


Paraver internals and details

Jesus Labarta
Jesus.Labarta@bsc.es

Index

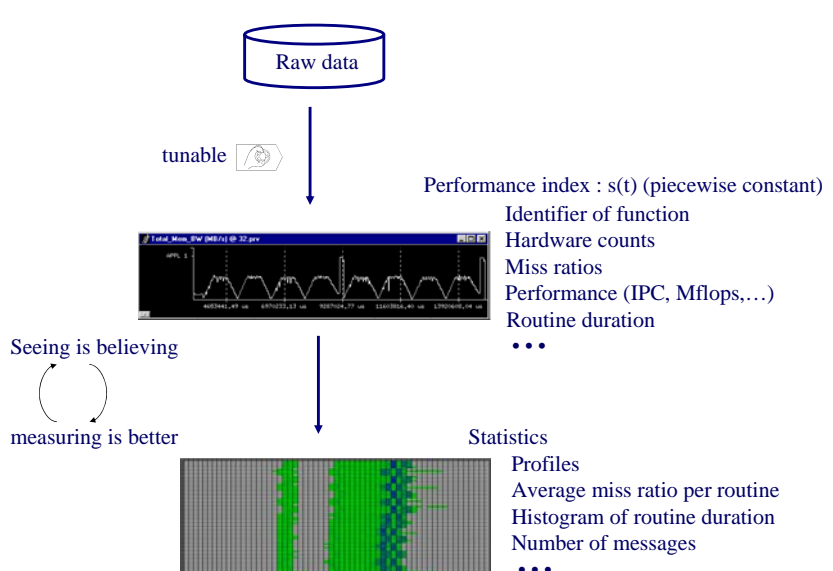
- Overview
- Semantic module
- Analysis module

overview




3

Paraver: Performance Data browser



Raw data

tunable 

Performance index : $s(t)$ (piecewise constant)


- Identifier of function
- Hardware counts
- Miss ratios
- Performance (IPC, Mflops,...)
- Routine duration
- ...

Seeing is believing

measuring is better

Statistics

- Profiles
- Average miss ratio per routine
- Histogram of routine duration
- Number of messages
- ...



4

Performance views

- Few types of views
 - Timelines
 - Textual
 - Statistics
 - Profiles
 - Histograms
- Configuration files
 - Capture knowledge on how to generate view from raw data
 - Loading configuration file pops up display or analysis window
 - Distribution
 - Set of basic view and stats provided with the tool.
 - Also useful for
 - Checkpointing of studies
 - Cooperative work
 - Bug reporting
 - Recording your own favorite views
 - Get used with the tool and to learn on the internals

5



Configuration files

The screenshot shows the Paraver application window with the 'Load windows' dialog box open. The dialog box has a 'TRACEFILE:' field containing 'bt.9.mpi.prv' and a 'CURRENT DIRECTORY:' field containing '/home/jesus/tutorials/IBM/cfgs'. Below these fields is a list of configuration files, with 'in MPI call' selected. To the right of the list is a 'DESCRIPTION' field containing text about the default call being 'Send'. The dialog box has 'Load' and 'Ok' buttons. Annotations with blue arrows point to various parts of the dialog box:

- 'Select trace to which to apply' points to the 'TRACEFILE:' field.
- 'Select directory' points to the 'CURRENT DIRECTORY:' field.
- 'List of configuration files in current directory' points to the list of files.
- 'Navigate through directory tree' points to the left and right arrow buttons below the list.
- 'Description of view of select file' points to the 'DESCRIPTION' field.

