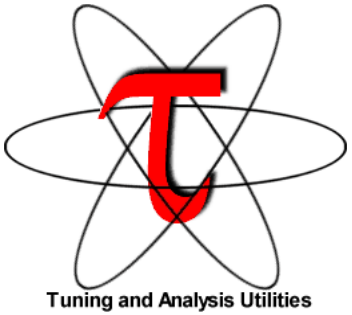
Scalability Study of S3D on Intrepid BGP using TAU

Wyatt Spear
tau-team@cs.uoregon.edu

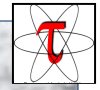


UNIVERSITY
OF OREGON

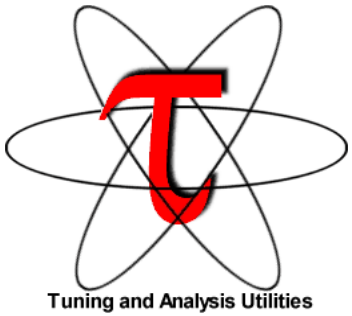
Acknowledgements

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

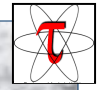
TAU Parallel Performance System



- **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

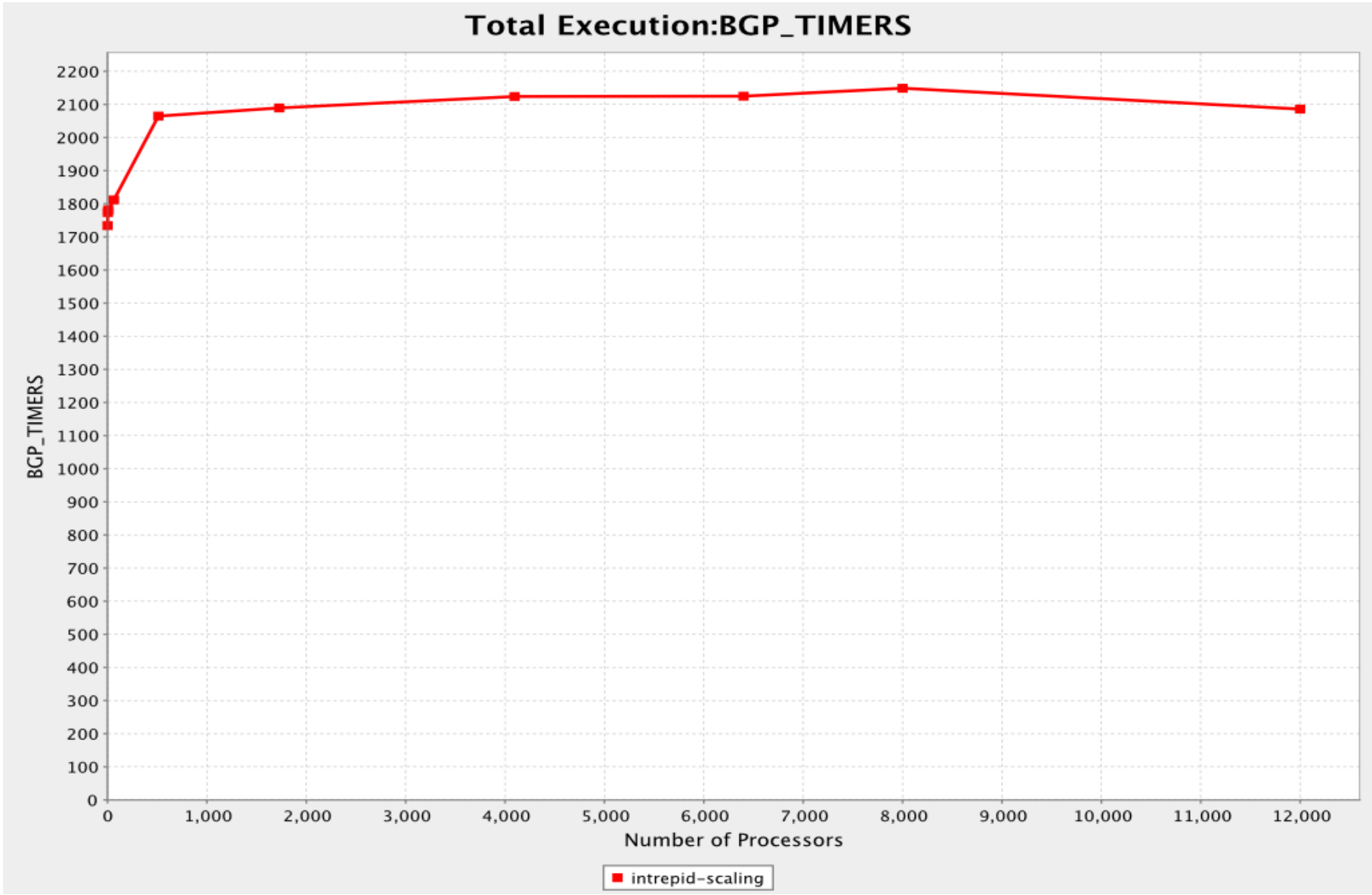


Scalability Study

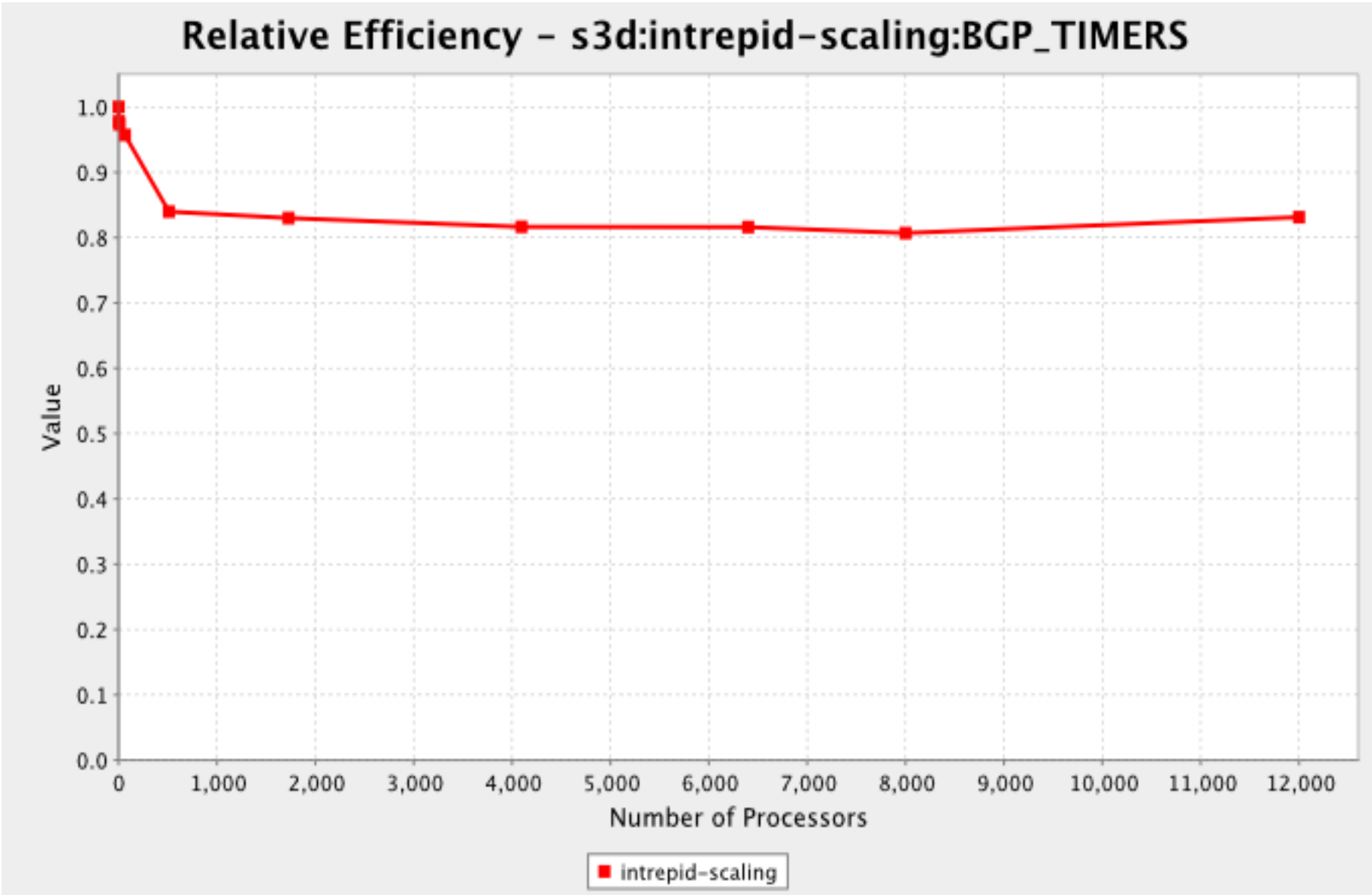


- C2H4 Benchmark
- Platform: Intrepid BGP
  - 1p
  - 4p
  - 64p
  - 512p
  - 1728p
  - 4096p
  - 8000p
  - 12000p
- Goal: to evaluate scaling properties of code regions
- Scalability of MPI operations

