

allinea



Leaders in parallel software development tools

Allinea DDT Intelligence Within

14 September 2012

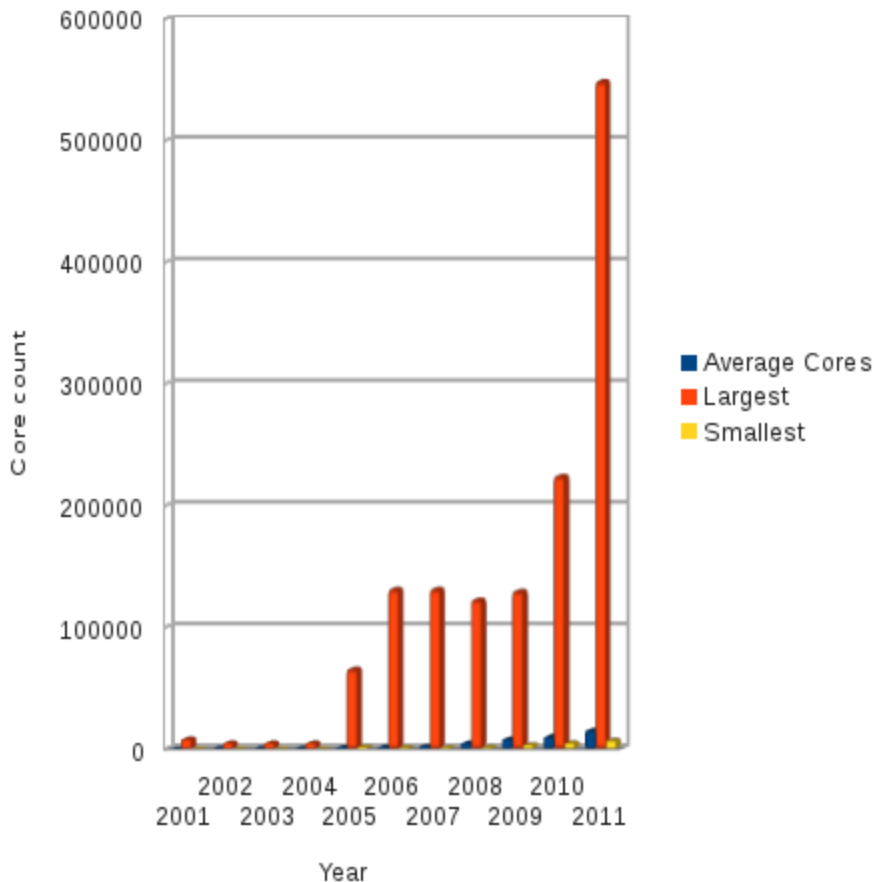
Agenda



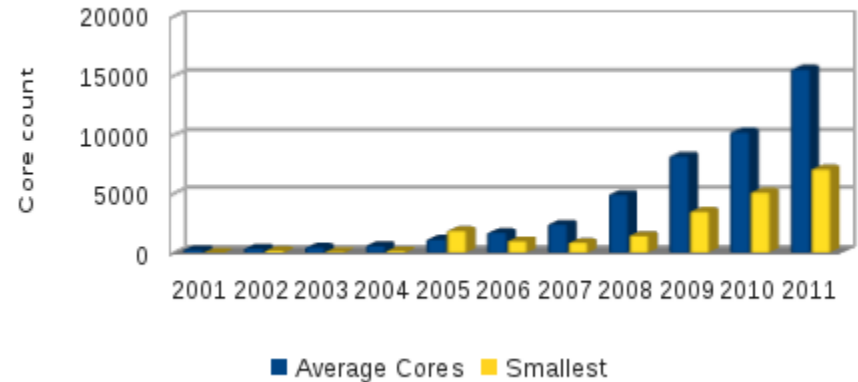
- Introduction
- Allinea DDT overview
- Conclusion

What is happening ? Extreme Machine Size

Growth in HPC core counts



HPC core counts



- Scientific progress requires more CPU hours
 - Maximum machine size is exploding
 - Average machine size grows exponentially

The Company



- Development tools company
 - Leading in HPC software tools market worldwide
 - Global customer base
 - Blue-chip engineering, government and academic research
- Allinea DDT
 - The leading debugger in parallel computing
 - Production use from desktop to extreme scale
 - World's only scalable debugger
 - Record holder for debugging software on largest machines
 - First at Petascale – and first for GPUs and ARM support!
- Allinea OPT
 - Profiling tool for parallel applications

Collaboration – National Labs



Partnership to develop Petascale debugger with NVIDIA support, with Cray and Caps Enterprise



Partnership on Full Scale debugging on IBM Blue Gene /P & /Q



Partnership with CEA French Atomic Energy Authority on scalable programming, CUDA and Alinea OPT



European partnership to develop techniques and solutions which address the exascale challenges

