*Exceptional service in the national interest*

Sandia National Laboratories

# The Importability of Performance Tools: The Good, the Bad, and the Ugly of PMUs

Jeanine Cook

Scalable Architectures

SNL ABQ

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

# Performance Tools

- Mostly talking about non-commercial tools
  - Tools that lots of us hack up to look at performance issues
    - MPI and node data
      - Interfaces like PAPI and mpip
    - Hoping not to have to learn intricacies of complicated performance tool
    - Tool doesn't quite do what we want it to

# PMUs and Performance Counters (1)

- PMUs on CPUs are all different
  - Implement different events
    - "Architected" PMUs provide subset of base events across model lines
  - Names of identical events can differ
  - Different number of physical counters
    - Divided amongst threads
  - Significant errata in PMU behavior
    - But info often not propagated to end user

- Event mapping tables
  - Connects event name and qualifiers to hardware register(s) configuration(s)
  - Some change frequently (Intel's https://download.01.org/perfmon/)
  - Some hardly get updated or are community supported (papi_presets)
  - Kernel contains some events with compile time mappings (perf list)
    - If hardware requires a change, no way to update map, either can't access event or may return strange results

# PMUs and Performance Counters (2)

- Documentation is spotty at best, inaccurate at worst
  - Hard to really be sure of precisely what you're counting without investigation, sometimes microbenchmarking
    - EX: Events commonly counted at issue, so many overcount due to speculation and re-issue (e.g, SNB and IVB FP events)

# PAPI Presets vs Native Events

- Presets offer portability… sort of
  - Event names consistent across platforms, but may map to different underlying native events, as best effort
    - Particular event may not actually count same behavior across systems
    - Particular event on single system may not count what event name reflects
      - SNB: PAPI_DP_OPS – Counts DP FP add and multiply, but not divide instructions (separate counter for that)
        - » Use papi_decode –a to see what presets count
        - » Decode libpfm events using showevtinfo from libpfm

# Takeaways

- If using performance counters directly
  - Don't use presets, use natives
  - Be sure you know what you're counting
- Performance tools (commercial and home-grown) access PMUs
  - Restricts cross-platform usage

An attempt to make PMU/performance counter use easier and cross-platform…

# Perfminer

Property of MinimalMetrics and SNL

# Perfminer: Performance Analysis

- Performance Analysis
  - First line support tool for performance investigations
  - Low-overhead, scalable with simple instrumentation
  - "Inside-out" performance metrics from CPU, Memory, Network, Filesystem, Power, with localizing information
    - Data is per-thread and aggregated to all levels
  - Immediate, efficient, visual performance feedback via a web page
    - Drill down methodology with click through to investigate further
      - Supports using other tools
  - Enables continuous performance regression of system and applications
  - Actionable feedback (future development)
    - Suggestions for further testing; automatic generation of test scripts for Perfminer and potentially other tools

# Perfminer: Performance Analysis

- Performance Analysis
  - Modular design
    - Collector, front-end, back-end
      - Java web-based, front-end
      - Python back end,
      - Schema-less NoSQL data store.
        - » Data is easily accessible for analytics via Excel, R, Matlab, Python, etc
      - Extensible analytics with Python
      - Elegant visualization using D3
  - No recompilation required
  - Easy to install (few dependencies) and trivial to use.
  - Full batch system integration (coming soon)

# Perfminer: Performance as a Service

- Continuous monitoring infrastructure
  - Can opt in or out
- Deliver a report through email with actionable feedback

