

Debugging at Scale on the Blue Gene/Q

Mira Performance Boot Camp
May 21, 2014

Ray Loy
Applications Performance Engineering
ALCF

Outline

- `bgq_stack`
- `coreprocessor`
- `gdb`
- Allinea DDT



Interpreting your job's .error

```
...
<Jan 23 06:54:21.008467> FE_MPI (Info) : Starting job 535016
<Jan 23 06:54:21.085232> FE_MPI (Info) : Waiting for job to terminate
<Jan 23 06:54:23.515642> BE_MPI (Info) : IO - Threads initialized
<Jan 23 06:54:23.537559> BE_MPI (Info) : I/O input runner thread terminated
<Jan 23 06:54:33.589347> BE_MPI (Info) : I/O output runner thread terminated
<Jan 23 06:54:33.644169> BE_MPI (Info) : Job 535016 switched to state TERMINATED
('T')
<Jan 23 06:54:33.644228> BE_MPI (Info) : Job successfully terminated - TERMINATE
D ('T')
<Jan 23 06:54:33.719443> FE_MPI (Info) : Job terminated normally
<Jan 23 06:54:33.719541> FE_MPI (Info) : exit status = (139)
<Jan 23 06:54:33.719747> BE_MPI (Info) : Starting cleanup sequence
<Jan 23 06:54:33.719788> BE_MPI (Info) : cleanupDatabase() - job already termina
ted / hasn't been added
<Jan 23 06:54:33.750097> BE_MPI (ERROR): The error message in the job record is
as follows:
<Jan 23 06:54:33.750147> BE_MPI (ERROR): "killed with signal 11"
<Jan 23 06:54:33.762069> BE_MPI (Info) : Destroying partition ANL-R00-M1-N04-64
<Jan 23 06:55:08.913586> BE_MPI (Info) : Partition ANL-R00-M1-N04-64 switched to
state FREE ('F')
<Jan 23 06:55:09.162052> FE_MPI (Info) : == FE completed ==
<Jan 23 06:55:09.162126> FE_MPI (Info) : == Exit status: 0 ==
<Jan 23 06:55:09.162203> SCHED_IF (Info) : mpirun result code: 0
<Jan 23 06:55:09.164395> SCHED_IF (Info) : job result code: 139
<Jan 23 06:55:09.184948> SCHED_IF (Info) : boot failure: False
<Jan 23 06:55:09.314332> SCHED_IF (Info) : resources associated with partition A
NL-R00-M1-N04-64 have been released
<Jan 23 06:55:09.334075> SCHED_IF (Info) : scheduler library unloaded
```



Lightweight core files

- Look for files
 - core.0, core.1, etc.
- Lightweight core files
 - One for each rank that failed
 - Contain stack backtrace in *address* form
 - Decode to symbolic (useful!) form
- Environment settings for core files
 - <https://www.alcf.anl.gov/resource-guides/vesta-debugging-core-files>



Lightweight Core File Example

```
+++PARALLEL TOOLS CONSORTIUM LIGHTWEIGHT COREFILE FORMAT version 1.0
+++LCB 1.0
Program: /gpfs/home/rloy/public/winter-workshop-2012/hellompi
Job ID : 535016
Personality:
  XYZT coordinates : 0,0,0,0
  MPI Rank       : 0
  DDR Size (MB)  : 2048
  Mode          : SMP
+++ID TGID 100, Core 0, Thread 1 State 40000000, Sched: 48000000
General Purpose Registers:
  r00=00000078 r01=02100ee0 r02=021087a0 r03=00000000 r04=02100ee0 r05=02101388 r06=021087a0 r07=02101388
[...]
Special Purpose Registers:
  lr=0117f6e8 cr=04002022 xer=00000000 ctr=00000000
[...]
Floating Point Registers
  f0=00000000 00000000 00000000 00000000 f1=00000000 00000000 00000000 00000000
[...]
Memory:
  Stack top   : 0x00000000
[...]
+++STACK
0x011c1434
---STACK
---ID
+++ID TGID 100, Core 0, Thread 5 State 00000000, Sched: 48000000 Running
[...]
```



Decoding Lightweight Core Files

- `bgq_stack [optional_exename] [corefile]`

```
-----  
+++ID TGID 100, Core 0, Thread 5 State 00000000, Sched: 48000000 Running
```

```
0x01001534  
foo  
/gpfs/home/rloy/public/winter-workshop-2012/hellompi.c:33
```

```
0x01001704  
main  
/gpfs/home/rloy/public/winter-workshop-2012/hellompi.c:88
```

```
0x01194a44  
generic_start_main  
../csu/libc-start.c:231
```

```
0x01194cb8  
__libc_start_main  
../sysdeps/unix/sysv/linux/powerpc/libc-start.c:137
```

```
0xffffffffc  
??  
?:0
```



Decoding Lightweight Core Files (2)

- What's this other stuff? (MPI threads)

