

VI-HPS



scalasca

Tutorial Exercise

NPB-MZ-MPI/BT

Brian Wylie & Markus Geimer
Jülich Supercomputing Centre
scalasca@fz-juelich.de
March 2011

0. Reference preparation for validation
 1. Program instrumentation: `skin`
 2. Summary measurement collection & analysis: `scan [-s]`
 3. Summary analysis report examination: `square`
 4. Summary experiment scoring: `square -s`
 5. Event trace collection & analysis: `scan -t`
 6. Event trace analysis report examination: `square`
- Configuration & customization
 - Instrumentation, Measurement, Analysis, Presentation

- Intermediate-level tutorial example
- Available in MPI, OpenMP & **hybrid OpenMP/MPI** variants
 - also MPI File I/O variants (collective & individual)
- Summary measurement collection & analysis
 - Automatic instrumentation
 - ▶ OpenMP, MPI & application functions
 - Summary analysis report examination
 - PAPI hardware counter metrics
- Trace measurement collection & analysis
 - Filter determination, specification & configuration
 - Automatic trace analysis report patterns
- Manual and PDT instrumentation
- Measurement configuration
- Analysis report algebra

- HPC sites have slightly different setups for installed tools, but it is typically based on “modules”
- The `scalasca` module(s) may be installed as part of the `UNITE` module which configures a *Unified Tool Environment*
 - NB: older non-UNITE modules may be the default!
 - Load the UNITE module first, if necessary
- Check which modules are available, and then load an appropriate module

```
% module avail scalasca
scalasca/1.3-gnu+impi          scalasca/1.3-intel+impi
scalasca/1.3-gnu+mpt           scalasca/1.3-intel+mpt
scalasca/1.3-gnu+mvapich2     scalasca/1.3-intel+mvapich2
% module load scalasca
```

- There may be multiple versions to choose from
- Depending on your compiler & MPI library combination, load a corresponding version of Scalasca

- Load the module

```
% module load UNITE
UNITE loaded
% module load scalasca
scalasca/1.3 loaded
```

- ... and run `scalasca` for brief usage information

```
% scalasca
Scalasca 1.3
Toolset for scalable performance analysis of large-scale applications
usage: scalasca [-v][[-n] {action}]
  1. prepare application objects and executable for measurement:
     scalasca -instrument <compile-or-link-command>      # skin
  2. run application under control of measurement system:
     scalasca -analyze <application-launch-command>      # scan
  3. interactively explore measurement analysis report:
     scalasca -examine <experiment-archive|report>      # square

-v: enable verbose commentary
-n: show actions without taking them
-h: show quick reference guide (only)
```

- Prefix compile/link commands in Makefile definitions (config/make.def) with the Scalasca instrumenter

```
MPIF77 = scalasca -instrument mpif77
FLINK = $(MPIF77)
FFLAGS = -O -fopenmp

bt-mz: $(OBJECTS)
    $(FLINK) $(FFLAGS) -o bt-mz $(OBJECTS)
.f.o:
    $(MPIF77) $(FFLAGS) -c $<
```

- or use PREP macro as customizable preparation preposition

```
MPIF77 = $(PREP) mpif77
```

- By default, PREP macro is not set and no instrumentation is performed for a regular “production” build
- Specifying a PREP value in the Makefile or on the make command line uses it as a preposition, e.g., for instrumentation
 - ▶ % make **PREP=“scalasca -instrument”** ...
scalasca -instrument mpif77 -O -fopenmp -c bt.f

- Return to root directory and clean-up

```
% make clean
```

- Re-build specifying Scalasca instrumenter as PREP

```
% make bt-mz CLASS=B NPROCS=4 PREP="scalasca -instrument"
cd BT-MZ; make CLASS=B NPROCS=4 VERSION=
gmake: Entering directory 'BT-MZ'
cd ../sys; cc -o setparams setparams.c
../sys/setparams bt-mz 4 B
scalasca -instrument mpif77 -c -O -fopenmp bt.f
...
scalasca -instrument mpif77 -c -O -fopenmp setup_mpi.f
cd ../common; scalasca -instrument mpif77 -c -O -fopenmp timers.f
scalasca -instrument mpif77 -O -fopenmp -o ../bin.scalasca/bt-mz_B.4 \
bt.o make_set.o initialize.o exact_solution.o exact_rhs.o \
set_constants.o adi.o define.o copy_faces.o rhs.o solve_subs.o \
x_solve.o y_solve.o z_solve.o add.o error.o verify.o setup_mpi.o \
../common/print_results.o ../common/timers.o
INFO: Instrumented executable for OMP+MPI measurement
gmake: Leaving directory 'BT-MZ'
```

- Run the application using the Scalasca measurement collection & analysis nexus prefixed to launch command

```
% cd bin.scalasca
% OMP_NUM_THREADS=4 scalasca -analyze mpiexec -np 4 ./bt-mz_B.4
S=C=A=N: Scalasca 1.3 runtime summarization
S=C=A=N: ./epik_bt-mz_B_4x4_sum experiment archive
S=C=A=N: Sun Mar 29 16:36:31 2009: Collect start
mpiexec -np 4 ./bt-mz_B.4
[00000]EPIK: Created new measurement archive ./epik_bt-mz_B_4x4_sum
[00000]EPIK: Activated ./epik_bt-mz_B_4x4_sum [NO TRACE] (0.006s)

    [... Application output ...]

[00000]EPIK: Closing experiment ./epik_bt-mz_B_4x4_sum
[00000]EPIK: 164 unique paths (178 max paths, 7 max frames, 0 unknown)
[00000]EPIK: Unifying... done (0.023s)
[00000]EPIK: Collating... done (0.049s)
[00000]EPIK: Closed experiment ./epik_bt-mz_B_4x4_sum (0.073s)
S=C=A=N: Sun Mar 29 16:36:45 2009: Collect done (status=0) 57s
S=C=A=N: ./epik_bt-mz_B_4x4_sum complete.
```

- Produces experiment directory ./epik_bt-mz_B_4x4_sum

- Interactive exploration with Scalasca GUI

```
% scalasca -examine epik_bt-mz_B_4x4_sum  
INFO: Post-processing runtime summarization result...  
INFO: Displaying ./epik_bt-mz_B_4x4_sum/summary.cube...
```

[GUI showing summary analysis report]


- The measurement archive directory ultimately contains
 - a copy of the execution output (epik.log)
 - a record of the measurement configuration (epik.conf)
 - the basic analysis report that was collated after measurement (epitome.cube)
 - the complete analysis report produced during post-processing (summary.cube.gz)

Analysis report exploration (opening view)

The screenshot displays the Cube 3.0 QT application window titled "Cube 3.0 QT: epik_bt_16_sum/summary.cube.gz". The interface is divided into three main panels, each with a dropdown menu set to "Absolute":