6



Timelines

- Each window displays one view
 - **Piecewise constant** function of time
 - One such function of time per object:
 - Thread, process, application, workload, CPU, node

$$S(t) = S_i, i \in [t_i, t_{i+1})$$

- Types of functions
 - Categorical
 - State, user function, outlined routine

$$S_i \in [0, n] \subset \mathbb{N}, \quad n <$$

- Logical
 - In specific user function, In MPI call, In long MPI call

$$S_i \in \{0, 1\}$$

- Numerical
 - IPC, L2 miss ratio, Duration of MPI call, duration of computation burst

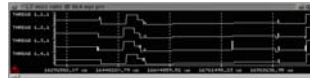
$$S_i \in \mathbb{R}$$

7

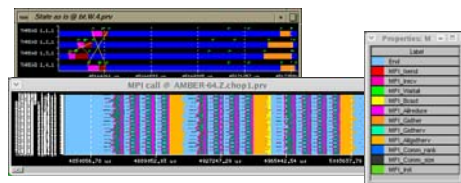


Timelines

- Representation
 - Function of time

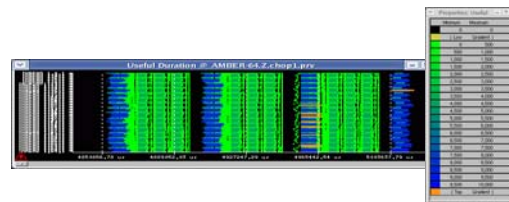


- Color encoding



- Gradient color
 - Light green → Dark blue

- Not null gradient
 - Black for zero value
 - Light green → Dark blue



8



Navigation

Annotations and actions shown in the screenshot:

- Hide lower panel
- Right click
- Zoom: select area with:
 - left click – middle click for zoom within same window
 - 2 left clicks for cloning zoomed window
- select area with 2 left clicks (within same or between different windows)
- Display animation
- Display as function or color encoded
- Draw flag icon for events
- Draw icon at receive points
- Draw icon at send points
- Draw lines for pt2pt communication
- Copy features from one window to another
 - Window size
 - Synchronize region of trace analyzed
 - Set same color gradient
 - Filter settings

Context menu items visible:

- Window Options
 - Undo Zoom
 - Redo Zoom
 - Redraw
 - Done
 - Zoom
 - Zoom XY
 - Timing
 - Copy
 - Paste X scale
 - Paste Y scale
 - Paste Filter
 - Paste window size
 - Scale
 - Color Type
 - Scroll Bar
 - Draw Mode
 - EXPORT Menu
 - Save As
 - Show Properties
- Paste All
- Paste Content's
- Paste Events
- Code Color
- Gradient Color
- Next Nat Gradient Color
- List
 - Maximum
 - Min = 0
 - Random
 - Average
- X axis
 - flow axis
 - Both axis
- Fit Max Y-Scale
- Fit Min Y-Scale
- Fit Both Y-Scale

Textual

- Appears when clicking on a timeline
- Textual display of
 - Semantic value
 - Events
 - Communications

Top dialog box (Object: THREAD 1.1.1, Click Time: 235381, Time Units: Milliseconds (ms)):

```

Running Duration: 0,13
User Event at 23496,42 TYPE 50003006 VALUE 1
Idle Duration: 0,01
User Event at 23496,42 TYPE 50003006 VALUE 0
Running Duration: 0,36
    
```

Bottom dialog box (TRACEFILE: NAS_BT_primary_misses.prv, Window Name: Global view, Object: THREAD 1.1.1, Click Time: 60870930, Time Units: Microseconds (us)):

```

Sched. and Fork/Join Duration: 15344
User Event at 60525522 Join (OMP) End
Sched. and Fork/Join Duration: 5
User Event at 60525527 Parallel function VALUE 0
User Event at 60525527 Parallel (OMP) End
Running Duration: 9
User Event at 60525536 Parallel function __mpd_compute_rts_2
User Event at 60525536 Parallel (OMP) Do/Sections
Sched. and Fork/Join Duration: 473
User Event at 60526009 Primary instruction cache misses VALUE 77
    
```

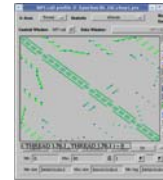
Bottom dialog box controls: Semantic Events Communication 5 All the burst Text Mode Repeat Save as Text

