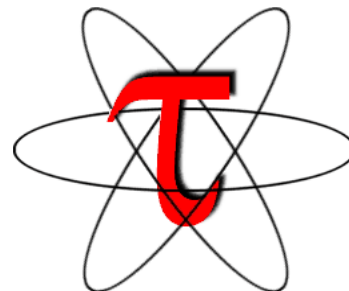


# *Scalability Study of S3D on Intrepid BGP using TAU*

Wyatt Spear

tau-team@cs.uoregon.edu



Tuning and Analysis Utilities



UNIVERSITY  
OF OREGON



# *Acknowledgements*

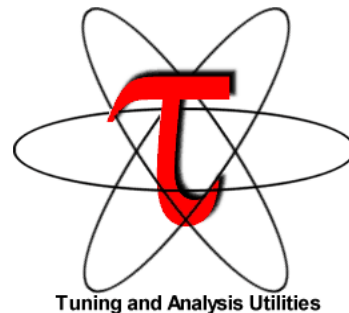


- ❑ Alan Morris [UO]
- ❑ Kevin Huck [UO]
- ❑ Sameer Shende [UO]
- ❑ Allen D. Malony [UO]
- ❑ Bronis R. de Supinski [LLNL]



# *TAU Parallel Performance System*

- ❑ **<http://tau.uoregon.edu/>**
- ❑ **Multi-level performance instrumentation**
  - Multi-language automatic source instrumentation
- ❑ **Flexible and configurable performance measurement**
- ❑ **Widely-ported parallel performance profiling system**
  - Computer system architectures and operating systems
  - Different programming languages and compilers
- ❑ **Support for multiple parallel programming paradigms**
  - Multi-threading, message passing, mixed-mode, hybrid

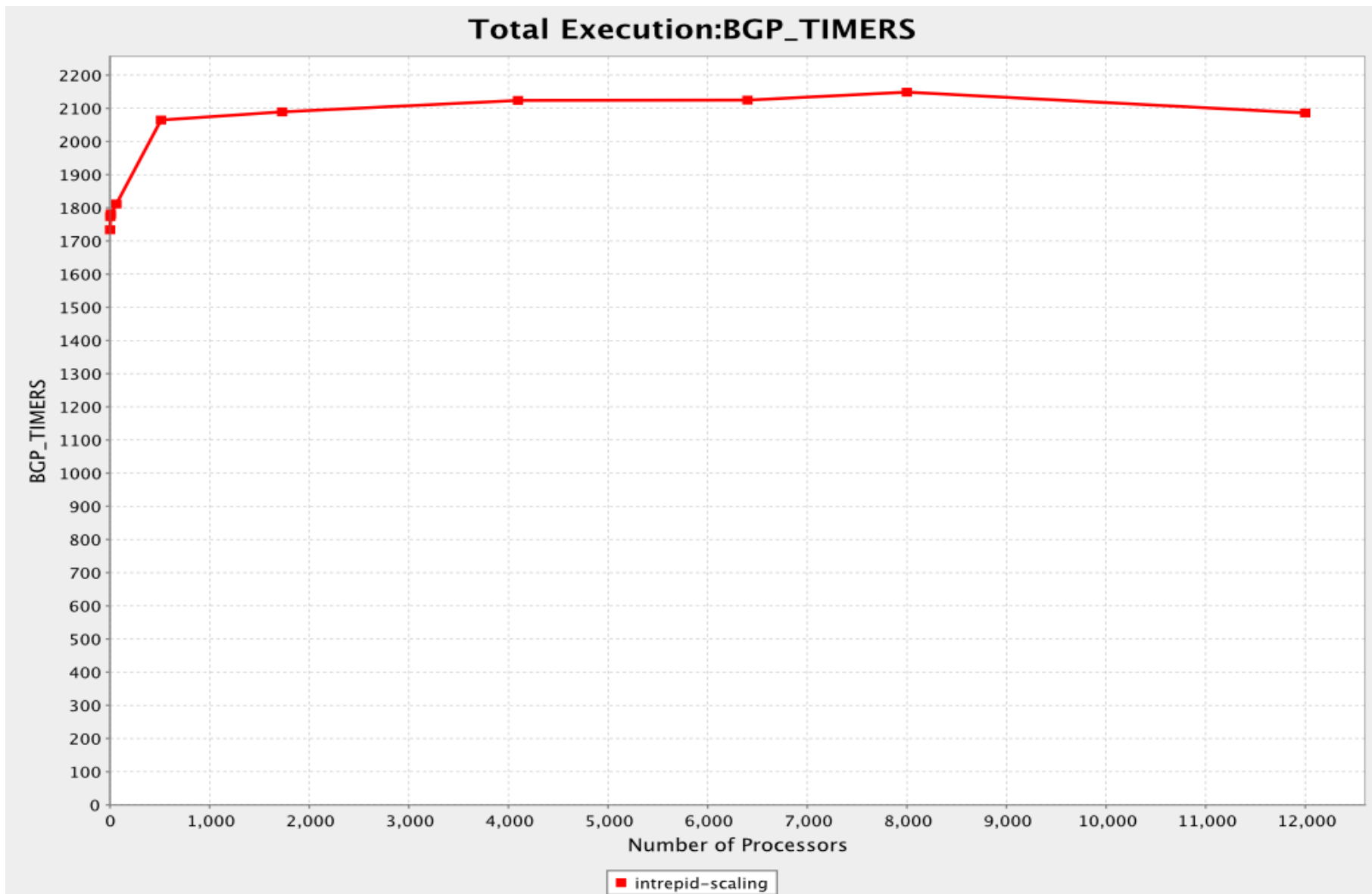


# *Scalability Study*



- ❑ C2H4 Benchmark
- ❑ Platform: Intrepid BGP
  - 1p
  - 4p
  - 64p
  - 512p
  - 1728p
  - 4096p
  - 8000p
  - 12000p
- ❑ Goal: to evaluate scaling properties of code regions
- ❑ Scalability of MPI operations