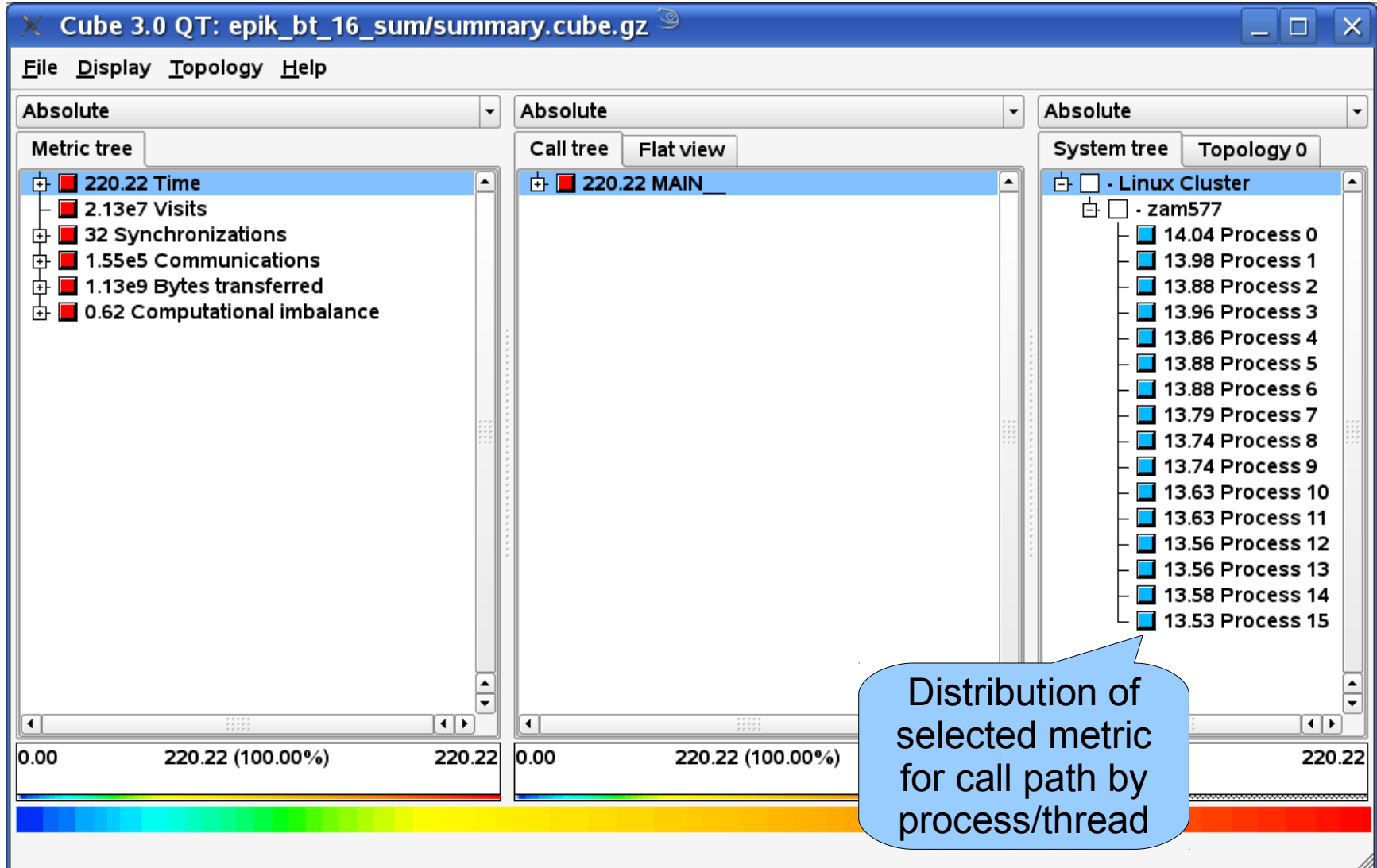
- Metric tree:** Shows a hierarchical list of metrics:
 - 220.22 Time
 - 2.13e7 Visits
 - 32 Synchronizations
 - 1.55e5 Communications
 - 1.13e9 Bytes transferred
 - 0.62 Computational imbalance
- Call tree:** Shows a single entry: 220.22 MAIN_.
- System tree:** Shows a single entry: 220.22 Linux Cluster.

An "About Cube 3.0 QT" dialog box is overlaid in the center, containing the following information:

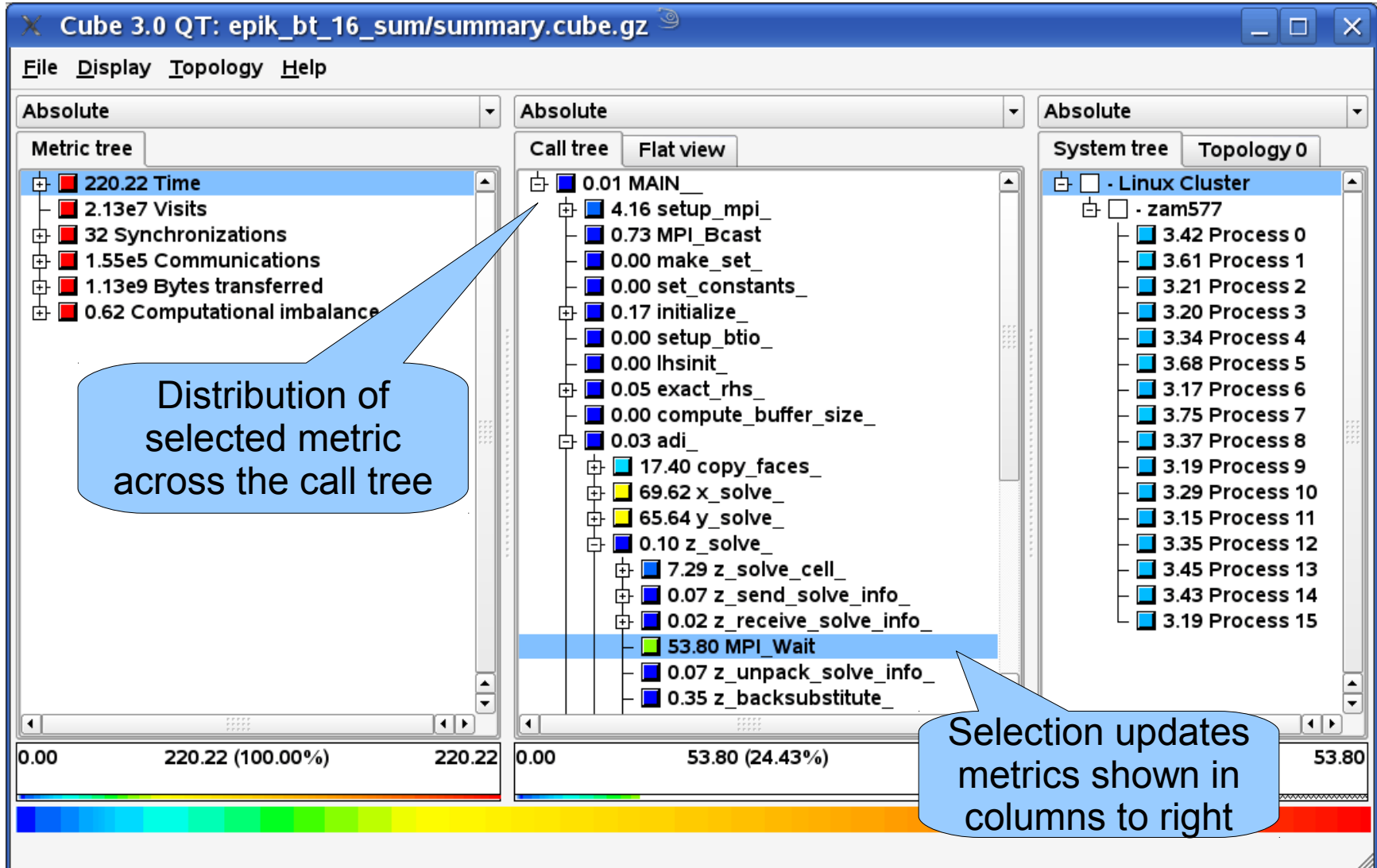
- This is Cube 3.0 QT.
-  (c) 2009 Juelich Supercomputing Centre, Forschungszentrum Juelich GmbH
- Home page: www.scalasca.org
- Technical support: scalasca@fz-juelich.de

At the bottom of each panel, there is a progress bar showing "0.00 220.22 (100.00%) 220.22". A color gradient bar is visible at the very bottom of the application window.