+++ID TGID 100, Core 0, Thread 1 State 40000000, Sched: 48000000

0x011c1434
clone
?:0

+++ID TGID 100, Core 1, Thread 2 State 40000000, Sched: 08000000

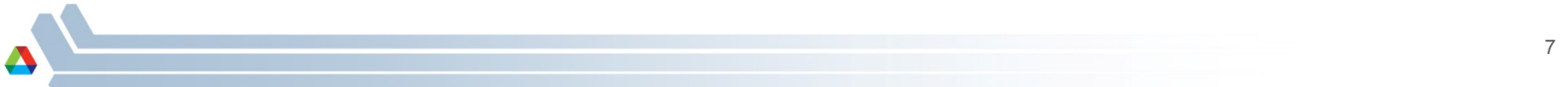
0x011c1434
clone
?:0

+++ID TGID 100, Core 2, Thread 3 State 40000000, Sched: 08000000

0x011c1434
clone
?:0

+++ID TGID 100, Core 3, Thread 4 State 40000000, Sched: 08000000

0x011c1434
clone
?:0



Outline

- bgq_stack
- coreprocessor
- gdb
- Allinea DDT



coreprocessor

- Useful when you have a large set of core files
 - Shows symbolic backtrace
 - Groups ranks that aborted in the same location together
 - *Can also attach to a running job to take snapshot*
- Location
 - BG/Q: coreprocessor.pl is in your default PATH
 - Attaching to running job does **not** require administrator
 - `coreprocessor -nogui -snapshot=<filename> -j=<jobid>`
- Scalability limit
 - Maximum 32K ranks
- Instructions:
 - BG/Q Application Developer Redbook
 - <http://www.redbooks.ibm.com/redpieces/abstracts/sg247948.html>



coreprocessor window

```
File  Control  Analyze  Filter  Sessions

Group Mode:  Stack Traceback (condensed)  Session 1 (MMC)

0 : Compute Node (128)
1 :   0xffffffffc (128)
2 :     __libc_start_main (32)
3 :       generic_start_main (32)
4 :         main (16)
5 :           Allgather (16)
6 :             PMPI_Allgather (16)
7 :               MPIDO_Allgather (8)
8 :                 MPIDO_Allreduce (8)
9 :                   MPID_Progress_wait (1)
10:                     DCMF_CriticalSection_cycle (1)
9 :                   MPID_Progress_wait (7)
10:                     DCMF_Messenger_advance (1)
11:                       DCMF::Queueing::Lockbox::Device::advance() (1)
10:                     DCMF_Messenger_advance (1)
11:                       DCMF::Queueing::Tree::Device::advance() (1)
10:                     DCMF_Messenger_advance (5)
11:                       DCMF::DMA::Device::advance() (2)
12:                         DCMF::DMA::RecFifoGroup::advance() (2)
13:                           DMA_RecFifoSimplePollNormalFifoById (2)
11:                         DCMF::DMA::Device::advance() (3)
7 :                   MPIDO_Allgather (8)
8 :                     MPIDO_Allreduce (8)
9 :                       MPIR_Allreduce (8)
10:                         MPIC_Sendrecv (8)
11:                           MPID_Progress_wait (8)
12:                             DCMF_Messenger_advance (8)
13:                               DCMF::Queueing::GI::Device::advance() (1)
13:                               DCMF::DMA::Device::advance() (3)
14:                               DCMF::DMA::RecFifoGroup::advance() (3)
15:                                 DMA_RecFifoSimplePollNormalFifoById (3)
```



Outline

- bgq_stack
- coreprocessor
- **gdb**
- Allinea DDT



gdb

- Can connect single gdb client to single rank of your job
- BG/Q
 - Limitations of CDTI (Common Debug and Tool Interface)
 - Each instance of gdb client counts as a “debug tool”
 - Only 4 tools may be connected to a job
 - Therefore at most 4 ranks can be examined
- Start a debug session using ***qsub -I*** (interactive job)
 - `qsub -I -q default -t 30 -n 64`
 - See Redbook for more info on starting gdb with runjob
- gdb can also load a compute-node **binary** corefile
- Generally a parallel debugger (e.g. DDT) will be more useful



Outline

- bgq_stack
- coreprocessor
- gdb
- Allinea DDT



Allinea DDT

- Licensing
 - 64K-process permanent license
 - Full machine development license available (contact support)
- Startup overview
 - Compile `-g -O0`
 - OMP code compile `-qsmp=omp:noauto:noopt`
 - Softenv key “+ddt”
- More details:
 - <http://www.alcf.anl.gov/user-guides/allinea-ddt>