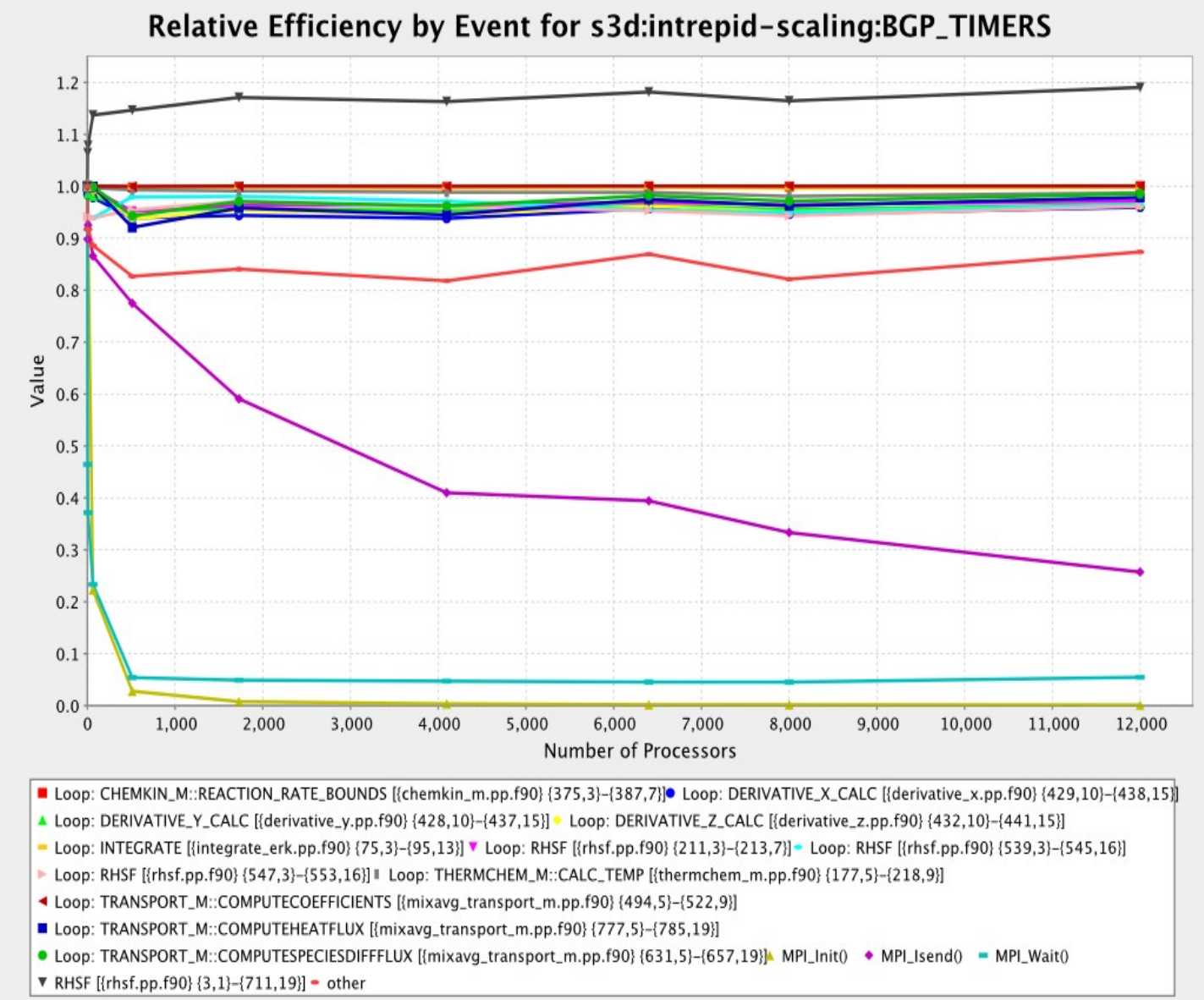
Total Execution Time



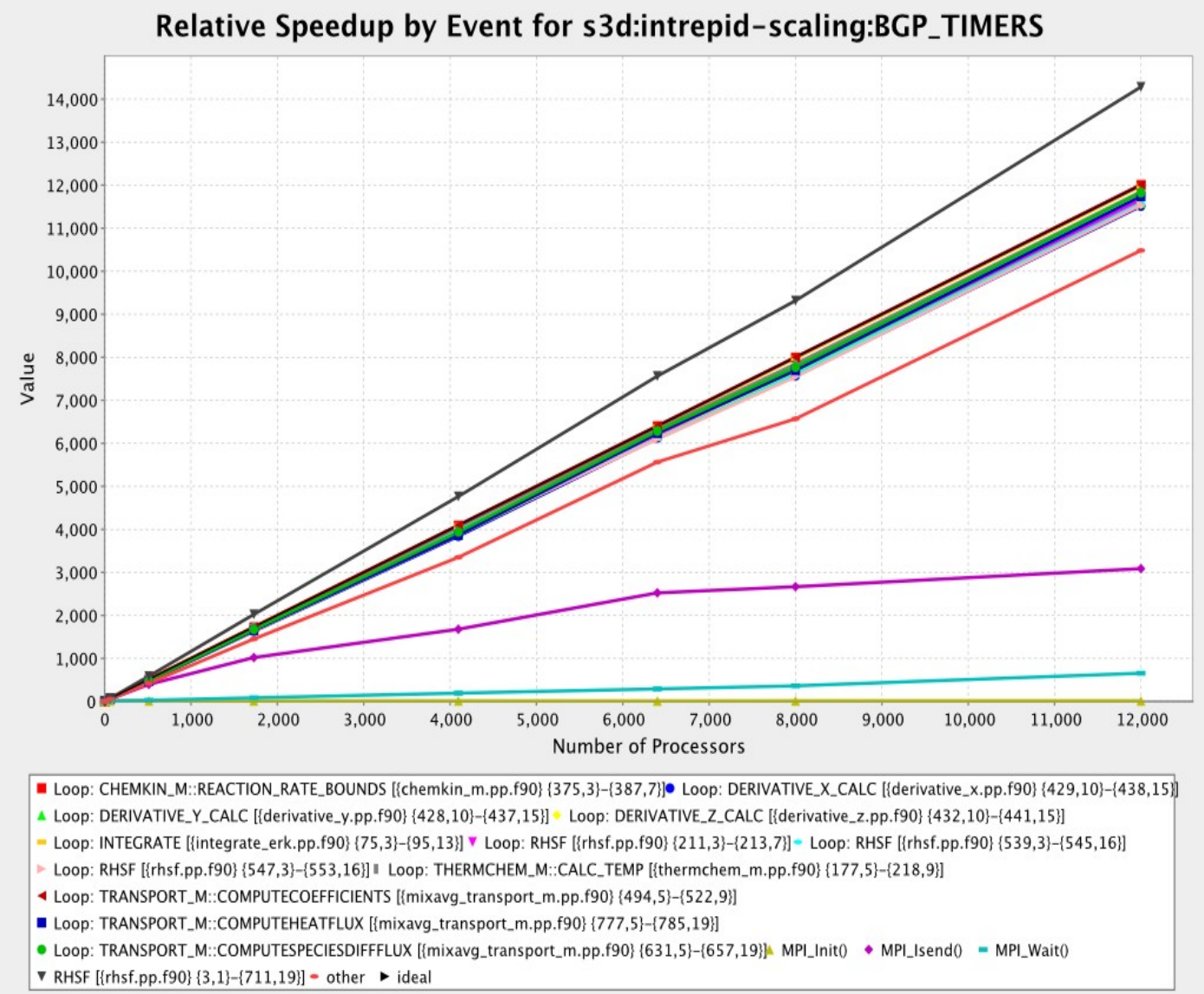
Relative Efficiency For S3D - Weak Scaling



