

Linaro Forge

Debugging and Optimising Parallel Codes

Rudy Shand - Field Application Engineer

Linaro Forge

Agenda

- 10:00 - 11:00 Lecture on Debugging with DDT
- 11:00 - 12:30 DDT Debugging Hands-on session
- 12:30 - 13:00 Break
- 13:00 - 14:00 Lecture on Profiling with MAP
- 14:00 - 15:30 MAP Profiling Hands-on session
- 15:30 - 16:00 Break
- 16:00 - 17:00 Try DDT / MAP with own codes

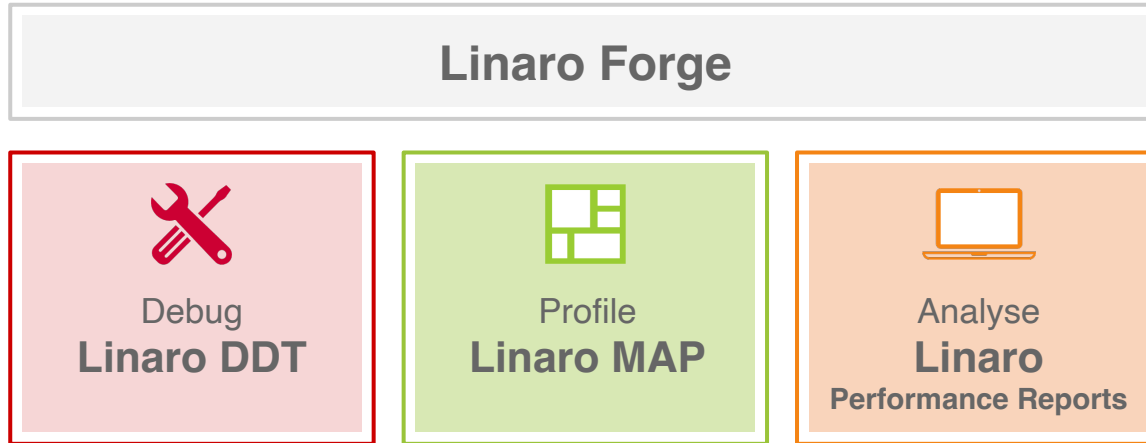
A little about myself

- Graduated from University of Reading
 - Cybernetics and Electronic Engineering
 - ML, Maths, Biology, Physics - No HPC
- Most working career in debuggers and performance tools
 - Arm DS-5 debugger and Streamline Performance Analyser
 - Compilers, Models, Embedded devices, mobile
 - In Embedded - both in a developer and quality role
- Joined the Arm Forge team (Now Linaro Forge)
 - Quality Lead / Field Application Engineer
 - 6 years in HPC, 11 years overall in debug and profiling tools



HPC Development Solutions from Linaro

Best in class commercially supported tools for Linux and high-performance computing (HPC)



Performance Engineering for any architecture, at any scale

Linaro Forge

An interoperable toolkit for debugging and profiling



The de-facto standard for HPC development

- Most widely-used debugging and profiling suite in HPC
- Fully supported by Linaro on Intel, AMD, Arm, Nvidia, AMD GPUs, etc.



State-of-the art debugging and profiling capabilities

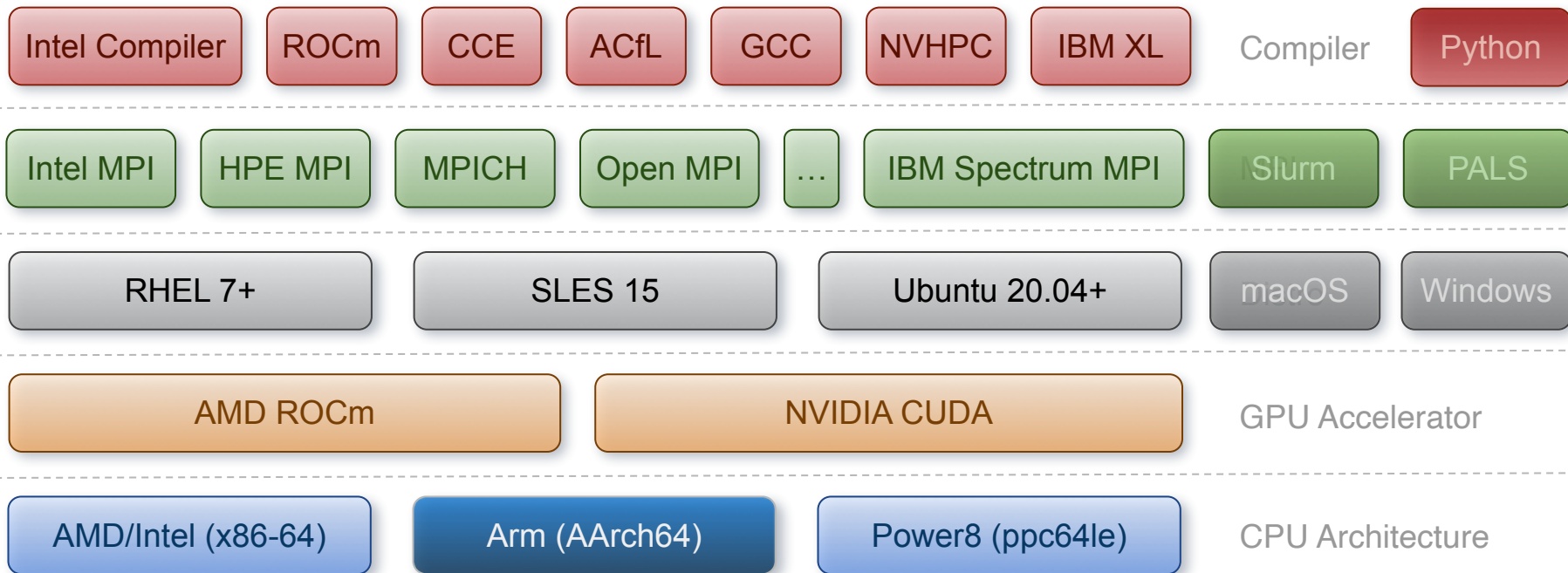
- Powerful and in-depth error detection mechanisms (including memory debugging)
- Sampling-based profiler to identify and understand bottlenecks
- Available at any scale (from serial to exascale applications)



Easy to use by everyone

- Unique capabilities to simplify remote interactive sessions
- Innovative approach to present quintessential information to users

Supported Platforms



Bug classification

- Crashes
 - One or more processes in application terminates
 - Most common and generally easiest to solve
- Hangs
 - Deadlocks - Stuck waiting for something that never happens
 - Livelocks - Making local progress, but no global progress
- Race conditions
 - One or more threads accessing the same data at the same time in non deterministic way
 - Shows up as incorrect answer or sometimes crashes



Linaro DDT Debugger Highlights

Tracepoint	Process	Values logged
hnone RD05	0% n/a	
hnone RD05	90% n/a	
hnone RD01	0% n/a	
hnone RD05	94% n/a	
hnone RD01	0% n/a	
hnone RD05	97% n/a	
hnone RD01	0% n/a	
hnone RD05	88% n/a	
hnone RD05	84% n/a	
hnone RD01	80% n/a	

The scalable print alternative

```

for (i = 0; i < SIZE_M; i++)
  for (j = 0; j < SIZE_M; j++)
    C[i][j] = 0;

for (i = 0; i < SIZE_M; i++)
  for (j = 0; j < SIZE_M; j++)
    for (k = 0; k < SIZE_Q; k++)
      C[i][j] += A[i][k] * B[k][j];

if (frunp)
  MPI_S
  MPI_B
}

printf("1/2\n");

if (argc
  for (i = 0; i < SIZE_M; i++)
    print ("");
    
```

Stop on variable change

```

hello.c:43: warning: 'Y' is of type 'void *'. When using void pointers in calculations, the behaviour is undefined.
43     else
44     {
45         test=1;
46     }
47 void func1()
48 {
49     void* i = (void*) 1;
50     while(i++ || i)
51         free((void*)i);
52 }
53 {
54     typeThree test;
55     typeThree t2;
56     int i;
57 }
58
    
```

Static analysis warnings on code errors

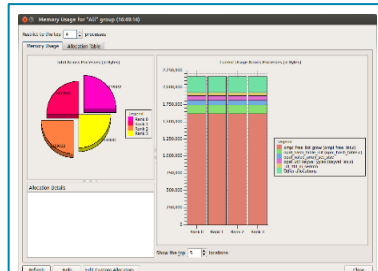
```

if (argv[i] && !strcmp(argv[i], "crash")) {
  argv[i] = 0;
  printf("this %c", (char) **argv[i]);
}
/* we shall see */
}

func1();
func2();
fprintf(stderr, "1\n");
beingWatched = 1;

test.anotherList.s
test.c = 'p';
beingWatched = 0;
    
```

Detect read/write beyond array bounds



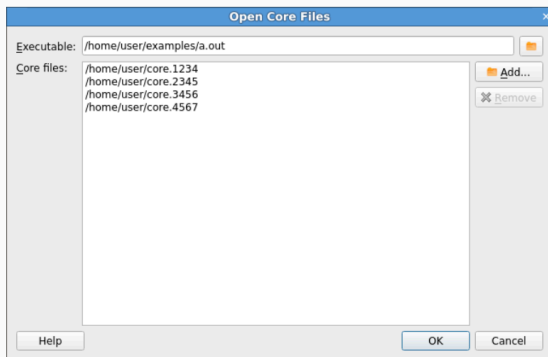
Detect stale memory allocations

Core files

You can open and debug one or more core files generated by your application.

