution start time:** Sun Feb 2 13:38:33 2014  
**System name and speed:** nid01665 2401 MHz  
**Wall Clock Time:** 290.822940 secs  
**High Memory:** 243.36 MBytes  
**MFLOPS (aggregate):** Not supported (see observation below)  
**I/O Read Rate:** 46.30 MBytes/Sec  
**I/O Write Rate:** 5.91 MBytes/Sec  

Table 1: Profile by Function Group and Function (top 10 functions shown)

| 100.0% | 28484.6 | -- | -- | Total |--------------------------------------------------|
| 61.8%  | 17598.4 | -- | -- | USER |--------------------------------------------------|
| 36.3%  | 10328.2 | 58.8 | 0.6% | decompmod_initdecomp |--------------------------------------------------|
| 29.6%  | 8432.1  | -- | -- | MPI |--------------------------------------------------|
| 9.0%   | 2571.0  | 129.0 | 4.8% | MPI_GATHERV |--------------------------------------------------|
Allinea MAP

- Allinea MAP is a parallel MPI profiler with GUI, small overhead.
- Reports: Memory usage, MPI usage, CPU time, CPU instructions, I/O, etc. as a function of time.
- Available on Hopper, Edison, and Carver.

Documentation:

http://www.nersc.gov/users/software/debugging-and-profiling/MAP/

http://www.allinea.com/products/map/
### Vampir and Vampirtrace

|----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
• **Using even the best tools can be tedious**
  – “Follow these 10 steps to perform the basic analysis of your program” – from a supercomputer center web site for a well-known tool

• **NERSC wants to enable easy access to information that can help you improve your parallel code**
  – *automatic* data collection
  – provide useful tools through the web

• **Efforts**
  – Work with vendors (e.g., CRAY ARU, Allinea Perf. Report)
  – IPM (MPI profiling, chip HW counters, memory used)
  – Accounting & UNIX resource usage
  – System-level I/O monitoring
  – User-level I/O profiling (Darshan)
## NERSC Completed Jobs

<table>
<thead>
<tr>
<th>#</th>
<th>Host</th>
<th>JobID</th>
<th>Job Name</th>
<th>User</th>
<th>Nds</th>
<th>Complete</th>
<th>Wall hrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>edison</td>
<td>782557</td>
<td>my_job</td>
<td>hdixit</td>
<td>11</td>
<td>02/17/14 13:14</td>
<td>0.160</td>
</tr>
<tr>
<td>1</td>
<td>edison</td>
<td>782431</td>
<td>cp_ref_big</td>
<td>hsinyu</td>
<td>13</td>
<td>02/17/14 13:14</td>
<td>0.849</td>
</tr>
<tr>
<td>2</td>
<td>edison</td>
<td>782591</td>
<td>tvsoi</td>
<td>wangyu</td>
<td>2</td>
<td>02/17/14 13:14</td>
<td>0.005</td>
</tr>
<tr>
<td>3</td>
<td>edison</td>
<td>782560</td>
<td>my_job</td>
<td>hdixit</td>
<td>11</td>
<td>02/17/14 13:13</td>
<td>0.131</td>
</tr>
<tr>
<td>4</td>
<td>edison</td>
<td>719618</td>
<td>nacl_big</td>
<td>carnevav</td>
<td>16</td>
<td>02/17/14 13:12</td>
<td>11.453</td>
</tr>
<tr>
<td>5</td>
<td>edison</td>
<td>782403</td>
<td>slab-Ni-pa...</td>
<td>[Full Name]</td>
<td>hdixit</td>
<td>11</td>
<td>02/17/14 13:12</td>
</tr>
<tr>
<td>6</td>
<td>edison</td>
<td>782590</td>
<td>NL_test</td>
<td>khkim</td>
<td>86</td>
<td>02/17/14 13:12</td>
<td>0.008</td>
</tr>
</tbody>
</table>

### Command Details

**APID**: 3140812

**Command**: gribmeanp.x

**Nodes**: 3

**Tasks**: 64

**Threads per Task**: 1

**Tasks per Node**: 24

**Max Task Mem (MB)**: 235

**Run Time (secs)**: 5

**Command Line**:

```
/opt/cray/alps/default/bin/aprun -n 64 -N 24
/scratch1/scratchdirs/whitaker/edison/newstuff/bin/gribmeanp.x
/scratch1/scratchdirs/whitaker/gfsenkf_t126_1999iau_bias
```