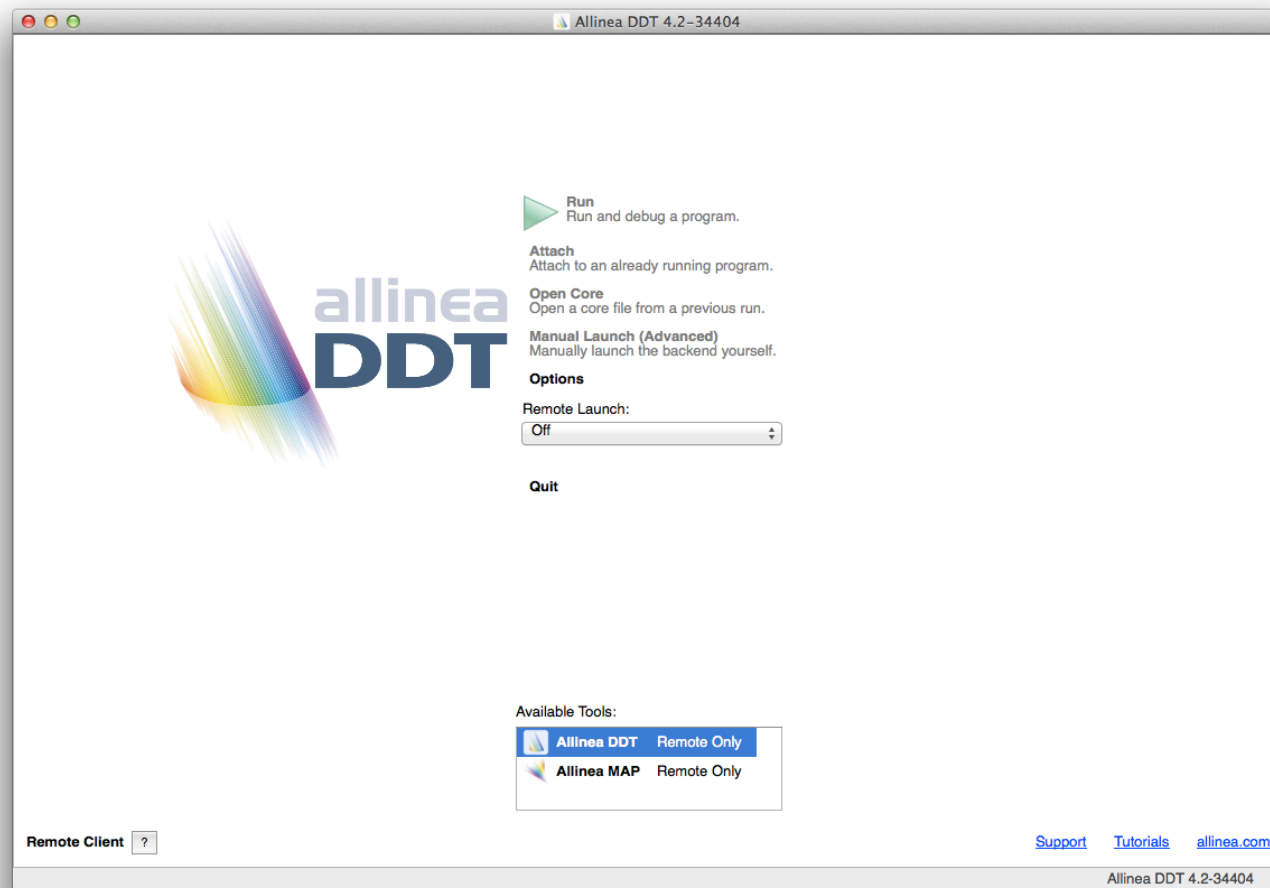
Allinea DDT startup

- Run using remote client (RECOMMENDED)
 - Download and install Mac or Windows "Remote client" from <http://www.allinea.com/products/download-allinea-ddt-and-allinea-map>
 - Optional: use ssh master mode so you only need log in once per session
 - Note: supported on Mac OS/X; not supported in Windows <= XP (? for >XP)
 - ~/.ssh/config
 - ControlMaster auto
 - ControlPath ~/.ssh/master-%r@%h:%p
- Run from login node
 - Need X11 server on your laptop and ssh -X forwarding
 - Run ddt and let it submit job through GUI



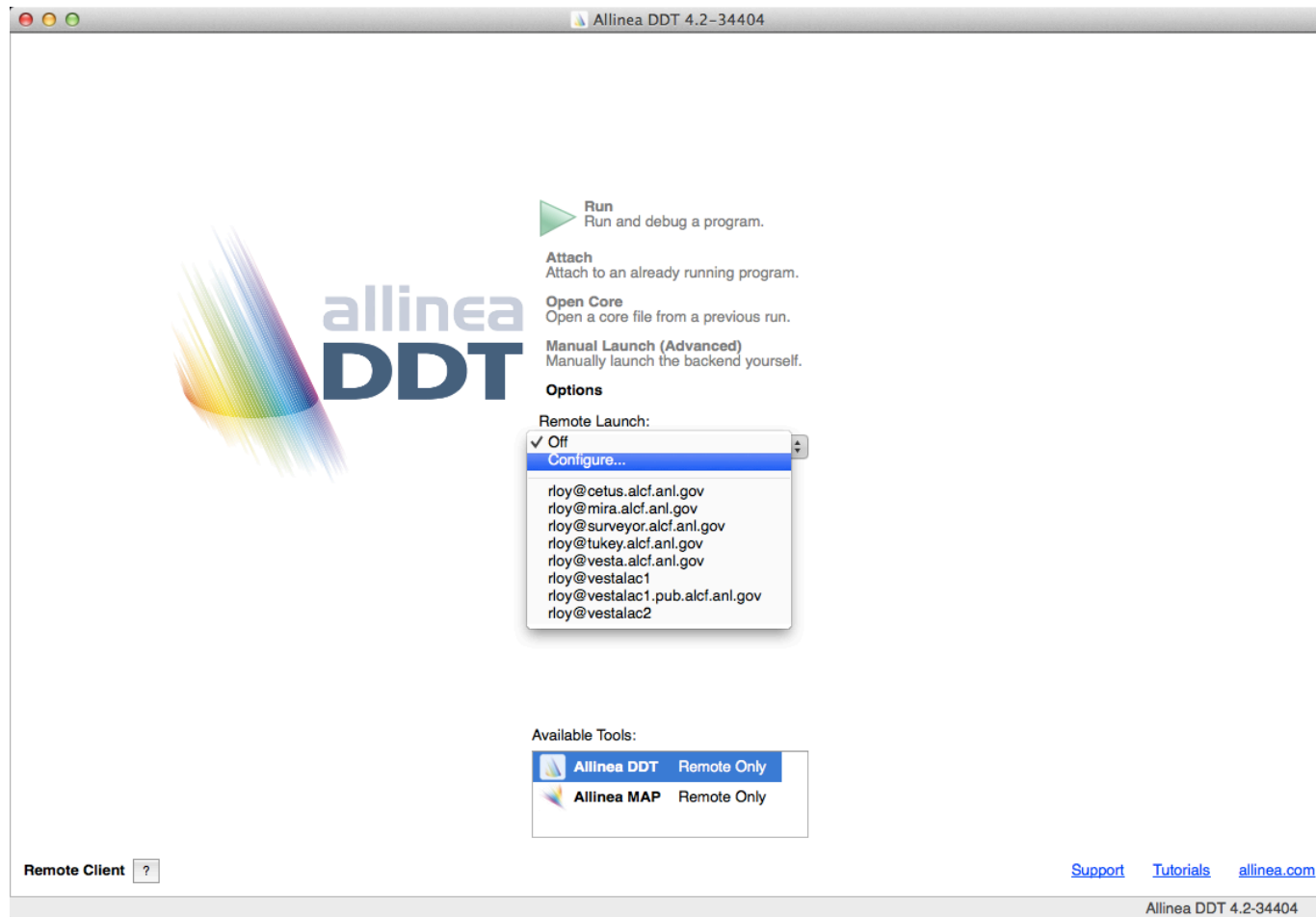
DDT Remote Client (0)