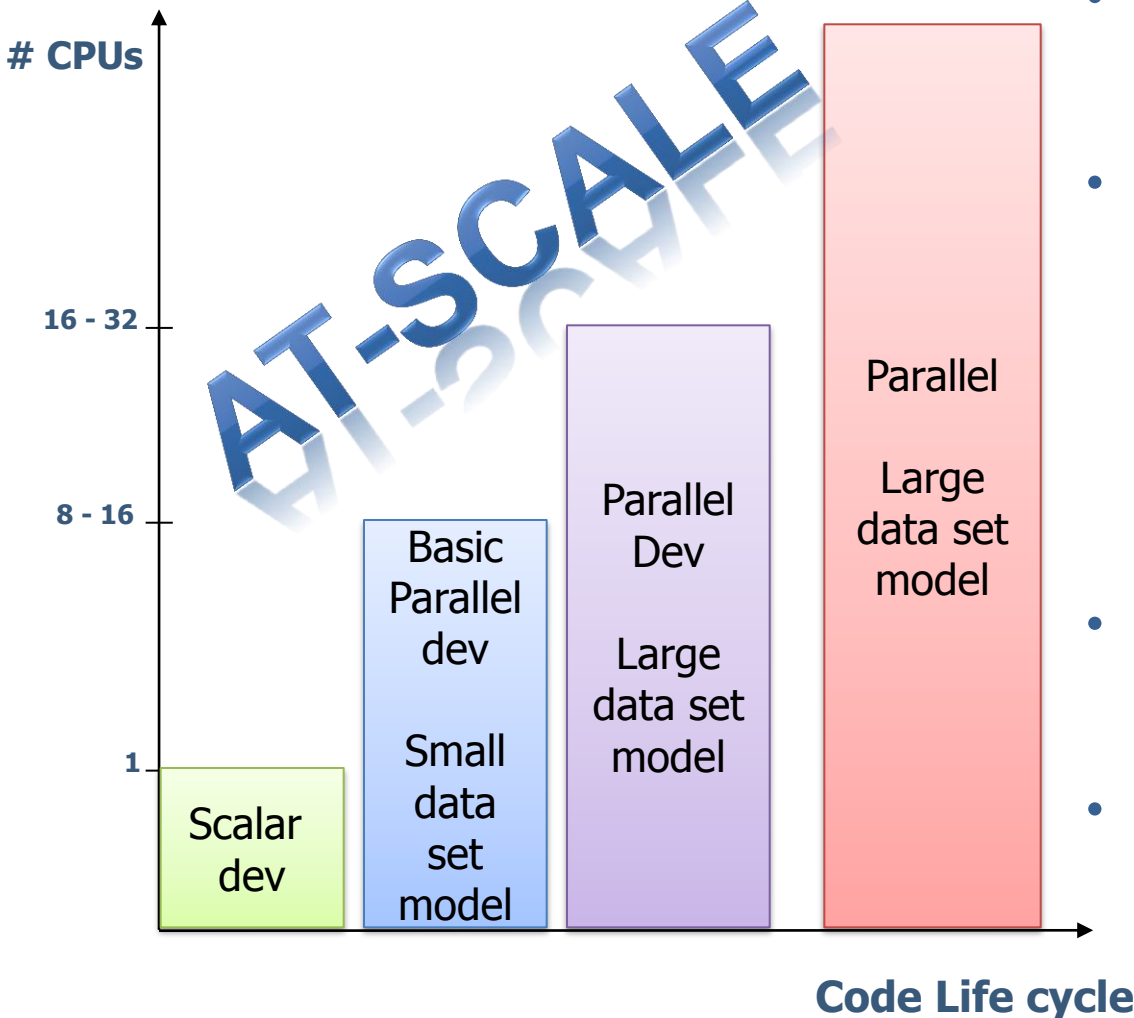


Partnership with BSC through its utilization of ARM technology to develop energy efficient HPC systems



allinea
DDT

Application Development Life Cycles



- Bugs are present at any stage
- Bugs get more and more tricky
 - More threads/processes
 - More data
 - Easy bugs already found
- Time to debug increases
- Debugging at scale : this is what you mostly do !

Example

```
$> mpirun -np 4 ./cpi.exe
```

```
Process 0 on dopey
```

```
Process 2 on dopey
```

```
Process 1 on dopey
```

```
Process 3 on dopey
```

```
pi is approximately 3.1416009869231249
```

```
Error is 0.0000083333333318
```

```
wall clock time = 0.012800
```



Happy !!

```
$> mpirun -np 8 ./cpi.exe
```

```
Process 5 on dopey
```

```
Process 1 on dopey
```

```
Process 4 on dopey
```

```
Process 6 on dopey
```

```
Process 3 on dopey
```

```
Process 0 on dopey
```

```
Process 2 on dopey
```

```
Process 7 on dopey
```



Deadlock !!

```
^Cmpirun: killing job...
```


How to debug At-Scale ?

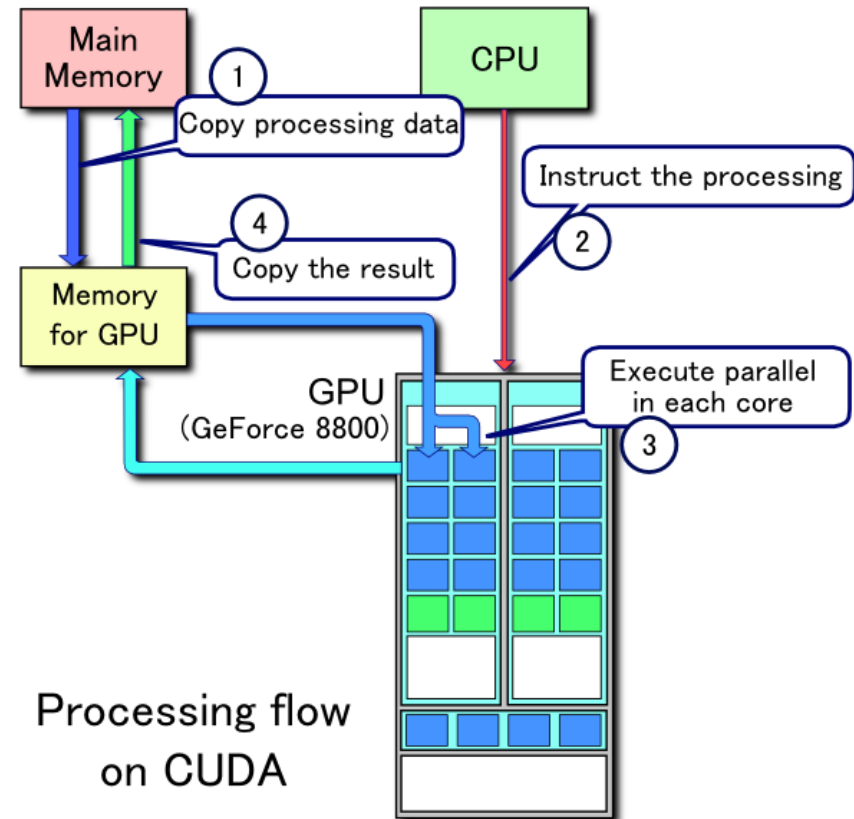


- Natural reflex : printf !
- No fully integrated solution : time consuming
 - Problems do not appear at smaller scale
 - Takes time to move the problems to a smaller size
 - Need to use multiple tools for multiple needs
 - More cores means more bugs...
 - ... And more debugging information available

What can we do... ?

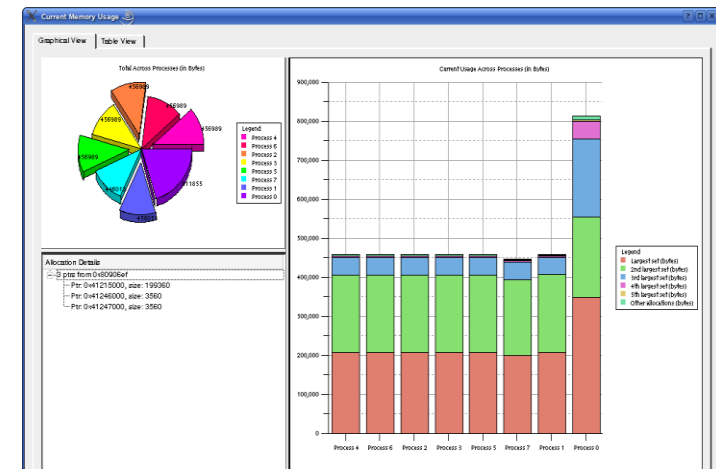
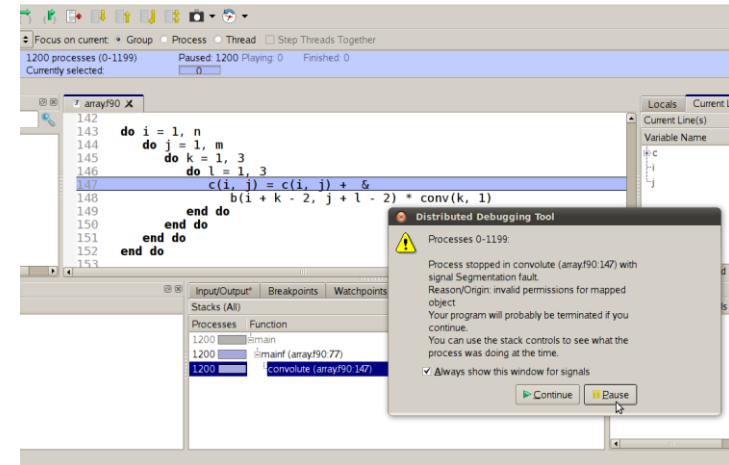
HPC's current challenge