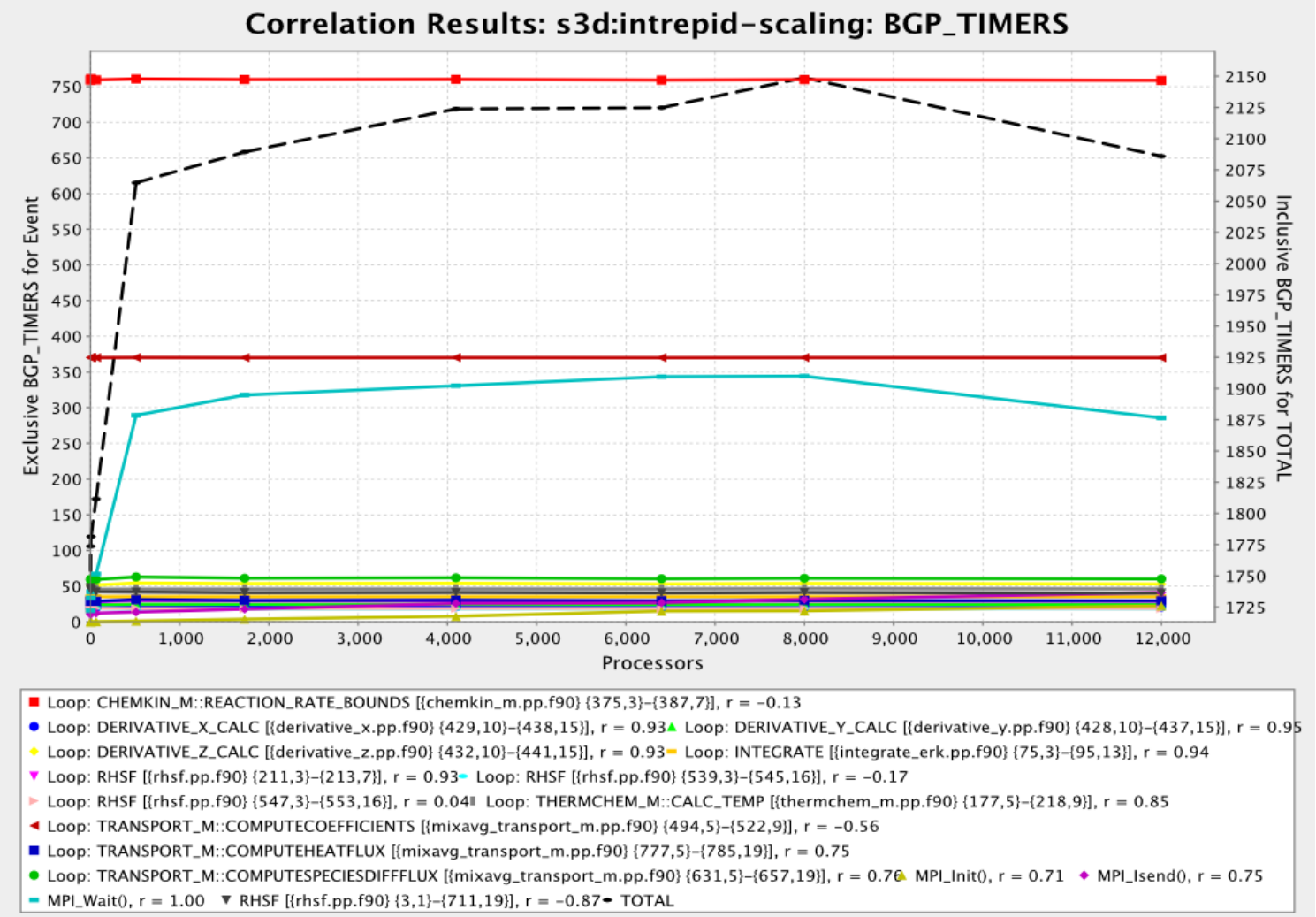
Relative Efficiency by Event



Relative Speedup by Event