# Job Info

PerfMiner

Admin Console | Contact

Show 10 entries

Filter:

| Job ID | Job Number | Owner | Nodes | Binary | Running Time | Finish Time |
|---|---|---|---|---|---|---|
| BOJY380439 | 1025821 | pjmucci | 4 | miniaero.exe | 0:00:23 | 2016-03-20 10:05:25 |
| XIAT133711 | 1025820 | pjmucci | 1 | stream-triad-omp-1M | 0:00:46 | 2016-03-20 10:05:48 |
| CMNG317917 | 1025818 | pjmucci | 10 | miniaero.exe | 0:00:01 | 2016-03-20 10:09:22 |
| UQFM259478 | 1025819 | pjmucci | 1 | stencil-27pt-50 | 0:00:03 | 2016-03-20 10:05:05 |
| OXBX917360 | 1025817 | pjmucci | 3 | miniaero.exe | 0:00:06 | 2016-03-20 10:04:42 |
| URGM160184 | 1025816 | pjmucci | 1 | stream-copy-omp-1M | 0:01:05 | 2016-03-20 10:05:41 |
| AYKC210492 | 1025815 | pjmucci | 1 | miniaero.exe | 0:00:04 | 2016-03-20 10:04:40 |
| NGOJ725228 | 1025814 | pjmucci | 1 | stream-copy-omp-2G | 0:02:39 | 2016-03-20 10:07:15 |
| WOPI548517 | 1025813 | pjmucci | 1 | stream-triad-1M | 0:00:46 | 2016-03-20 10:05:21 |
| FJGL475723 | 1025812 | pjmucci | 8 | lulesh-opt | 0:04:27 | 2016-03-20 10:09:03 |

Showing 31 to 40 of 125 entries

Previous 1 2 3 4 5 ... 13 Next

## JOB DETAILS

RAW DATA | CLOCK SCALING | JOB OUTPUT | MODULES

**Job Number :** 1025844
**Job Owner :** pjmucci
**Job Script :** miniaero-sbatch-3d_sod_parallel_test_big.sh
**Job Path :** ...tests/3D_Sod_Parallel_Big/miniaero.exe
**Job Time :** 0:26:23.075943
**mpirun Time :** 0:26:16.289926
**Start Time :** 2016-03-20 11:03:40
**Finish Time :** 2016-03-20 11:30:03

**Nodes Allocated :** 16
**Nodes Used :** 8
**MPI Ranks Used :** 32 (map)
**Threads :** 1

# Job Info

## JOB DETAILS

RAW DATA | CLOCK SCALING | JOB OUTPUT | MODULES

| | | | |
|---|---|---|---|
| **Job Number :** | 1025844 | **Nodes Allocated :** | 16 |
| **Job Owner :** | pjmucci | **Nodes Used :** | 8 |
| **Job Script :** | miniaero-sbatch-3d_sod_parallel_test_big.sh | **MPI Ranks Used :** | 32 (map) |
| | | **Threads :** | 1 |
| **Job Path :** | ...tests/3D_Sod_Parallel_Big/miniaero.exe | | |
| **Job Time :** | 0:26:23.075943 | | |
| **mpirun Time :** | 0:26:16.289926 | | |
| **Start Time :** | 2016-03-20 11:03:40 | | |
| **Finish Time :** | 2016-03-20 11:30:03 | | |

```
"
N=${N:-32}
Nmo=$(( N - 1 ))
#slot | hwthread | core | socket (default) | numa | board | node
policy=${policy:-"--map-by socket:PE=8"}

echo "Running parallel 3D sod test using $policy and $N processors."
mpirun $policy -np $N ../../make/src/miniaero.exe
diff=0
for i in `seq 0 $Nmo`;
do
../tools/numeric_text_diff results.$i results.$i.gold > diff.$i.txt
diff=$(($diff + $?))
done
if [ $diff -gt 0 ];
```