Procedure

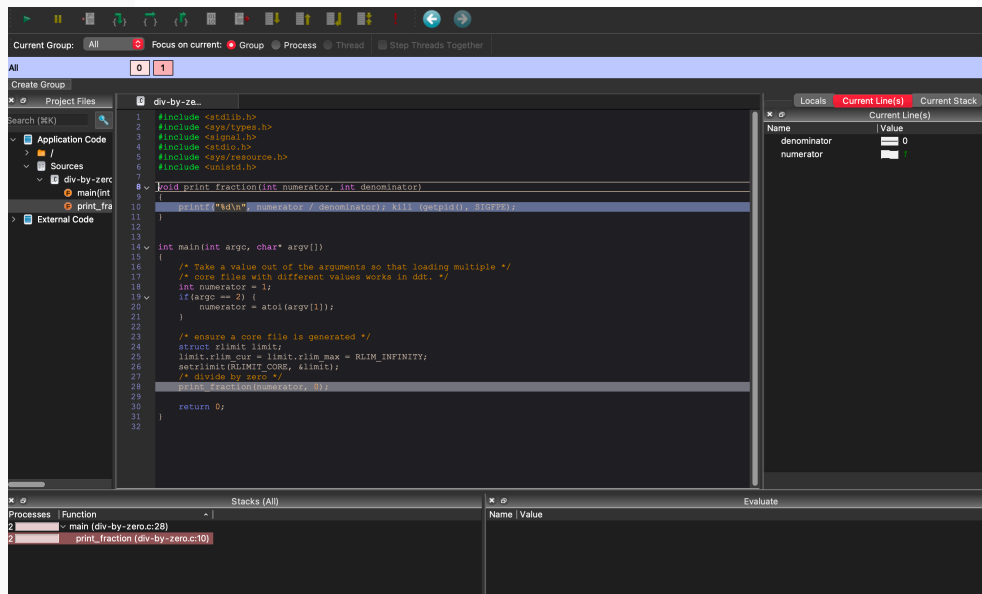
1. On the Welcome page click **Open Core Files**. The **Open Core Files** window opens.



2. Select an executable and a set of core files, then click **OK** to open the core files and start debugging them.

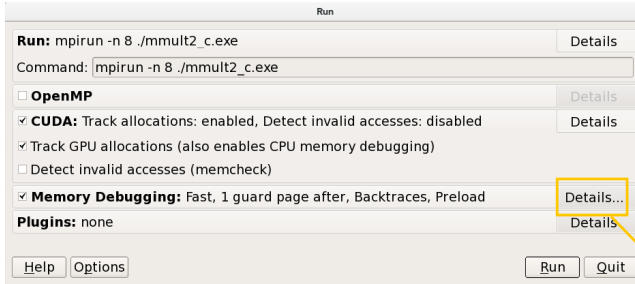
Note

While Linaro DDT is in this mode, you cannot play, pause, or step, because there is no process active. You are, however, able to evaluate expressions and browse the variables and stack frames saved in the core files.

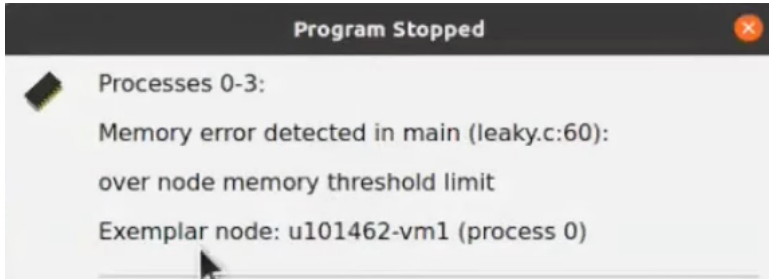
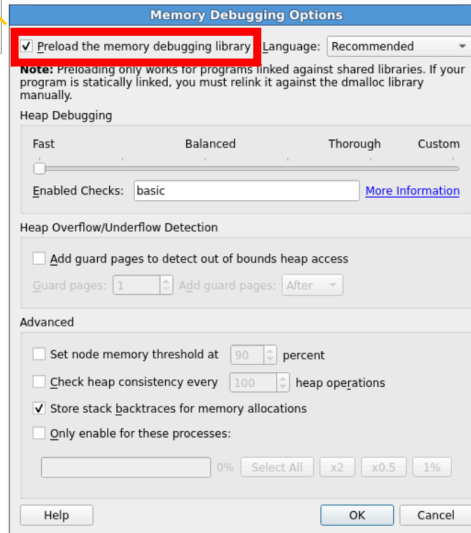


- View core files for CPU's
- View core files for GPU's

Memory debugging menu in Linaro DDT



When manual linking is used,
untick "Preload" box



Multi-dimensional Array Viewer

What does your data look like at runtime?

View arrays

- On a single process
- Or distributed on many ranks
- Display the array values from `tables[0:11][0:11]`

Use metavariables to browse the array

- Example: `$i` and `$j`
- Metavariables are unrelated to the variables in your program
- The bounds to view can be specified
- Visualise draws a 3D representation of the array

Data can also be filtered

- “Only show if”: `$value>0` for example `$value` being a specific element of the array

The screenshot shows the Multi-Dimensional Array Viewer interface. At the top, the Array Expression is set to `tables[$i][$j]`. Below this, there are controls for Distributed Array Dimensions (set to None), Staggered Array (unchecked), and Range of `$i` and `$j` (both from 0 to 11). The Display options are set to Rows and Columns. There are buttons for Evaluate, Cancel, and a checkbox for Align Stack Frames. Below the controls, there is a Data Table tab and a Statistics tab. The Data Table tab is active, showing a 12x12 grid of values. A 3D visualization of the array is shown in the bottom right corner, with axes labeled 'row', 'col', and 'value'. The 3D plot shows a surface with a peak at the top right corner, corresponding to the highest values in the array.

i \ j	0	1	2	3	4	5	6	7	8	9	10	11
0	0	1	2	3	4	5	6	7	8	9	10	11
1	2	4	6	8	10	12	14	16	18	20	22	24
2	3	6	9	12	15	18						
3	4	8	12	16	20	24						
4	5	10	15	20	25	30						
5	6	12	18	24	30	36						
6	7	14	21	28	35	42						
7	8	16	24	32	40	48						
8	9	18	27	36	45	54						
9	10	20	30	40	50	60						
10	11	22	33	44	55	66						

DDT: Production-scale debugging

Isolate and investigate faults at scale

Who misbehaved?

- Merge stacks from processes and threads
- Sparklines comparing data across processes
- Which MPI rank

Where is the problem?

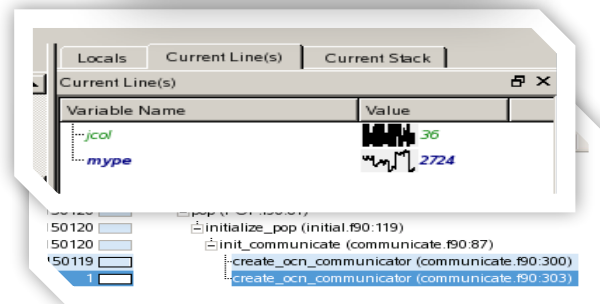
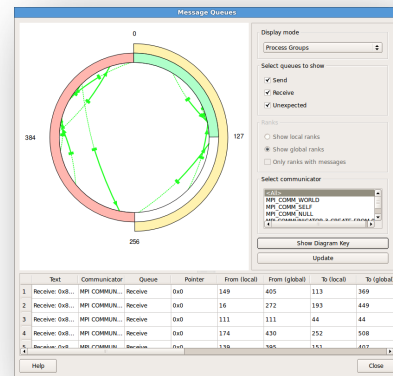
- Integrated source code editor
- Dynamic data structure visualization

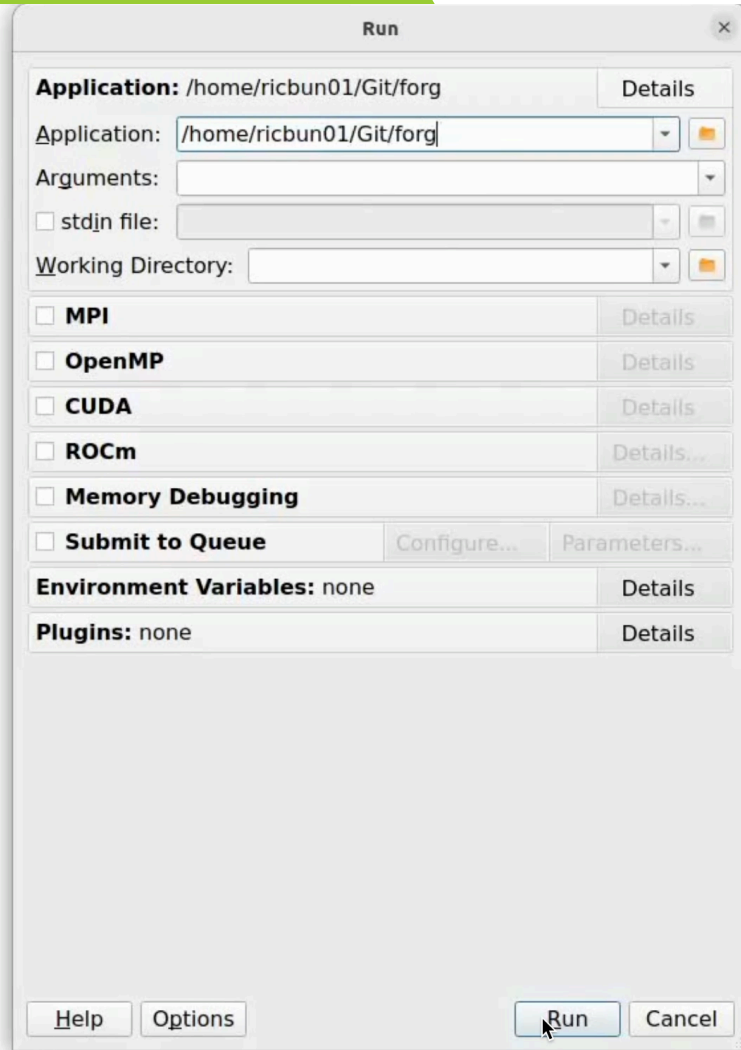
How did it happen?