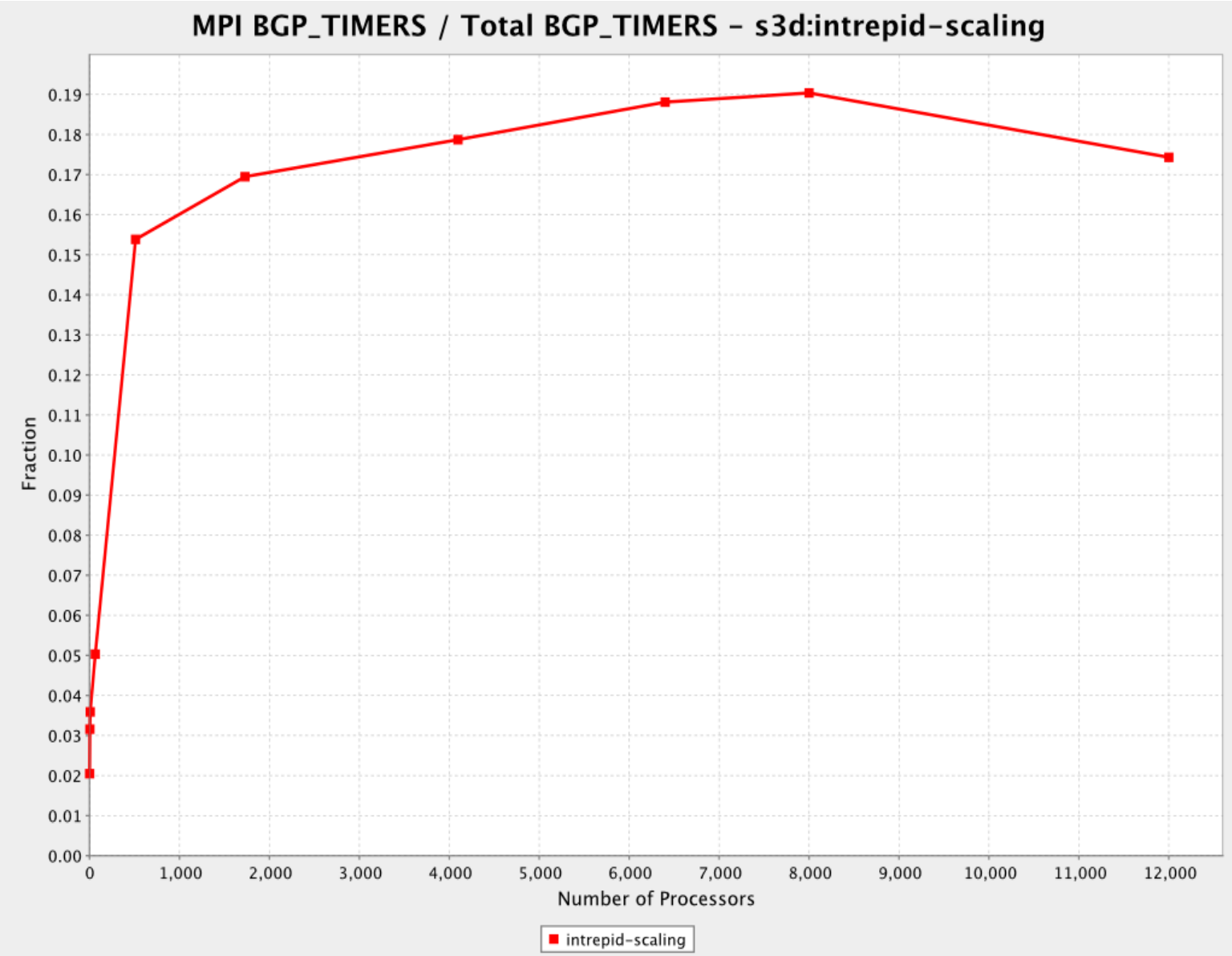


Data Mining: Event Correlation to Total Time

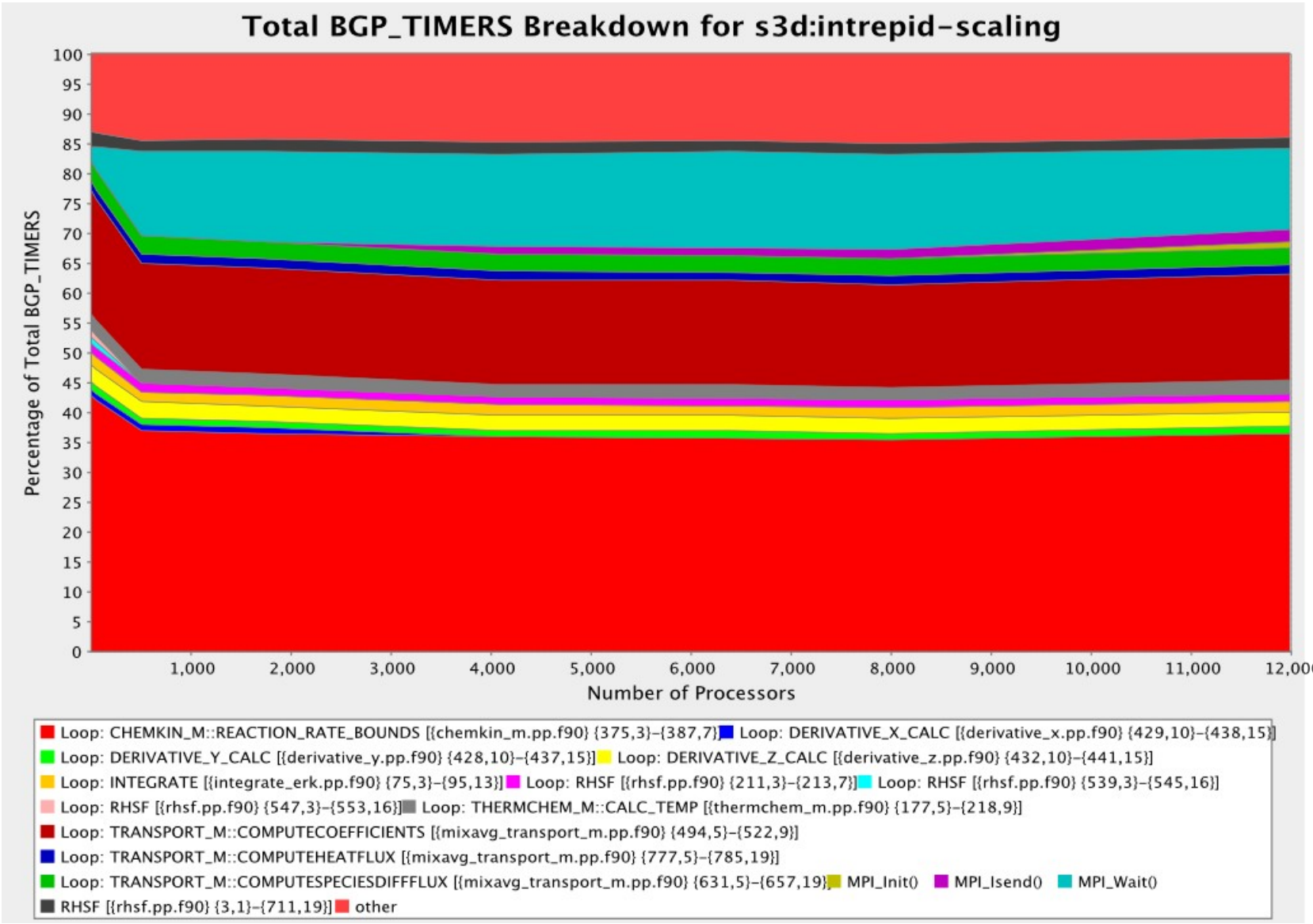