PerfMiner

```
Running parallel 3D sod test using --map-by socket:PE=8 and 32 processors.
script: miniaero-sbatch-3d_sod_parallel_test_big.sh
mpiP: Found MPIP environment variable [-y -k 0 -e -p -q -f slurm-1025844.perfminer]
mpiP: Set the callsite stack traceback depth to [0].
mpiP: Set the output directory to [slurm-1025844.perfminer].
mpiP:
mpiP: mpiP: mpiP V3.4.2 (Build Feb 17 2016/04:46:45)
mpiP: Direct questions and errors to mpip-help@lists.sourceforge.net
mpiP:

... Face creation time: 1.00 seconds ...

... Extract BC face and delete ghost time: 0.00 seconds ...
Start setup communication.

... Setup Communcation function time: 0.00 seconds ...

... Rest of setup communication time: 1.00 seconds ...
End setup communication.

... Setup time: 3.00 seconds ...
```

| | |
|---|---|
| 0 | zlib/1.2.8<br>binutils/2.25.0<br>mpc/1.0.1<br>mpfr/3.1.2<br>gmp/6.0.0<br>isl/0.12.2<br>gcc/4.9.3<br>openmpi/1.10.0/gcc/4.9.3<br>python/2.7.9<br>mmperftools<br>perfminer<br>minimalmetrics |
| 1 | zlib/1.2.8<br>binutils/2.25.0<br>mpc/1.0.1<br>mpfr/3.1.2<br>gmp/6.0.0<br>isl/0.12.2<br>gcc/4.9.3<br>openmpi/1.10.0/gcc/4.9.3<br>python/2.7.9<br>mmperftools<br>perfminer<br>minimalmetrics |

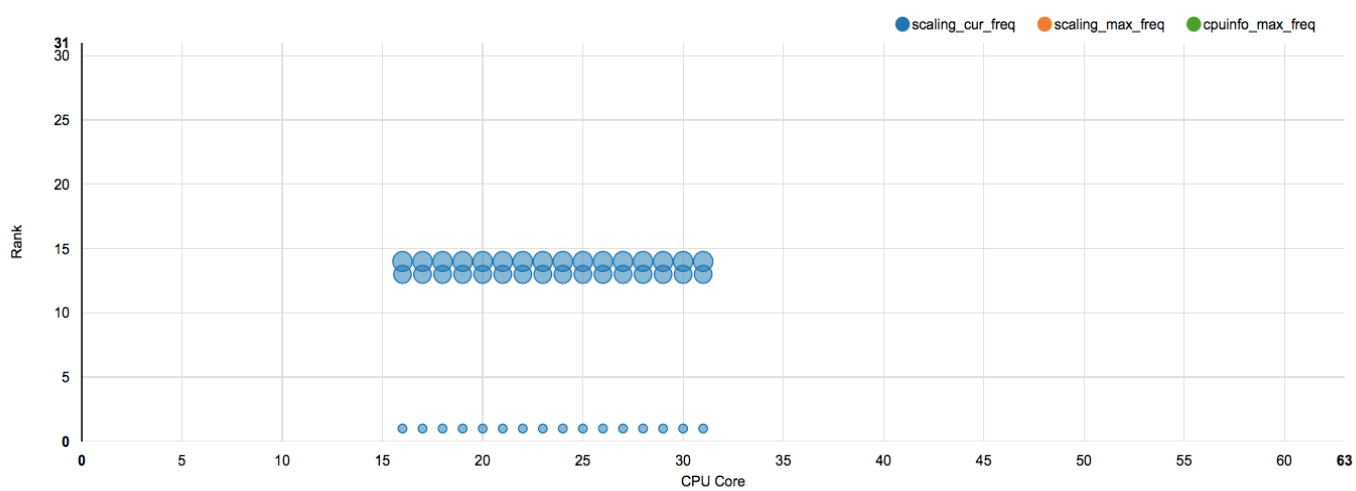
# Job Info

## JOB DETAILS

RAW DATA | CLOCK SCALING | JOB OUTPUT | MODULES

| | | | |
|---|---|---|---|
| **Job Number :** | 1025844 | **Nodes Allocated :** | 16 |
| **Job Owner :** | pjmucci | **Nodes Used :** | 8 |
| **Job Script :** | miniaero-sbatch-3d_sod_parallel_test_big.sh | **MPI Ranks Used :** | 32 (map) |
| | | **Threads :** | 1 |
| **Job Path :** | ...tests/3D_Sod_Parallel_Big/miniaero.exe | | |
| **Job Time :** | 0:26:23.075943 | | |
| **mpirun Time :** | 0:26:16.289926 | | |
| **Start Time :** | 2016-03-20 11:03:40 | | |
| **Finish Time :** | 2016-03-20 11:30:03 | | |

