Introduction to Analysis with Paraver

MPI

Timelines: Navigation and Basic concepts

Let us go through some basic navigation and analysis functionalities.

- *Load trace*: From the main menu, select *File → Load Trace...*, and select the file Iberia-128-CA.chop1.1it.shifted.prv.
    - After the file is loaded, from the main menu, select *File → Load Configuration...*, and select mpi_call.cfg. A display window will appear with the timeline of which MPI call is being executed at each point in time by each process. The horizontal axis represents time, from the start time of the application at the left of the window to the end time at the right. For every thread, the colors represent the MPI call or black when doing user level computation outside of MPI.
    - Moving the mouse over the display window you will see at the bottom of the window the name of the MPI call (the label correspoding to the pointed color)
    - *Info Panel*: Double clicking anywhere inside the window will open the *Info Panel* with the textual information of the function value in the selected point. Right click inside the window and select the option *Info Panel* to hide it.
    - *Zoom*: Click with the Left Button of the mouse to select the starting time of the zoomed view, drag the mouse over the area of interest, and release the mouse to select the end time of the zoomed view.
    - *Undo Zoom* and *Redo Zoom* commands are available on the Right Button menu. You can do and undo several levels of zooming.
    - The *Control-Zoom* option will let you select a subset of threads. This is useful when analyzing runs with many processes and you want to concentrate on a few of them. Hold down the "*Control*" or "*CTRL*" key on the keyboard, and using the mouse, identify a rectangular area by clicking on top left corner of the desired area with the left mouse button and dragging and releasing the button on the bottom right corner.
    - *Flags*: Right-click on the window, and select the *View → Event Flags* checkbox. Alternatively, you can toggle the flag button on the *View* tab on the *Info Panel*. Flags appear at the entry and exit points to the MPI calls. Depending on the scale, displaying flags may help differentiate whether there is one or many MPI calls at a given zoom level.
    - To measure time between any two points in the trace: Use the Shift-Zoom combination to activate the timing. The time and the interval between the two selected points of the trace is displayed in the *Timing* tab of the *Info Panel*.
- Load configuration file useful_IPC.cfg**.** From the main menu, select *File → Load Configuration...*, and select the configuration file specified.
    - This configuration file shows a timeline with the instructions per cycle (IPC) achieved by in each interval of useful computation. Right-click on the window, select *Info Panel*, and select the *Color* tab to see the actual coloring scheme and scale for this window. The IPC function of time for each process is represented as a gradient between light green representing a low IPC value and dark blue representing a high IPC value. This view shows in black in the regions where processes are in MPI in order to let us focus on the actual useful computation parts.

  - Display as a function of time: Use the *Control-Zoom* to select a few processes for the whole duration of the trace. Select the *View* option in the window menu (right click) and un-select the *Function Line With Color* checkbox. You will see a display of the IPC for the selected processes as a function of time. Both the color and function of time view present the same information. The color scheme is inherently more scalable.
  - Textual Display: Click with the Left Button on any point in the window. It will list in textual form the actual value of IPC at the point selected (an vicinity if time scale too coarse). The text display will be in the *What/Where* tab of the *Info Panel*.
  - Y scale: you can change the vertical scale for the function of time representation as well as for the color encoding. To manually control the scale, use the *Semantic Maximum* and *Semantic Minimum* fields in the Main Window. Just keep in mind that values outside the specified range will be truncated in the function of time display and will be assigned a different color (orange) in the color display. To automatically fit the scale to the whole dynamic range of the function, inside the window, right-click on the window, and select *Fit Semantic Scale → Fit Both* (or *Fit Maximum* or *Fit Minimum*).
  - Synchronize windows: In Paraver every timeline window represents a single metric (MPI call, useful IPC,....) for all selected processes and time span. It is possible to synchronize two timelines by making them display the exact same processes and time span. For doing so, just right-click and select *Copy* on the source (reference) window and then on the target window right-click and select *Paste → Size* and *Paste → Time*. Both windows will then be of the same size and represent different views (metrics) for the same part of the trace. If you put one above the other there is a one to one correspondence between points in vertical. The *Paste Time* lets you copy only the time scale from the first window to the target one. The *Paste Default Special* is a shortcut for copying size, time and objects (threads). The *Paste Semantic Scale* lets you apply to the target window the same vertical scale the source window had.
- Load the configuration L2missratio.cfg**.** This configuration file shows a timeline with the L2 cache miss ratio (L2 cache misses per 1000 instructions) in each interval between MPI events. You can observe that the L2 cache miss ratio is higher in the communication phases than in the long computation phases. The red triangle in the lower left corner warns that the Y scale is not enough to display the whole dynamic range of L2 cache miss ratios. This reflects in some areas appearing as orange (above the Semantic Minimum value). You can automatically set the Semantic Minimum and Maximum values to have a linear gradient display for the whole range by right-clicking and selecting *Fit Semantic Scale → Fit Both*. In this case the original semantic scale does show the difference between the computation phases. You can go back by manually changing the *Semantic Maximum* value in the Main Window to 0.5.
- Load useful_duration.cfg**.** This configuration file shows a timeline where the color represents the duration of a computation burst between an exit form MPI and the next entry. The function is valued to 0 (black) in the regions where processes are in MPI. This view gives a good perception of where are the major computation phases, and their balance across processors.
- Load p2p_size.cfg to see the message sizes for each message sent along the time axis.