r = 1 implies direct correlation

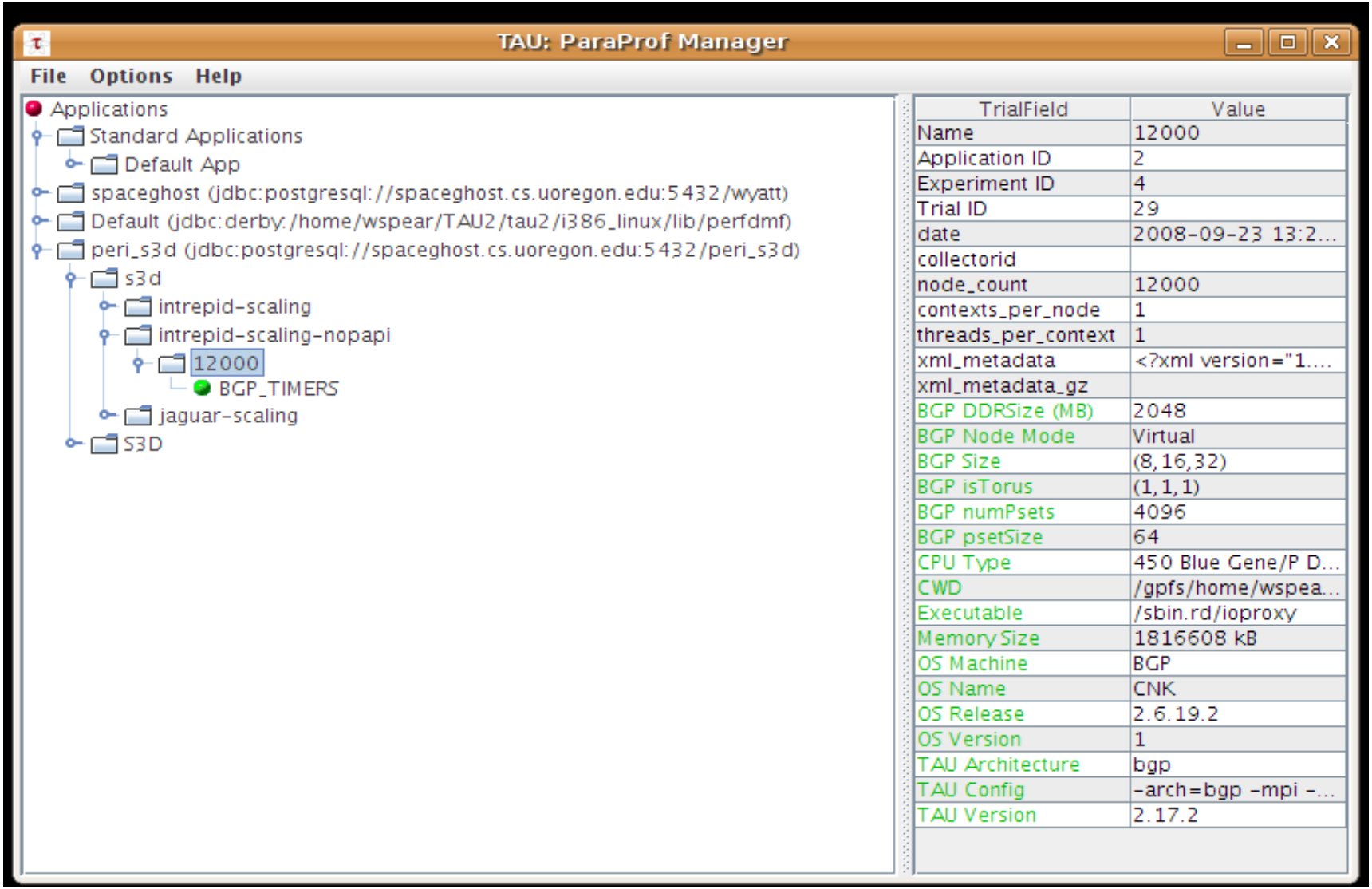
MPI Scaling (Total time in MPI/Total Time)