Analysis report exploration (system tree)



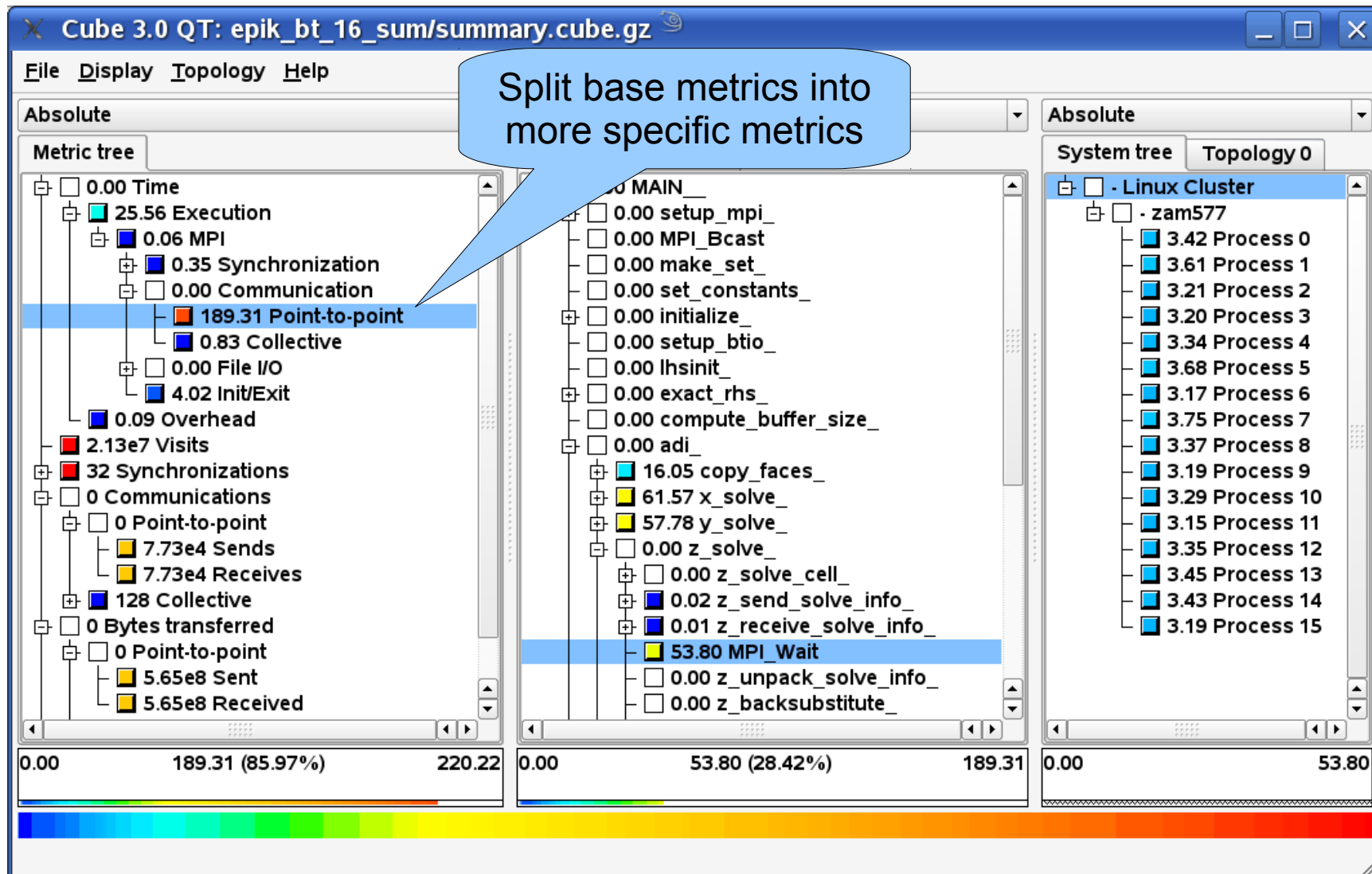
Analysis report exploration (call tree)



Distribution of selected metric across the call tree

Selection updates metrics shown in columns to right

Analysis report exploration (metric tree)



Analysis report exploration (source browser)

```

c -----
c   in our terminology stage is the number of the cell in the y-dir
c   i.e. stage = 1 means the start of the line stage=ncells means e
c -----
c   do stage = 1,ncells
c     c = slice(3,stage)
c     isize = cell_size(1,c) - 1
c     jsize = cell_size(2,c) - 1
c     ksize = cell_size(3,c) - 1
c -----
c   set last-cell flag
c -----
c     if (stage .eq. ncells) then
c       last = 1
c     else
c       last = 0
c     endif
c
c     if (stage .eq. 1) then
c -----
c     This is the first cell, so solve without receiving data
c -----
c       first = 1
c       call lhsz(c)
c       call z_solve_cell(first,last,c)
c     else
c -----
c     Not the first cell of this line, so receive info from
c     processor working on preceeding cell
c -----
c       first = 0
c       call z_receive_solve_info(recv_id,c)
c -----
c     overlap computations and communications
c -----
c       call lhsz(c)
c -----
c     wait for completion
c -----
c     call mpi_wait(send_id,r_status,error)
c     call mpi_wait(recv_id,r_status,error)
c -----
c     install C'(kstart+1) and rhs'(kstart+1) to be used in this cell
c -----
c       call z_unpack_solve_info(c)
c       call z_solve_cell(first,last,c)
c     endif
c
c     if (last .eq. 0) call z_send_solve_info(send_id,c)
c   enddo

```

The screenshot displays a performance analysis tool with two main panels. The left panel, titled 'Flat view', shows a hierarchical tree of execution times for various tasks. The right panel, titled 'System tree', shows a tree of process execution times for a 'Linux Cluster'.

Flat view (Left Panel):

- 0.00 MAIN_
- 0.00 setup_mpi_
- 0.00 MPI_Bcast
- 0.00 make_set_
- 0.00 set_constants_
- 0.00 initialize_
- 0.00 setup_btio_
- 0.00 lhsinit_
- 0.00 exact_rhs_
- 0.00 compute_buffer_size_
- 0.00 adi_
- 16.05 copy_faces_ (highlighted)
- 61.57 x_solve_ (highlighted)
- 57.78 y_solve_ (highlighted)
- 0.00 z_solve_ (expanded)
- 0.00 z_solve_cell_ (expanded)
- 0.02 z_send_solve_info_ (expanded)
- 0.01 z_receive_solve_info_ (expanded)
- 53.80 MPI Wait (highlighted)
- 0.00 z_unpack_solve_info_ (expanded)
- 0.00 z_backsubstitute_ (expanded)

System tree (Right Panel):

- Linux Cluster
- zam577
 - 3.42 Process 0
 - 3.61 Process 1
 - 3.21 Process 2
 - 3.20 Process 3
 - 3.34 Process 4
 - 3.68 Process 5
 - 3.17 Process 6
 - 3.75 Process 7
 - 3.37 Process 8
 - 3.19 Process 9
 - 3.29 Process 10
 - 3.15 Process 11
 - 3.35 Process 12
 - 3.45 Process 13
 - 3.43 Process 14
 - 3.19 Process 15

Summary (Bottom):

53.80 (28.42%)	189.31	0.00	53.80
----------------	--------	------	-------

- If you made it this far, you successfully used Scalasca to
 - *instrument* the application
 - *analyze* its execution with a summary measurement, and
 - *examine* it with the interactive analysis report explorer GUI
- ... revealing the call-path profile annotated with
 - Time metrics (including MPI & OpenMP times)
 - Visit counts
 - MPI message statistics (sends/receives, bytes sent/received)
 - Computational imbalance
- ... but how **good** was the measurement?
 - The measured execution produced the desired valid result
 - however, the execution took rather longer than expected!
 - ▶ even when ignoring measurement start-up/completion, therefore
 - ▶ it was probably dilated by instrumentation/measurement overhead

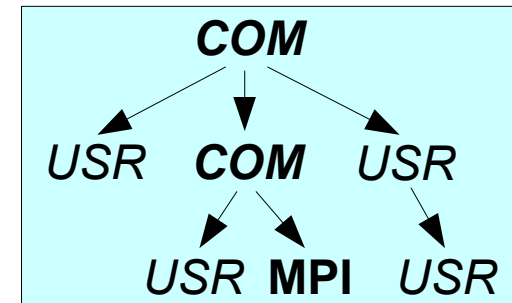
- Report scoring as textual output

```
% scalasca -examine -s epik_bt-mz_B_4x4_sum
cube3_score -r ./epik_bt-mz_B_4x4_sum/summary.cube
Reading ./epik_bt-mz_B_4x4_sum/summary.cube... done.
Est. aggregate size of event trace (total_tbc): 39,231,218,072 bytes
Est. size of largest process trace (max_tbc): 2,632,541,576 bytes
(When tracing set ELG_BUFFER_SIZE to avoid intermediate flushes or
 reduce requirements using filter file listing names of USR regions.)

INFO: Score report written to ./epik_bt-mz_B_4x4_sum/epik.score
```