- GPUs – a rival to traditional processors
 - AMD, Intel, ARM
 - NVIDIA, OpenCL, CUDA
- A big challenge for HPC developers
 - Data transfer
 - Several memory levels
 - Grid/block layout and thread scheduling
 - Synchronization
- New languages, compilers, potential standards



Allinea DDT In a Nutshell

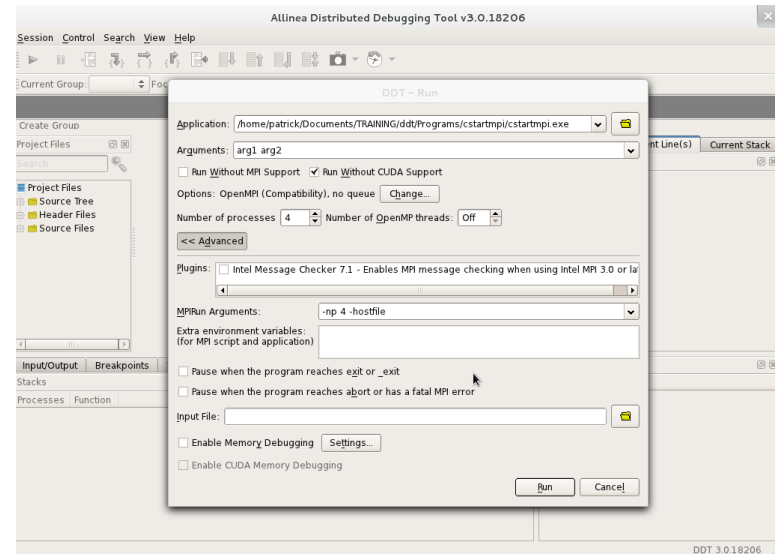
- Graphical debugger designed for:
 - C/C++, Fortran, UPC, CUDA
 - Multithreaded code
 - Single address space
 - Multiprocess code
 - Interdependent or independent processes
 - GPU codes
 - Hybrid software
 - Any mix of the above
- Managing concurrency
 - Emphasizing differences
 - Collective control
- Strong feature set
 - Memory debugging
 - Data analysis



Allinea DDT

Gather, Sort and Display information

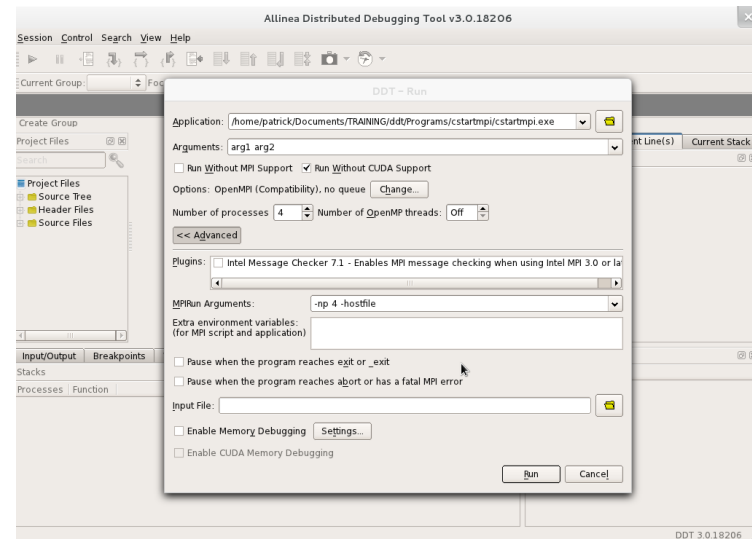
- User and administrator friendly
- Flow control
- Data monitoring
- Many environments



Allinea DDT

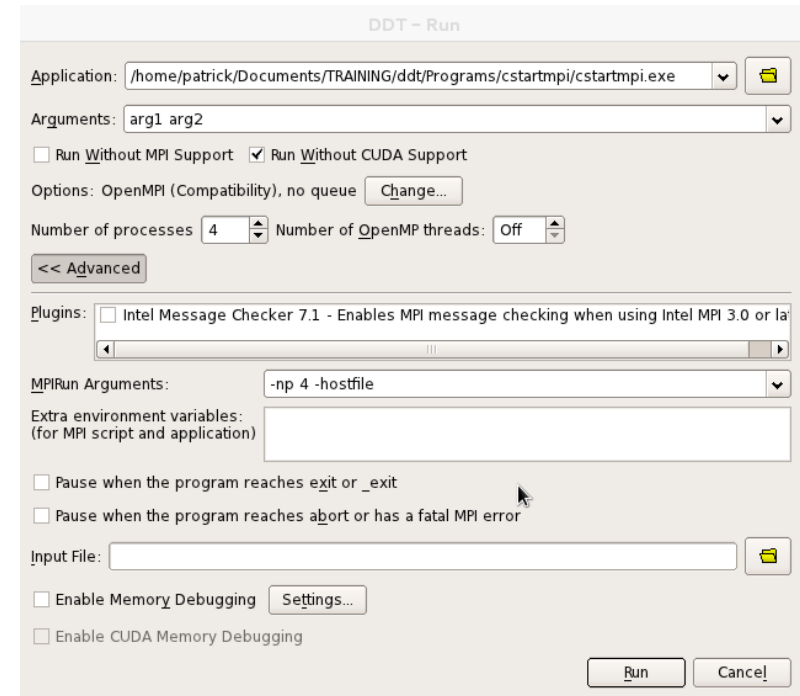
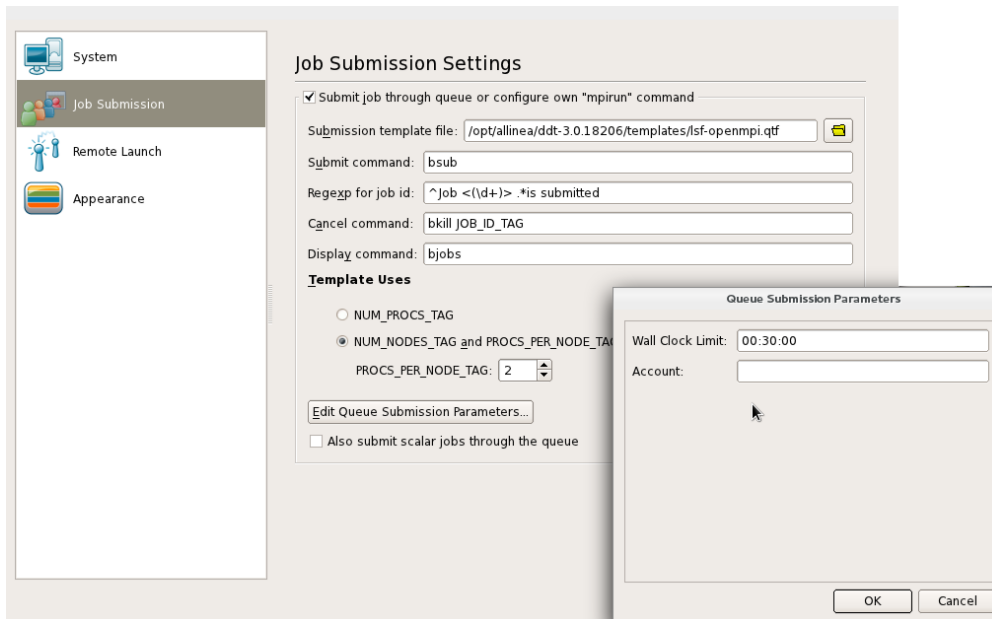
Gather, Sort and Display information