**Node List**: 2628-2630
IPM: An Easy to Use Performance Tool

Just load the module, relink, and run.

```
# host   : s05601/006035314C00_AIX
# start  : 11/30/04/14:35:34
# stop   : 11/30/04/14:36:00
# mpi_tasks : 32 on 2 nodes
# gbytes : 6.65863e-01 total
mpi_tasks : 32 on 2 nodes
wallclock : 29.975184 sec
comm : 27.72

gbytes : 6.65863e-01 total
gflop/sec : 2.33478e+00 total

[total]  [avg]  min   max
wallclock  953.272  29.7897  29.6092  29.9752
user       837.25   26.1641  25.71    26.92
system     60.6     1.89375  1.52     2.59
mpi        264.267  8.25834  7.73025  8.70985
%comm      27.7234  25.8873  29.3705
gflop/sec  2.33478  0.0729619 0.0722048 0.0745817
gbytes     0.665863 0.0208082 0.0195503 0.0237541

PM_FPU0_CMPL  2.28827e+10  7.15084e+08  7.07373e+08  7.30171e+08
PM_FPU1_CMPL  1.70657e+10  5.33304e+08  5.28487e+08  5.42882e+08
PM_FPU_FMA    3.00371e+10  9.3866e+08  9.27762e+08  9.62547e+08
PM_INST_CMPL  2.78819e+11  8.71309e+09  8.20981e+09  9.21761e+09
PM_LD_CMPL    1.25478e+11  3.92118e+09  3.74541e+09  4.11658e+09
PM_ST_CMPL    7.45961e+10  2.33113e+09  2.21164e+09  2.46327e+09
PM_TLB_MISS   2.45894e+08  7.68418e+06  6.98733e+06  2.05724e+07
PM_CYC        3.0575e+11  9.55467e+09  9.36585e+09  9.62227e+09

MPI_Send     188.386  639616  71.29    19.76
MPI_Wait     69.5032  639616  26.30    7.29
MPI_Irecv    6.34936  639616  2.40     0.67
MPI_Barrier  0.0177442 32  0.01    0.00
MPI_Reduce   0.00540609 32  0.00    0.00
MPI_Comm_rank 0.00465156 32  0.00    0.00
MPI_Comm_size 0.000145341 32  0.00    0.00
```
IPM Data on NERSC Web Site

NERSC job details


Task distribution of IPM Summary Statistics for Jobstep 619349

<table>
<thead>
<tr>
<th>Metric</th>
<th>Sum</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>CV (%)</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregate Floating Point Operations (Flop x 10**9)</td>
<td>3.011e+02</td>
<td>1.470e-01</td>
<td>4.968e-03</td>
<td>3.36e+00</td>
<td>1.395e-01</td>
<td>2.161e-01</td>
</tr>
<tr>
<td>GFlop/sec</td>
<td>6.147e-01</td>
<td>3.002e-04</td>
<td>1.008e-05</td>
<td>3.36e+00</td>
<td>2.847e-04</td>
<td>4.411e-04</td>
</tr>
<tr>
<td>Maximum Memory Usage (Bytes)</td>
<td>4.101e+02</td>
<td>2.002e-01</td>
<td>9.806e-03</td>
<td>4.80e+00</td>
<td>1.78e-01</td>
<td>2.448e-01</td>
</tr>
<tr>
<td>Time Spent in MPI Routines (sec)</td>
<td>1.228e+00</td>
<td>5.958e-02</td>
<td>4.984e+00</td>
<td>3.81e+00</td>
<td>5.177e+02</td>
<td>6.801e+02</td>
</tr>
<tr>
<td>Wallclock Time (sec)</td>
<td>1.003e+00</td>
<td>4.898e-02</td>
<td>6.428e+00</td>
<td>1.31e-02</td>
<td>4.899e+02</td>
<td>4.927e+02</td>
</tr>
</tbody>
</table>

CV = Coefficient of Variance = (Standard Deviation / Mean)

The MPI rank is the sum of the column and row indices in the table.