## FREQUENCY SCALING




© Minimal Metrics LLC 2015

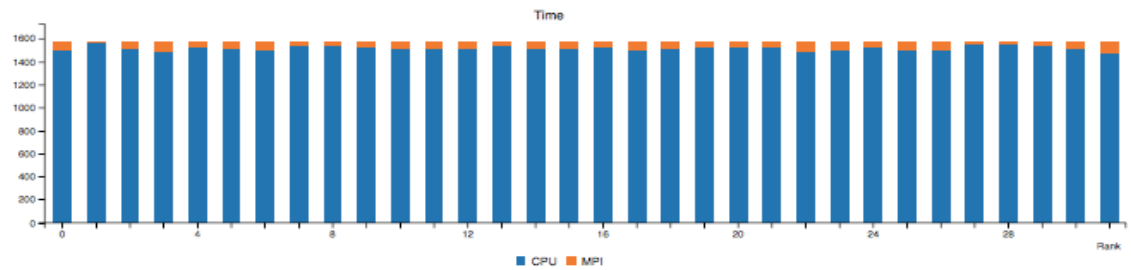
# Metric Correlation

## JOB DETAILS

RAW DATA | CLOCK SCALING | JOB OUTPUT | MODULES

| | | | |
|---|---|---|---|
| Job Number : | 1025844 | Nodes Allocated : | 16 |
| Job Owner : | pjmucci | Nodes Used : | 8 |
| Job Script : | miniaero-sbatch-3d_sod_parallel_test_big.sh | MPI Ranks Used : | 32 (map) |
| Job Path : | ...tests/3D_Sod_Parallel_Big/miniaero.exe | Threads : | 1 |
| Job Time : | 0:26:23.075943 | | |
| mpirun Time : | 0:26:16.289926 | | |
| Start Time : | 2016-03-20 11:03:40 | | |
| Finish Time : | 2016-03-20 11:30:03 | | |