- Region/callpath classification

- MPI (pure MPI library functions)
- OMP (pure OpenMP functions/regions)
- USR (user-level source local computation)
- COM (“combined” USR + OpenMP/MPI)
- ANY/ALL (aggregate of all region types)

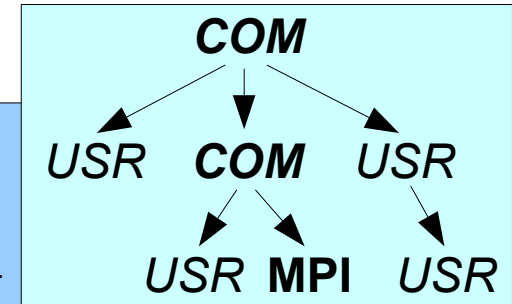


- Score report breakdown by region

```
% less epik_bt-mz_B_4x4_sum/epik.score
```

flt	type	max_tbc	time	% region	
	ANY	2632541576	871.73	100.00	(summary) ALL
	MPI	73064	13.27	1.52	(summary) MPI
	OMP	5186496	496.36	56.94	(summary) OMP
	COM	1087824	3.15	0.36	(summary) COM
	USR	2626194144	358.88	41.17	(summary) USR
USR		841575744	109.22	12.53	matmul_sub_
USR		841575744	168.61	19.34	binvcrhs_
USR		841575744	68.95	7.91	matvec_sub_
USR		37120680	5.14	0.59	binvrhs_
USR		37120680	4.10	0.47	lhsinit_
USR		29960856	2.85	0.33	exact_solution_
COM		308736	0.82	0.09	copy_x_face_
COM		308736	0.81	0.09	copy_y_face_
OMP		283008	2.07	0.24	!\$omp parallel @exch_qbc.f:204
OMP		283008	2.02	0.23	!\$omp parallel @exch_qbc.f:215
OMP		283008	2.16	0.25	!\$omp parallel @exch_qbc.f:244
OMP		283008	2.01	0.23	!\$omp parallel @exch_qbc.f:255

...



- Summary measurement analysis score reveals
 - Total size of event trace would be almost 40GB
 - Maximum trace buffer size would be over 2.5GB per thread
 - ▶ smaller buffer would require flushes to disk during measurement resulting in substantial perturbation
 - 99.76% of the trace requirements are for USR regions
 - ▶ purely computational routines never found on COM call-paths common to communication routines
 - These USR regions contribute around 40% of total time
 - ▶ however, much of that is very likely to be measurement overhead for a few frequently-executed small routines
- Advisable to tune measurement configuration
 - Specify an adequate trace buffer size
 - Specify a filter file listing (USR) regions not to be measured

- Report scoring with prospective filter listing USR regions

```
% scalasca -examine -s -f bt.filt epik_bt-mz_B_4x4_sum
cube3_score -r -f bt.filt ./epik_bt-mz_B_4x4_sum/summary.cube
Applying filter "./bt.filt":
Estimated aggregate size of event trace (total_tbc): 16,852,888 bytes
Estimated size of largest process trace (max_tbc): 1,053,304 bytes

INFO: Score report written to ./epik_bt-mz_B_4x4_sum/epik.score_bt.filt
```

```
% less epik_bt-mz_B_4x4_sum/epik.score_bt.filt
```

flt	type	max_tbc	time	% region
+	FLT	2626190016	358.88	41.17 (summary) FLT
*	ANY	6351584	512.85	58.83 (summary) ALL-FLT
-	MPI	73064	13.27	1.52 (summary) MPI-FLT
-	OMP	5186496	496.36	56.94 (summary) OMP-FLT
*	COM	1087824	3.15	0.36 (summary) COM-FLT
*	USR	4152	0.00	0.00 (summary) USR-FLT

```
% cat bt.filt
# bt-mz filter
matmul_sub_
binvrhs_
matvec_sub_
binvrhs_
lhsinit_
exact_solution_
timer_*
```

Filtered routines marked with '+'

+	USR	841575744	109.22	12.53	matmul_sub_
+	USR	841575744	168.61	19.34	binvrhs_
+	USR	841575744	68.95	7.91	matvec_sub_
+	USR	37120680	5.14	0.59	binvrhs_
+	USR	37120680	4.10	0.47	lhsinit_
+	USR	29960856	2.85	0.33	exact_solution_

...

- Rename former measurement archive directory, set new filter configuration and re-run the measurement

```
% mv epik_bt-mz_B_4x4_sum epik_bt-mz_B_4x4_sum.nofilt
% export EPK_FILTER=bt.filt
% OMP_NUM_THREADS=4 scalasca -analyze mpiexec -np 4 ./bt-mz_B.4
S=C=A=N: Scalasca 1.3 runtime summarization
S=C=A=N: ./epik_bt-mz_4x4_sum experiment archive
S=C=A=N: Sun Mar 29 16:58:34 2009: Collect start
mpiexec -np 4 ./bt-mz_B.4
[00000.0]EPIK: Created new measurement archive ./epik_bt-mz_B_4x4_sum
[00000.0]EPIK: EPK_FILTER "bt.filt" filtered 10 of 113 functions
[00000.0]EPIK: Activated ./epik_bt-mz_B_4x4_sum [NO TRACE] (0.071s)

[... Application output ...]

[00000.0]EPIK: Closing experiment ./epik_bt-mz_B_4x4_sum
[00000.0]EPIK: 134 unique paths (148 max paths, 7 max frames, 0 unkns)
[00000.0]EPIK: Unifying... done (0.014s)
[00000.0]EPIK: Collating... done (0.059s)
[00000.0]EPIK: Closed experiment ./epik_bt-mz_B_4x4_sum (0.075s)
S=C=A=N: Sun Mar 29 16:58:41 2009: Collect done (status=0) 36s
S=C=A=N: ./epik_bt-mz_B_4x4_sum complete.
```