Views: Statistics

- 2D
 - Profiles
 - Histograms
 - Correlations
 - Communication Patterns



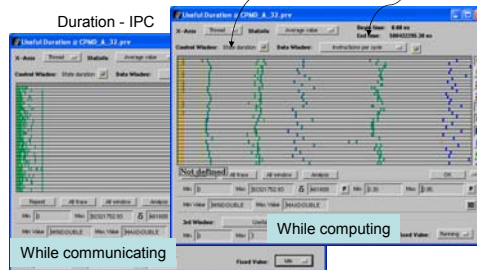
MPI calls profile



Control window

Data window

- 3D

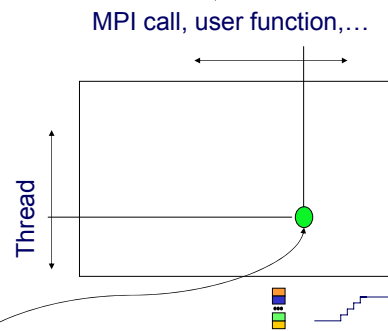


11



How to read profiles

Columns determined by categorical **Control window**



Value is a statistic (i.e. #occurrences, avg,) computed on **Data window**

12



How to read histograms

Columns correspond to bins of values of a numeric **Control window**

duration, instructions, BW, IPC,
...

Thread

Value is a statistic (i.e. #occurrences, avg,) computed on **Data window**



2D analysis module

Region analyzed

Display whole table / cell text

Translate (.pcf)

Transpose

Color/not cells

Hide null columns

Show/hide lower panel

Fit color encoding

Min and max to dynamic range of statistic

Activate 3D analysis

Bin definition

- Fit bin size
To cover the whole control window dynamic range and generate at most 20 columns
- Color encoding

max

min



2D analysis module

The screenshot shows the '2D analysis module' interface. The main window displays a 2D heatmap with a color scale from blue to red. The X-axis is labeled 'Semantic' and the Y-axis is labeled '% Time'. The window title is 'MPI call profile' and the file name is 'MPI-call-632.y.chop1.prv'. The 'Control Window' is set to 'MPI call' and the 'Data Window' is set to 'MPI call'. The 'Begin time' is 7980429.98 ns and the 'End time' is 8338274.75 ns. The interface includes a toolbar with icons for 'Repeat', 'All trace', 'All window', 'Analyze', and 'OK'. Below the toolbar are input fields for 'Min Value' (MINDOUBLE) and 'Max Value' (MAXDOUBLE). A context menu is open over the heatmap, showing options: 'Copy table', 'Paste scale', 'Copy gradient scale', 'Paste gradient scale', 'Copy trace', 'Paste order', 'Paste Objects', 'New Analyzer', 'Zoom ...', 'Refresh', 'Calculate All', 'Sort by ...', 'n, Ctrl+n', 'Save as text', 'Save to CFG', and 'Show Properties'. Annotations with arrows point to various parts of the interface:

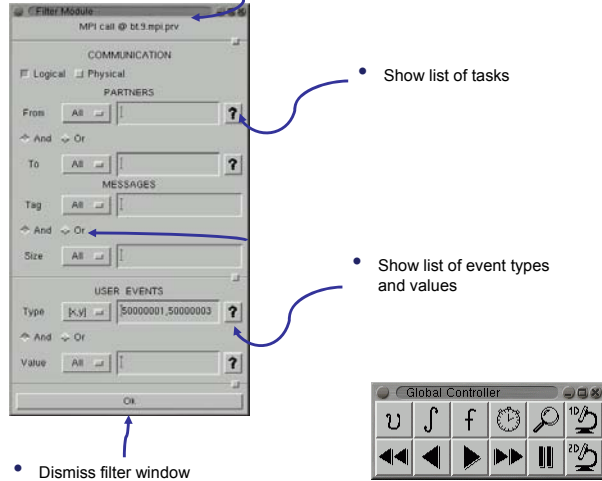
- 'Open Control window' points to the 'Control Window' dropdown.
- 'Open Data window' points to the 'Data Window' dropdown.
- 'Display whole table colored / text for each cell' points to the heatmap.
- 'Generate a timeline, derived from control window by zeroing values outside range selected by clicking in the table' points to the heatmap.
- 'Copy/paste analyzed time interval to from other 2D/timeline' points to the 'Copy table' menu item.
- 'Copy/paste color scale between 2Ds' points to the 'Paste scale' menu item.
- 'Generate ASCII file with table data' points to the 'Save as text' menu item.
- 'Save configuration file' points to the 'Save to CFG' menu item.
- 'Right click' points to the context menu.
- 'Text for cursor' points to the 'Repeat' button.
- 'Repeat the analysis (after changing statistics or data window)' points to the 'Repeat' button.
- 'Repeat the analysis for the whole trace' points to the 'All trace' button.
- 'Perform a new analysis (selecting new control window and interval)' points to the 'Analyze' button.