- **User and administrator friendly**
 - Get started easily
 - Fast, reliable, simple and intuitive GUI interface
 - Offline debugging
- Flow control
- Data monitoring
- Many environments



User and administrator friendly Getting started

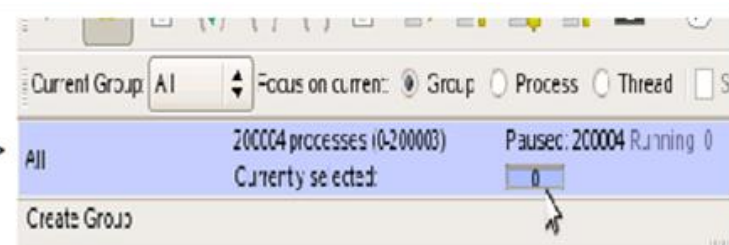
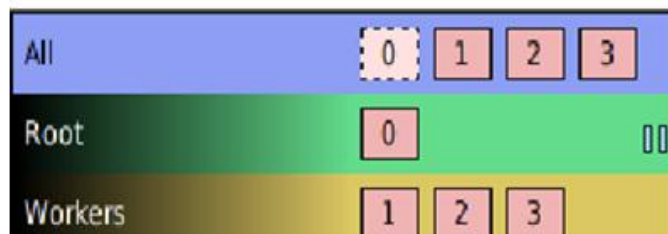
- Quick creation of runs
- Well integrated in workload schedulers
 - SLURM, LSF, SGE, PBS,.....



User and administrator friendly Intuitive GUI interface

- Intelligent GUI that adapts to the environment
 - From workstation to large scale clusters

| Stacks (All) | |
|--------------|---|
| Processes | Function |
| 150120 | _start |
| 150120 | __libc_start_main |
| 150120 | main |
| 150120 | pop (POP.f90:81) |
| 150120 | initialize_pop (initial.f90:119) |
| 150120 | init_communicate (communicate.f90:87) |
| 150119 | create_ocn_communicator (communicate.f90:300) |
| 1 | create_ocn_communicator (communicate.f90:303) |



User and administrator friendly Intuitive GUI interface

- Intelligent GUI that adapts to the environment
 - Even in CUDA environments

The screenshot shows a CUDA development IDE interface. At the top, there are radio buttons for 'Process', 'Thread', and 'Step Threads Together'. Below that, there are spinners for 'Block' and 'Thread' (each set to 0), a 'Go' button, and 'Grid size: 8x1' and 'Block size: 64x1x1'. The main editor shows a C++ file named 'prefix.cu' with the following code:

```
49 int x = threadIdx.x + blockIdx.x * BLOCK_SIZE;
50
51 if (x < length)
52     out[x] = in[x];
53
54 __syncthreads();
55
56 for ( int i = 1; i < BLOCK_SIZE; i <= 1)
57 {
58     if (threadIdx.x + i < BLOCK_SIZE && x + i < length)
59     {
60
61
62
63
64
65
66
67
68
```

A tooltip is displayed over line 58, containing the following text:

```
On this line:
1 Process: rank 0
1 Thread (Process 0): #2

512 GPU threads:
<<<(0,0),(0,0,0)>>> ... <<<(7,0),(63,0,0)>>> (512 threads)
```

On the right side, there is a 'Locals' panel with a table of variables:

| Variable Name | Value |
|---------------|----------|
| x | 0 |
| out | 0x100800 |
| length | 500 |
| in | 0x100000 |
| | 1 |

At the bottom right, the type is shown as '@register int'.

User and administrator friendly Offline debugging

- Using a workload scheduler
 - Machines are available when the scheduler decides (by night ?)
 - Can be tricky to get a big cluster exactly when the developer wants it
- Offline debugging : printf replacement
 - Tracepoints and offline debugging
 - Job runs without debugger interface and record variables
- Worlds first scalable batch debugger
 - Set tracepoints, breakpoints, and run !
 - Memory debugging errors, crashes
 - Reports in HTML or plain text

Allinea DDT Off-line Log

Messages Tracepoints Output

Messages

| Type | Time | Processes | Message |
|------|--------------|-----------|--|
| | 08:53:45.437 | n/a | Launching program /home/david/Work/HEAD/code/ddt/libexec/ddt.bin. |
| | 08:53:48.154 | n/a | DDT could not find valid debug information in one or more of your processes. Source files, local variables and other features may be unavailable or inaccurate. Please check you are using the correct debug interface and have compiled your code with debug information. |
| | 08:53:48.436 | 0 | Memory error detected in FcPatternReference from /usr/lib/x86_64-linux-gnu/libfontconfig.so.1. Thread 1 attempted to dereference a null pointer or execute an SSE instruction with an incorrectly aligned memory address. Tip: Use the stack list and the local variables to explore your program's current state and identify the source of the error. Stack List: Threads.Function 1. main 1. QApplication::QApplication 1. QApplicationPrivate::construct 1. QApplicationPrivate::x11_apply_settings 1. FcInit 1. FcInitLoadConfigAndFonts 1. FcInitLoadConfig 1. FcConfigParseAndLoad 1. XML_ParseBuffer 1. 0x00007ffff0808e6b 1. 0x00007ffff08072e2 1. 0x00007ffff080a77e 1. 0x00007ffff0809e5 1. FcConfigParseAndLoad 1. FcConfigParseAndLoad 1. FcConfigParseAndLoad 1. XML_ParseBuffer 1. 0x00007ffff0808e6b 1. 0x00007ffff08072e2 1. 0x00007ffff080a77e 1. 0x00007ffff0809e5 1. FcConfigParseAndLoad 1. FcConfigParseAndLoad 1. FcValueSave 1. FcValueSave 1. FcValueSave 1. FcPatternReference |
| | 08:53:48.477 | n/a | Every process in your program has terminated. |

Messages Tracepoints Output

Tracepoints

Allinea DDT and Offline Debugging Using the CLI

- General Options :

- offline <YYY> : activate offline mode
 - n <XXX> : the number of MPI processes
 - memory : enable memory debugging
 - ddtsession : the session file to use (if exists)

- Create breakpoints : -break-at LOCATION[,N:M:P]

Example:

```
$> ddt -offline myjob.html -break-at main.c:22,-:2:-,x myapp
```

- Create tracepoints : -trace-at LOCATION[,N:M:P],VAR1,VAR2...