Table Columns: 64:2
IPM Examples

Time spent by each task in $\text{MPI\_Recv}$ as a percentage of the maximum value.

The MPI rank represented by each cell in the table is the sum of the cell's column and row indices.

Table Columns: 32 32

Time vs. MPI Rank for $\text{MPI\_Recv}$
NERSC will be installing a 30-petaflop/s Intel KNL-based Cray system in the 2016 time frame, named after American Biochemist Gerty Cori.

- Over 9,300 single-socket nodes in the system with each node > 3TeraFLOPS/s theoretical peak performance.
- Cray Aries high speed "dragonfly" topology interconnect.
- Cabinets, Liquid cooling.
- Lustre filesystem with > 430 GB/sec I/O bandwidth and 28 PB of disk capacity; 1-2 TB/sec Burst Buffer bandwidth.
• Single socket node with NUMA
• Greater than 60 cores per node with up to 4 hardware threads each
• AVX512 Vector pipelines with a hardware vector length of 512 bits (eight double-precision elements)
• Better performance per watt than previous generation Xeon Phi™ systems and 3X single-thread performance
• 64-128 GB of DRAM memory per node
• On-package, high-bandwidth memory, up to 16GB capacity with bandwidth projected to be 5X that of DDR4
NESAP Tools Partners

- HPCToolkit
  - [http://hpctoolkit.org/](http://hpctoolkit.org/)
- TAU
  - [http://www.cs.uoregon.edu/research/tau/home.php](http://www.cs.uoregon.edu/research/tau/home.php)
- Open|Speedshop
- PAPI
- MAP (Allinea)
- Vampir
  - [https://www.vampir.eu/](https://www.vampir.eu/)
- PerfExpert
  - [https://www.tacc.utexas.edu/research-development/tacc-projects/perfexpert](https://www.tacc.utexas.edu/research-development/tacc-projects/perfexpert)
- Scalasca
  - [http://www.scalasca.org/](http://www.scalasca.org/)
- VTUNE
- TotalView
- DDT (Allinea)
Challenges with debugging and profiling on Cori

- Debugging and performance tools need to scale up to thousands (or more?) of threads and tasks
- NESAP participants will want to know
  - Hotspots
  - Memory usage, bandwidth and errors with on-package memory
  - Cache utilization
  - MPI communication performance - same interconnect fabrics for more powerful nodes
  - Threading performance
  - I/O statistics
  - ...

NERSC
Summary

- Debugging and Parallel Code Optimization can be hard
- Tools can help
  - See NERSC web pages for recommendations
  - Use the ones that work for you
- Be aware of some of the more common errors and best practices
- Look for outliers in parallel programs
- Refer to NERSC web pages for details
  - [http://www.nersc.gov/users/training/courses/CS267/](http://www.nersc.gov/users/training/courses/CS267/)
Job Physical Topology
Users can see the system-wide I/O activity while their job ran to look for contention.
**IPM Examples**

Time spent by each task in **MPI_Allreduce** as a percentage of the maximum value.

The MPI rank represented by each cell in the table is the sum of the cell’s column and row indices.

**Table Columns:**

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 |
| 0 | 100 | 100 | 100 | 100 | 99 | 99 | 100 | 100 | 99 | 99 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

---

Time vs. MPI Rank for **MPI_Allreduce**
Task distribution of Maximum Memory Usage (GB/proc) - as a percentage of maximum

The MPI rank is the sum of the column and row indices in the table.
Parallel Tools for the Masses

• Using even the best tools can be tedious
  – “Follow these 10 steps to perform the basic analysis of your program”
Vampir w/ CUDA
Compiler runtime bounds checking

Out of bounds reference in source code for program “flip”

allocate(put_seed(random_size))
...
bad_index = random_size+1
put_seed(bad_index) = 67

Intel compiler:

```bash
ftn -c -g -Ktrap=fp -check bounds flip.f90
ftn -c -g -Ktrap=fp -check bounds printit.f90
ftn -o flip flip.o printit.o -g

% qsub -I -qdebug -lmppwidth=16
% cd $PBS_O_WORKDIR
%
% aprun -n 16./flip

forrtl: severe (408): fort: (2): Subscript #1 of the array SEED has value 3 which is greater than the upper bound of 2
```
• For a list of compiler options, see the man pages for the individual compilers
  – man pgcc | pgCC | pgf90
  – man icc | icpc| ifort
  – man gcc | g++ | gfortran