- Scoring of new analysis report as textual output

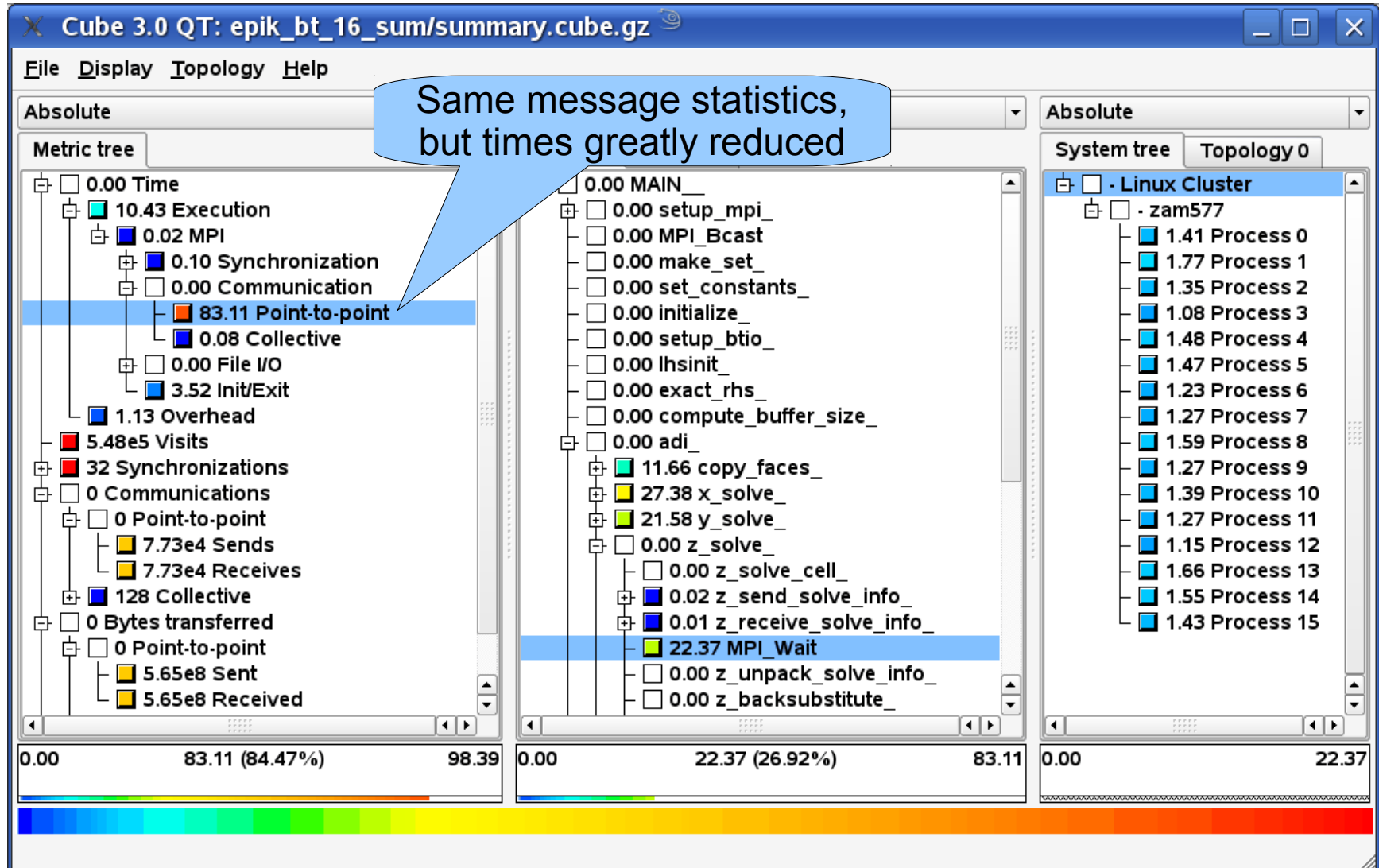
```
% scalasca -examine -s epik_bt-mz_B_4x4_sum
INFO: Post-processing runtime summarization result...
cube3_score ./epik_bt-mz_B_4x4_sum/summary.cube
Estimated aggregate size of event trace (total_tbc): 83,920,952 bytes
Estimated size of largest process trace (max_tbc): 6,351,584 bytes
...

INFO: Score report written to ./epik_bt_W_16_sum/epik.score

flt  type          max_tbc          time          % region
     ANY          6351584         531.69       100.00 (summary) ALL
     MPI           73064           13.27         2.50 (summary) MPI
     OMP          5186496         515.11        96.88 (summary) OMP
     COM          1087824          3.22          0.61 (summary) COM
     USR           4152             0.00          0.00 (summary) USR
```

- Significant reduction in runtime (measurement overhead)
 - Not only reduced time for USR regions, but OMP reduced too!
- Further measurement tuning (filtering) may be appropriate
 - e.g., “timer_*” filters timer_start_, timer_read_, etc.

Summary analysis report exploration (filtered)



- Re-run the application using Scalasca nexus with “-t” flag

```
% OMP_NUM_THREADS=4 scalasca -analyze -t mpiexec -np 4 ./bt-mz_B.4
S=C=A=N: Scalasca trace collection and analysis
S=C=A=N: ./epik_bt-mz_B_4x4_trace experiment archive
S=C=A=N: Sun Apr 5 18:50:57 2009: Collect start
mpiexec -np 4 ./bt-mz_B.4
[00000.0]EPIK: Created new measurement archive ./epik_bt-mz_B_4x4_trace
[00000.0]EPIK: EPK_FILTER "npb.filt" filtered 10 of 113 functions
[00000.0]EPIK: Activated ./epik_bt-mz_B_4x4_trace [10000000 bytes] (0.051s)

    [... Application output ...]

[00000.0]EPIK: Closing experiment ./epik_bt-mz_B_4x4_trace [0.069GB] (max 18466028)
[00000.0]EPIK: Flushed 6351570 bytes to file ./epik_bt-mz_B_4x4_trace/ELG/00000
[00000.0]EPIK: 134 unique paths (148 max paths, 7 max frames, 0 unknowns)
[00000.0]EPIK: Unifying... done (0.021s)
[00003.0]EPIK: Flushed 6351570 bytes to file ./epik_bt-mz_B_4x4_trace/ELG/00003
...
[00001.0]EPIK: Flushed 6351570 bytes to file ./epik_bt-mz_B_4x4_trace/ELG/00001
[00000.0]EPIK: 1flush=0.001GB@2.582MB/s, Pflush=0.015GB@35.458MB/s
[00000.0]EPIK: Closed experiment ./epik_bt-mz_B_4x4_trace (0.178s)
S=C=A=N: Sun Apr 5 18:51:05 2009: Collect done (status=0) 41s
[. continued ...]
```

- Separate trace file per MPI rank written straight into new experiment directory ./epik_bt-mz_B_4x4_trace

- Continues with automatic (parallel) analysis of trace files

```
S=C=A=N: Sun Apr 5 18:51:05 2009: Analyze start
mpiexec -np 4 scout.hyb ./epik_bt-mz_B_4x4_trace
SCOUT Copyright (c) 1998-2009 Forschungszentrum Juelich GmbH

Analyzing experiment archive ./epik_bt-mz_B_4x4_trace

Reading definitions file ... done (0.563s).
Reading event trace files ... done (0.495s).
Preprocessing ... done (0.134s).
Analyzing event traces ... done (2.186s).
Writing CUBE report ... done (0.160s).

Total processing time : 3.737s
Max. memory usage : 47.504MB

S=C=A=N: Sun Apr 5 18:51:09 2009: Analyze done (status=0) 4s
S=C=A=N: ./epik_bt-mz_B_4x4_trace complete.
```

- Produces trace analysis report in experiment directory

```
% scalasca -examine epik_bt-mz_B_4x4_trace
INFO: Post-processing runtime summarization result...
INFO: Post-processing trace analysis report ...
INFO: Displaying ./epik_bt-mz_B_4x4_trace/trace.cube...
```

[GUI showing trace analysis report]

- Scalasca analysis reports can be viewed with *ParaProf* for a multitude of interactive 2D & 3D graphical profiles

```
% paraprof epik_bt-mz_B_4x4_trace_PAT_RT_HWPC_0/trace.cube.gz
```

- Scalasca traces can be viewed directly with *Vampir7* for interactive timeline and communication matrix displays

```
% vampir epik_bt-mz_B_4x4_trace_PAT_RT_HWPC_0/epik.esd
```

- Scalasca traces can also be merged and then converted to the formats of other analysis and visualization tools

```
% elg_merge epik_bt-mz_B_4x4_trace_PAT_RT_HWPC_0  
% elg2prv epik_bt-mz_B_4x4_trace_PAT_RT_HWPC_0  
% wxparaver epik_bt-mz_B_4x4_trace_PAT_RT_HWPC_0/epik.prv
```

- Trace merging and conversion are both done serially and therefore only practical for relatively small traces.
- External tools can often manage to analyze traces that Scalasca's automatic trace analyzer can't handle

TAU: ParaProf Manager

File Options Help

- Applications
 - Standard Applications
 - Default App
 - Default Exp
 - epik_bt-mz_B_4x4_trace_PAT_RT_HWPC_0/trace.cube.gz

TAU: ParaProf: Function Data Window: epik_bt-mz_B_4x4_trace

File Options Windows Help

Name: main => MAIN_ => adi_ => z_solve_ => !\$omp parallel @z_solve.f:43 => !\$omp do @z_solve.f:52
 Metric Name: Time
 Value: Exclusive
 Units: seconds

9.609	node 1, thread 2
9.547	node 1, thread 0
9.54	node 1, thread 1
9.118	node 3, thread 0
9.118	node 3, thread 2
9.104	node 3, thread 1
9.057	node 2, thread 1
9.037	node 2, thread 2
9.025	node 2, thread 0
9.019	node 0, thread 1
8.995	node 0, thread 0
8.977	node 0, thread 2
8.636	mean
7.477	node 1, thread 3
6.911	node 2, thread 3
6.851	node 3, thread 3
6.788	node 0, thread 3
0.971	std. dev.