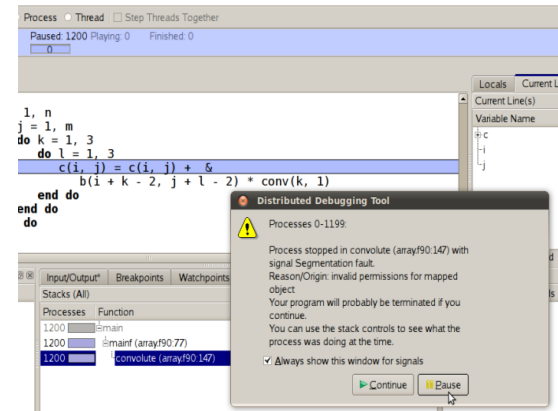
Example:

```
$> ddt -offline myjob.html -trace-at trisol.F90:14, -:-:3,x myapp
```

Allinea DDT

Gather, Sort and Display information

- User and administrator friendly
- Flow control
 - Static analysis
 - Control progress at scale
 - Understand deadlocks
 - Start investigation
- Data monitoring
- Many environments



Flow control

Static analysis

- Fix those errors before they bite !
- Static analysis
 - Integrated with cppcheck
 - Also includes ftncheck

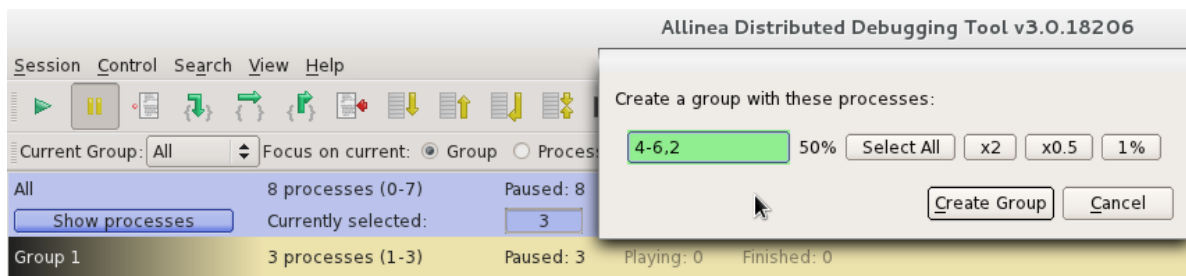
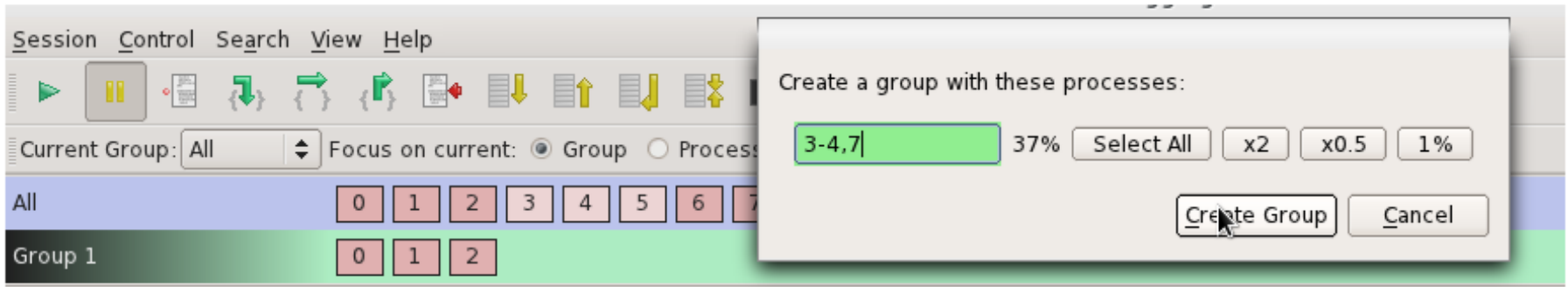
```
29
30 threads = calloc(sizeof(pthread_t), nthreads);
31 ids = calloc(sizeof(int), nthreads);
32
33 init_mutex();
34
35 pthread_mutex_lock(mutley);
36 for (i = 0; i < nthreads; ++i) {
37     ids[i] = i;
38     pthread_create (threads + i, NULL, &thread,
39 }
40 pthread_mutex_unlock(mutley);
41 for (i = 0; i < nthreads; ++i)
42     pthread_join (threads[i], NULL);
43
44 return 0;
45 }
46 void *q)
47
48
49 volatile int busy = 0;
50 volatile int locker = 0; /* to be amended by
51 int i, j;
52 double k = 1;
53 int tid = *(int*) q;
54
55 usleep(rand() % 31);
56
```

error Memory leak: threads
error Memory leak: ids

Flow control

Controlling progress at scale

- Bulk control is essential for multicore debugging
 - Group processes together
 - Play, step, reach breakpoints... Based on groups
 - Change interleaving order by stepping/playing selectively



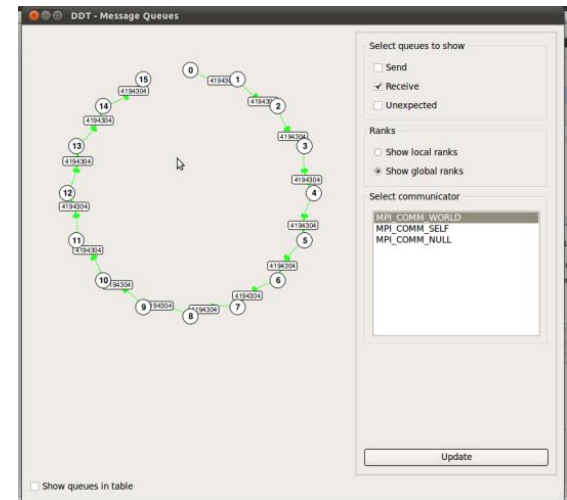
Flow control

Resolving MPI issues

```
10 MPI_Recv(&received, 1, MPI_INT, MPI_ANY_SOURCE, MPI_ANY_TAG, MPI_COMM_WORLD, &status);
11 void passItOn()
12 {
13     int from;
14     int to;
15     int *buffer;
16     MPI_Status status;
17     MPI_Recv(&received, 1, MPI_INT, MPI_ANY_SOURCE, MPI_ANY_TAG, MPI_COMM_WORLD, &status);
18     from = (myid + numprocs - 1) % numprocs;
19     to = (myid + 1) % numprocs;
20     buffer = malloc(sizeof(int)*BUFSIZE);
21
22     MPI_Recv(buffer, BUFSIZE, MPI_INT, from, 0, MPI_COMM_WORLD, &status);
23     received = received + 1;
24
25     send_buffer = malloc(sizeof(int)*BUFSIZE);
26     memcpy(send_buffer, buffer, sizeof(int)*BUFSIZE);
27     MPI_Send(send_buffer, BUFSIZE, MPI_INT, to, 0, MPI_COMM_WORLD);
28     free(send_buffer);
29 }
30
31
32 int main(int argc, char** argv)
33 {
34     MPI_Init(&argc, &argv);
35     MPI_Recv(&received, 1, MPI_INT, MPI_ANY_SOURCE, MPI_ANY_TAG, MPI_COMM_WORLD, &status);
```