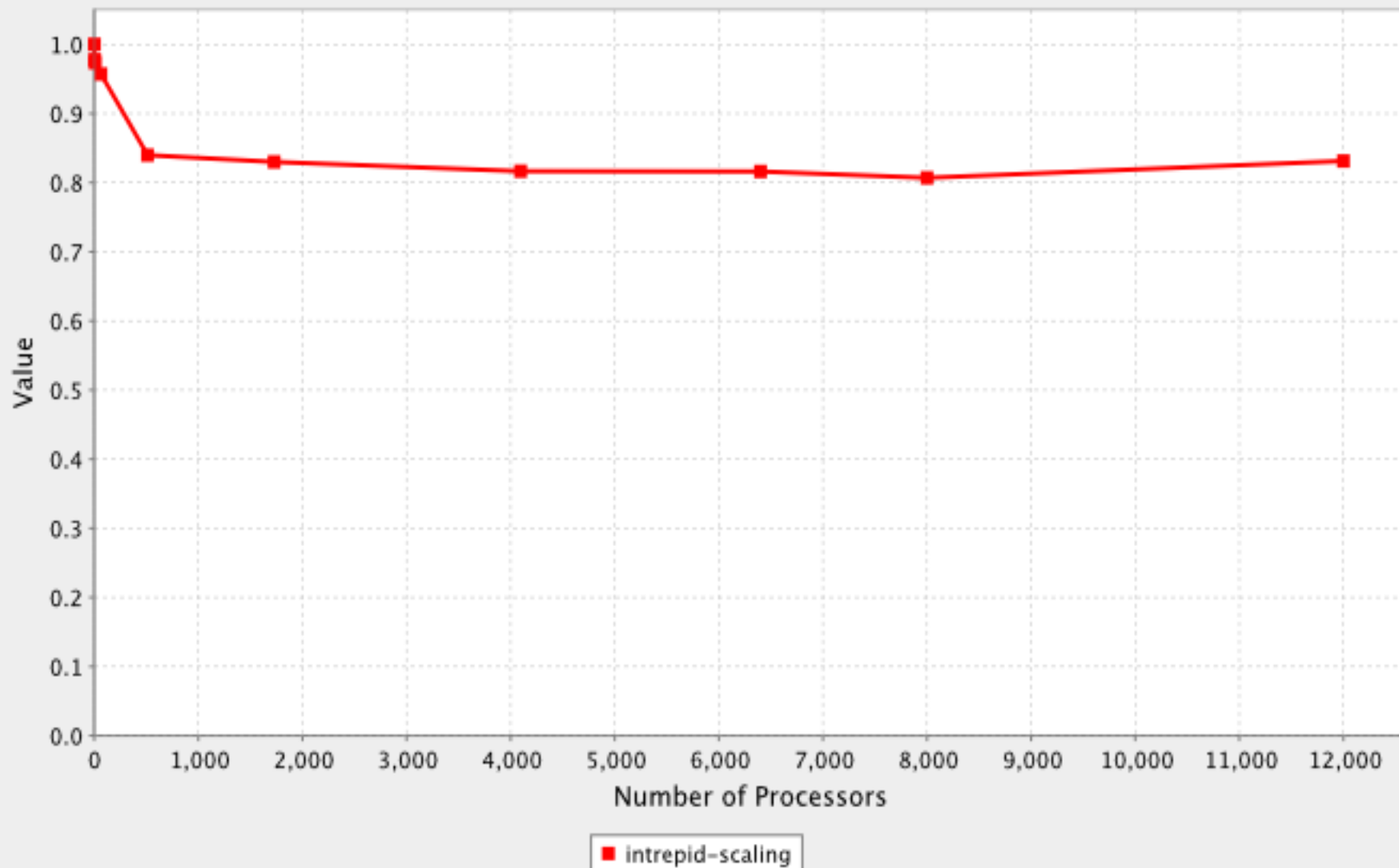
# Total Execution Time



# Relative Efficiency For S3D - Weak Scaling

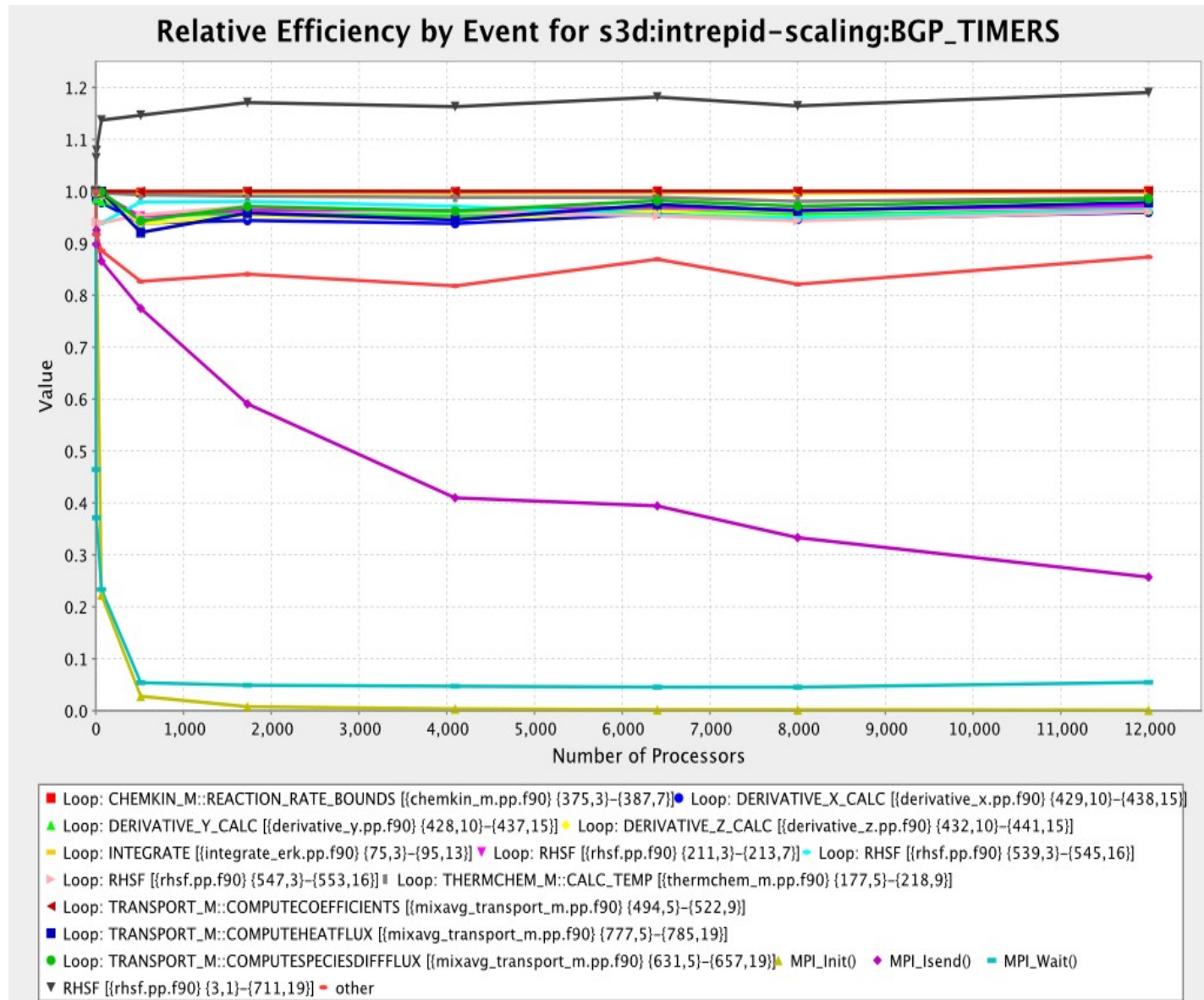


Relative Efficiency - s3d:intrepid-scaling:BGP\_TIMERS

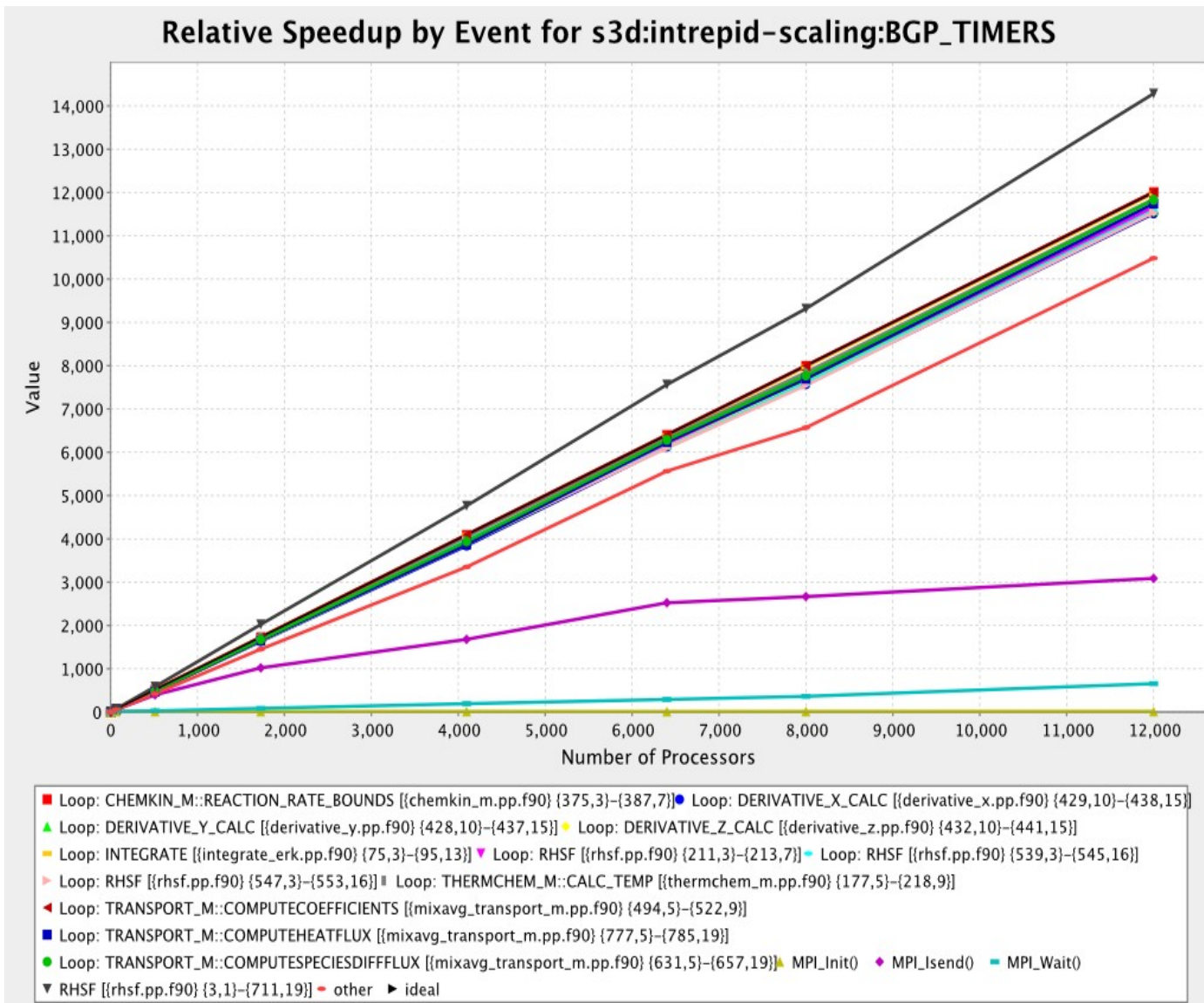




# Relative Efficiency by Event



# Relative Speedup by Event

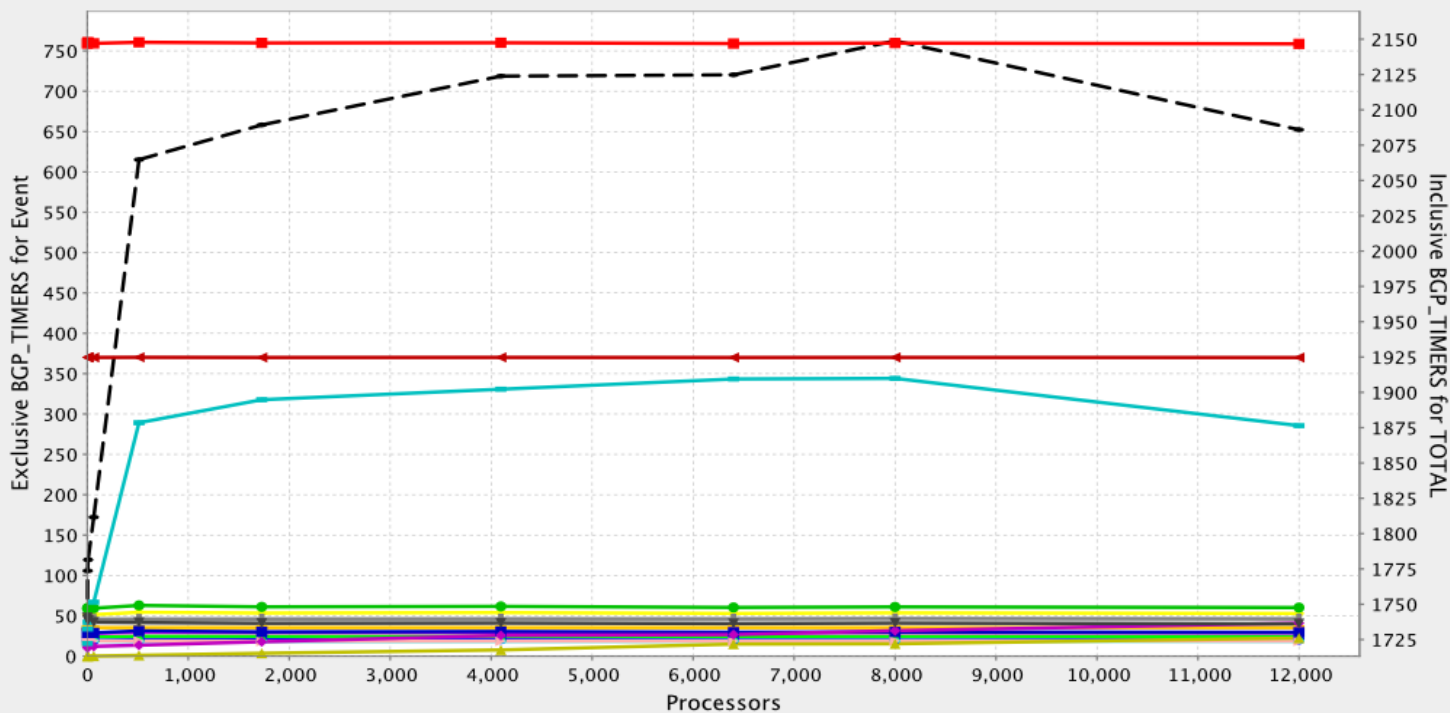




# Data Mining: Event Correlation to Total Time



Correlation Results: s3d:intrepid-scaling: BGP\_TIMERS



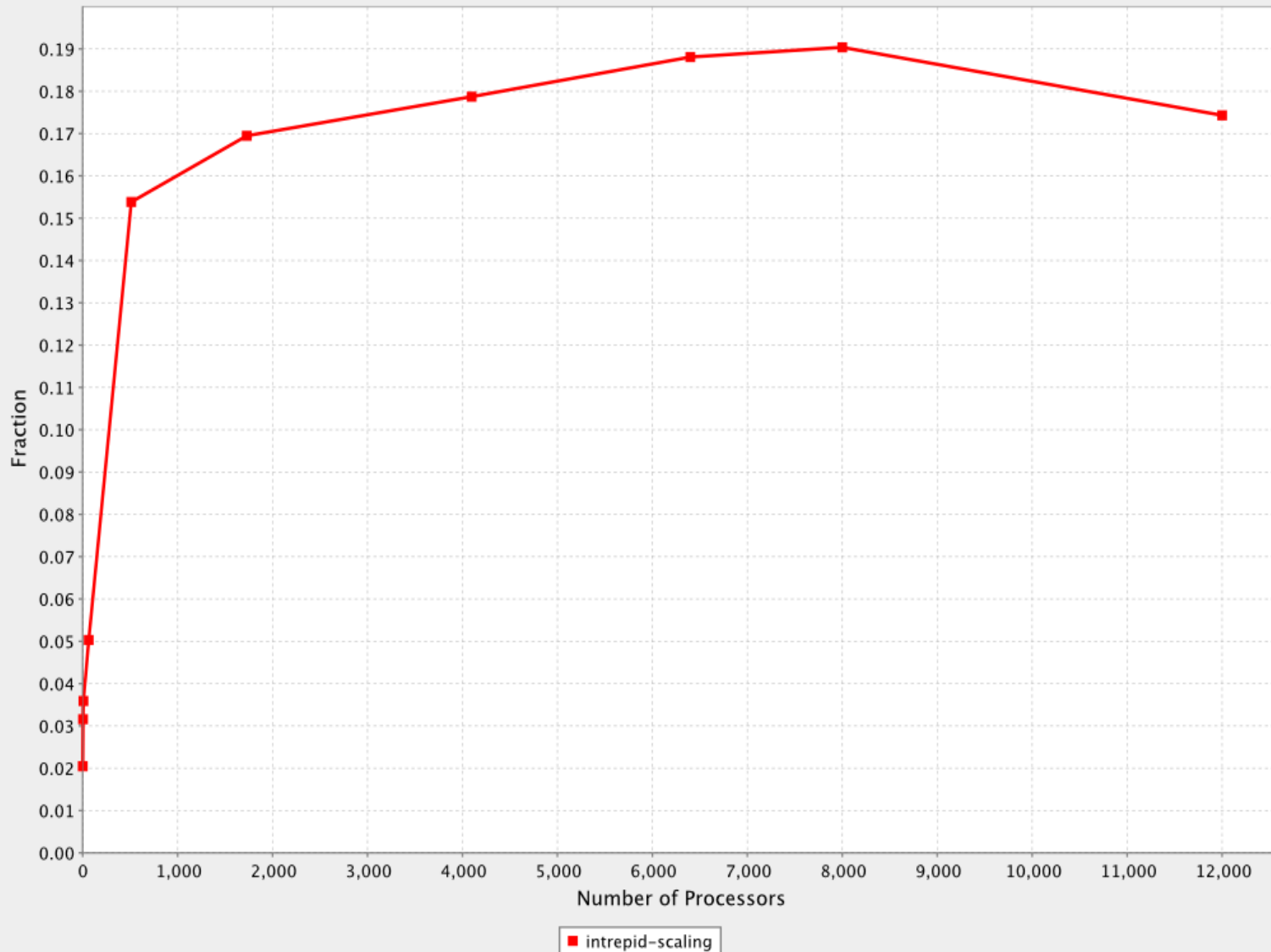