TAU: ParaProf: epik_bt-mz_B_4x4_trace_PAT_RT_HWPC_0/trace.cube.gz

File Options Windows Help

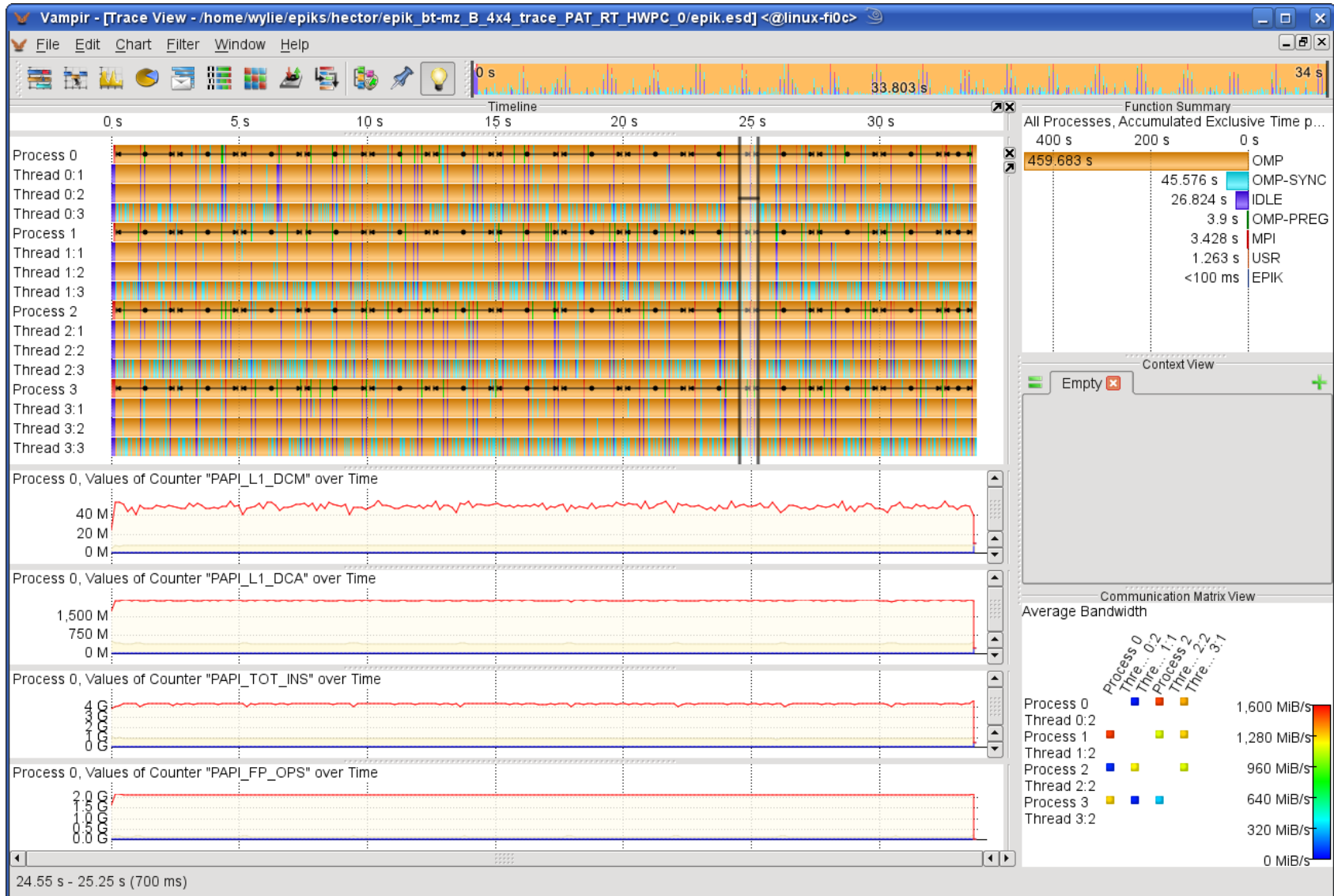
Metric: Time
Value: Exclusive

Thread	Std. Dev.	Mean
node 0, thread 0		
node 0, thread 1		
node 0, thread 2		
node 0, thread 3		
node 1, thread 0		
node 1, thread 1		
node 1, thread 2		
node 1, thread 3		
node 2, thread 0		
node 2, thread 1		
node 2, thread 2		
node 2, thread 3		
node 3, thread 0		
node 3, thread 1		
node 3, thread 2		
node 3, thread 3		

main => MAIN_ => adi_ => z_solve_ => !\$omp parallel @z_solve.f:43 => !\$omp do @z_solve.f:52
 Exclusive Time: 9.118 seconds
 Inclusive Time: 9.118 seconds
 Calls: 3216.0

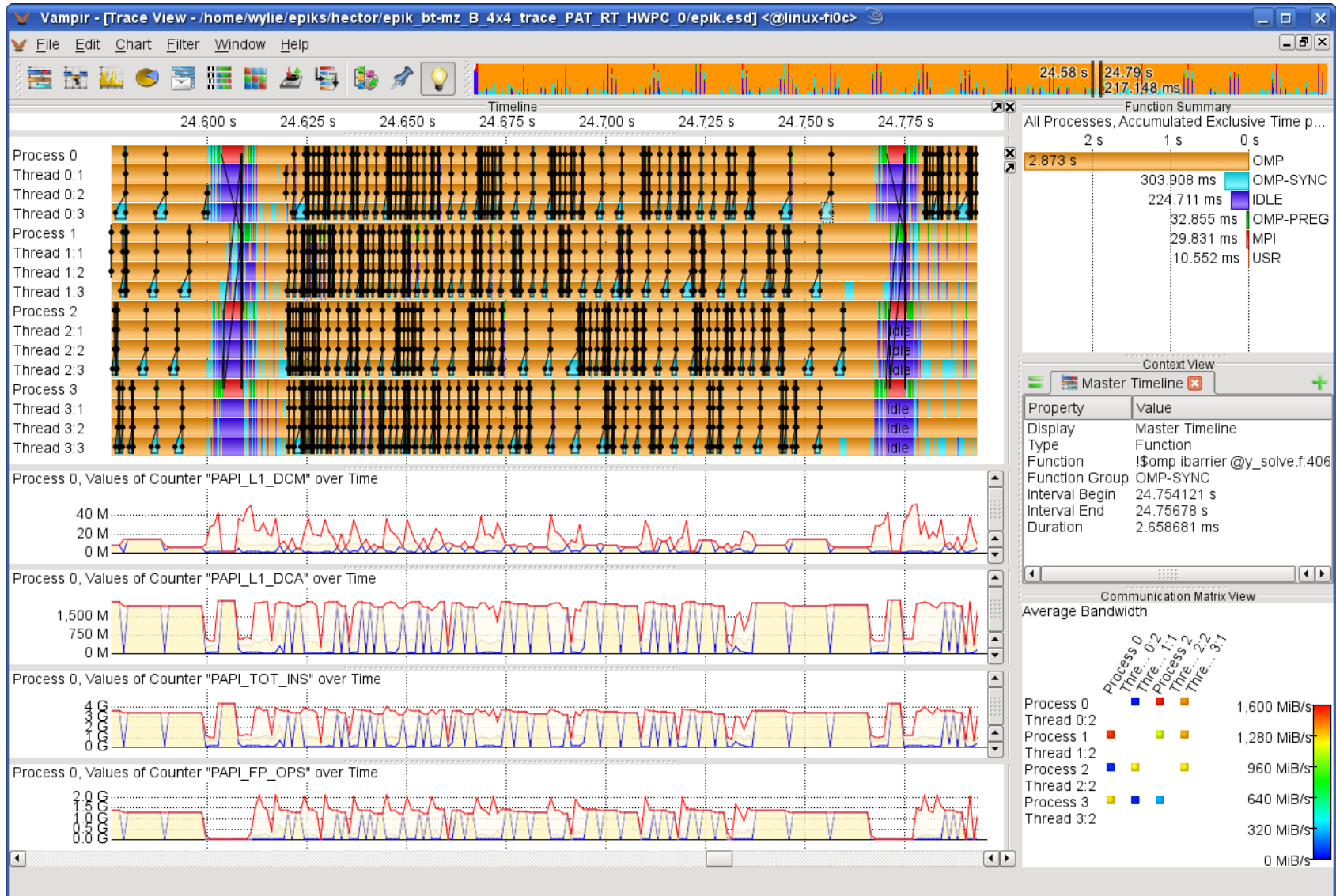
% paraprof epik_bt-mz_B_4x4_trace_PAT_RT_HWPC_0/trace.cube.gz