Total Runtime Breakdown by Events

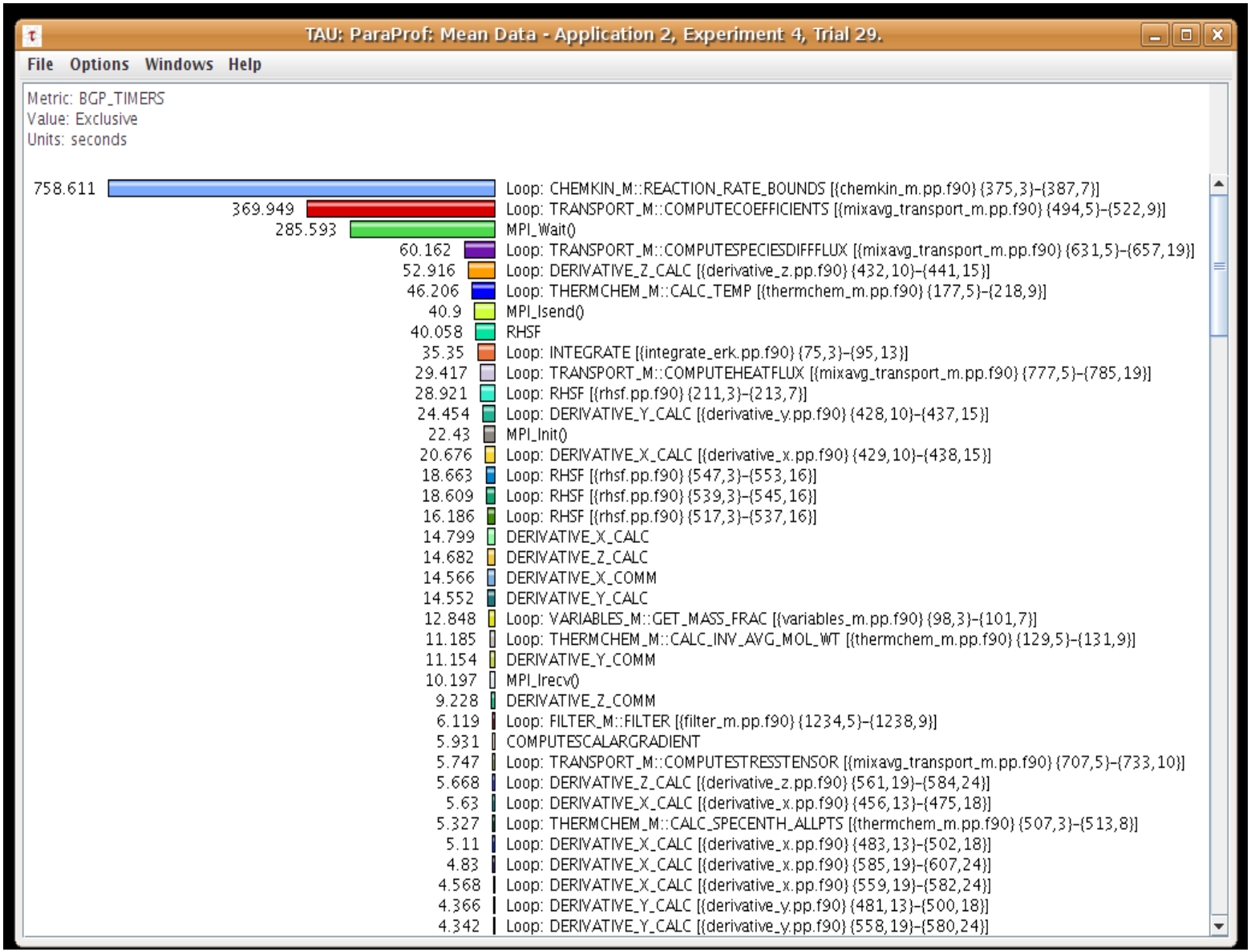


ParaProf: 12000 core job

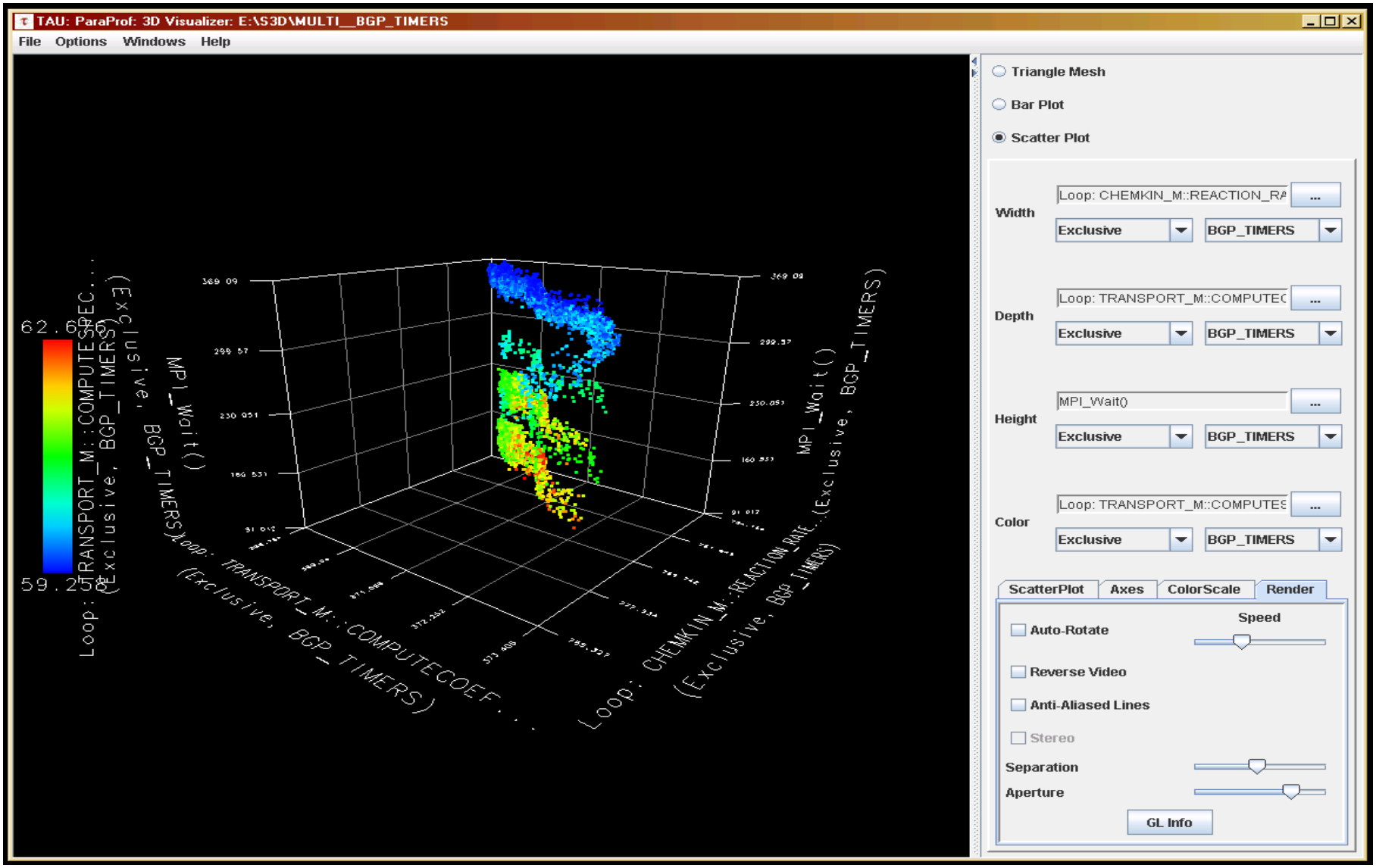


TrialField	Value
Name	12000
Application ID	2
Experiment ID	4
Trial ID	29
date	2008-09-23 13:2...
collectorid	
node_count	12000
contexts_per_node	1
threads_per_context	1
xml_metadata	<?xml version="1....
xml_metadata_gz	
BGP DDRSize (MB)	2048
BGP Node Mode	Virtual
BGP Size	(8,16,32)
BGP isTorus	(1,1,1)
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