- We can see messages
 - MPI standard exists for debugging message queues
- Integrates with MPI correctness tools
 - Check correctness of the messages
 - Intel MPI Checker ; Marmot

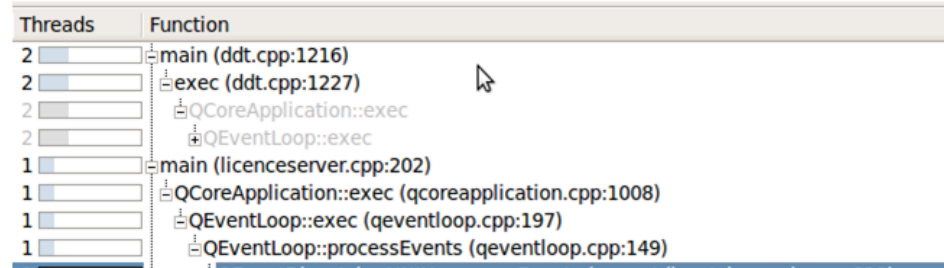
- Find cycles or blockage in message queue display
 - Parallel stacks
 - Variables
- More details than examining variables and processes alone



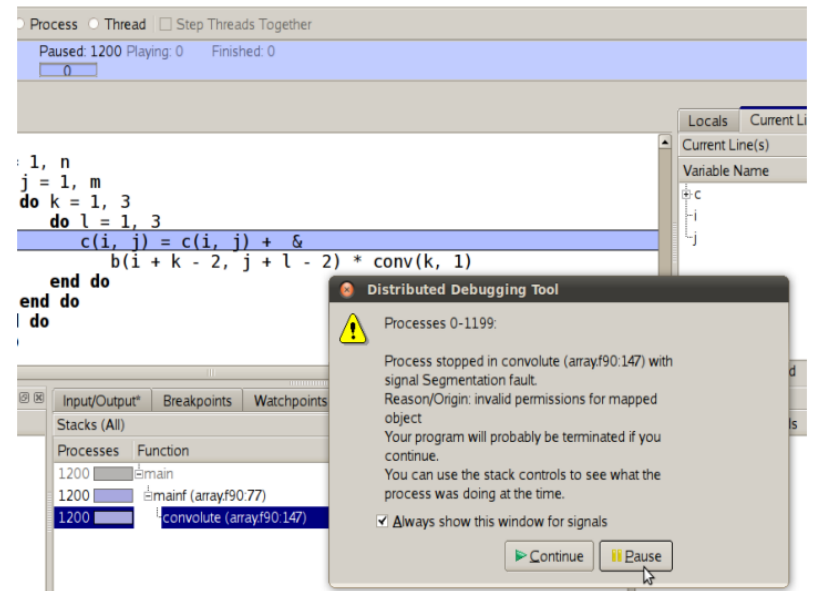
Flow control

Understanding what happens

- Application crashes
 - Threads/processes can be anywhere
 - Impossible to scroll through them individually
- Finding where processes crashed is essential
 - Allinea DDT merges stacks from processors and threads into a tree
 - Common faults patterns instantly evident
 - Divergence, deadlocks...
 - Information scalable without overload



| Threads | Function |
|---------|--|
| 2 | main (ddt.cpp:1216) |
| 2 | exec (ddt.cpp:1227) |
| 2 | QCoreApplication::exec |
| 2 | QEventLoop::exec |
| 1 | main (licenceserver.cpp:202) |
| 1 | QCoreApplication::exec (qcoreapplication.cpp:1008) |
| 1 | QEventLoop::exec (qeventloop.cpp:197) |
| 1 | QEventLoop::processEvents (qeventloop.cpp:149) |



Process Thread Step Threads Together
Paused: 1200 Playing: 0 Finished: 0

```
1, n  
j = 1, m  
do k = 1, 3  
do l = 1, 3  
c(i, j) = c(i, j) + &  
b(i + k - 2, j + l - 2) * conv(k, 1)  
end do  
end do  
do
```

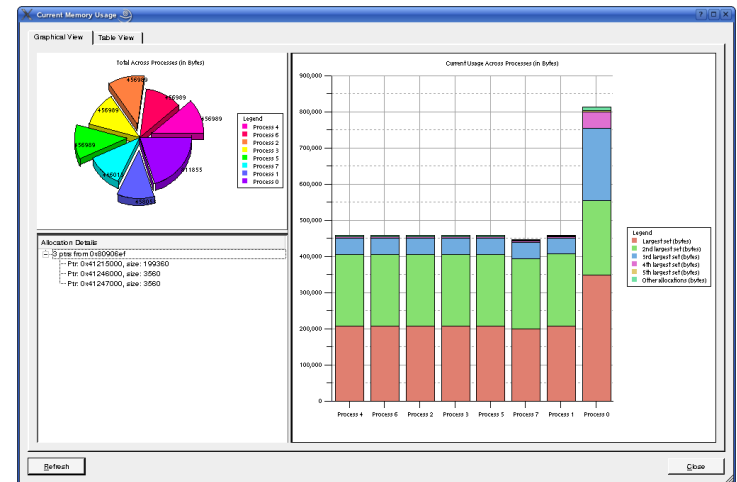
Stacks (All)
Processes Function
1200 main
1200 mainf (arrayf90.77)
1200 convolve (arrayf90.147)

Distributed Debugging Tool
Processes 0-1199:
Process stopped in convolve (arrayf90.147) with signal Segmentation fault.
Reason/Origin: invalid permissions for mapped object
Your program will probably be terminated if you continue.
You can use the stack controls to see what the process was doing at the time.
 Always show this window for signals
Continue Pause

Allinea DDT

Gather, Sort and Display information

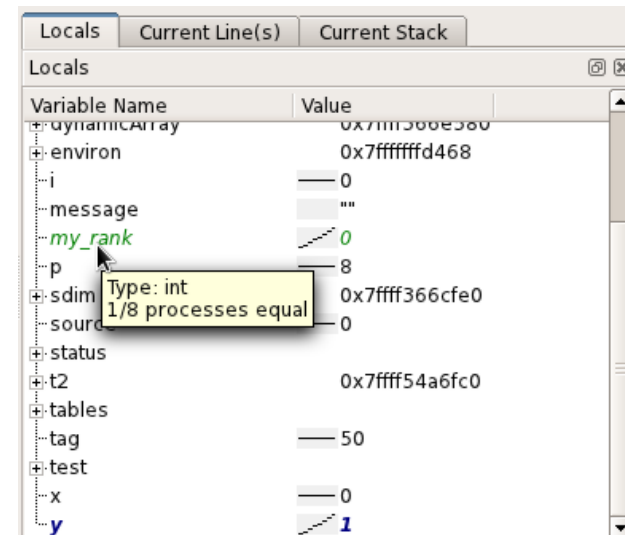
- User and administrator friendly
- Flow control
- Data monitoring
 - Monitor variables
 - Detect memory errors
 - Check the calculation data
- Many environments