15

Semantic module internals

Filter module

- Communications that pass through the filter
- events that pass through the filter
- Display window to which applies
- Show list of tasks
- Show list of event types and values
- Dismiss filter window



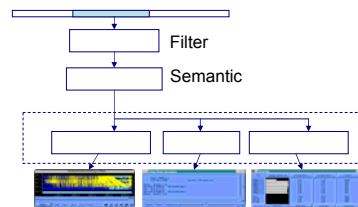
17

Basic functions of time

- The filter module presents a subset of the trace to the semantic module. Each thread is described by
 - A sequence of events $Ev_i, i \in N$, states $St_i, i \in N$ and communications $C_i, i \in N$
 - For each event let $T(Ev_i)$ be its time and $V(Ev_i)$ its value
 - For each state let $T_s(St_i)$ be its start time $T_e(St_i)$ its stop time and $V(St_i)$ its value
 - For each Communication let $T(C_i)$ be its time, $Sz(C_i)$ its size and $Dir(C_i) \in \{send, recv\}$
- Semantic module builds

$$s(t) = S(i), t \in [t_i, t_{i+1}), i \in N$$

Function of time Series of values



18

From events to functions

- Different possibilities

- Last event value $S(i) = V(EV_i)$

- Next event value $S(i) = V(EV_{i+1})$

- Average Next Event Value $S(i) = \frac{V(EV_{i+1})}{T(EV_{i+1}) - T(EV_i)}$

- Interval btw. Events $S(i) = T(EV_{i+1}) - T(EV_i)$

- Incoming bytes $S_i = S_{i-1} + InComm_{i,size}$

- Bytes btw. events

- Bytes sent/received

$$S(i) = \sum_j Sz(C_j), j | T(C_j) \in [T(EV_i), T(EV_{i+1}))$$

19



The power of maths: Composition

- $S'(t) = f(S(t))$

$$S' = f \circ S$$

- Sign $S'(t) = \text{sign}(S(t))$

- 1-sign $S'(t) = 1 - \text{sign}(S(t))$

- Select range $S'(t) = S(t) \in [a, b] ? S(t) : 0$

- Sign ° Is equal $S'(t) = \text{sign}(S(t) = a ? S(t) : 0)$

- Delta $S'(t) = S_{i+1} - S_i$

- Stacked value

20



Semantic module

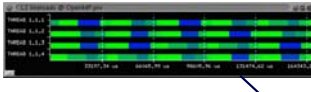
- Derived windows

- Point wise operation

$$S = \alpha * S^a <op> \beta * S^b$$

- $<op>$: +, -, *, /, ...