- Parse diagnostic messages
- Trace variables through execution

Why did it happen?

- Unique “Smart Highlighting”
- Experiment with variable values





Starting a debug session

Current Group: All | Focus on current: Group Process Thread Step Threads Together

All | 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

Create Group

Project Files | Fortran Modules

Project Files

Search (Ctrl+K)

- Application Code
 - /
 - Headers
 - Sources
 - hello.f
 - forge_constants
 - FUNC1(): INTEGER
 - hellof77
 - show_consts
 - SUB1
- External Code

hello.f x

```

19 SUBROUTINE SUB1 ()
20   INTEGER test, FUNC1
21   test = FUNC1()
22   IF (test.EQ.1) THEN
23     test = 0
24   ELSE
25     test = -1
26   END IF
27 END
28
29 PROGRAM hellof77
30   include 'mpif.h'
31
32   INTEGER i, my_rank, p, source, dest, tag, x, y, beingwatched, ierr, my_size
33   CHARACTER message*21
34   CHARACTER messagefirst
35   INTEGER status(MPI_STATUS_SIZE)
36   INTEGER, ALLOCATABLE :: domain(:)
37   INTEGER stat
38
39   CALL MPI_INIT(ierr)
40   CALL MPI_COMM_SIZE(MPI_COMM_WORLD, my_size, ierr)
41   CALL MPI_COMM_RANK(MPI_COMM_WORLD, my_rank, ierr)
42
43   IF (my_size.EQ.8) THEN
44     IF (my_rank.EQ.5) THEN
45       CALL MPI_SEND(message, 400, MPI_CHARACTER, dest, tag, MPI_COMM_W
46     CORLD, ierr)
47     END IF
48   END IF
49
50   ALLOCATE(domain(my_rank*100000))
51   DO i = 1, SIZE(domain)
52     domain(i) = 2*i + 1
53   END DO
54
55   message = "Hello From Me      !"
56
57   PRINT *, "My rank is ", my_rank, "!"
58
59   CALL SUB1()
60
61   beingwatched = 1

```

Locals | Current Line(s) | Current Stack

Locals

Name	Value
mpi_argv_null	
mpi_argvs_null	
mpi_bottom	0
mpi_errcodes_i...	
mpi_in_place	0
mpi_status_ign...	
mpi_statuses_i...	
mpi_unweighted	
mpi_weights_e...	
beingwatched	0
dest	0
domain	<not allocated>
i	-933977151
ierr	0
message	^000\000\000\000
my_rank	0
my_size	0
source	0
stat	0
tag	0
ompi_release_v...	4
ompi_minor_ve...	1
ompi_major_ve...	3
ompi_comm_ty...	6
ompi_comm_ty...	7
ompi_comm_ty...	0
ompi_comm_ty...	5
ompi_comm_ty...	4
ompi_comm_ty...	3
ompi_comm_ty...	1
ompi_comm_ty...	9
ompi_comm_ty...	10
ompi_comm_ty...	2
ompi_comm_ty...	11
ompi_comm_ty...	8
mpi_wtime_is_g...	3
mpi_win_unified	0
mpi_win_size	8
mpi_win_senar	1

Input/Output | Breakpoints | Watchpoints | Stacks (All) | Tracepoints | Tracepoint Output | Logbook

Stacks (All)

Processes	Threads	Function
20	20	hellof77 (hello.f:40)
20	40	progress engine

Evaluate

Name	Value
------	-------

Current Group: All | Focus on current: Group | Process | Thread | Step Threads Together

All | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19

Create Group

Project Files | Fortran Modules

Project Files

Search (Ctrl+K)

- Application Code
 - /
 - Headers
 - Sources
 - hello.f
 - forge_constants
 - FUNC1(): INTEGER
 - hellof77
 - show_consts
 - SUB1
 - External Code

```

1  c123456
2
3  module forge_constants
4      real, parameter :: pi = 3.14
5      real, parameter :: e = 2.71
6
7  contains
8      subroutine show_consts()
9          print*, "Pi = ", pi, "e = ", e
10         end subroutine show_consts
11     end module forge_constants
12
13     INTEGER FUNCTION FUNC1 ()
14         INTEGER my_int, your_int
15         my_int=2
16         your_int=3
17         FUNC1=my_int*your_int
18     END
19
20     SUBROUTINE SUB1 ()
21         INTEGER test,FUNC1
22         test=FUNC1()
23         IF (test.eq.1) THEN
24             test=0
25         ELSE
26             test=-1
27         END IF
28     END
29
30     PROGRAM hellof77
31         include 'mpif.h'
32
33         INTEGER i,my_rank,p,source,dest,tag,x,y,beingwatched,ierr,my_size
34         CHARACTER message*21
35         CHARACTER messagefirst
36         INTEGER status(MPI_STATUS_SIZE)
37         INTEGER, ALLOCATABLE :: domain(:)
38         INTEGER stat
39
40         CALL MPI_INIT(ierr)
41         CALL MPI_COMM_SIZE(MPI_COMM_WORLD, my_size, ierr)
42         CALL MPI_COMM_RANK(MPI_COMM_WORLD, my_rank, ierr)
43
44         IF (my_size.eq.8) THEN

```

Locals | Current Line(s) | Current S

Locals

Name	Value
mpi_argv_null	
mpi_argv_null	
mpi_bottom	0
mpi_errcodes_i...	
mpi_in_place	0
mpi_status_ign...	
mpi_statuses_i...	
mpi_unweighted	
mpi_weights_e...	
beingwatched	0
dest	0
domain	<not allocate
i	-933977151
ierr	0
message	'0001000100'
my_rank	0
my_size	0
source	0
stat	0
tag	0
mpi_release_v...	4
mpi_minor_ve...	1
mpi_major_ve...	3
mpi_comm_ty...	6
mpi_comm_ty...	7
mpi_comm_ty...	0
mpi_comm_ty...	5
mpi_comm_ty...	4
mpi_comm_ty...	3
mpi_comm_ty...	1
mpi_comm_ty...	9
mpi_comm_ty...	10
mpi_comm_ty...	2
mpi_comm_ty...	11
mpi_comm_ty...	8
mpi_wtime_is_g...	3
mpi_win_unified	0
mpi_win_size	8
mpi_win_spar...	1

Input/Output | Breakpoints | Watchpoints | Stacks (All) | Tracepoints | Tracepoint Output | Logbook

Breakpoints

Processes	Threads	File	Line	Function	Condition	Start After	Trigger Ever	Stop After	Full path
-----------	---------	------	------	----------	-----------	-------------	--------------	------------	-----------

Evaluate

Name	Value
------	-------

Current Group: All | Focus on current: Group Process Thread Step Threads Together

All | 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

Create Group

Project Files | Fortran Modules

Project Files

Search (Ctrl+K)

- Application Code
 - /
 - Headers
 - Sources
 - hello.f
 - forge_constants
 - FUNC1(): INTEGER
 - hellof77
 - show_consts
 - SUB1
 - External Code

```

46  cORLD,ierr)
47      END IF
48  END IF
49
50  ALLOCATE(domain(my_rank*100000))
51  DO i = 1, SIZE(domain)
52      domain(i) = 2*i + 1
53  END DO
54
55  message="Hello From Me      !"
56
57  PRINT *, "My rank is ",my_rank, "!"
58
59  CALL SUB1()
60
61  beingwatched=1
62  tag=0
63
64  IF (my_rank.ne.0) THEN
65      PRINT *, "Greetings from process ",my_rank, "!"
66      PRINT *, "Sending message from ",my_rank, "!"
67      dest=0
68      CALL MPI_Send(message,21,MPI_CHARACTER,dest,tag,MPI_COMM_WORLD
69  c,ierr)
70      beingwatched=beingwatched-1
71  ELSE
72      message="Hello from my process"
73      DO source=1,(my_size-1)
74          PRINT *, "waiting for message from ",source
75          CALL MPI_Recv(message,21,MPI_CHARACTER,source,tag,MPI_COMM_WORLD
76  c,stat,ierr)
77          PRINT *, "Message recieved: ",message
78          ! beingwatched=beingwatched+1
79      END DO
80  END IF
81
82  beingwatched=12
83  CALL MPI_Finalize(ierr)
84  beingwatched=0
85  PRINT *, "All done...",my_rank
86  END
87

```

Locals | Current Line(s) | Current Stack

Name	Value
mpi_argv_null	
mpi_argv5_null	
mpi_bottom	0
mpi_errcodes_i...	
mpi_in_place	0
mpi_status_ign...	
mpi_statuses_i...	
mpi_unweighted	
mpi_weights_e...	
beingwatched	1
dest	0
domain	
i	1
ierr	0
message	'Hello From Me
my_rank	0
my_size	20
source	0
stat	0
tag	0
mpi_release_v...	4
mpi_minor_ve...	1
mpi_major_ve...	3
mpi_comm_ty...	6
mpi_comm_ty...	7
mpi_comm_ty...	0
mpi_comm_ty...	5
mpi_comm_ty...	4
mpi_comm_ty...	3
mpi_comm_ty...	1
mpi_comm_ty...	9
mpi_comm_ty...	10
mpi_comm_ty...	2
mpi_comm_ty...	11
mpi_comm_ty...	8
mpi_wtime_js_g...	3
mpi_win_unified	0
mpi_win_size	8
mpi_win_senar	1