- Loop: CHEMKin\_M::REACTION\_RATE\_BOUNDS [{chemkin\_m.pp.f90} {375,3}-{387,7}], r = -0.13
- Loop: DERIVATIVE\_X\_CALC [{derivative\_x.pp.f90} {429,10}-{438,15}], r = 0.93 ▲ Loop: DERIVATIVE\_Y\_CALC [{derivative\_y.pp.f90} {428,10}-{437,15}], r = 0.95
- ◆ Loop: DERIVATIVE\_Z\_CALC [{derivative\_z.pp.f90} {432,10}-{441,15}], r = 0.93 ▬ Loop: INTEGRATE [{integrate\_erk.pp.f90} {75,3}-{95,13}], r = 0.94
- ▼ Loop: RHSF [{rhsf.pp.f90} {211,3}-{213,7}], r = 0.93 ▬ Loop: RHSF [{rhsf.pp.f90} {539,3}-{545,16}], r = -0.17
- ▶ Loop: RHSF [{rhsf.pp.f90} {547,3}-{553,16}], r = 0.04 ▬ Loop: THERMCHEM\_M::CALC\_TEMP [{thermchem\_m.pp.f90} {177,5}-{218,9}], r = 0.85
- ◀ Loop: TRANSPORT\_M::COMPUTECOEFFICIENTS [{mixavg\_transport\_m.pp.f90} {494,5}-{522,9}], r = -0.56
- Loop: TRANSPORT\_M::COMPUTEHEATFLUX [{mixavg\_transport\_m.pp.f90} {777,5}-{785,19}], r = 0.75
- Loop: TRANSPORT\_M::COMPUTESPECIESDIFFLUX [{mixavg\_transport\_m.pp.f90} {631,5}-{657,19}], r = 0.76 ● MPI\_Init0, r = 0.71 ◆ MPI\_Isend0, r = 0.75
- MPI\_Wait0, r = 1.00 ▼ RHSF [{rhsf.pp.f90} {3,1}-{711,19}], r = -0.87 ▶ TOTAL