L2 Line Loads

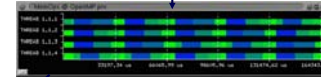


Loads

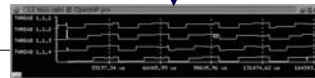
Stores



Mem Ops



x100



L2 miss ratio

21



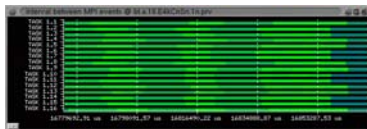
Semantic module

- Derived windows

- Point wise operation

$$S = \alpha * S^a <op> \beta * S^b$$

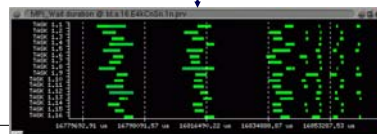
- $<op>$: +, -, *, /, ...



Interval between MPI events



In MPI call



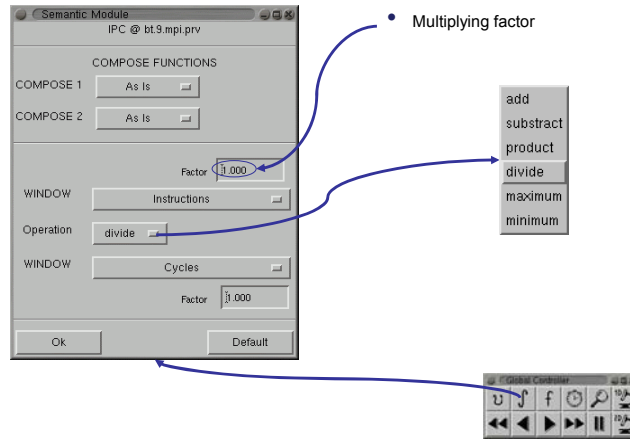
MPI call duration

22



Semantic module: derived windows

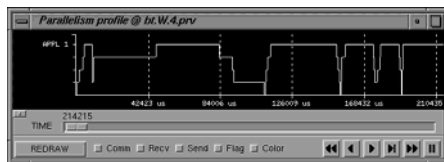
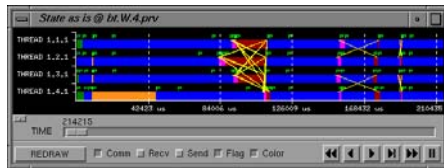
- How to build expression



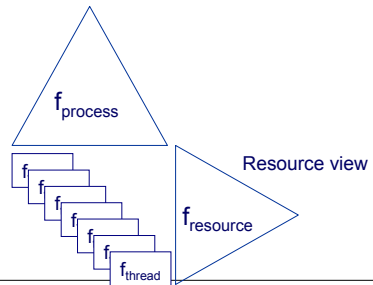
23

Semantic module

- Angle:
 - Process model
 - Thread, task, application, workload
 - Resource model
 - CPU, node, system



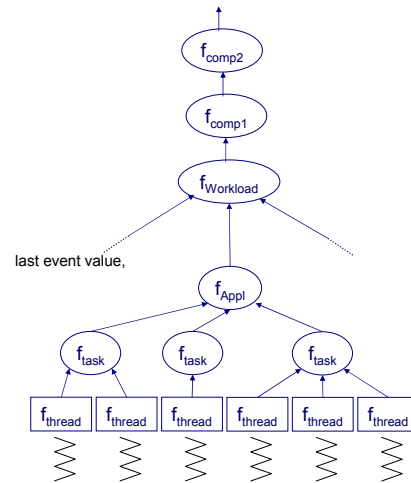
Process view



24

Semantic module: process model view

- Semantic value: $S(t)$
- $S = f_{comp2} \circ f_{comp1} \circ f_{Workload} \circ f_{Application} \circ f_{task} \circ S_{thread}$
- Semantic functions
 - f_{comp2}, f_{comp1} : sign, mod, div, in range, select range
 - $f_{Application}, f_{Workload}$: add, average, max, select
 - f_{task} : add, average, max, select
 - S_{thread} : in state, useful, given state,
- next event value,
- average next event value
- interval between events, ...