GUI looks just like the regular version



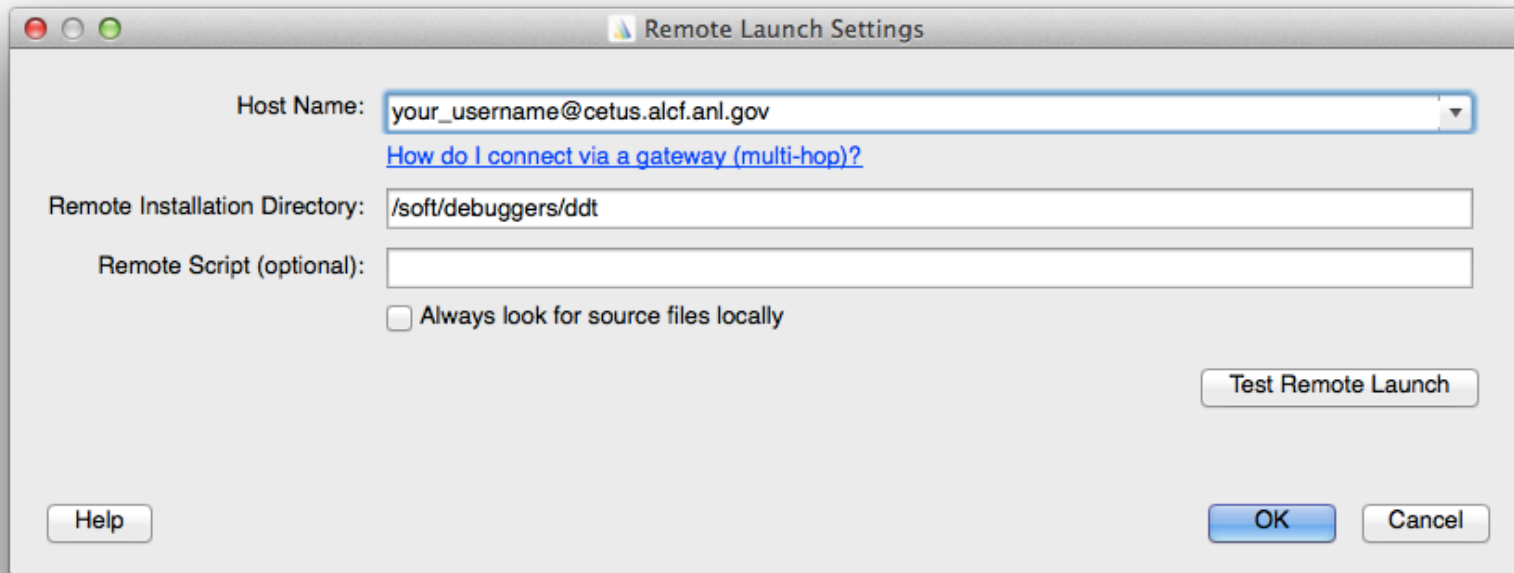
DDT Remote Client (1)

Select "configure" to add a new remote host



DDT Remote Client (2)

Note: this remote installation directory is the default version of DDT, corresponding to +ddt
Click "Test Remote Launch" to verify