• Use your favorite search engine to find vendor manuals on line
Other Debugging Tips

• Try different compilers
  – Diagnostic messages and language spec compliances differ

• Look for memory corruption
  – Bad memory reference in one place (array out of bounds) can make code crash elsewhere
  – It might appear that you’re crashing on a perfectly valid line of code

• Check the arguments to your MPI calls

• Call the NERSC Consultants (800-66-NERSC or 510 486-8600)
Identifying Targets for Optimization

• Hardware Event Counters
  – Special registers count events on processor
  – E.g. number of floating point instructions
  – Many possible events
  – Only a few can be recorded at a time (~4 counters)
  – Can give you an idea of how efficiently you are using the processor hardware
Apprentice Basic View

- Can select new (additional) data file and do a screen dump
- Worthless
- Useful
- Can select other views of the data
- Can drag the “calipers” to focus the view on portions of the run
• You will have a homework assignment using TAU
  – %module load tau
  – Define paths in Makefile
  – Modify header file to define TAU macros
  – Add macro calls to the code
  – Compile and submit to batch queue
  – Use pprof to produce readable output

• Good reference
Users Want (Need?) Tools

- Users are asking for tools because HPC systems and programming models are changing
- More and more components to worry about
  - CPU (caches, FPUs, pipelining, ...)
  - Data movement to main memory, GPU memory, levels of cache
  - I/O
  - Network (message passing)
  - CPU Threads (OpenMP)
  - GPU performance
Questions to You

• What tools do you use?
• What tools do you want?
• What would you like centers to support?
• Can you get to exascale without tools?
What I Want in a Tool

• Let the users help themselves
• Work for everyone all (most of?) the time
• Easy to use
• Useful
• Easy to interpret the results
• Affordable & ubiquitous
• Simple, supplement existing complex tools
  – Point the way for a “deeper dive” in problem areas
• Integrated Performance Monitoring
• Developed by David Skinner at NERSC
• MPI profiling, hardware counter metrics, IO profiling (?)
• IPM requires no code modification & no instrumented binary
• IPM uses hooks already in the MPI library to intercept your MPI calls and wrap them with timers and counters
• **How it works (user perspective)**
  
  – `% module load IPM`*
  
  – Run program as normal
  
  – Look at results on the web

• **It’s that easy!**
  
  – And extremely low overhead, so IPM is examining your production code

* (As long as your system supports dynamic load libs)
IPM Examples

Click on the metric you are want.
• **Sampling**
  - Regularly interrupt the program and record where it is
  - Build up a statistical profile of time spent in various routines
  - Concentrate first on longest running sections or routines

• **Tracing**
  - Insert hooks into program to record and time program events (logging)
  - Reasonable for sequential programs
  - Unwieldy for large parallel programs (too much data!)
Using Apprentice

- Optional visualization tool for Cray’s perftools data
- Use it in a X Windows environment
- Uses a data file as input (XXX.ap2) that is prepared by pat_report

  app2 [--limit_per_pe tags] XXX.ap2
• PAPI (Performance API) provides a standard interface for use of the performance counters in major microprocessors

• Predefined actual and derived counters supported on the system
  – To see the list, run ‘papi_avail’ on compute node via aprun:
    ```
    qsub -I -lmppwidth=24
    module load perftools
    aprun -n 1 papi_avail
    ```

• AMD native events also provided; use ‘papi_native_avail’:
  ```
  aprun -n 1 papi_native_avail
  ```
• Tuning and Analysis Utilities
• Fortran, C, C++, Java performance tool
• Procedure
  – Insert macros
  – Run the program
  – View results with pprof
• More info than gprof
  – E.g. per process, per thread info; supports pthreads
• http://acts.nersc.gov/tau/index.html
What IPM measures

- IPM “only” gives a high-level, entire-program-centric view
- Still, very valuable guidance
  - Shows whole-run info per MPI task, OpenMP thread, (CUDA under development)
  - Many pieces of data in one place
- Reveals what many users don’t know about their code
  - High-water memory usage (per task)
  - Load balance
  - Call imbalance
  - MPI time
  - I/O time