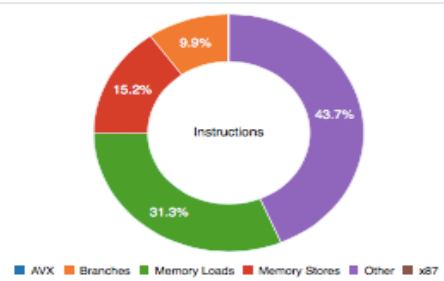
## JOB OVERVIEW




## JOB PERFORMANCE DETAILS

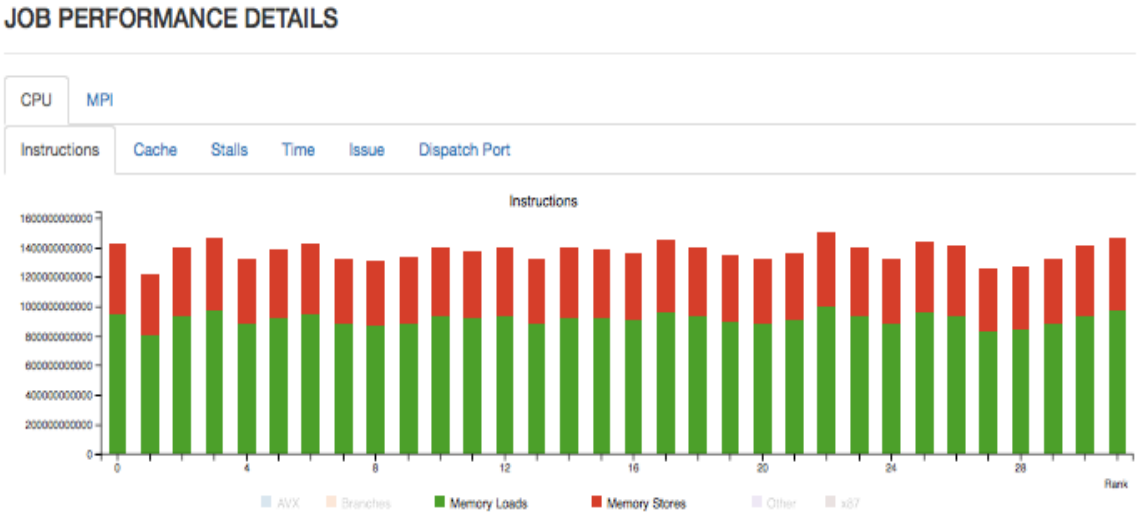
CPU | MPI

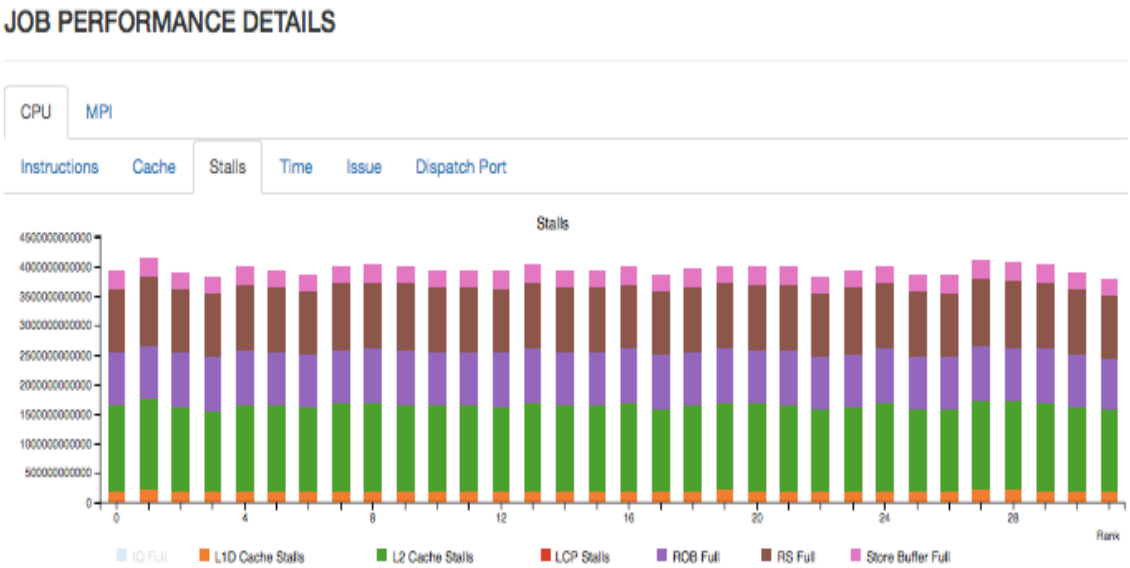
Instructions | Cache | Stalls | Time | Issue | Dispatch Port