Input/Output | Breakpoints | Watchpoints | Stacks (All) | Tracepoints | Tracepoint Output | Logbook

Stacks (All)

Processes	Threads	Function
19	19	hellof77 (hello.f:65)
1	1	hellof77 (hello.f:72)
20	40	progress_engine

Evaluate
Name Value

Current Group: All | Focus on current: Group | Process | Thread | Step Threads Together

All: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

Root: 0

Leaves: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

Create Group

Project Files | Fortran Modules

Project Files

Search (Ctrl+K)

Application Code

- /
- Headers
- Sources
 - hello.f
 - forge_constants
 - FUNCI() : INTEGER
 - hello77
 - show_consts
 - SUB1
- External Code

```

50  ALLOCATE (domain(my_rank*100000))
51  DO i = 1, SIZE(domain)
52      domain(i) = 2*i + 1
53  END DO
54
55  message="Hello From Me      !"
56
57  PRINT *, "My rank is ", my_rank, "!"
58
59  CALL SUB1()
60
61  beingwatched=1
62  tag=0
63
64
65  IF (my_rank.ne.0) THEN
66      PRINT *, "Greetings from process ", my_rank, "!"
67      PRINT *, "Sending message from ", my_rank, "!"
68      dest=0
69      CALL MPI_Send(message, 21, MPI_CHARACTER, dest, tag, MPI_COMM_WORLD
70  c, ierr)
71      beingwatched=beingwatched-1
72  ELSE
73      message="Hello from my process"
74      DO source=1, (my_size-1)
75          PRINT *, "waiting for message from ", source
76          CALL MPI_Recv(message, 21, MPI_CHARACTER, source, tag, MPI_COMM_WORLD
77  c, stat, ierr)
78          PRINT *, "Message received: ", message
79          ! beingwatched=beingwatched+1
80      END DO
81  END IF
82
83  beingwatched=12
84  CALL MPI_Finalize(ierr)
85  beingwatched=0
86  PRINT *, "All done...", my_rank
87  END

```

Locals | Current Line(s) | Current Stack | Raw Command

Locals

Name	Value
mpi_argv_null	
mpi_argvs_null	
mpi_bottom	0
mpi_errcodes_i...	
mpi_in_place	0
mpi_status_ign...	
mpi_statuses_i...	
mpi_unweighted	
mpi_weights_e...	
beingwatched	1
dest	0
domain	
i	1
ierr	0
message	'Hello from my process'
my_rank	0
my_size	20
source	0
stat	0
tag	0
ompi_release_v...	4
ompi_minor_ve...	1
ompi_major_ve...	3
ompi_comm_ty...	6
ompi_comm_ty...	7
ompi_comm_ty...	0
ompi_comm_ty...	5
ompi_comm_ty...	4
ompi_comm_ty...	3
ompi_comm_ty...	1
ompi_comm_ty...	9
ompi_comm_ty...	10
ompi_comm_ty...	2
ompi_comm_ty...	11
ompi_comm_ty...	8

Input/Output* | Breakpoints | Watchpoints | Stacks (All) | Tracepoints | Tracepoint Output | Logbook

Stacks (All)

Processes	Threads	Function
19	19	hellof77 (hello.f:66)
1	1	hellof77 (hello.f:73)
20	40	progress_engine

Evaluate

Name	Value
------	-------

GPU Debugging

The screenshot displays a GPU debugging interface. At the top, there's a menu bar (File, Edit, View, Control, Tools, Window, Help) and a toolbar with various icons. Below that, a 'Focus on current' section has radio buttons for 'Process', 'Thread', and 'Step Threads Together'. A 'Threads' panel shows three threads: 1, 2, and K4. The main editor shows C++ code for a matrix multiplication kernel. The code includes a loop over 'k' and a calculation of 'temp'. The current line is highlighted. To the right, a 'GPU Devices' panel shows 'vega20' with 2 devices, 2400 threads, and 240 cores. At the bottom, a 'Kernel Progress View' shows a green progress bar for the 'MatrixMul...' kernel. An 'Evaluate' panel shows the values of variables: i=82, wA=128, wB=128, and temp=1.27999914.

```
19     int i = blockIdx.y * blockDim.y + threadIdx.y;
20     int j = blockIdx.x * blockDim.x + threadIdx.x;
21
22     for( int k = 0; k < wA; k++)
23     {
24         temp += A[ i * wA + k ] * B[k* wB +
25
26         C[ i * wB + j ] = temp;
27
28         __syncthreads();
29     }
30
31 __global__ void MatrixMulHIPShared(float *C, float *A, float *B, int wA, int wB, int wC, int hA, int hB, int hC, int blockRow, int blockCol, int blockDimX, int blockDimY, int blockDimZ, int threadsPerBlock)
32 {
33     // Block row and column
34     int blockRow = blockIdx.y;
```

- Support both AMD and Nvidia GPUs
- Debug simultaneously on GPU and CPU
- Look and feel exactly the same
- Main Features work in GPU
- Key (additional) GPU features:
 - Kernel Progress View
 - GPU thread in parallel stack view
 - GPU Thread Selector
 - GPU Device Pane
- For NVIDIA's nvcc compiler, kernels must be compiled with the -g -G flags
- ROCm GPU Debugging requires rocdbg to be available in your environment.
- For the hipcc compiler, kernels must be compiled with the -g flag

Python Debugging

- Debug Features

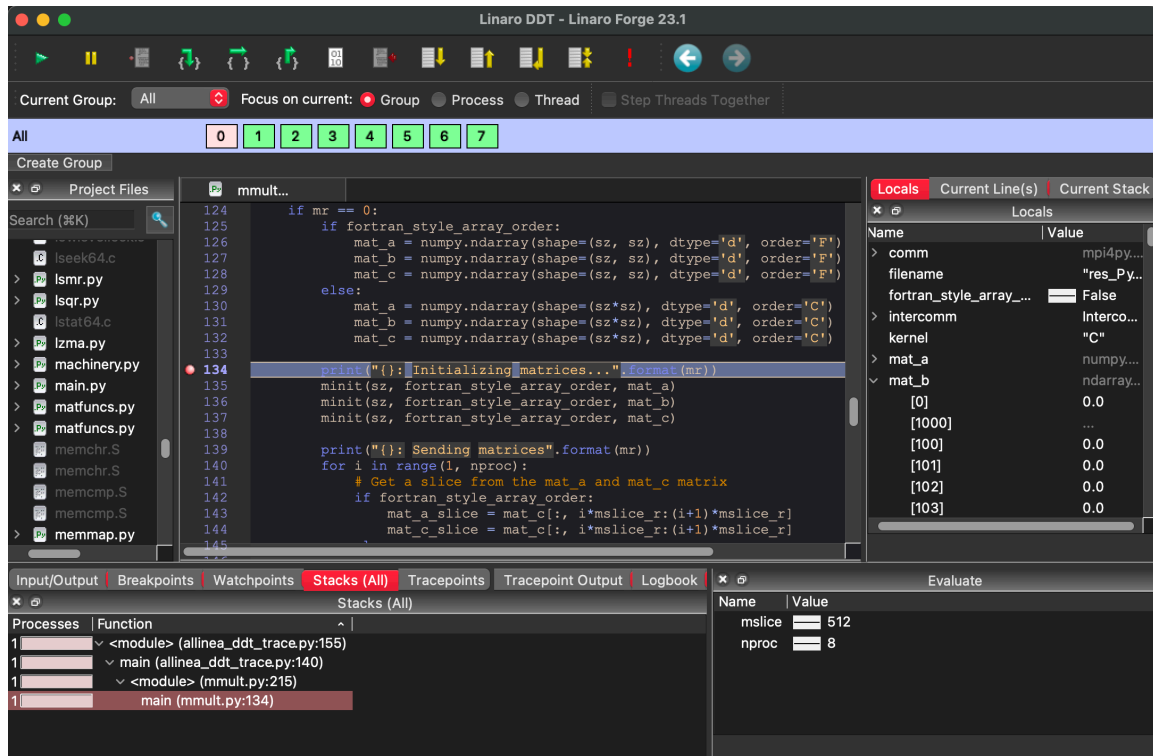
- Sparklines for Python variables
- Tracepoints
- MDA viewer
- Mixed language support

- Improved Evaluations:

- Matrix objects
- Array objects
- Pandas DataFrame
- Series objects

- Python Specific:

- Stop on uncaught Python exception
- Show F-string variables in “Current Line” display
- Mpi4py, NumPy, SciPy



```
ddt --connect mpirun -n 8 python3  
%allinea_python_debug% ./mmult.py
```

Run DDT in offline mode

Run the application under DDT and halt or report when a failure occurs

You can run the debugger in non-interactive mode

- For long-running jobs / debugging at very high scale
- For automated testing, continuous integration...