Data monitoring

Monitor variables

- Developers need to see data
 - Too many variables to trawl them manually
 - Too many tasks or thread to display them at the same time
- Intelligent data management inside the debugger
 - Automatic monitoring of the data
 - Subtle highlights differences
 - Sparklines and smart display
- Even more detailed analysis if needed
 - Cross process comparison
 - Historical values of variables



Data monitoring

Smart displays : tracepoints

- “printf” is still mostly used but serious drawbacks:
 - Need to recompile the code
 - Information randomly printed on screen (depending on interleaving)
 - One line per process
- Scalable and advanced printf :
 - First step to reconcile printf and GUI
 - Information sorted by steps
 - Merged by groups of processes
 - No information overload
 - Possibility to filter the printed values
 - Save output for offline analysis

Tracepoint Output

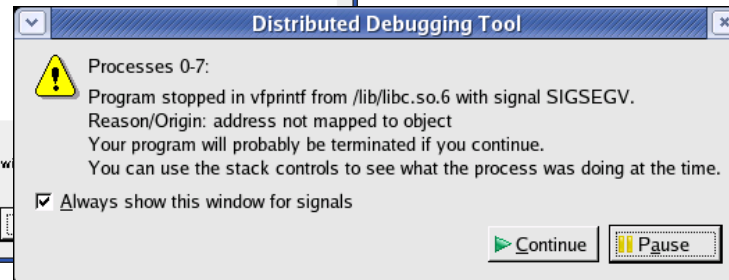
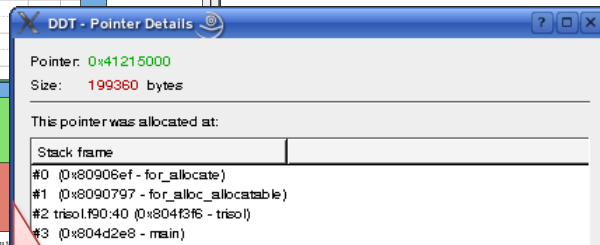
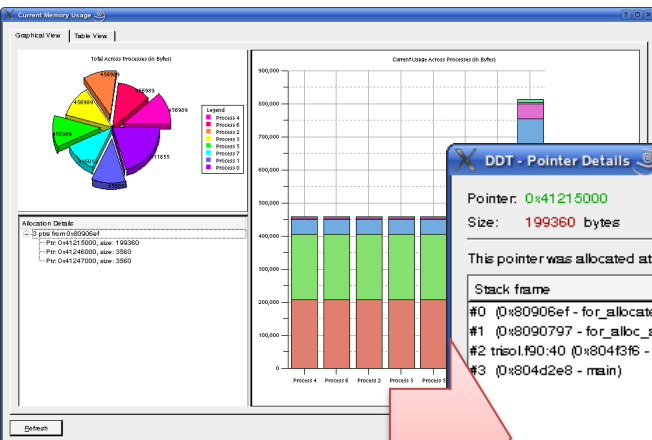
| Tracepoint | Processes | |
|----------------|--------------|---------|
| cstartmpi.c:98 | 4, ranks 0-3 | x: — 0 |
| cstartmpi.c:98 | 4, ranks 0-3 | x: — 10 |
| cstartmpi.c:98 | 4, ranks 0-3 | x: — 20 |
| cstartmpi.c:98 | 4, ranks 0-3 | x: — 30 |
| cstartmpi.c:98 | 4, ranks 0-3 | x: — 40 |
| cstartmpi.c:98 | 4, ranks 0-3 | x: — 50 |

Only show lines containing:

Data monitoring

Memory debugging

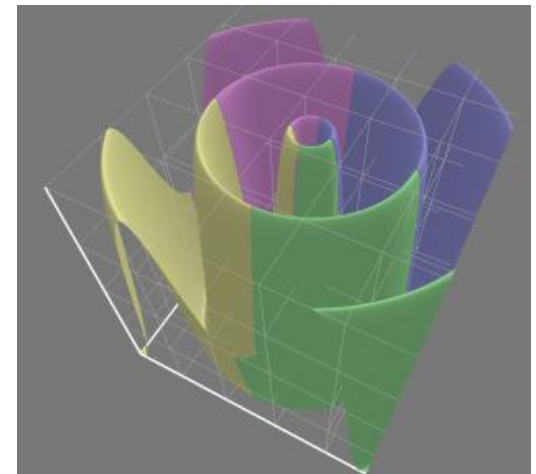
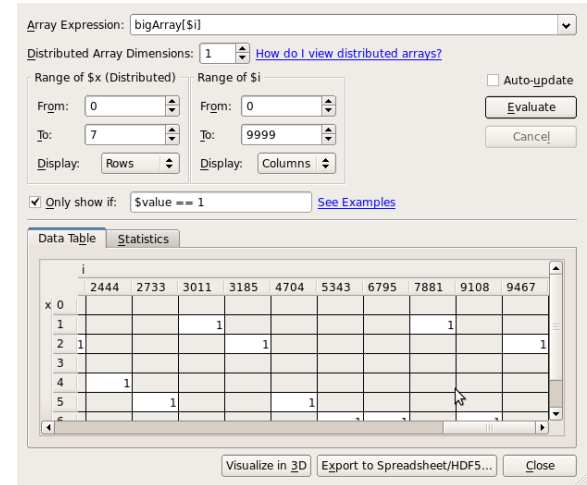
- Random errors are sneaky
 - Can't fix a bug that doesn't repeat
 - Often caused by memory issues
- Memory debugging can force the bug
 - Better to happen every time, than only during product demos
- Memory debugging :
 - Places agent between memory library and user process
 - Communicates problems to the debuggers
 - Monitors usage : detect memory leaks
 - Automatically protects ends of arrays
 - Trigger instant stop on touching memory
 - Many classes of errors can be checked
 - Also has CUDA support



Data monitoring

Searching haystacks

- Arrays are the building blocks of HPC
 - Largest jobs accumulate TB of data
 - Usually 2GB or 4GB per core
- Integrated visualization tool
 - Search data across all tasks or threads
 - Data displayed on a picture as requested by the user
 - Export at runtime



Data Monitoring Unified Parallel C

- New programming models
 - Support for Cray UPC
 - Support for Cray Co-Array Fortran