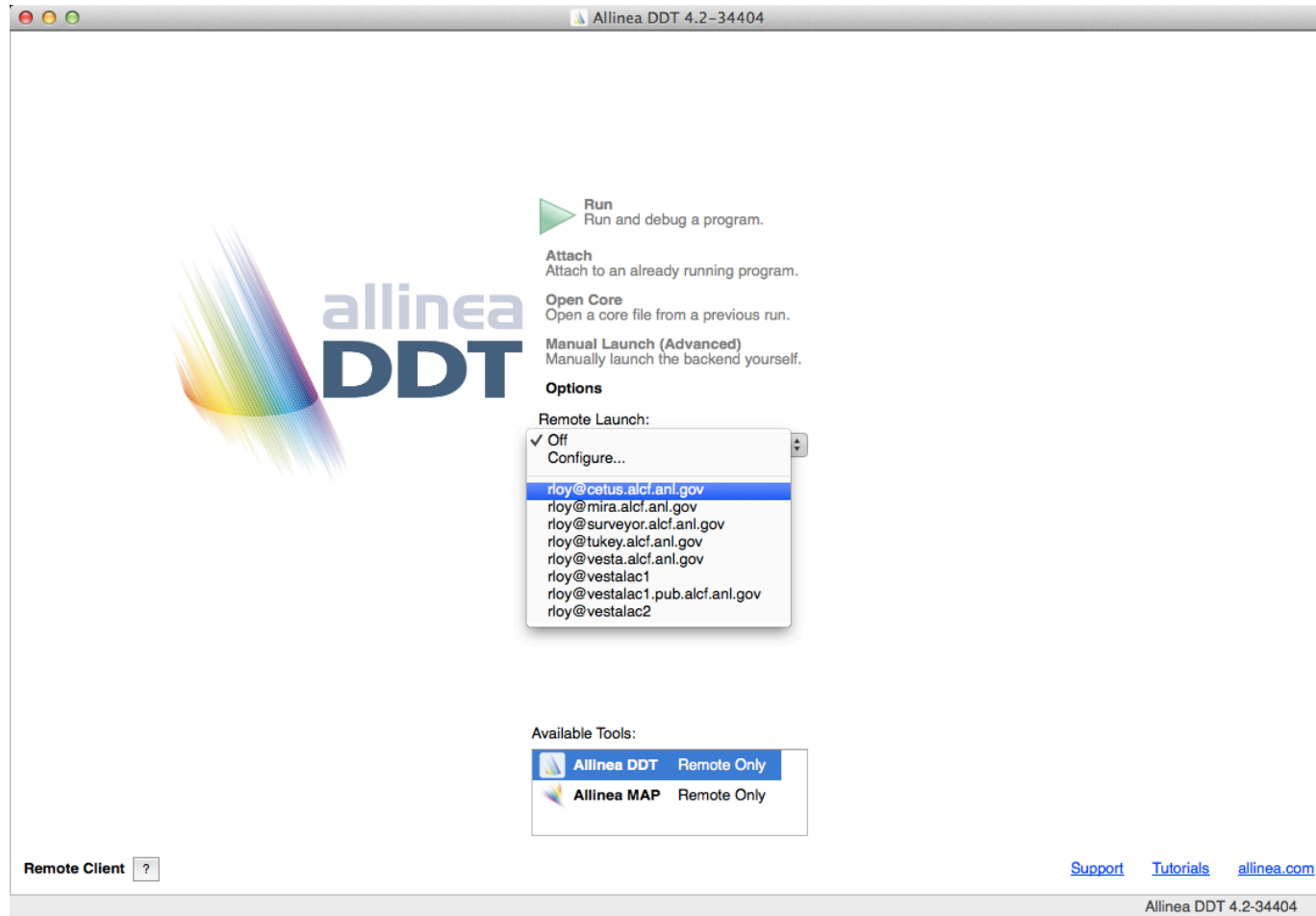
A screenshot of a macOS-style dialog box titled "Remote Launch Settings". The dialog has a light gray background and standard window controls (red, yellow, green buttons) in the top-left corner. It contains the following fields and controls:

- Host Name:** A text field with the value "your_username@cetus.alcf.anl.gov". Below it is a blue hyperlink that reads "[How do I connect via a gateway \(multi-hop\)?](#)".
- Remote Installation Directory:** A text field containing the path "/soft/debuggers/ddt".
- Remote Script (optional):** An empty text field.
- Always look for source files locally:** An unchecked checkbox.
- Buttons:** At the bottom, there are three buttons: "Help" on the left, and "Test Remote Launch", "OK", and "Cancel" on the right.



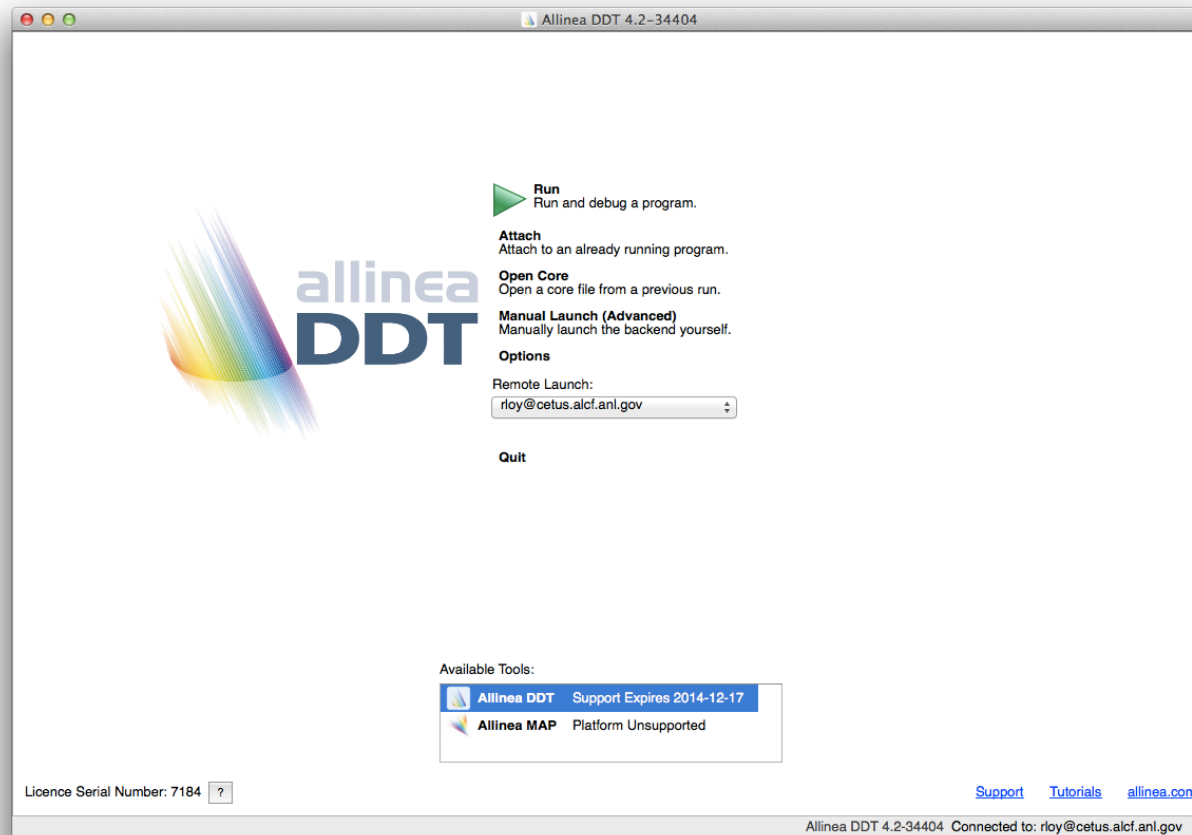
DDT Remote Client (3)