$r = 1$  implies direct correlation

# MPI Scaling (Total time in MPI/Total Time)



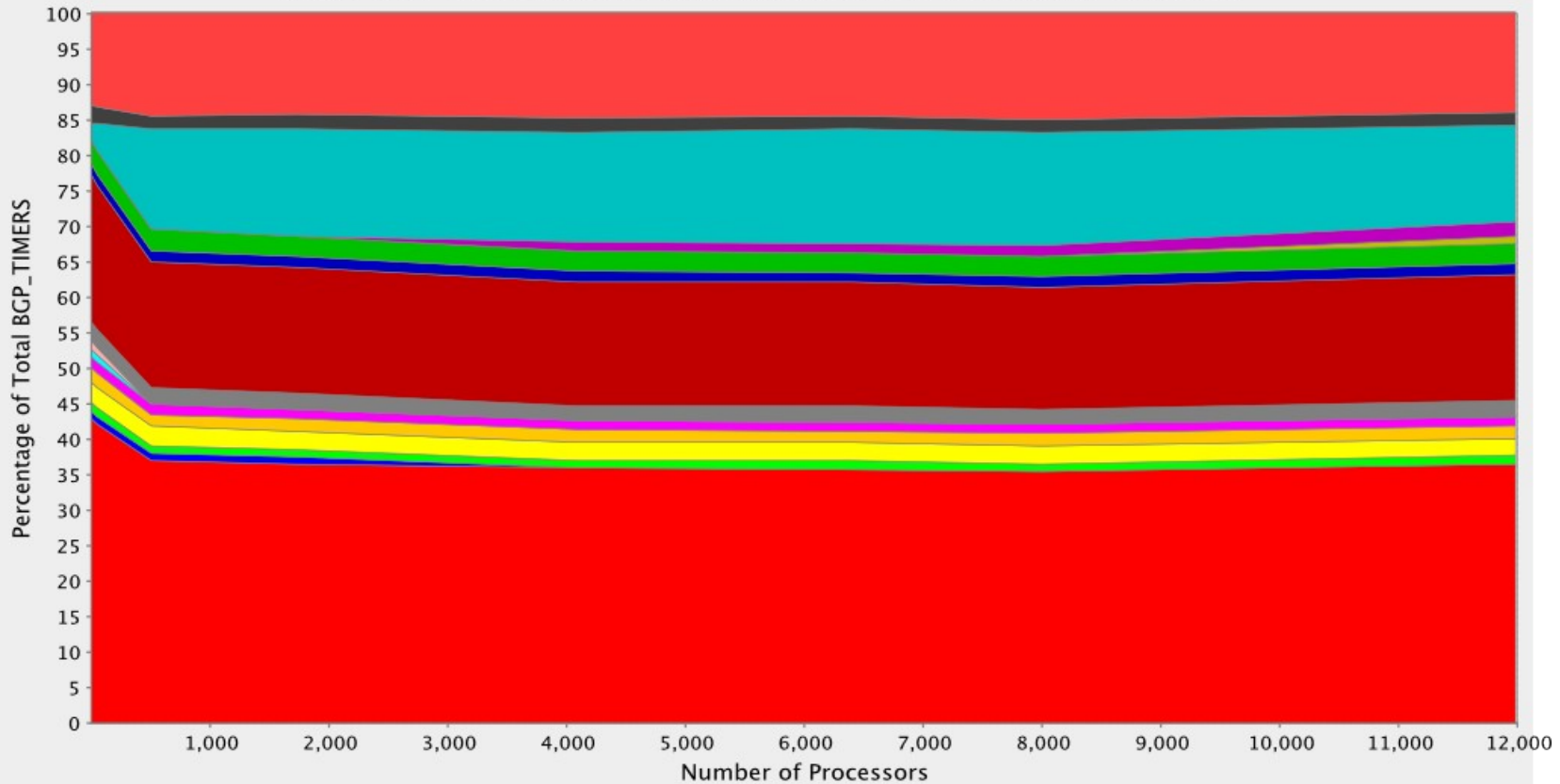
MPI BGP\_TIMERS / Total BGP\_TIMERS - s3d:intrepid-scaling





# Total Runtime Breakdown by Events

## Total BGP\_TIMERS Breakdown for s3d:intrepid-scaling



- Loop: CHEMKIN\_M::REACTION\_RATE\_BOUNDS [{"chemkin\_m.pp.f90"} {375,3}–{387,7}]
- Loop: DERIVATIVE\_Y\_CALC [{"derivative\_y.pp.f90"} {428,10}–{437,15}]
- Loop: DERIVATIVE\_Z\_CALC [{"derivative\_z.pp.f90"} {432,10}–{441,15}]
- Loop: INTEGRATE [{"integrate\_erk.pp.f90"} {75,3}–{95,13}]
- Loop: RHSF [{"rhsf.pp.f90"} {211,3}–{213,7}]
- Loop: RHSF [{"rhsf.pp.f90"} {539,3}–{545,16}]
- Loop: RHSF [{"rhsf.pp.f90"} {547,3}–{553,16}]
- Loop: THERMCHEM\_M::CALC\_TEMP [{"thermchem\_m.pp.f90"} {177,5}–{218,9}]
- Loop: TRANSPORT\_M::COMPUTECOEFFICIENTS [{"mixavg\_transport\_m.pp.f90"} {494,5}–{522,9}]
- Loop: TRANSPORT\_M::COMPUTEHEATFLUX [{"mixavg\_transport\_m.pp.f90"} {777,5}–{785,19}]
- Loop: TRANSPORT\_M::COMPUTESPECIESDIFFFLUX [{"mixavg\_transport\_m.pp.f90"} {631,5}–{657,19}]
- MPI\_Init()
- MPI\_Isend()
- MPI\_Wait()
- RHSF [{"rhsf.pp.f90"} {3,1}–{711,19}]
- other