Profiles

The above analysis went directly to the detailed timeline, but a less detailed averaged statistic can often be sufficient to identify problems and gives a summarized view of the behavior of an application. Paraver provides one mechanism to obtain such profiles for the desired region of a trace. We call it the *2D Analyzer* as it is a very flexible mechanism to generate tables of summarized statistics. Let's use it:

- Load configuration file mpi_stats.cfg. A table pops up with one row per thread and one column per MPI call. The first column corresponds to the time outside MPI. Each entry in the table tells the percentage of time the corresponding thread has been inside the specific call.
- The table shows the global perception of the profile for all MPI calls and processes. Click on the magnifying glass at the top of the icons column on the right of the window to see the numerical values. You can switch between numerical and zoomed out display by clicking again on the icon.
- To see a different statistic change the *Statistic selector* in the Main Window (expand the *Statistics* section in the *Window Properties*, if necessary). Interesting options at this time may be:
  - *Time*: to get the accumulated time each process has spent in each MPI call.
  - *#Bursts*: To count the number of invocations to each call.
  - *Average Burst Time*: to compute the average duration of the call.
  - All the above statistics are computed based on a single timeline window, which we call the **C***ontrol Window* and which can be popped up by clicking on the control window icon in the top left corner of the window. In this example, you will see that it is the same MPI call view we had loaded before. The values of the control window determine to which column is a given statistic accumulated/accounted.
- The statistic inside a cell can actually be performed on a different window, that we call the **D***ata Window*. For example, if you select the *Average Value* statistic and you select as the data window the Instructions per cycle window we loaded before, the entry will report the average IPC within the specific MPI call. To change the Data Window, expand the *Data* section (if necessary) in the Main Window, and change the *Window* value by clicking on the value and selecting a window. Change it back to the MPI Call window and to the %Time statistic.
- To apply the analysis to a subset of the trace, zoom on any of the timelines to the time region you are interested on. Right-click and select *Copy* on this window and right-click and select *Paste → Time* on the table. The analysis will be repeated just for the selected time interval.
- If you want to focus only on the actual MPI call columns (discarding the first column) change the *Control: Minimum* on the Main Window. If the whole table appears as green, you can rescale the gradient coloring scheme by right-clicking in the window and selecting *Autofit Data Gradient*.
- Load 2dp_MPIcallerLine.cfg. This profile shows one column per source line from where an MPI call is made. It shows the average duration of the calls from that line. You can change the statistic to *% Time* to see the percentage of time spent at each calling line or to *# Bursts* to see how many calls from that line were made.
- You can change the statistic to *Average Value* and then select the p2p size view as *Data* Window. You will see which lines sent larger messages (the data window only reports the size of point to point messages).
- You can click on the Open Control Window icon on the top left corner of the Profile Window to show a time line of from where was MPI called. Right-click and use the *Info Panel* view to see the coloring encoding (on the *Colors* tab).
- You can click on the *Open Filtered Control Window* icon from the top of the Profile Window, and then left-click and drag to select one or more columns in the 2D table. A timeline will be created showing when in time the different processes called MPI from that specific line.

Histograms

The same mechanism used to compute profiles can be used to compute histograms. To get a histogram of continuous valued metrics:

- Load configuration file 2dh_useful_duration.cfg. A table pops up with one row per thread. The X axis represents bins of a histogram. In this case, every pixel represents a bin of 2000 microseconds (*Delta* value in the *Control* section of the Main Window). The pixels at the left of the table represent short durations, those at the right represent large duration. Every pixel is colored with the percentage of time that process spent in a computation phase of the length corresponding to the pixel column. The color encoding is light green for a low percentage, dark blue for a high percentage (low and high are defined by the Minimum and Maximum values in the “Control” section of the Main Window). The pixel is colored in grey when there is no value in the corresponding range

- Ideally on a balanced SPMD application you would expect vertical lines, representing the different computation bursts are of exactly the same duration for all processes. You can see that in this example theres is a certain variance between processors (vertical bands rather than lines).
- If you move the cursor over the colored pixels, the bottom of the window will show the actual range represented by the column of the pixel and the actual value (percentage in this case) for that pixel.
- If you want to see the numerical values of the table click on the magnifying glass at the top of the column of icons at the right of the window. You will see the individual columns, for each of the the range of durations it represents is show at the top.
- It is possible to change the statistic so that the actual value represented is total time, average duration of the bursts or average value of other metric (such as “IPC”). For example, in the *Statistics* section of the Main Window, you can change the statistic selector to *Average value* and select *Instructions Per Cycle* as the data window. You can change the range of the coloring scheme in the *Statistics: Minimum Gradient* and *Statistics: Maximum Gradient* fields in the Main Window. For example, if you put a 0.7 in the *Statistics: Minimum Gradient* field in the Main Window, you will be able to clearly differentiate regions with less than that IPC (light orange), regions just above 0.7 IPC (light green) and regions close to the Max IPC (dark blue, 0.9 in this case). This mechanism is very useful to analyze correlations between metrics. You will directly get this view by loading 2dc_ud_ipc.cfg.
- The histogram on instructions correlated to cache misses can be seen by loading 2dc_I_L2mr.cfg. In this histogram of instructions (columns now represent bins of number of instructions executed in a computation burst between MPI calls). You can see that there are also some vertical bands indicating that there is actually computational load imbalance. Curiously, the vertical bands are not equally separated in the instruction histogram and the duration histogram, corresponding to variations in IPC.
- If you load 3dh_msgsize_p2pcall.cfg**,** the histogram is now of the message size in point to point calls. This window actually shows the histogram for a specific MPI call, MPI_Isend in this case as shown at the *3D* section of the Main Window. The Statistic (color) represented is the number of messages of that size sent. You can see that there are more small than large messages.
- You can change the *Plane* selector in the *3D* section of the Main Window to see the histogram of message sizes for a different MPI call.