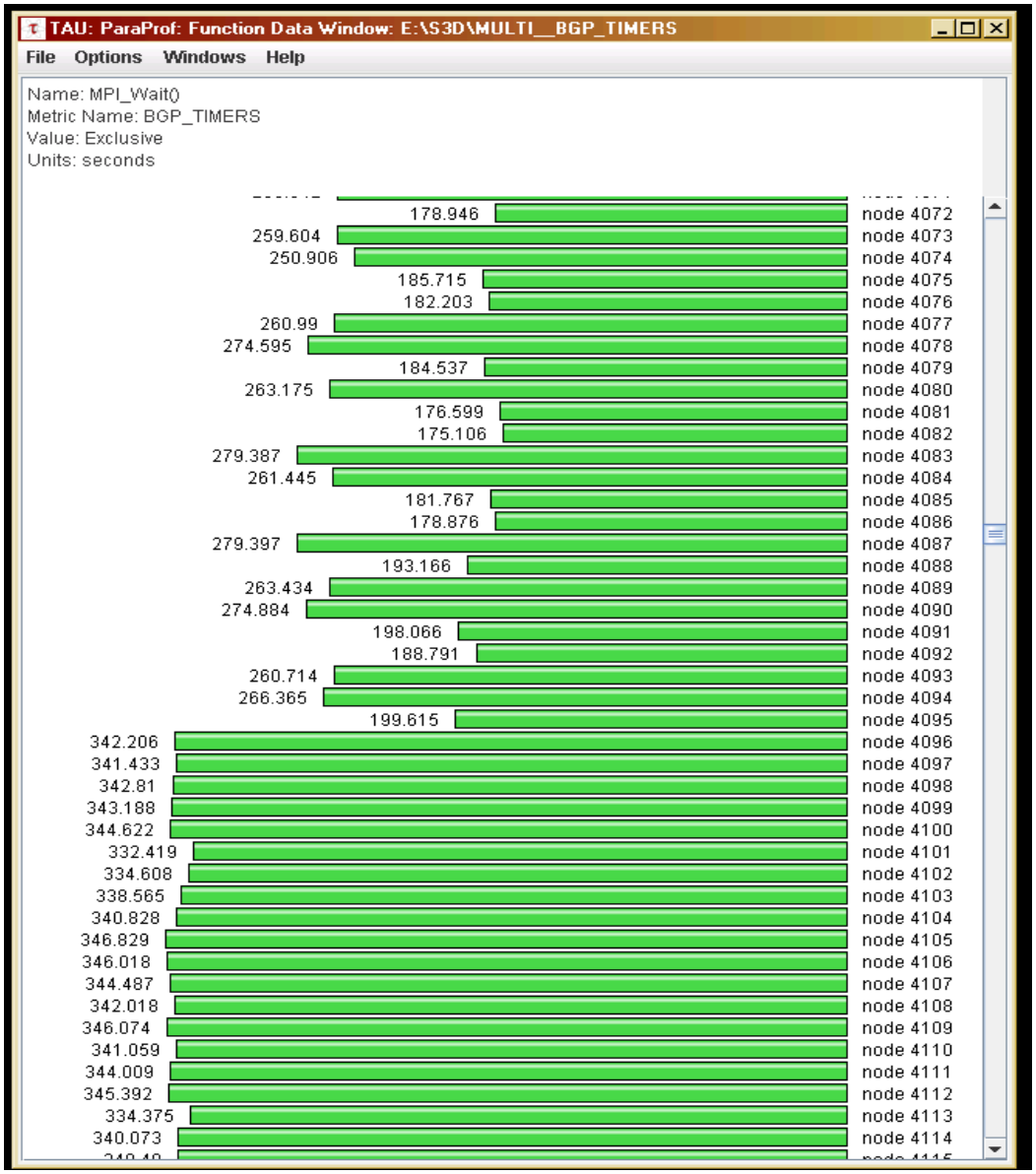
ParaProf: Mean across all nodes



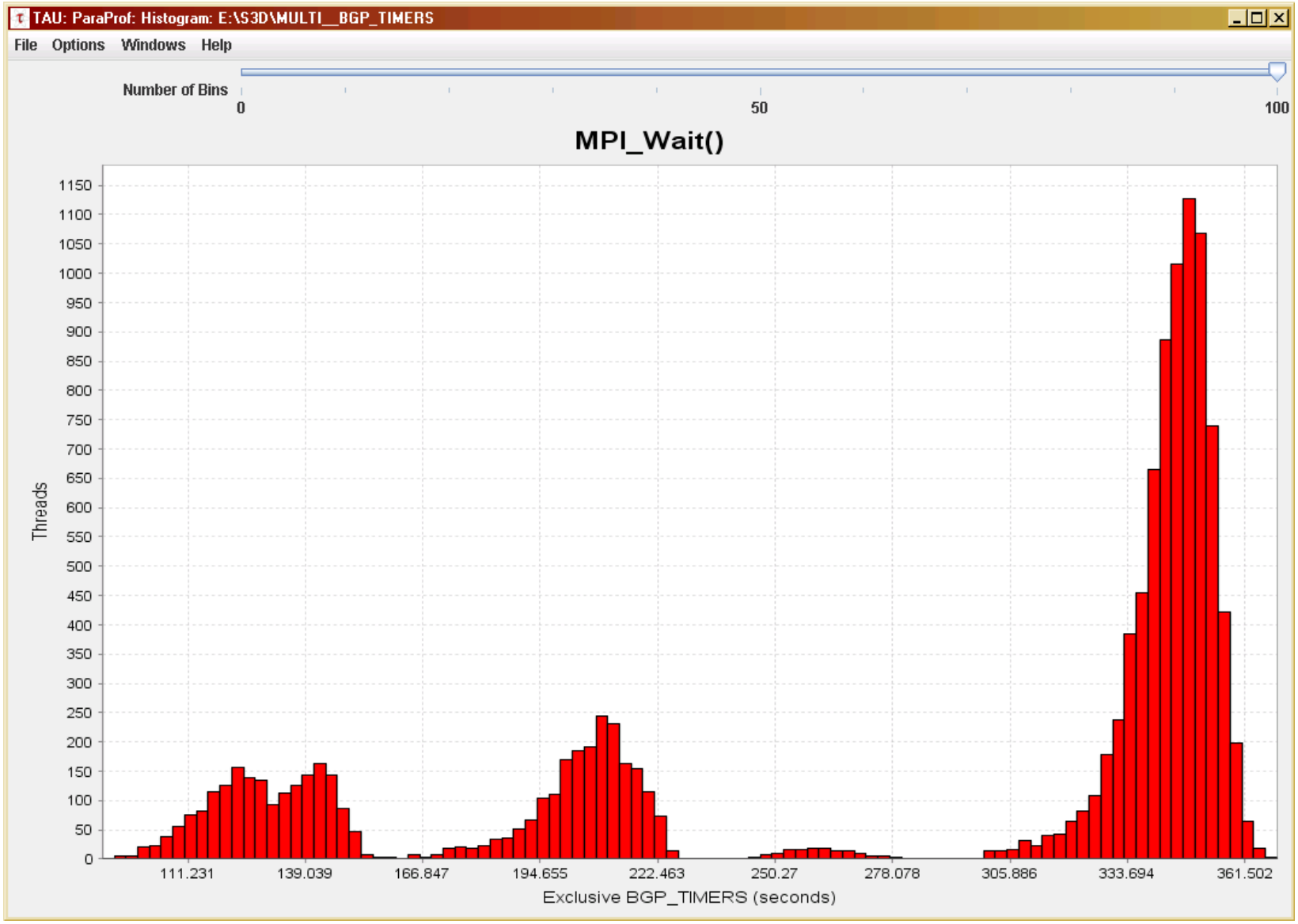
ParaProf: 3D Correlation Cube: MPI_Wait!



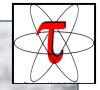
ParaProf: MPI_Wait variation!



ParaProf: MPI_Wait Histogram

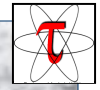


S3D - Building with TAU



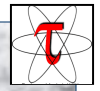
- Change name of compiler in build/make.XT3
  - **ftn=> tau_f90.sh**
  - **cc => tau_cc.sh**
- Set TAU compilation variables
  - **-tau_makefile=/home/wspear/bin/tau-2.17.2/bgp/lib/Makefile.tau-bgptimers-multiplecounters-mpi-papi-pdt**
    - Choose callpath, PAPI counters, MPI profiling, PDT for source instrumentation
  - **-tau_options="-optPreProcess -optTauSelectFile=/home/wspear/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