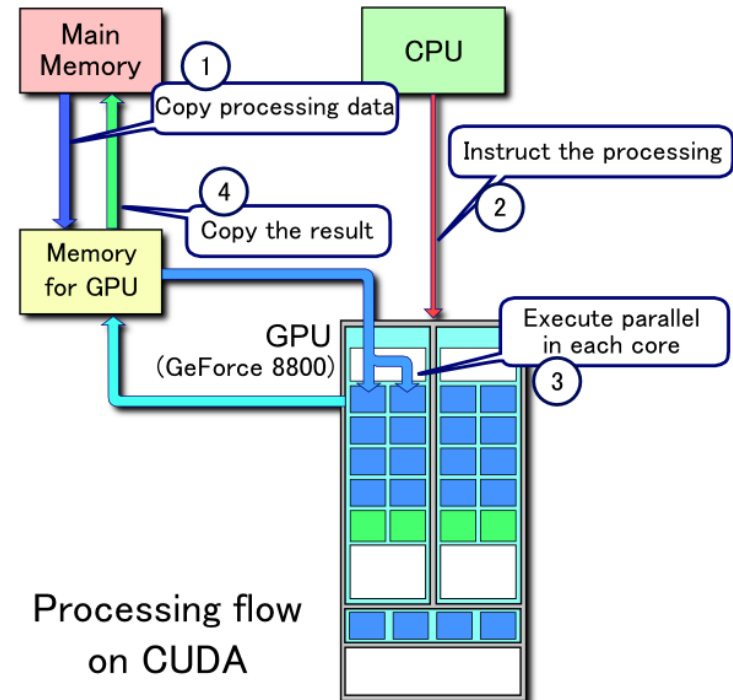
| Evaluate | Input/Output * | Breakpoints | Watchpoint |
|--------------------------------|---|-------------|------------|
| Evaluate | | | |
| Expression | Value | | |
| <code>&array[11999]</code> | <code>(shared [25] int *) (0x744dcc,0x77,0x18)</code> | | |

Type: shared [25] int *

Allinea DDT

Gather, Sort and Display information

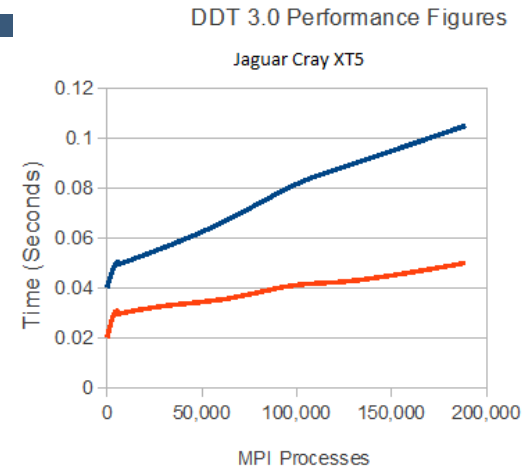
- User and administrator friendly
- Flow control
- Data monitoring
- Many environments
 - Debugging at scale
 - GPU Debugging



Allinea DDT and Scalability

Simplifies debugging for everyone

- At scale : new problems appear
 - Small and medium environments : multiple tools
 - More cores means more bugs...
 - ... And more debugging information available
- How can the design of Allinea DDT help your debugging whatever the size of your cluster (workstation or cluster) ?
- High-end architecture (in any Allinea DDT version)
 - Data consolidation (merged and sorted)
 - Highlight differences
 - Very low footprint



Allinea DDT and GPU

Successful add-on

- Built on vendors low level efforts
 - Nvidia cuda-gdb, compiler
 - Cray compiler
- Execution model is unusual
 - GUI supports blocks and grids
 - Support 32 thread units (warps)
- Mixed GPU/CPU in one interface
 - Interaction with CPUs
 - Easy to switch between contexts (stacks, threads, data...)
 - Support multiple nodes


The screenshot displays the Allinea DDT interface. At the top, it shows 'Current Group: All' and 'Focus on current: Group'. Below this, there are controls for 'Create Group' and 'CUDA Threads (Process 0, simpleMPIKernel)' with 'Block' and 'Thread' settings. The 'Project Files' pane on the left shows a tree view with 'simpleMPI.cu' selected. The main editor shows the source code of 'simpleMPI.cu', with lines 39-42 highlighted: `39 __global__ void simpleMPIKernel(float * input, float * output)`, `40 {`, `41 int tid = blockIdx.x * blockDim.x + threadIdx.x;`, and `42 output[tid] = sqrt(input[tid]);`. The 'Stacks (All)' pane at the bottom shows a table of execution stacks:

| Processes | Threads | GPU Thread | Function |
|-----------|---------|------------|-----------------------------------|
| 8 | 8 | 0 | main (simpleMPI.cpp:92) |
| 8 | 8 | 172032 | simpleMPIKernel (simpleMPI.cu:40) |
| 8 | 8 | 169984 | simpleMPIKernel (simpleMPI.cu:39) |
| 8 | 8 | 1792 | simpleMPIKernel (simpleMPI.cu:41) |
| 8 | 8 | 256 | simpleMPIKernel (simpleMPI.cu:42) |

The bottom status bar indicates '8 Processes: ranks 0-7'.


Many environments Directives support

```
5 # 29 "basic.c"
6 void hmpp_codelet_myFunc(int n, int A[n], const int B[n])
7 {
8 #pragma hmppcg gridify(i)
9 # 7 "<preprocessor>"
10 # 32 "basic.c"
11 for (int i = 0 ; i < n ; ++i)
12 {
13     A[i] += B[i] + 1;
14 }
15 }
16
17
18 /* end of extracted source code
19
```




On this line:
1 Process: rank 0
Kernel 1: 32 GPU threads
<<<(0,0),(0,0,0)>> ... <<<(0,0),(31,0,0)>> (32 threads)

```
1 program main
2 integer, parameter :: n = 1000
3 real, dimension(n) :: input
4 real, dimension(n) :: result
5 integer :: i
6 do i = 1,n
7 input(i) = i*4.0
8 enddo
9 !$omp acc_region
10 !$omp acc loop
11 do i = 1,n
12 result(i) = input(i) * 4.0
13 enddo
14 !$omp
15 !$omp
16 print
17 end program
```



On this line:
1 Process: rank 0
Kernel 1: 32 GPU threads
<<<(0,0),(0,0,0)>> ... <<<(0,0),(31,0,0)>> (32 threads)

```
#pragma acc region
{
for( i = 0; i < n; ++i ) r[i] = a[i]*2.0f;
}
/* compute on the host to compare */
for( i = 0; i < n; ++i ) e[i] = a[
/* check the results */
for( i = 0; i < n; ++i )
assert( r[i] == e[i] );
```



On this line:
1 Process: rank 0
1 Thread (Process 0): #1

- Wide range of partnerships
 - Support the environments you use for hybrid development
 - You swiftly benefit from the latest updates
 - [Read our latest white papers available with PGI and CAPS](#)

Many environments

Directives support - OpenACC

PGI[®]



CRAY
CAPS

- OpenACC : New parallel programming standard led by
 - CAPS Enterprise, Cray, Nvidia and PGI
- Fully supported within Alinea DDT
 - CAPS and PGI : still adapting their software to the new standard
 - Same mechanisms and features for CUDA programs
 - First implementation to be supported: Cray OpenACC
- You benefit from our partnerships

Summary



- Bugs happen at all stage of the development and during the entire life of the code
- An intelligent debugger is now available to fix bugs quickly
 - Other methods have limited success and issues at scale
 - Intelligence makes debugging easier and faster for the developer
- Allinea DDT scales in both performance and interface
 - Breaking all records and making problems now manageable

allinea



Leaders in parallel software development tools

Thank you

Your contacts :

- Technical Support team :
- Sales team :

support@allinea.com

sales@allinea.com

www.allinea.com