Now that it is defined, select remote machine



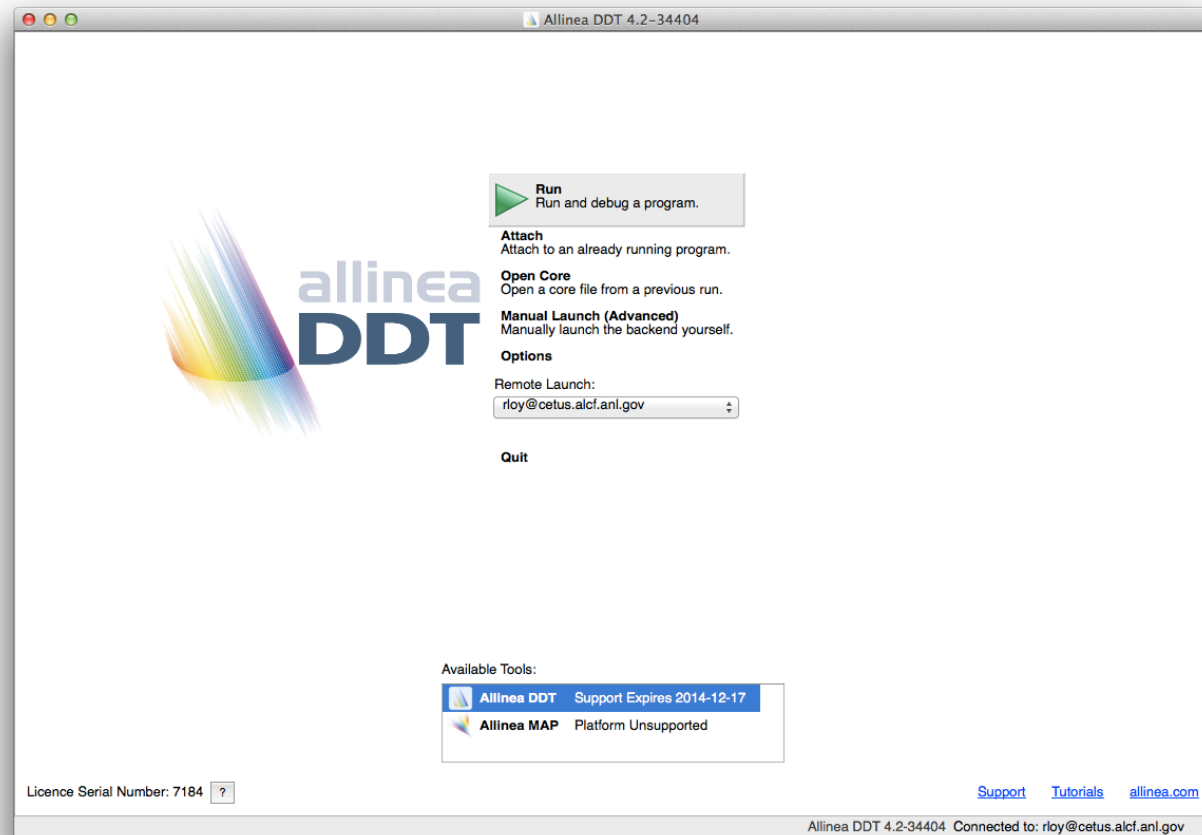
DDT (4)

Connected (note License info in lower left corner)
From this point, remote GUI works same as local



DDT (5)