To do so, use following arguments:

- `$ ddt --offline --output=report.html mpirun ./jacobi_omp_mpi_gnu.exe`
 - `--offline` enable non-interactive debugging
 - `--output` specifies the name and output of the non-interactive debugging session
 - Html
 - Txt
 - Add `--mem-debug` to enable memory debugging and memory leak detection

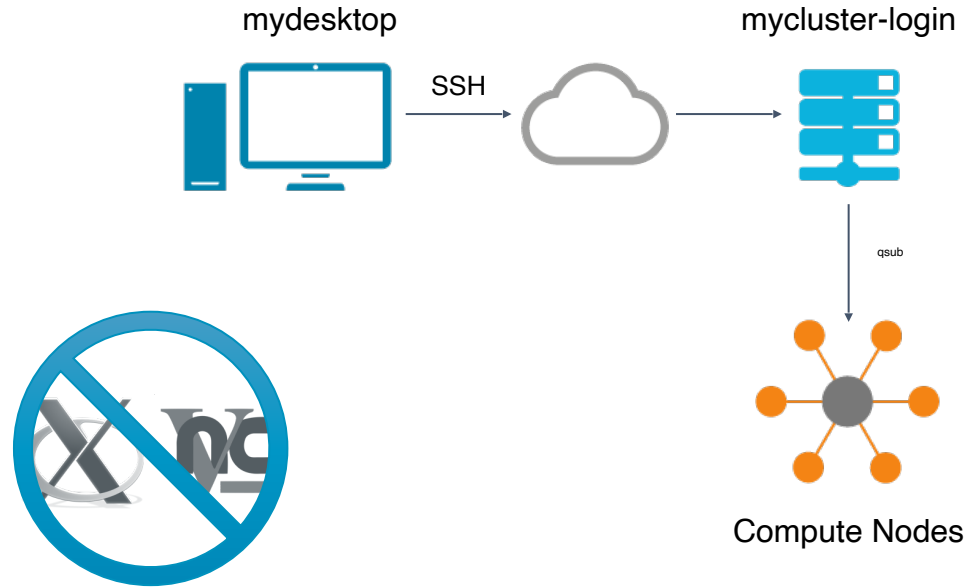
```
ddt --offline -o jacobi_omp_mpi_gnu_debug.txt \  
--trace-at _jacobi.F90:83,residual \  
srun ./jacobi_omp_mpi_gnu.exe
```

Report output

12		0:08.188	0-3	Process stopped at breakpoint in update (wave.c:216).																																																																								
13				<p>Additional Information</p> <p>▼ Stacks</p> <table border="1"><thead><tr><th>Processes</th><th>Threads</th><th>Function</th><th>Source</th><th>Variables</th></tr></thead><tbody><tr><td>0-3</td><td>4</td><td>main (wave.c:334)</td><td>▶ iterations = update(left, right);</td><td>▶ Rank 0, thread 1</td></tr><tr><td>0-3</td><td>4</td><td>update (wave.c:216)</td><td>▶ values[j] = newval[j];</td><td>▼ Rank 0, thread 1</td></tr><tr><td></td><td></td><td></td><td></td><td><table border="1"><thead><tr><th>Name</th><th>Value</th></tr></thead><tbody><tr><td>i</td><td>0</td></tr><tr><td>iterations</td><td>1</td></tr><tr><td>j</td><td>101</td></tr><tr><td>left</td><td>-2 (from -2 to 2)</td></tr><tr><td>now</td><td><aggregate value></td></tr><tr><td>right</td><td>1 (from -2 to 3)</td></tr><tr><td>stop</td><td>0</td></tr></tbody></table></td></tr><tr><td>0-3</td><td>8</td><td>progress_engine</td><td></td><td></td></tr><tr><td>0-3</td><td>8</td><td>opal_libevent2022_event_base_loop (event.c:1630)</td><td></td><td>▶ Rank 0, thread 2</td></tr><tr><td>0-3</td><td>4</td><td>poll_dispatch (poll.c:165)</td><td></td><td>▶ Rank 0, thread 2</td></tr><tr><td>0-3</td><td>4</td><td>poll</td><td></td><td></td></tr><tr><td>0-3</td><td>4</td><td>epoll_dispatch (epoll.c:407)</td><td></td><td>▶ Rank 0, thread 3</td></tr><tr><td>0-3</td><td>4</td><td>epoll_wait</td><td></td><td></td></tr></tbody></table> <p>▶ Current Stack</p> <p>▼ Evaluate</p> <table border="1"><thead><tr><th>Name</th><th>Value</th></tr></thead><tbody><tr><td>3*j*j</td><td>30603</td></tr><tr><td>j</td><td>101</td></tr></tbody></table>	Processes	Threads	Function	Source	Variables	0-3	4	main (wave.c:334)	▶ iterations = update(left, right);	▶ Rank 0, thread 1	0-3	4	update (wave.c:216)	▶ values[j] = newval[j];	▼ Rank 0, thread 1					<table border="1"><thead><tr><th>Name</th><th>Value</th></tr></thead><tbody><tr><td>i</td><td>0</td></tr><tr><td>iterations</td><td>1</td></tr><tr><td>j</td><td>101</td></tr><tr><td>left</td><td>-2 (from -2 to 2)</td></tr><tr><td>now</td><td><aggregate value></td></tr><tr><td>right</td><td>1 (from -2 to 3)</td></tr><tr><td>stop</td><td>0</td></tr></tbody></table>	Name	Value	i	0	iterations	1	j	101	left	-2 (from -2 to 2)	now	<aggregate value>	right	1 (from -2 to 3)	stop	0	0-3	8	progress_engine			0-3	8	opal_libevent2022_event_base_loop (event.c:1630)		▶ Rank 0, thread 2	0-3	4	poll_dispatch (poll.c:165)		▶ Rank 0, thread 2	0-3	4	poll			0-3	4	epoll_dispatch (epoll.c:407)		▶ Rank 0, thread 3	0-3	4	epoll_wait			Name	Value	3*j*j	30603	j	101
Processes	Threads	Function	Source	Variables																																																																								
0-3	4	main (wave.c:334)	▶ iterations = update(left, right);	▶ Rank 0, thread 1																																																																								
0-3	4	update (wave.c:216)	▶ values[j] = newval[j];	▼ Rank 0, thread 1																																																																								
				<table border="1"><thead><tr><th>Name</th><th>Value</th></tr></thead><tbody><tr><td>i</td><td>0</td></tr><tr><td>iterations</td><td>1</td></tr><tr><td>j</td><td>101</td></tr><tr><td>left</td><td>-2 (from -2 to 2)</td></tr><tr><td>now</td><td><aggregate value></td></tr><tr><td>right</td><td>1 (from -2 to 3)</td></tr><tr><td>stop</td><td>0</td></tr></tbody></table>	Name	Value	i	0	iterations	1	j	101	left	-2 (from -2 to 2)	now	<aggregate value>	right	1 (from -2 to 3)	stop	0																																																								
Name	Value																																																																											
i	0																																																																											
iterations	1																																																																											
j	101																																																																											
left	-2 (from -2 to 2)																																																																											
now	<aggregate value>																																																																											
right	1 (from -2 to 3)																																																																											
stop	0																																																																											
0-3	8	progress_engine																																																																										
0-3	8	opal_libevent2022_event_base_loop (event.c:1630)		▶ Rank 0, thread 2																																																																								
0-3	4	poll_dispatch (poll.c:165)		▶ Rank 0, thread 2																																																																								
0-3	4	poll																																																																										
0-3	4	epoll_dispatch (epoll.c:407)		▶ Rank 0, thread 3																																																																								
0-3	4	epoll_wait																																																																										
Name	Value																																																																											
3*j*j	30603																																																																											
j	101																																																																											
14		0:11.009	0-3	Play																																																																								

The Forge GUI and where to run it

DDT provides a powerful GUIs that can be run in a variety of configurations.



Hands on Setup

Remote System

Host coolmuc2

Hostname lxlogin1.lrz.de

user <username>

/lrz/sys/courses/hlin1w23/linaro/linaro-forge-training.tar.gz

module load ddt/23.1.1

Local Machine

Install Forge <https://www.linaroforge.com/downloadForge>

[Forge userguide](#)

Remote connection to CoolMUC-2

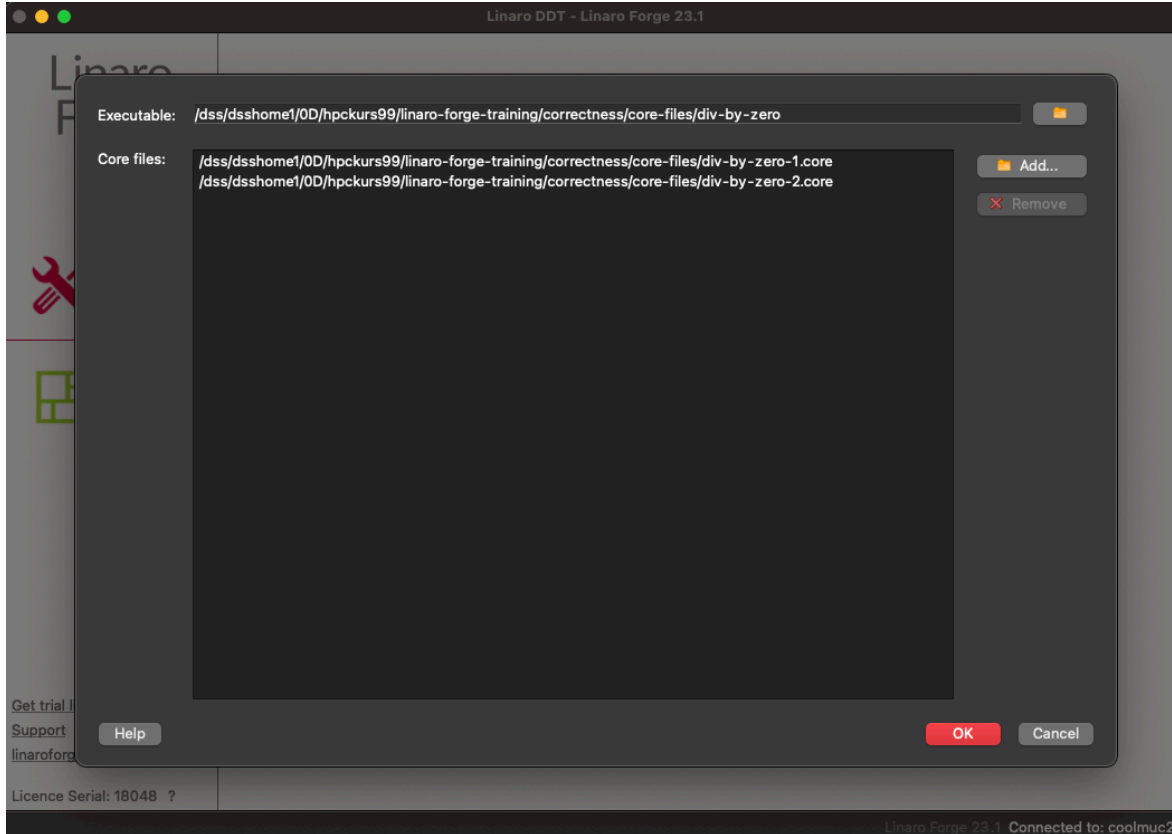
The screenshot shows the Linaro Forge IDE interface. On the left, there are logos for Linaro Forge, Linaro DDT, and Linaro MAP. The main area displays a sidebar with options: RUN (Run and debug a program.), ATTACH (Attach to an already running program.), OPEN CORE (Open a core file from a previous run.), and MANUAL LAUNCH (ADVANCED) (Manually launch the backend yourself.). Below these is an 'OPTIONS' section with a 'Remote Launch:' button highlighted by a red box, which says 'Configure...'. At the bottom left, there are links for 'Get trial licence', 'Support', and 'linaroforge.com', along with a 'Remote Client ?' icon.