# ParaProf: 12000 core job



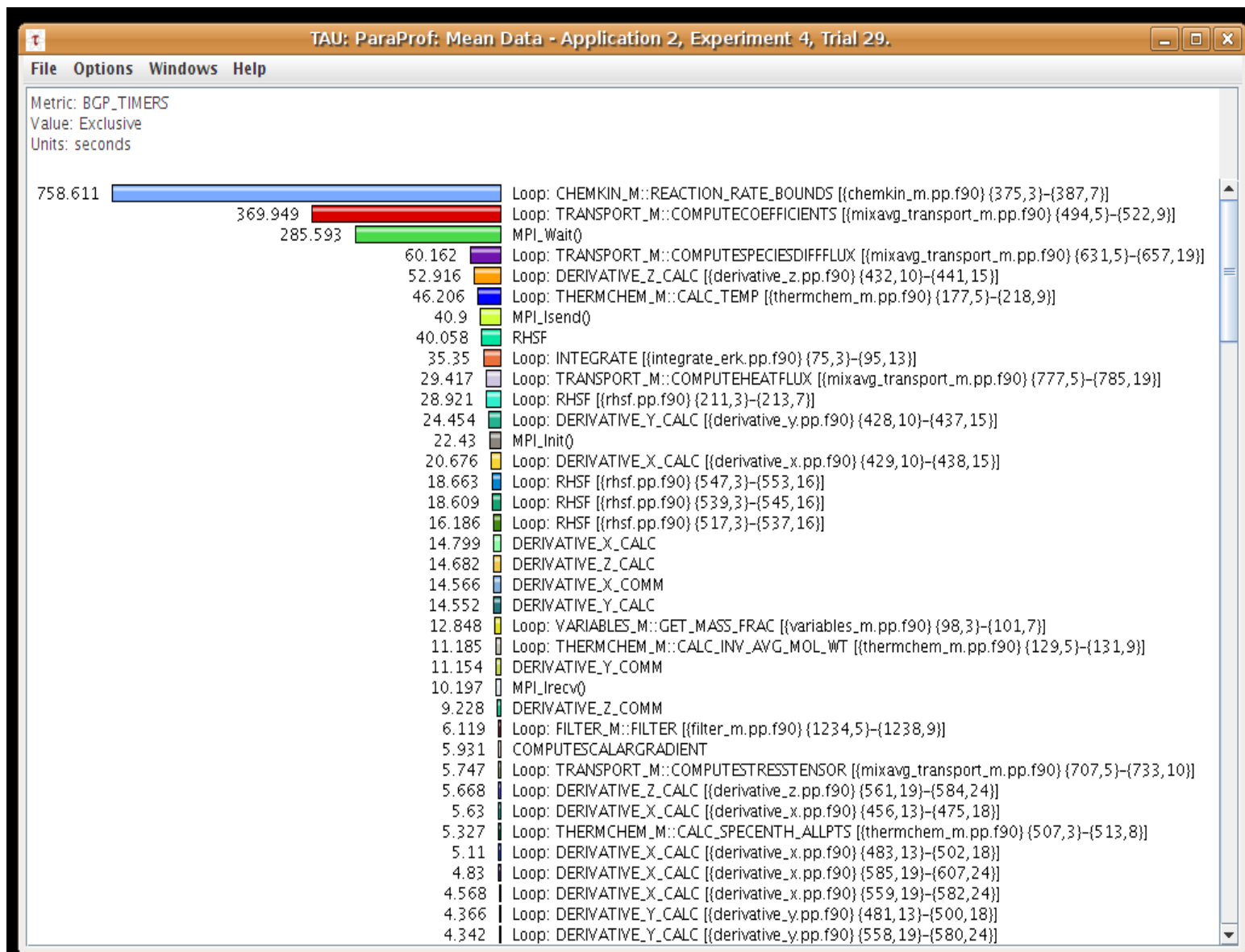
The screenshot shows the TAU: ParaProf Manager interface. On the left is a tree view of applications, and on the right is a table of trial fields.

**Applications Tree:**

- Applications
  - Standard Applications
    - Default App
  - spaceghost (jdbc:postgresql://spaceghost.cs.uoregon.edu:5432/wyatt)
  - Default (jdbc:derby:/home/wspear/TAU2/tau2/i386\_linux/lib/perfdmf)
  - peri\_s3d (jdbc:postgresql://spaceghost.cs.uoregon.edu:5432/peri\_s3d)
    - s3d
      - intrepid-scaling
      - intrepid-scaling-nopapi
        - 12000
          - BGP\_TIMERS
      - jaguar-scaling
    - S3D