Vampir visual trace exploration (overview)

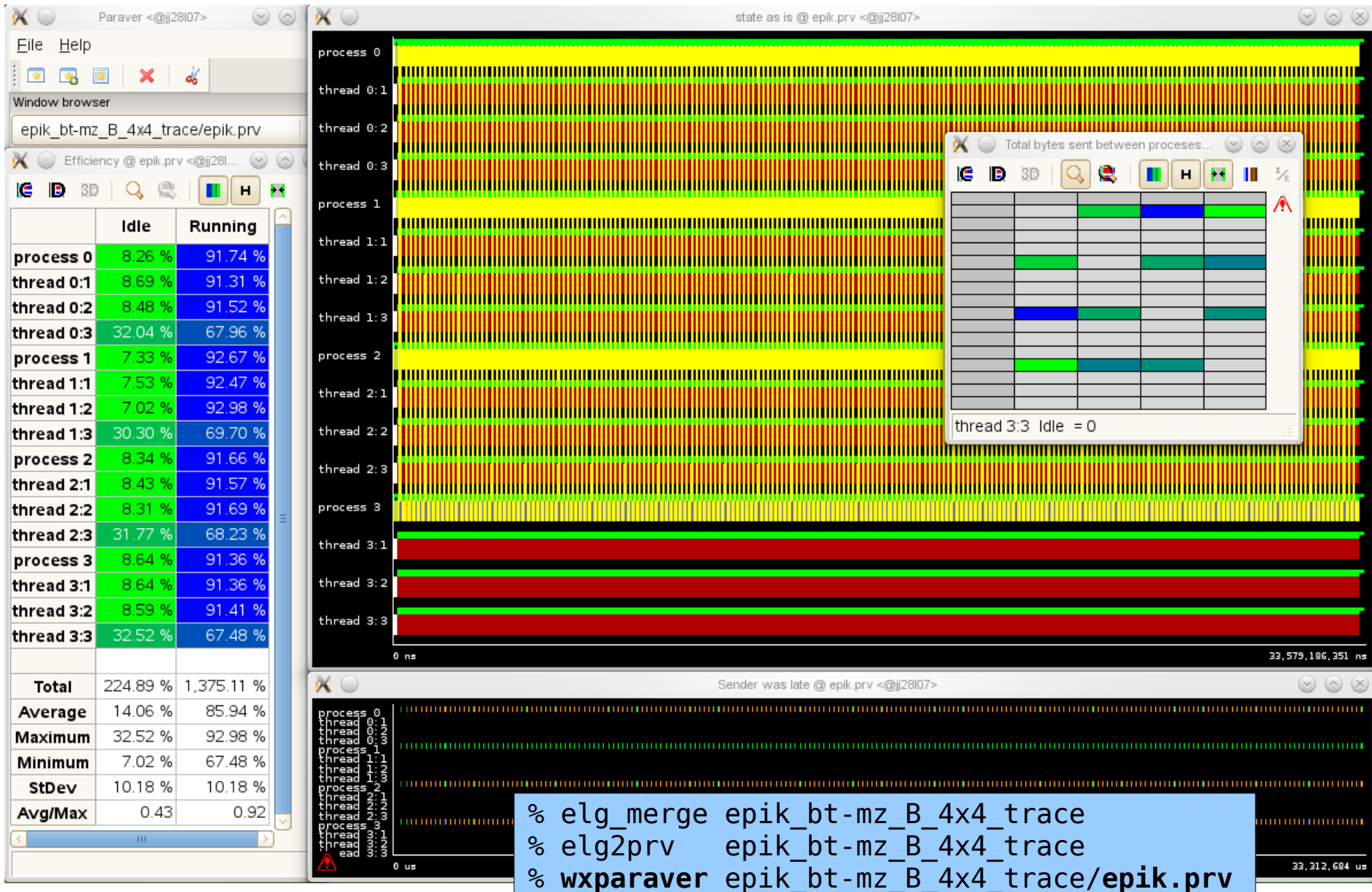


% vampir7 epik_bt-mz_B_4x4_trace_PAT_RT_HWPC_0/epik.esd

Vampir visual trace exploration (zoom)



Trace conversion & analysis with Paraver



- Traces can easily become extremely large and unwieldy
 - size is proportional to number of processes/threads (*width*), duration (*length*) and detail (*depth*) of measurement
- Traces containing intermediate flushes are of little value
 - uncoordinated flushes result in cascades of distortion
 - ▶ reduce size of trace such that it fits in available buffer space
- Traces should generally be written to a parallel filesystem
 - /work or /scratch are typically provided for this purpose
- Moving large traces between filesystems often impractical
 - however, systems with more memory can analyse larger traces
 - ▶ alternatively, run trace analyzer with undersubscribed nodes
- Traces can be archived or deleted after analysis completed to recover storage space
 - Scalasca binary trace data is stored in the ELG subdirectory

- Consult quick reference guide for further information

```
% scalasca -h  
Scalasca 1.3 – quick reference guide  
pdfview /UNITE/packages/scalasca/1.3/doc/manuals/QuickReference.pdf
```

[PDF viewer showing quick reference guide]

- CUBE GUI provides context sensitive help and on-line metric descriptions (including problem diagnosis hints)
- EPIK archive directories contain analysis report(s), measurement collection & analysis logs, etc.
- Instrumentation, measurement, analysis & presentation can all be extensively customized
 - covered in more detailed presentation
- Visit www.scalasca.org or mail scalasca@fz-juelich.de

0. Reference preparation for validation
 1. Program instrumentation: `skin`
 2. Summary measurement collection & analysis: `scan [-s]`
 3. Summary analysis report examination: `square`
 4. Summary experiment scoring: `square -s`
 5. Event trace collection & analysis: `scan -t`
 6. Event trace analysis report examination: `square`
- General usage/help: `scalasca [-h]`
 - Instrumentation, measurement, analysis & presentation can all be extensively customized
 - covered in more detailed presentation
 - Visit www.scalasca.org or mail scalasca@fz-juelich.de

- Prepares application objects & executables for measurement
 - **skin = scalasca -instrument**
 - **skin [options] <compile-or-link-command>**
 - ▶ defaults to automatic instrumentation of USR routines by compiler
 - available for most compilers, but not all
 - when not desired, disable with -comp=none
 - ▶ for OpenMP, includes source-level pre-processing of directives
 - ▶ for MPI, links wrappers to PMPI library routines
 - [-pdt] pre-process sources with PDTToolkit (when available)
 - ▶ configurable instrumentation of specified routines (or all by default)
 - Manual instrumentation activation
 - ▶ offers complementary program structure information for analyses via user-provided annotations (e.g., phases, loops, ...)
 - ▶ [-user] enable EPIK user instrumentation API macros
 - ▶ [-pomp] enable processing of POMP region pragmas/directives³⁴