The 'Remote Launch Settings' dialog box is open, showing the following configuration:

- Connection Name: CoolMUC2
- Host Name: coolmuc2
- Remote Installation Directory: //rz/sys/tools/ddt/23.1.1 (highlighted with a red box)
- Remote Script: Optional
- Private Key: Optional
- Always look for source files locally:
- KeepAlive Packets: Enable
- Interval: 60 seconds
- Proxy through login node:

Buttons at the bottom of the dialog include 'Help', 'Test Remote Launch', 'OK', and 'Cancel'.

Explore a core file



Hands on session

System Info

<https://doku.lrz.de/coolmuc-2-11484376.html>

CoolMUC-2: 812 nodes:

- 28-core Intel Hazwell processor per node
- 64GB DDR4 memory per node
- cm2_tiny partition

<https://doku.lrz.de/running-parallel-jobs-on-the-linux-cluster-11484078.html>

Interactive Session:

- `module load salloc_conf/cm2_tiny`
- `salloc -J linaro-hands-on --partition=cm2_tiny --time 00:30:00 --reservation=hlin1w23`

Scripting:

- `<linaro-forge-training>/slurm-coolmuc2.qtf`
- `<linaro-forge-training>/submit-job.sh`

Hands on session

Build and run debug examples

```
# Use default Intel modules
```

```
# build deadlock, simple and split programs  
cd <linaro-forge-training>/correctness/debug  
make -f Makefile
```

```
# run simple example with ddt  
ddt --connect mpiexec -n 4 ./simple
```

```
# offline-debugging  
sbatch submit-job.sh
```

Linaro Performance tools

Characterize and understand the performance of HPC application runs



Commercially supported
by Linaro

Gather a rich set of data

- Analyses metric around CPU, memory, IO, hardware counters, etc.
- Possibility for users to add their own metrics



Accurate and
Astute insight

Build a culture of application performance & efficiency awareness

- Analyses data and reports the information that matters to users
- Provides simple guidance to help improve workloads' efficiency



Relevant advice
to avoid pitfalls

Adds value to typical users' workflows

- Define application behaviour and performance expectations
- Integrate outputs to various systems for validation (eg. continuous integration)
- Can be automated completely (no user intervention)

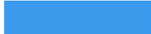
Linaro Performance Reports

A high-level view of application performance with “plain English” insights

Command: `mpiexec.hydra -host node-1,node-2 -map-by socket -n 16 -ppn 8 ./Bin/low_freq/../../Src//hydro -i ./Bin/low_freq/../../Src//Input/input_250x125_corner.nml`
Resources: 2 nodes (8 physical, 8 logical cores per node)
Memory: 15 GiB per node
Tasks: 16 processes, OMP_NUM_THREADS was 1
Machine: node-1
Start time: Thu Jul 9 2015 10:32:13
Total time: 165 seconds (about 3 minutes)
Full path: Bin/../../Src

Summary: hydro is **MPI-bound** in this configuration

Compute 20.6% 

MPI 63.2% 

I/O 16.2% 

Time spent running application code. High values are usually good. This is **very low**; focus on improving MPI or I/O performance first

Time spent in MPI calls. High values are usually bad. This is **high**; check the MPI breakdown for advice on reducing it


Time spent in filesystem I/O. High values are usually bad. This is **average**; check the I/O breakdown section for optimization advice


I/O

A breakdown of the 16.2% I/O time:

Time in reads 0.0% | 

Time in writes 100.0% 

Effective process read rate 0.00 bytes/s | 

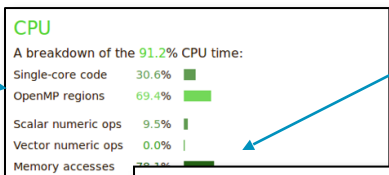
Effective process write rate 1.38 MB/s 

Most of the time is spent in **write operations** with a very low effective transfer rate. This may be caused by contention for the filesystem or inefficient access patterns. Use an I/O profiler to investigate which write calls are affected.

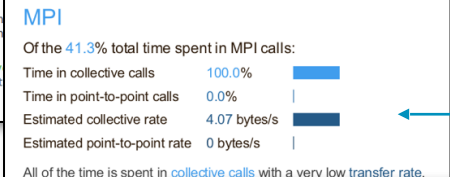
Linaro Performance Reports Metrics

Lowers expertise requirements by explaining everything in detail right in the report

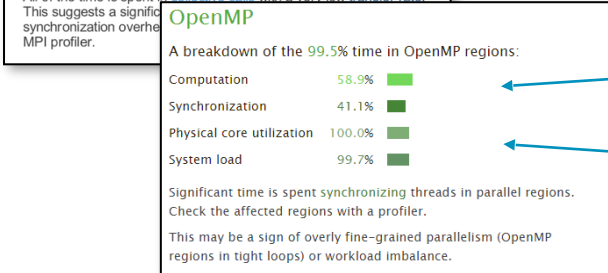
Multi-threaded parallelism



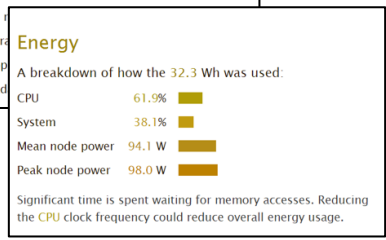
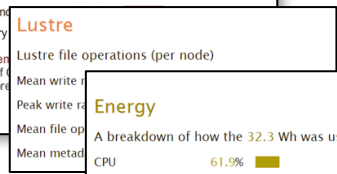
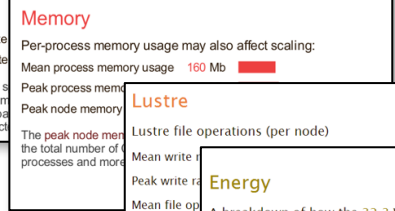
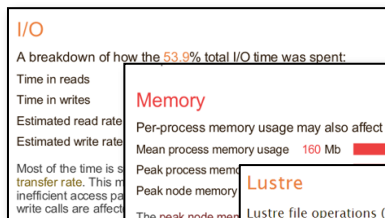
SIMD parallelism



Load imbalance



OMP efficiency
System usage



The Performance Roadmap

Optimizing high performance applications

Improving the efficiency of your parallel software holds the key to solving more complex research problems faster.

This pragmatic, 9 Step best practice guide, will help you identify and focus on application readiness, bottlenecks and optimizations one step at a time.

Bugs

- Correct application

Analyze before you optimize

- Measure all performance aspects. You can't fix what you can't see.
- Prefer real workloads over artificial tests.

I/O

- Discover lines of code spending a long time in I/O.
- Trace and debug slow access patterns.

Workloads

- Detect issues with balance.
- Slow communication calls and processes. Dive into partitioning code.

Communication

- Track communication performance. Discover which communication calls are slow and why.

Memory

- Reveal lines of code bottlenecked by memory access times.
- Trace allocation and use of hot data structure

Vectorization

- Understand numerical intensity and vectorization level.
- Hot loops, unvectorized code and GPU performance revealed

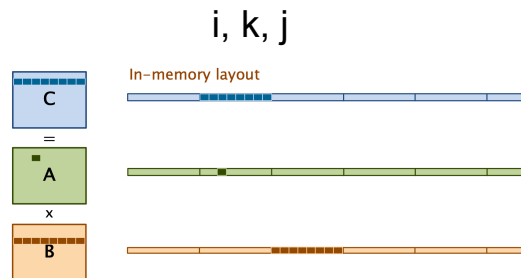
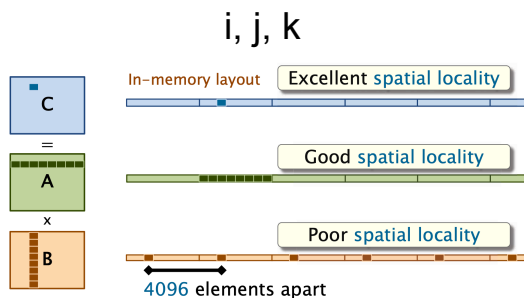
Cores

- Discover synchronization overhead and core utilization
- Synchronization-heavy code and implicit barriers are revealed

Verification

- Validate corrections and optimal performance

Performance Improvement



Think,



code,

```

i, j, k
for (int i = 0; i < n; ++i) {
  for (int j = 0; j < n; ++j) {
    for (int k = 0; k < n; ++k) {
      C[i][j] += A[i][k] * B[k][j];
    }
  }
}
    
```



run, run, run...