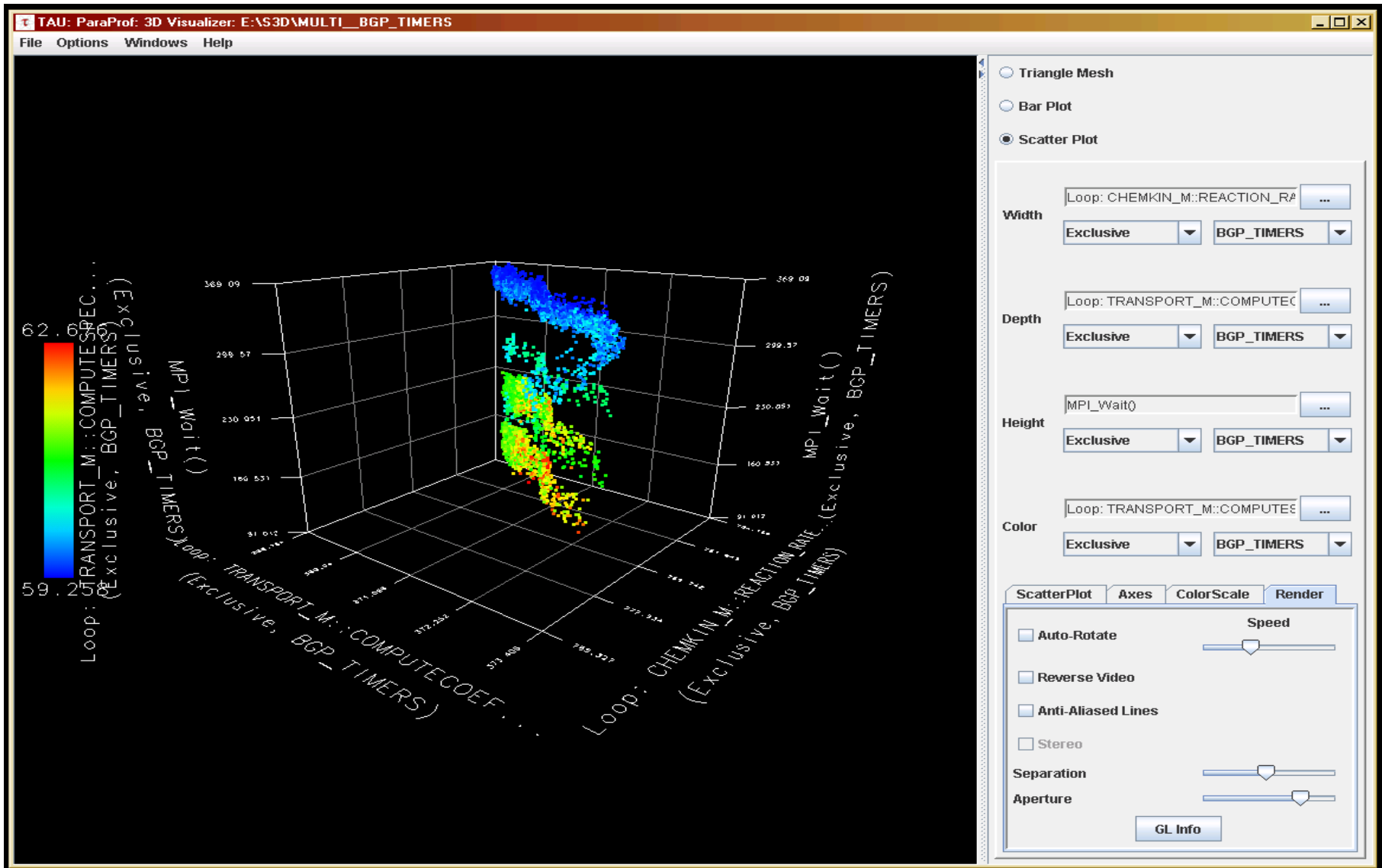
TrialField	Value
Name	12000
Application ID	2
Experiment ID	4
Trial ID	29
date	2008-09-23 13:2...
collectorid	
node_count	12000
contexts_per_node	1
threads_per_context	1
xml_metadata	<?xml version="1....
xml_metadata_gz	
BGP DDRSize (MB)	2048
BGP Node Mode	Virtual
BGP Size	(8, 16, 32)
BGP isTorus	(1, 1, 1)
BGP numPsets	4096
BGP psetSize	64
CPU Type	450 Blue Gene/P D...
CWD	/gpfs/home/wspea...
Executable	/sbin.rd/ioproxy
Memory Size	1816608 kB
OS Machine	BGP
OS Name	CNK
OS Release	2.6.19.2
OS Version	1
TAU Architecture	bgp
TAU Config	-arch=bgp -mpi -...
TAU Version	2.17.2