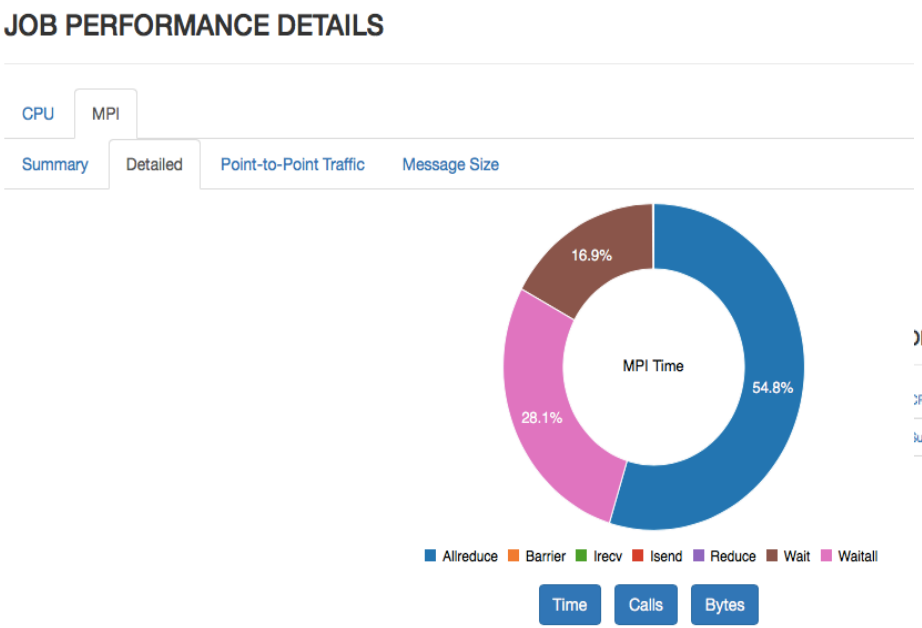
# Metric Correlation


# Metric Correlation



# MPI


JOB PERFORMANCE DETAILS
CPU
MPI
Summary
Detailed
Point-to-Point Traffic
Message Size
16.9%
MPI Time
54.8%
28.1%
Allreduce
Barrier
Irecv
Isend
Reduce
Wait
Waitall
Time
Calls
Bytes

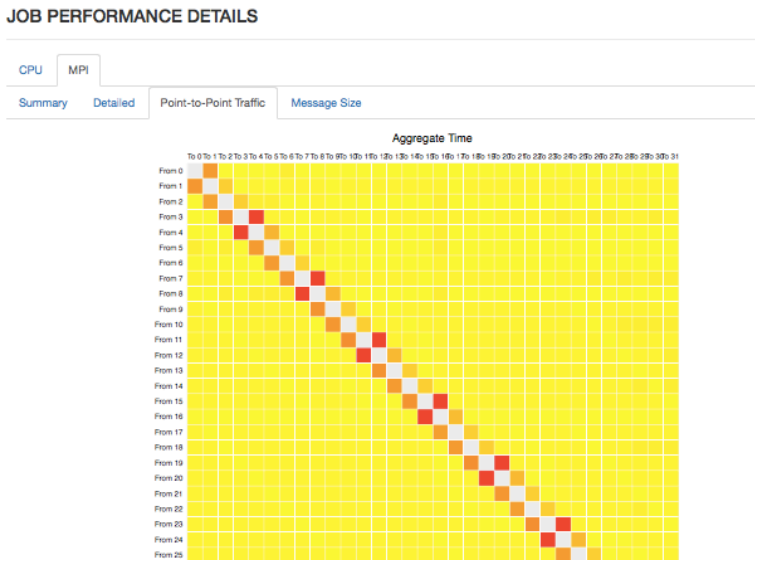

Summary
Detailed
Point-to-Point Traffic
Message Size
Aggregate Time
Time
Calls
Bytes