...to test and measure many different implementations

```

i, k, j
for (int i = 0; i < n; ++i) {
  for (int k = 0; k < n; ++k) {
    for (int j = 0; j < n; ++j) {
      C[i][j] += A[i][k] * B[k][j];
    }
  }
}
    
```

Loop order (outer to inner)	Running time (s)
i, j, k	1155.77
i, k, j	177.68
j, i, k	1080.61
j, k, i	3056.63
k, i, j	179.21
k, j, i	3032.82

MAP Capabilities

MAP is a sampling based scalable profiler

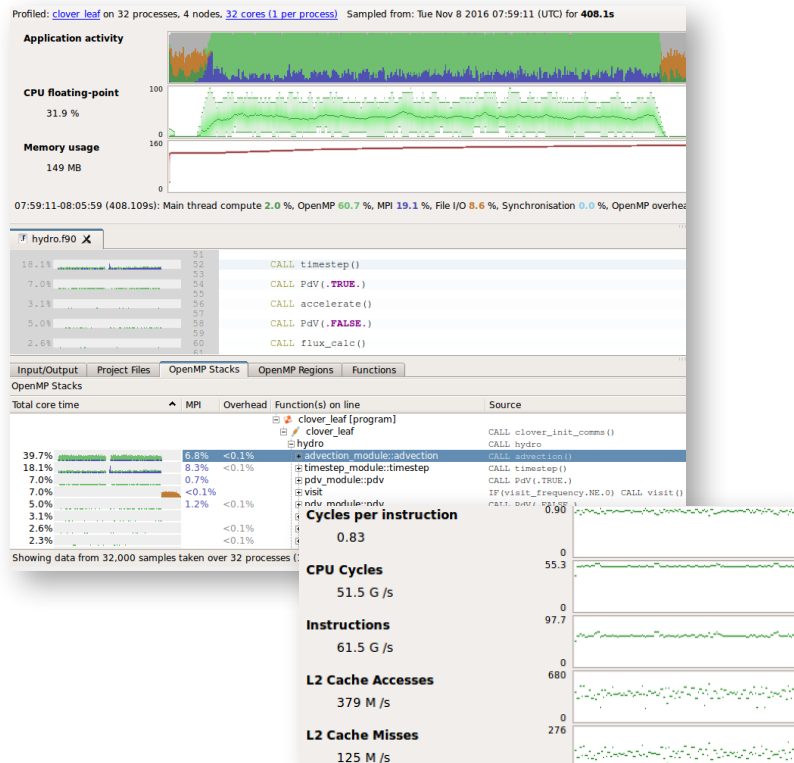
- Built on same framework as DDT
- Parallel support for MPI, OpenMP, CUDA
- Designed for C/C++/Fortran

Designed for 'hot-spot' analysis

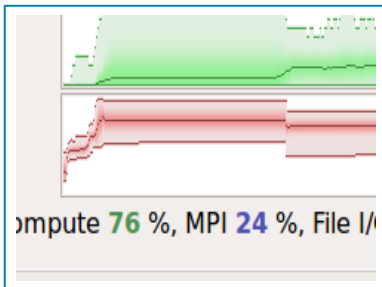
- Stack traces
- Augmented with performance metrics

Adaptive sampling rate

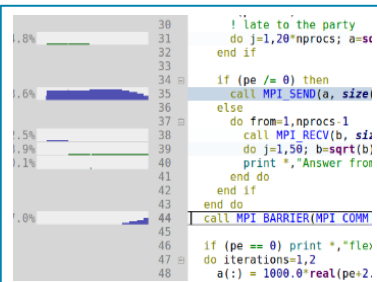
- Throws data away - 1,000 samples per process
- Low overhead, scalable and small file size



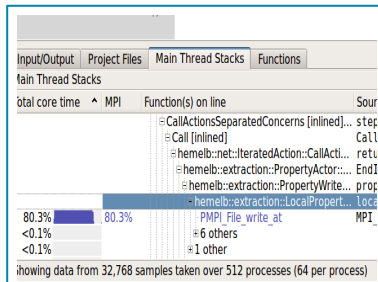
Linaro MAP Source Code Profiler Highlights



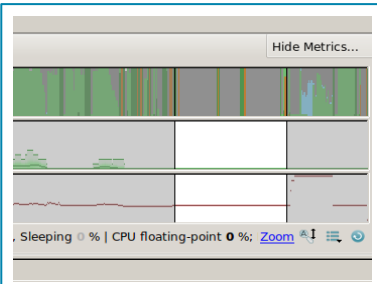
Find the peak memory use



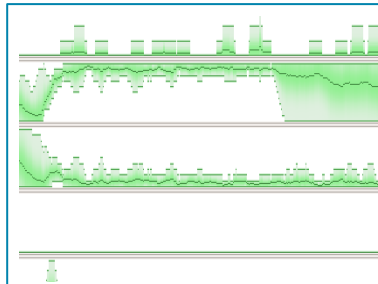
Fix an MPI imbalance



Remove I/O bottleneck



Make sure OpenMP regions make sense



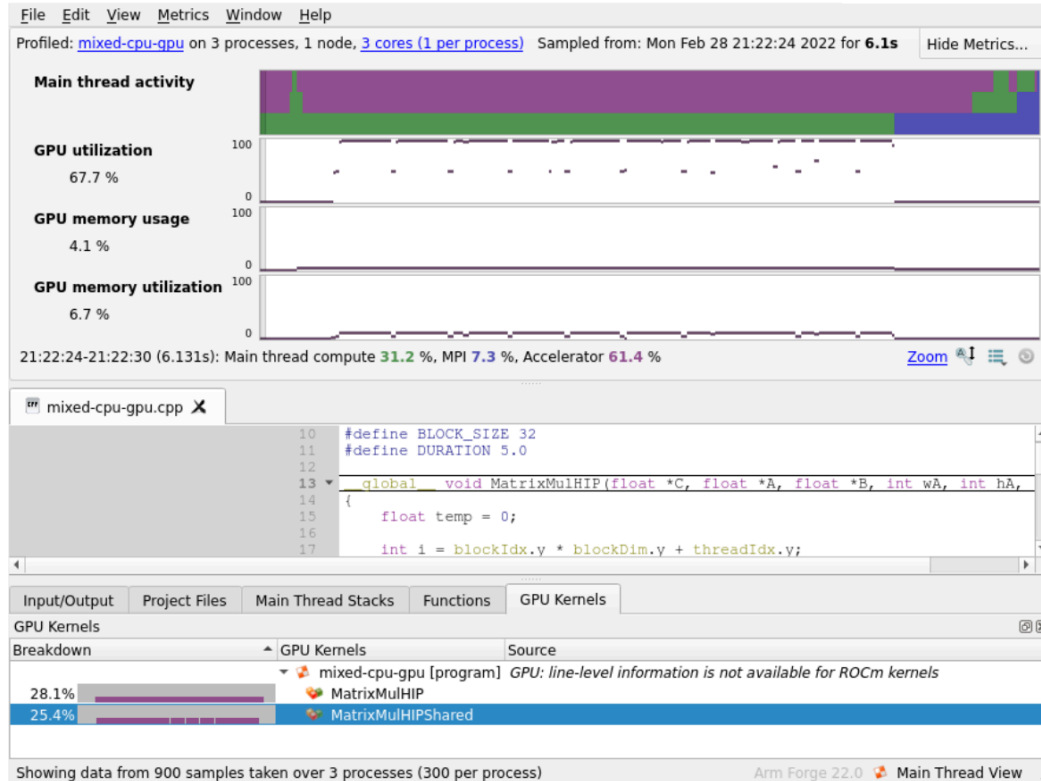
Improve memory access

```
{
  mmult(size, nproc, mat a, mat
  res += A[i*size+k]*B[k*size+j]
}

MPI_Finalize();
mwrite(size, mat c, filename)
```

Restructure for vectorization

ROCm AMD GPU Profiling



Profile

- Ran for 6s, taking 300 samples per process
- Able to bring up metadata of the profile
- Mixed CPU [green] / GPU [purple] application
- CPU time waiting for GPU Kernels [purple]
- GPU Kernels graph indicating Kernel activity

GUI information

- GUI is consistent across platforms
- Zoom into main thread activity
- Ranked by highest contributors to app time