# ParaProf: Mean across all nodes



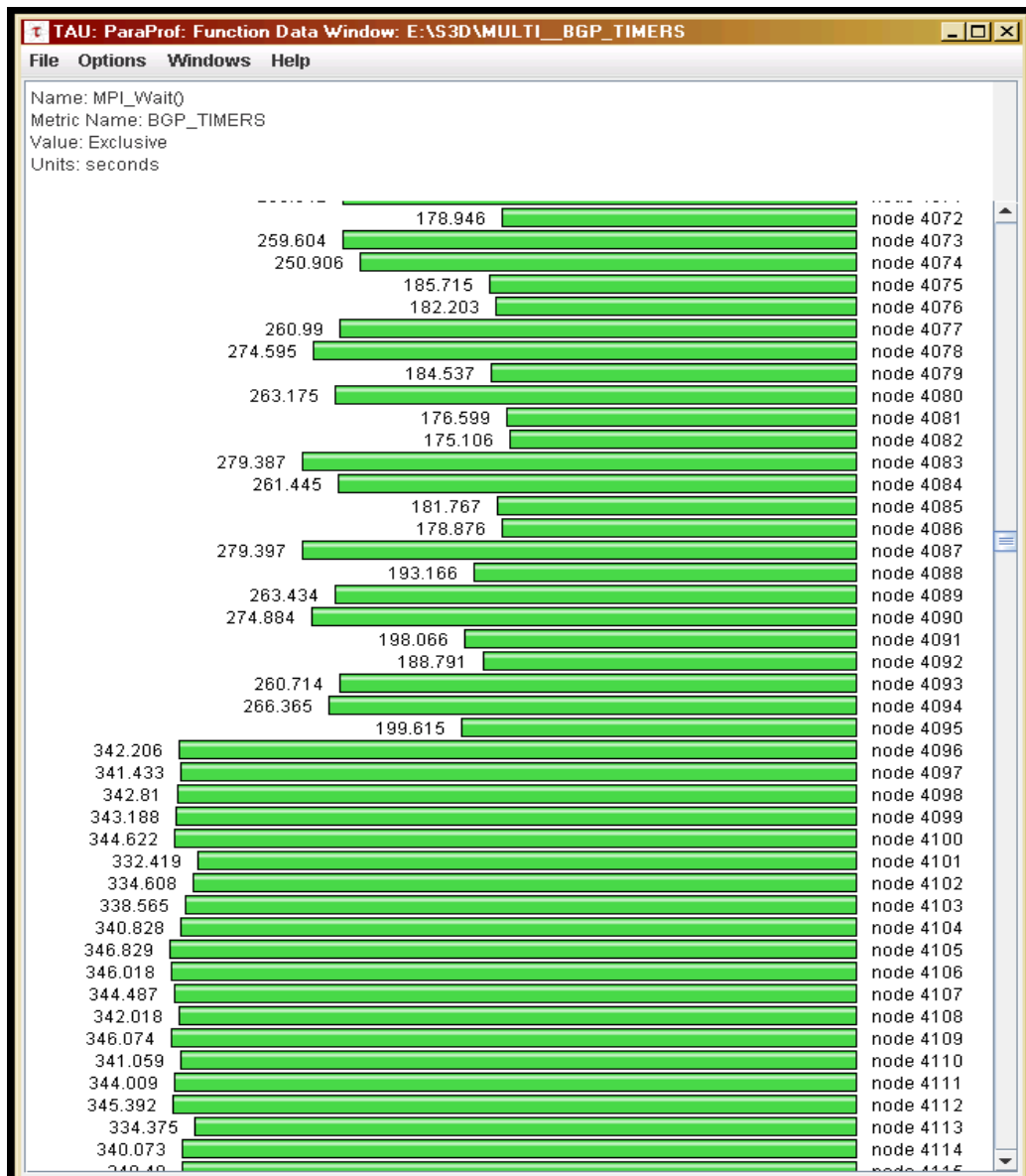


# ParaProf: 3D Correlation Cube: MPI\_Wait!

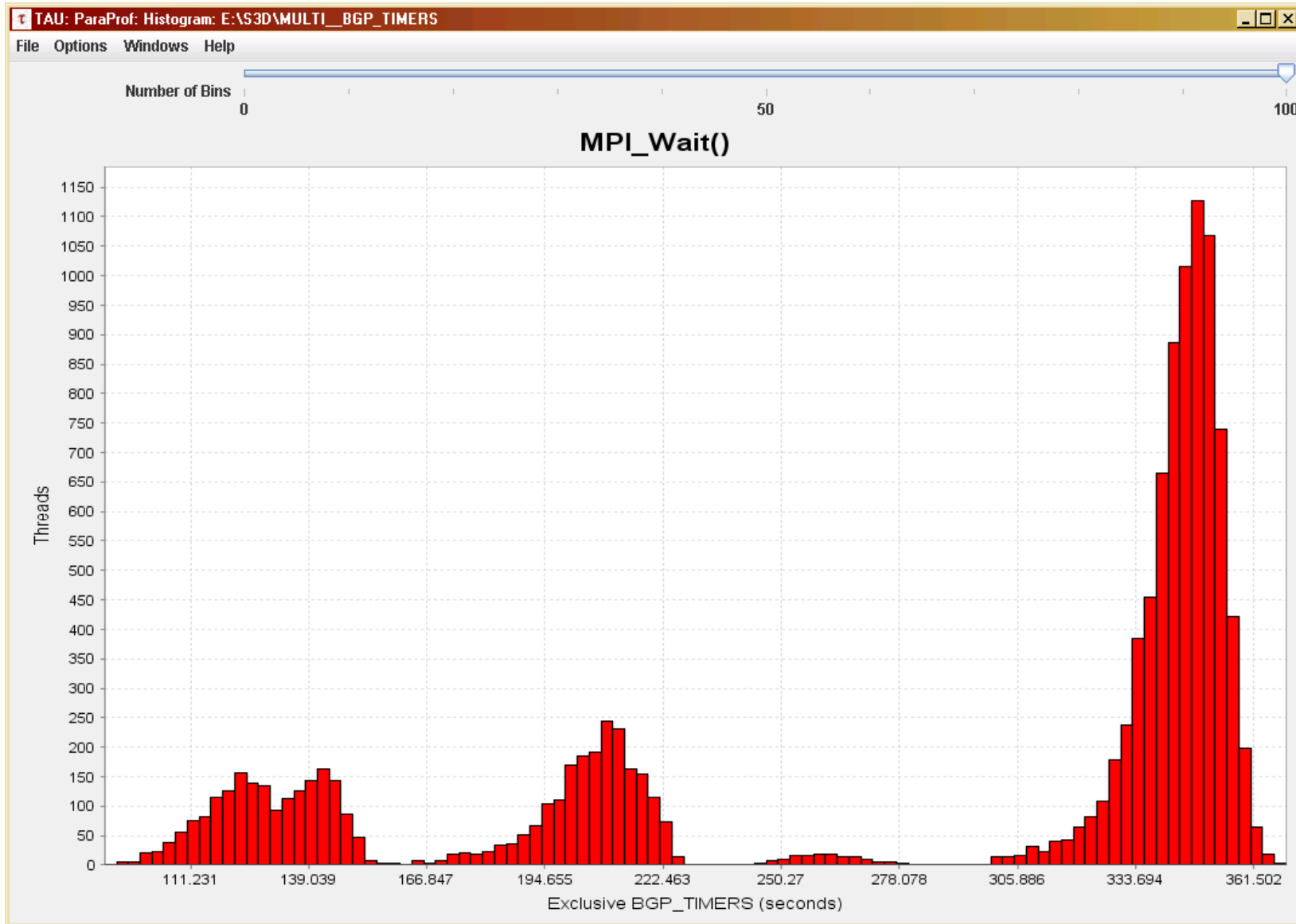




# ParaProf: MPI\_Wait variation!



# ParaProf: MPI\_Wait Histogram







# *S3D - Building with TAU*

- ❑ Change name of compiler in build/make.XT3
  - `ftn=> tau_f90.sh`
  - `cc => tau_cc.sh`
- ❑ Set TAU compilation variables
  - `-tau_makefile=/home/wsppear/bin/tau-2.17.2/bgp/lib/Makefile.tau-bgptimers-multiplecounters-mpi-papi-pdt`
    - Choose callpath, PAPI counters, MPI profiling, PDT for source instrumentation
  - `-tau_options="-optPreProcess -optTauSelectFile=/home/wsppear/sel.txt"`
    - Selective instrumentation file eliminates instrumentation in lightweight routines
    - Pre-process Fortran source code using `cpp` before compiling
- ❑ Specify runtime environment variables for instrumentation control and event PAPI counter selection:
  - `COUNTER1=BGP_TIMERS:`
  - `COUNTER2=PAPI_NATIVE_PNE_BGP_PU0_FPU_ADD_SUB_1`
  - `COUNTER3=PAPI_NATIVE_PNE_BGP_PU0_FPU_MULT_1`
  - `COUNTER4=PAPI_NATIVE_PNE_BGP_PU0_FPU_FMA_2`
  - `COUNTER5=PAPI_NATIVE_PNE_BGP_PU0_FPU_DIV_1`
  - `COUNTER6=PAPI_NATIVE_PNE_BGP_PU0_FPU_ADD_SUB_2`
  - `COUNTER7=PAPI_NATIVE_PNE_BGP_PU0_FPU_MULT_2`
  - `COUNTER8=PAPI_NATIVE_PNE_BGP_PU0_FPU_FMA_4`
  - `TAU_THROTTLE=1`



# *Concluding Discussion*

- ❑ Collected data for S3D up to 12k cores
- ❑ Observed behavioral differences between CNL and BGP
- ❑ Observed bimodal behavior in MPI\_Wait
- ❑ Other metrics of interest
  - Memory
  - Flop/S
- ❑ Issues with PAPI counters on BGP

# *Support Acknowledgements*



- Department of Energy (DOE)
  - Office of Science
  - LLNL, LANL, ORNL, ASC, ANL
  - PERI