Click "Run" to start a debugging session

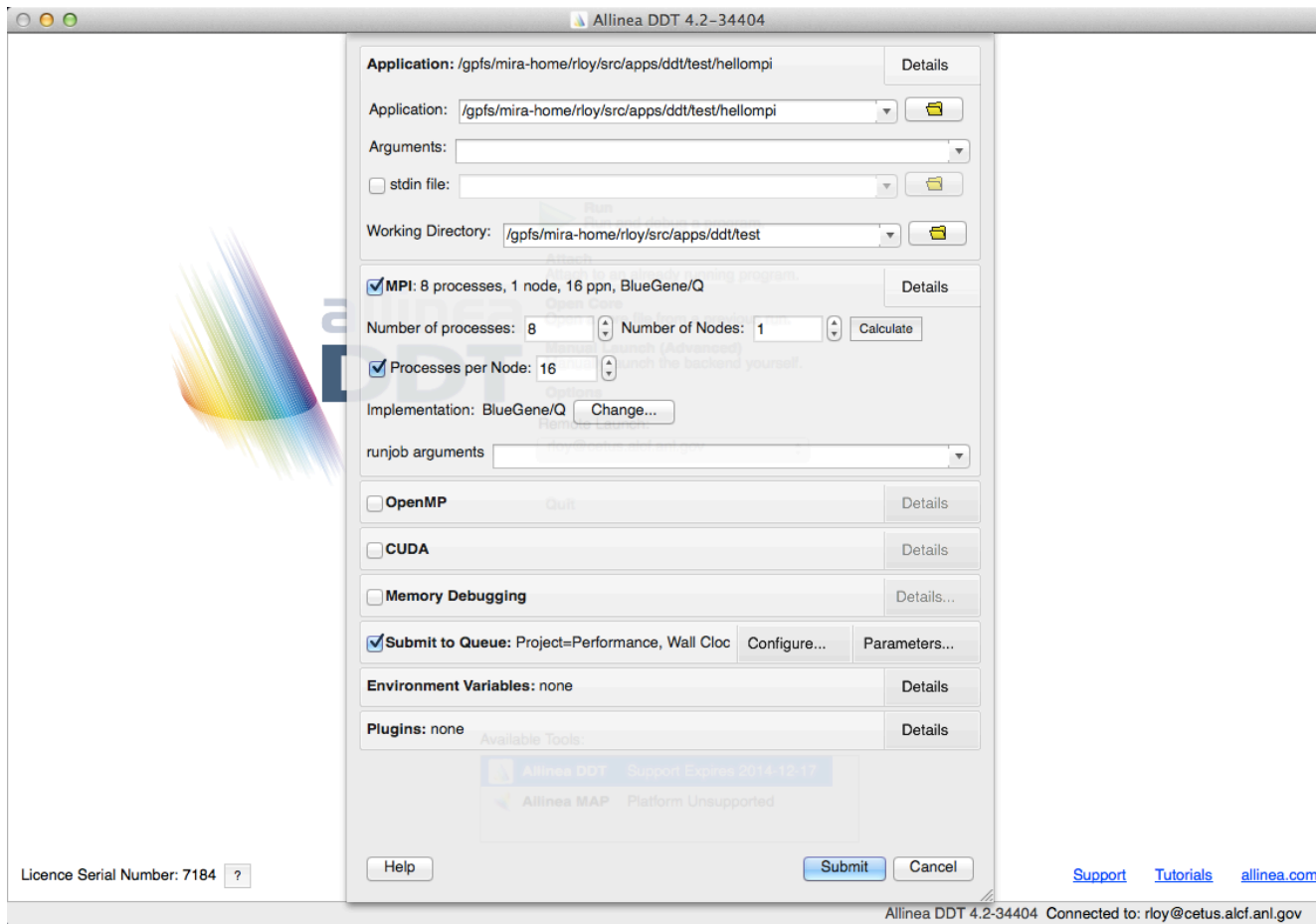


DDT (6)

Remember to set working directory

Important! Enable the checkbox "Submit to Queue"

- click "Configure" and "Parameters" for additional settings



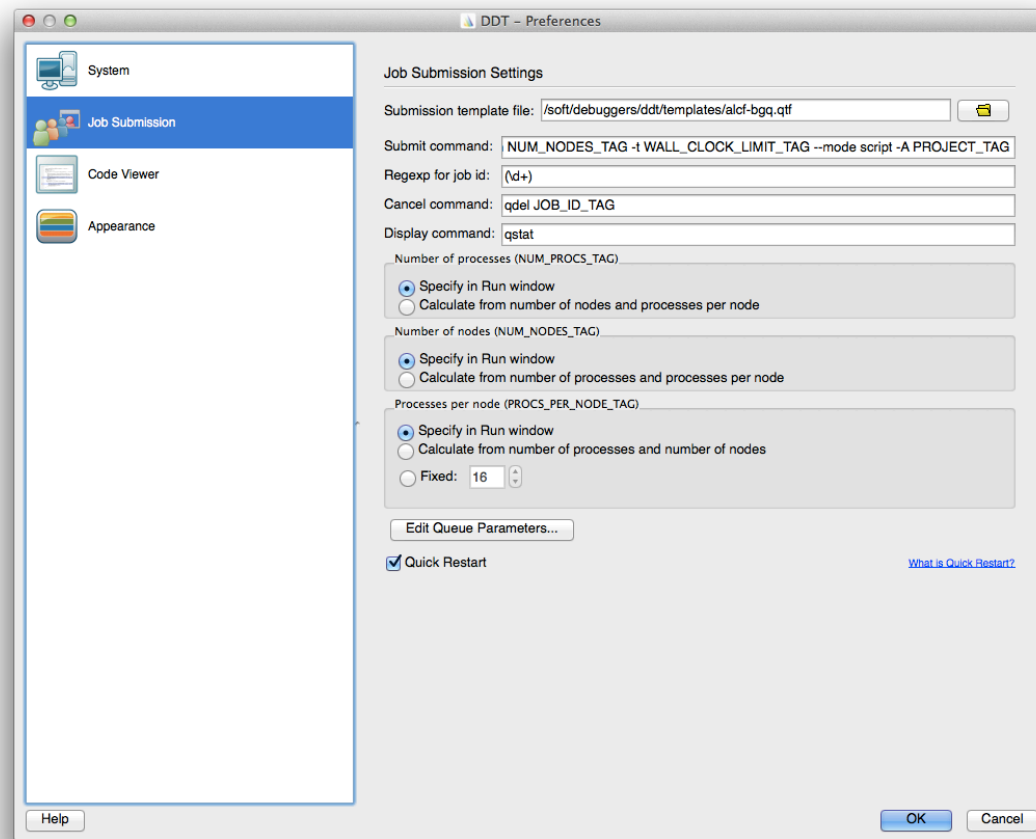
The screenshot shows the Alinea DDT 4.2-34404 application window. The interface is divided into several sections for configuring a job submission. The top section includes fields for Application, Arguments, stdin file, and Working Directory. Below this, there are checkboxes for MPI, OpenMP, CUDA, and Memory Debugging. The MPI section is expanded, showing settings for Number of processes (8), Number of Nodes (1), and Processes per Node (16). The Submit to Queue checkbox is checked, and the Project is set to Performance, Wall Cloc. The Environment Variables and Plugins sections are also visible. At the bottom, there is a License Serial Number field and a Help button. The status bar at the bottom right indicates the connection to roly@cetus.alcf.anl.gov.