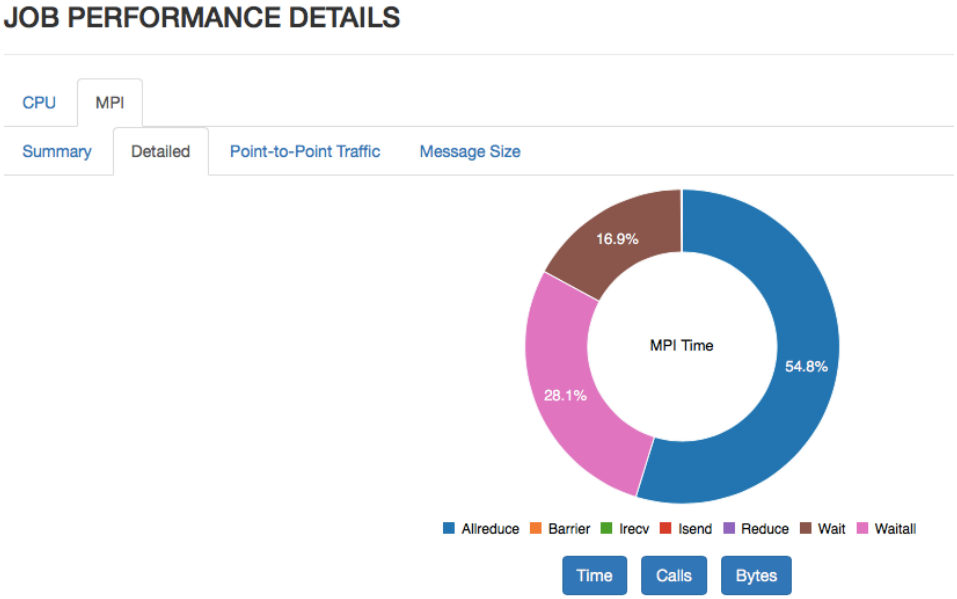
# THE END!

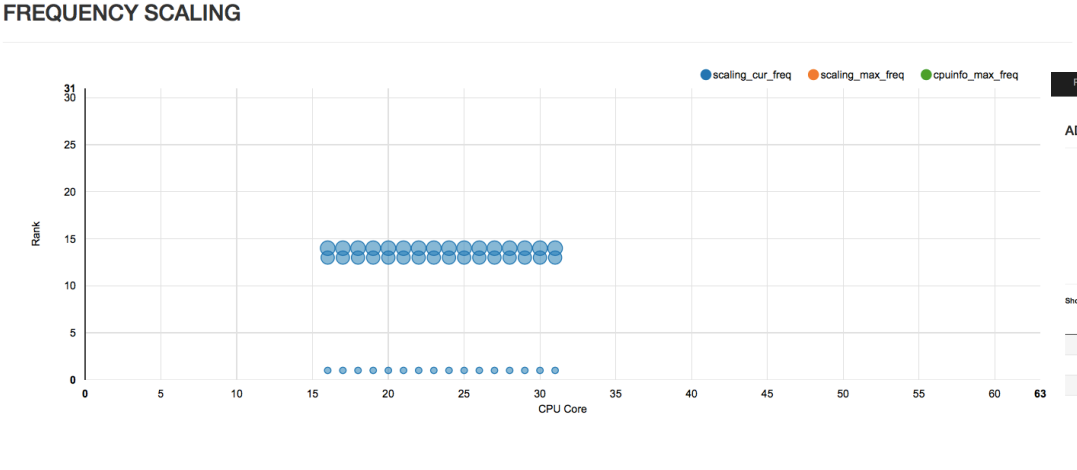
jeacook@sandia.gov

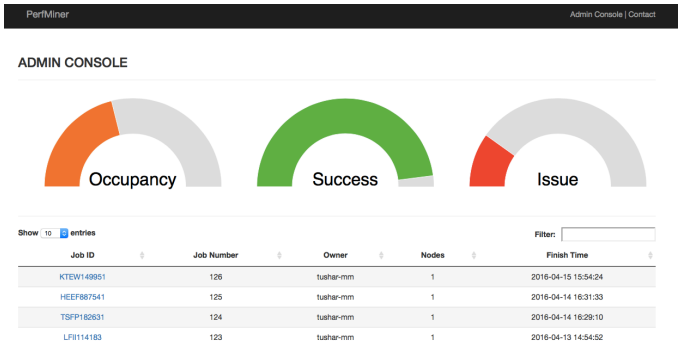
phil@minimalmetrics.com

# MPI, Outlier Detection, Productivity




© Minimal Metrics LLC 2015