- Automatic source instrumentation using PDTToolkit [-pdt]
 - only available if configured when Scalasca installed
- By default, instruments all routines in source file
 - source routines are automatically instrumented by compiler, therefore use **-comp=none** to avoid duplicate instrumentation
- Selective instrumentation of specified routines
 - -optTauSelectFile=<filename>
 - TAU-format plain text specification file
 - ▶ list names of source files and routines to include/exclude from instrumentation, using wildcard patterns
 - unsupported TAU instrumentation features are silently ignored
 - ▶ refer to TAU/PDToolkit documentation for details
 - ▶ refer to Scalasca User Guide for known limitations

- List routines with their PDT names one per line

```
% cat config/inst.pdt
# instrumentation specification for PDT
BEGIN_EXCLUDE_LIST
  BINVCRHS
  MATVEC_SUB
  MATMUL_SUB
  BINVRHS
  EXACT_SOLUTION
  TIMER_#
END_EXCLUDE_LIST
```

- ... and specify file when instrumenting

```
% make bt CLASS=W NPROCS=16 PREP="scalasca -inst -comp=none -pdt \  
-optTauSelectFile=$PWD/config/inst.pdt"
```

- PDT and EPIK user instrumentation macros expand to additional statements in program source files
 - this should be unproblematic, except for fixed-format Fortran where the default line-length limit (72 characters) may be exceeded and result in curious compilation errors
 - Fortran compilers allow extended source lines via special compile flags, e.g.,
 - ▶ CCE: -N132
 - ▶ GNU: -ffixed-line-length-none
 - ▶ Intel/Pathscale: -extend-source
 - ▶ PGI: -Mextend
 - For BT example therefore need to adjust FFLAGS

```
% make bt CLASS=W NPROCS=16 PREP="scalasca -inst -comp=none -pdt" \  
FFLAGS="-03 -ffixed-line-length-none"
```

- EPIK user instrumentation API
 - #include “epik_user.h”
 - EPIK_USER_REG(epik_solve, “<<Solve>>”)
 - EPIK_USER_START(epik_solve)
 - EPIK_USER_END(epik_solve)
- Can be used to mark initialization, solver & other phases
 - Annotation macros ignored by default
 - Instrumentation enabled with “-user” flag to instrumenter
 - Also available for Fortran
 - ▶ #include “epik_user.inc” and use C preprocessor
- Appear as additional regions in analyses
 - Distinguishes performance of important phase from rest

- In NPB3.3-MPI/BT compare `bt.f` & `bt_epik.F`
 - the `.F` suffix indicates that it should be preprocessed
 - ▶ otherwise could specify some obscure compiler flags
- EPIK user API `#include`'d at the top
 - `#include "epik_user.inc"`
- EPIK user instrumentation macros register & mark phases
`"<<INIT>>"`, `"<<STEP>>"`, `"<<FINI>>"`
- within the main routine `"<<MAIN>>"`
- Edit `BT/makefile` to set: `MAIN = bt_epik.F`
- Instrument specifying `-user` and extended source lines

```
% make bt CLASS=W NPROCS=16 PREP="scalasca -inst -comp=none -user" \  
FFLAGS="-03 -ffixed-line-length-none"
```

- Runs application under control of measurement system to collect and analyze an execution experiment
 - **scan = scalasca -analyze**
 - **scan [options] <application-launch-command>**
 - ▶ e.g., scan [options] [\$MPIEXEC [mpiexec-options]] [target [args]]
 - **[-s]** collect summarization experiment [default]
 - **[-t]** collect event traces and then analyze them automatically
 - Additional options
 - ▶ **[-e title]** specify experiment archive (directory) name: epik_*title*
 - ▶ **[-f filter]** specify file listing routines to ignore during measurement
 - ▶ **[-m metric1:metric2:...]** include hardware counter metrics
 - ▶ **[-n]** preview scan and perform checks but don't execute
 - ▶ **[-q]** quiesce (disable most) measurement collection
 - ▶ **[-a]** (re-)analyze a previously collected trace experiment

- Via `./EPIK.CONF` file

```
EPK_FILTER=smg2000.filt
EPK_MPI_ENABLED=CG:COLL:ENV:IO:P2P:RMA:TOPO
ELG_BUFFER_SIZE=40000000
```

- Via environment variables

```
% export EPK_FILTER=smg2000.filt
% export EPK_MPI_ENABLED=CG:COLL:ENV:IO:P2P:RMA:TOPO
% export ELG_BUFFER_SIZE=40000000
```

- Via command-line flags (partially)

```
% scalasca -analyze -f smg2000.filt ...
```

- To show current/default configuration

```
% epik_conf
```

- Actual Scalasca measurement configuration saved in experiment archive as `epik.conf`

- Generally, the SCAN nexus will correctly parse execution command lines, but occasionally you may need to help it
- MPI launcher arguments may need to be explicitly separated from the target application with a double-dash

```
% scalasca -analyze mpirun -np 16 -- a.exe arg
```

- Unusual MPI launcher options may need to be quoted

```
% scalasca -analyze mpirun -np 16 "-verbose 2" a.exe arg
```

- (On most systems `-verbose` doesn't take an argument)

- Explicitly specify the instrumented target executable name when using imposition commands/scripts

```
% export SCAN_TARGET=a.exe  
% scalasca -analyze imposter.sh i.arg a.exe arg  
% scan -t mpirun -np 16 imposter.sh i.arg a.exe arg
```

- (*dplace*, *omplace* and *numactl* are common imposters)

- Prepares and presents measurement analysis report(s) for scoring and/or interactive exploration
 - **square = scalasca -examine**
 - **square [options] <experiment-archive|report>**
 - ▶ e.g., square epik_*title*
 - Post-processes intermediate measurement analysis reports
 - Launches GUI and presents default analysis report (if multiple reports available)
 - ▶ trace analysis given precedence over summary analysis
 - ▶ select other reports via File/Open menu
 - [-s] skip display and produce textual score report (epik.score)
 - ▶ estimates total trace size and maximum rank trace size
 - ▶ breakdown of USR vs. MPI/OMP vs. COM region requirements
 - ▶ add [-f test.filt] to test effect of a prospective filter file

- Extracting a sub-tree from an analysis report

```
% cube3_cut -r '<<SMG.Solve>>' epik_smg2000_12_trace/trace.cube  
Writing cut.cube... done.
```

- Calculating difference of two analysis reports

```
% cube3_diff epik_bt_9_trace/trace.cube epik_bt_16_trace/trace.cube  
Writing diff.cube... done.
```

- Combining two or more related analysis reports

```
% cube3_merge trace/trace.cube HWC1/summary.cube HWC2/summary.cube  
Writing merge.cube... done.
```

- Additional algebra utilities for calculating mean, etc.
 - Default output of `cube3_utility` is a new report `utility.cube`
- Further utilities for report scoring & statistics
- Run utility with “-h” (or no arguments) for brief usage info

- Example set of experiments collected with and w/o HWC

```
% ls -ld epik_*  
epik_bt_B_16_sum_PAT_RT_HWPC_0/  
epik_bt_B_16_sum_PAT_RT_HWPC_1/  
epik_bt_B_16_sum_PAT_RT_HWPC_7/  
epik_bt_B_16_sum_PAT_RT_HWPC_8/  
epik_bt_B_16_trace/
```

- Ensure that each is post-processed

```
% for epik in epik_* ; do scalasca -examine -s $epik ; done
```

- Merge the HWC summary reports into the non-HWC report

```
% cube3_merge -o HWC_combo.cube \  
epik_bt_B_16_trace/trace.cube epik_bt_B_16_sum_*/summary.cube  
Writing HWC_combo.cube... done.
```

- Metrics are merged as they are encountered in reports
 - already defined metrics are not modified by later versions
- Since measurements with HWCs have higher overhead, include a non-HWC measurement first

DON'T PANIC!

- Remember the Scalasca User Guide is your friend
- And if you need more advice, <mailto:scalasca@fz-juelich.de>