Python Profiling

19.0 adds support for Python

- Call stacks
- Time in interpreter

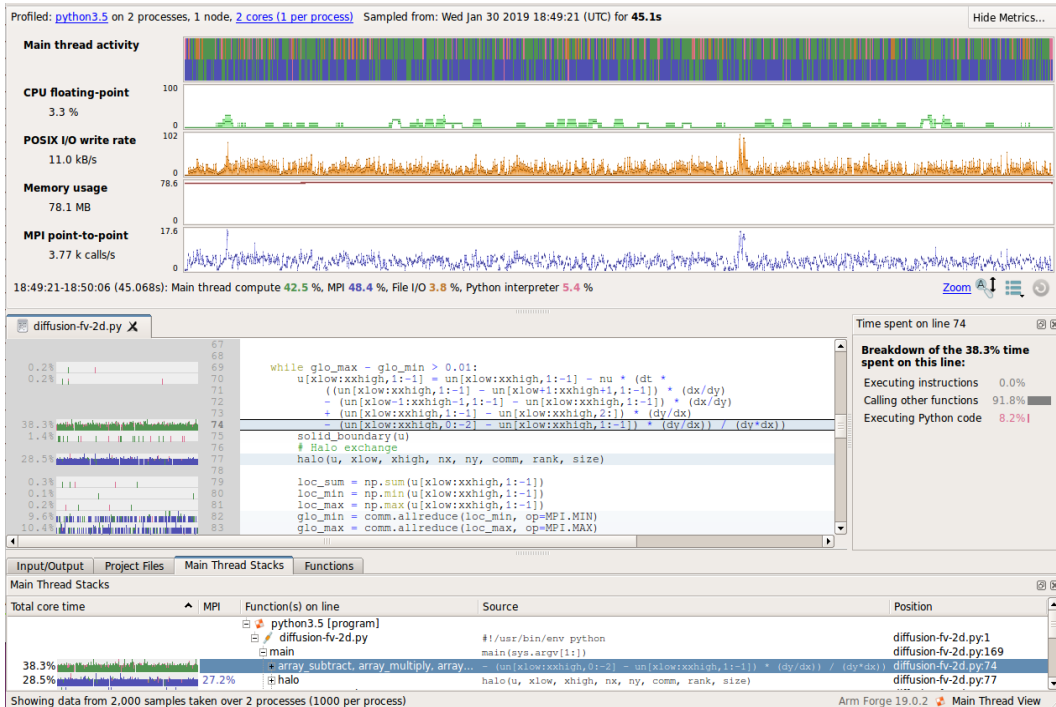
Works with MPI4PY

- Usual MAP metrics

Source code view

- Mixed language support

Note: Green as operation is on numpy array, so backed by C routine, not Python (which would be pink)

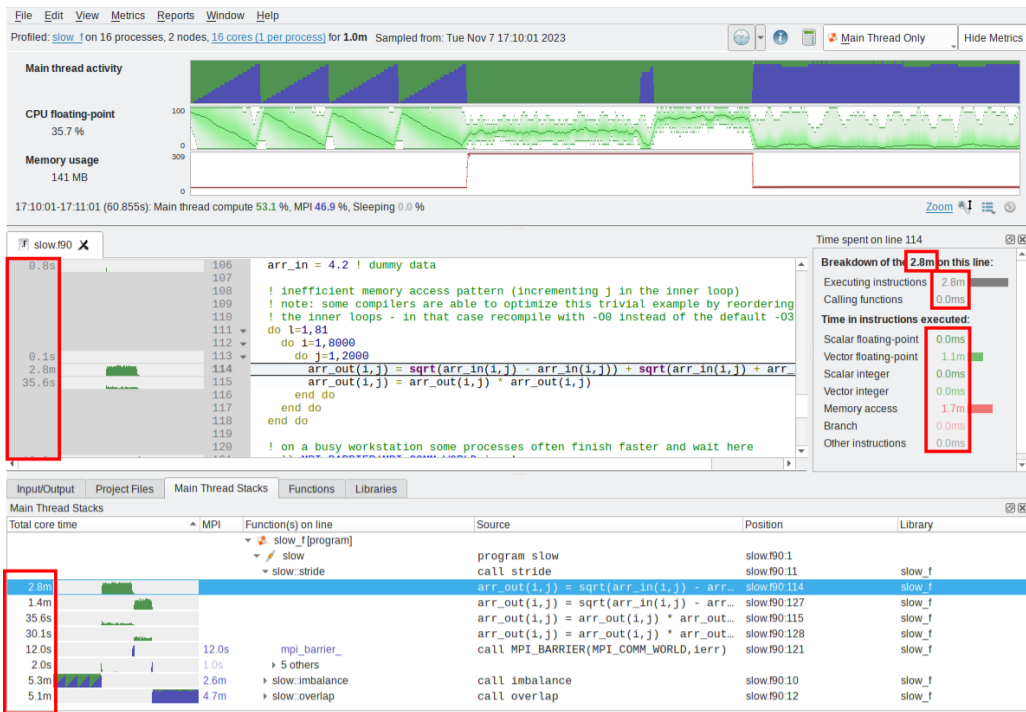


map --profile jsrun -n 2 python3 ./diffusion-fv-2d.py

A large, stylized graphic composed of low-poly triangles. The top part is a diamond shape, and the bottom part is a flame or fire shape. The colors transition from dark purple on the left to bright orange on the right.

Linaro Forge

Toggle percentage-time and core-time in MAP

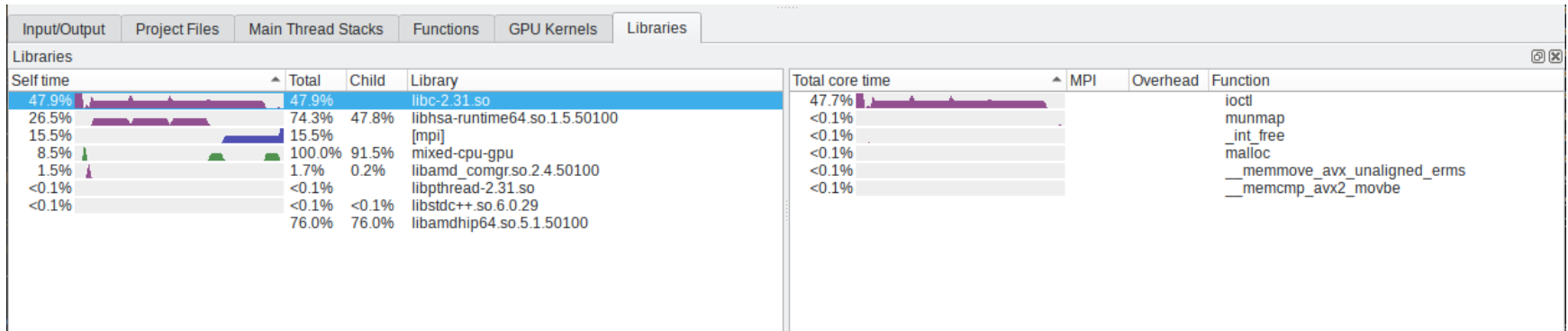


Use for direct comparisons between runs at the same scale (process/core counts).

- Easily determine if a change has made a portion of code faster, slower, or largely unchanged.
- Performance report automatically includes both percentage-time and core time
- Core-time is an estimation, but should be very close to the application run time

Libraries tab in MAP

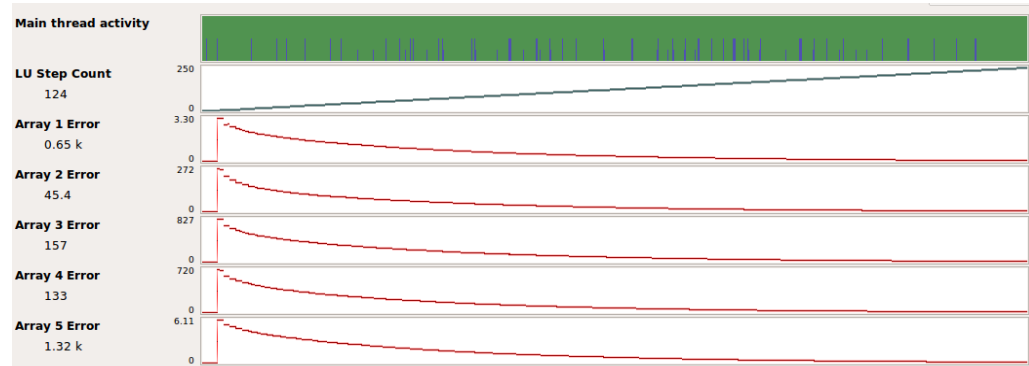
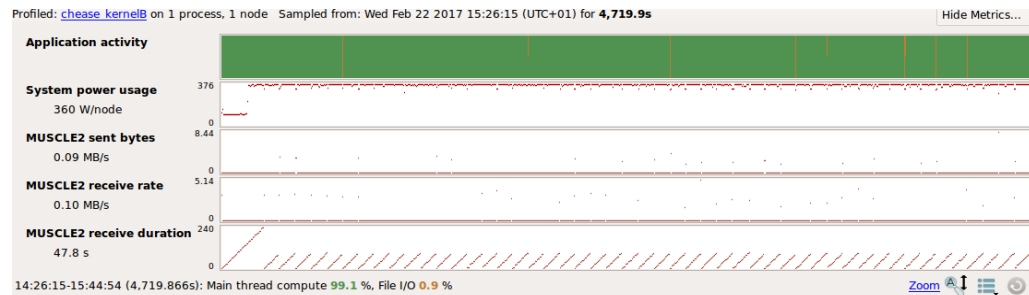
- List time spent in shared libraries (left)
- List entry point functions into the selected library (right)



Use to identify the libraries that would benefit the most from optimisation or replacement (e.g. alternative maths library or memory management implementation).

Custom metric example: MUSCLE2 & LU error terms

<https://github.com/arm-hpc/custom-metrics>



- Customized application instrumentation, eg, NPB LU
- Record error terms of solve
- Plot over time and step count for optimisation

Matrix Multiplication example

Build and run matrix multiplication example

https://docs.linaroforge.com/23.1.1/html/forge/worked_examples_appendix/mmult/analyze.html

Build C and Fortran Examples

```
export MPIF90=mpif90
make -f mmult.makefile
```

Build Python Examples

```
module load python
python -m venv run-mmult
. run-mmult/bin/activate
pip3 install numpy==1.23.5 scipy mpi4py
make -f mmult_py.makefile
```

Debug using UI

```
ddt --connect mpirun -n 8 ./mmult_c -s 3072
ddt --connect python3 %allinea_python_debug% ./mmult.py -s 3072
```

Offline profile

```
sbatch submit-job.sh
```



Thank you

rudy.shand@linaro.org

Linaro Forge