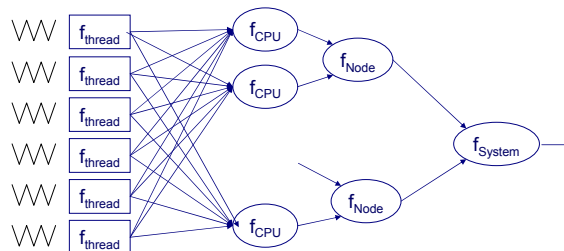


25



Semantic module: resource view

- $Sf_{resource} = f_{comp2} \circ f_{comp1} \circ f_{System} \circ f_{Node} \circ f_{CPU} \circ S_{thread}$
- Semantic functions
 - f_{System} : add, average, max, select
 - f_{Node} : add, average, max, select
 - f_{CPU} : active thread, select
 - S_{thread} : in state, useful, given state, next event value, thread_id



26



Semantic module: process model view

The screenshot displays the 'Semantic Module' interface with the following components:

- Left Panel (List of Properties):**
 - State
 - Useful
 - State Sign
 - State As Is
 - Given State
 - In State
 - Not In State
 - Event
 - Last Evt Type
 - Last Evt Val
 - Next Evt Type
 - Next Evt Val
 - Avg Next Evt Val
 - Avg Last Evt Val
 - Given Evt Val
 - In Evt Val
 - Int. Between Evt
 - Not In Evt Val
 - In Evt Range
 - Comm.
 - Last Tag
 - Comm Size
 - Comm Partner
 - Last Send Duration
 - Next Recv Duration
 - Object
 - Application ID
 - Task ID
 - Thread ID
 - Cpu ID
 - Node ID
 - In Thread ID
 - In Task ID
 - In Appl ID
- Top Center Panel (As Is List):**
 - Sign
 - 1-Sign
 - Mod+1
 - Mod
 - Div
 - Prod
 - Subs
 - Select Range
 - Is In Range
 - Is Equal
 - Is Equal (Sign)
 - Stacked Val
 - In Stacked Val
 - Nesting level
- Right Panel (Adding List):**
 - Adding
 - Adding Sign
 - Average
 - Maximum
 - Minimum
 - Thread i
 - Activity
 - In Activity
- Central Dialog Box (COMPOSE FUNCTIONS):**
 - COMPOSE 1: As Is
 - COMPOSE 2: 1-Sign
 - PROCESS MODEL
 - WORKLOAD: Adding
 - APPL: Adding
 - TASK: Thread i
 - THREAD: Last Evt Val
- Bottom Right Panel (Global Controller):** A set of navigation and control icons.

27



2D/3D internals

28



3D/2D analysis module

- Single flexible quantitative analysis mechanism
- Let
 - cw_1 and cw_2 two views we will call control views
 - dw a view we will call data window

For each window w

$$S_{th}^w(t) = S_{th}^w(i), t \in [t_i^w, t_{i+1}^w)$$

- For each control window we define a set of bins

$$bin_j^{cw} = [range_j^{cw}, range_{j+1}^{cw}) \quad range_{j+1}^{cw} = range_j^{cw} + delta^{cw}$$

- And the discriminator functions

$$\delta_j^{cw}(t) = ((S^{cw}(t) \in bin_j^{cw}) ? 1 : 0)$$

$$\delta_{j,k}(t) = \delta_j^{cw}(t) * \delta_k^{cw}(t)$$

Identify regions with cw's within the (j,k) bin

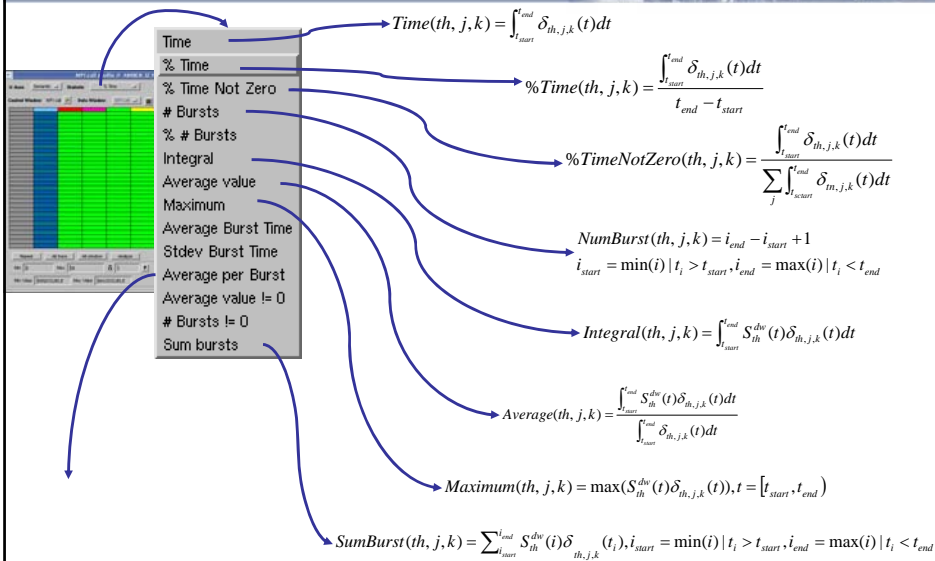
- The 3D analysis module computes a cube (or plane in the case of 2D) of statistics

$$M(thread, j, k) = statistic(S_{th}^{dw}(t) * \delta_{th,j,k}(t))$$

- Where the statistic can represent the average value, the number of intervals,....

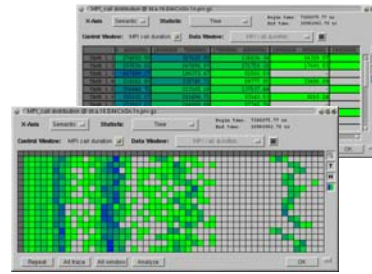
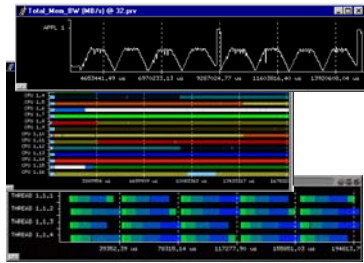


2D analysis module



Representation module

Functions of time



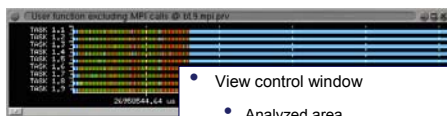
Seeing is believing

measuring is better

31

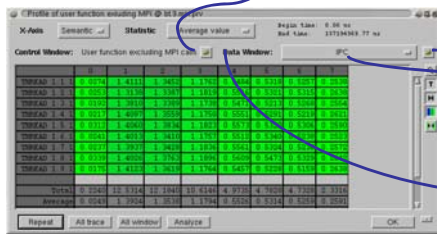


2D analysis



View control window
Analyzed area

View data window



Select data window

- Time
- % Time
- % Time Not Zero
- # Bursts
- % # Bursts
- Integral
- Average value
- Maximum
- Average Burst Time
- Stdev Burst Time
- Average per Burst
- Average value != 0

- win_1
- User_function
- Outside MPI call
- User function excluding MPI calls
- L2_lineloads
- Loads
- Stores
- MemOps
- L2_miss_ratio
- Instructions
- Cycles
- IPC
- Stores_per_ms
- Loads_per_ms
- DataBW (MB/s)
- L2_lineloads_per_ms
- L2_writebacks_per_ms
- L2_Mem_BW (MB/s)
- CPU/Memory BW ratio

- Perform an analysis
 - On same area
 - On whole trace (same CW)
 - On whole CW
 - Selecting new CW

Select statistic

32



2D analysis module

- Region analyzed
- Whole table / cell text
- Translate (.pcf)
- Transpose
- Color/not cells
- Hide null columns
- Show/hide lower panel
- Fit color encoding
Min and max to dynamic range of statistic
- Color encoding
 - max
 - min
- Fit bin size
To cover the whole control window
dynamic range and generate at most
20 columns
- Bin definition
- Text for cursor