Application: /gpts/mira-home/roy/src/apps/ddt/test/hellompi
Arguments:
stdin file:
Working Directory: /gpts/mira-home/roy/src/apps/ddt/test
MPI: 8 processes, 1 node, 16 ppn, BlueGene/Q
Number of processes: 8 Number of Nodes: 1 Calculate
Processes per Node: 16
Implementation: BlueGene/Q Change...
runjob arguments
OpenMP
CUDA
Memory Debugging
Submit to Queue: Project=Performance, Wall Cloc Configure... Parameters...
Environment Variables: none
Plugins: none
Available Tools:
Alinea DDT Support Expires 2014-12-17
Alinea MAP Platform Unsupported
Licence Serial Number: 7184 ?
Help Submit Cancel
Support Tutorials allinea.com
Alinea DDT 4.2-34404 Connected to: roly@cetus.alcf.anl.gov

DDT (6.1)

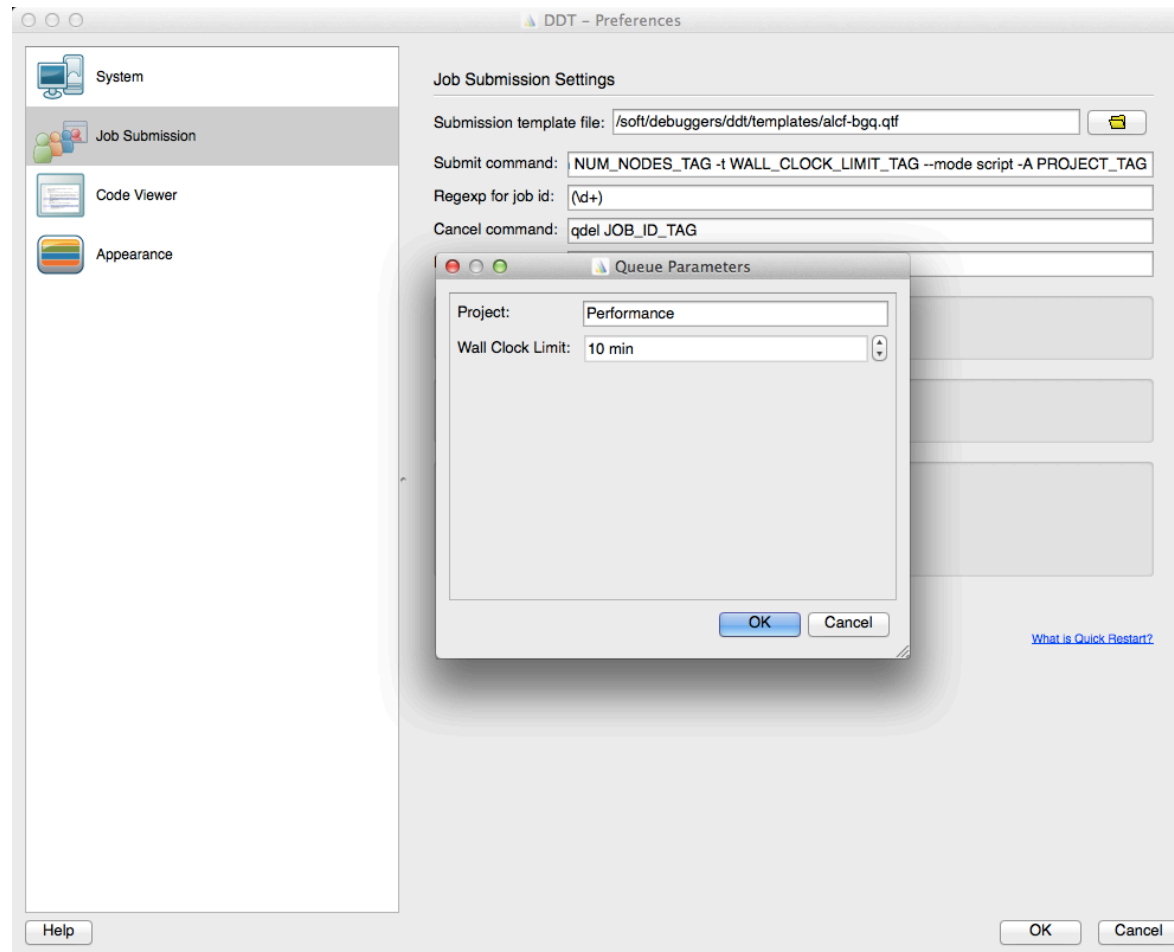
Job submission tab

Use submission template: `/soft/debuggers/ddt/templates/alcf-bgq.qtf`



DDT (6.2)

Remember to set your project



DDT (7)

Job must go through queue

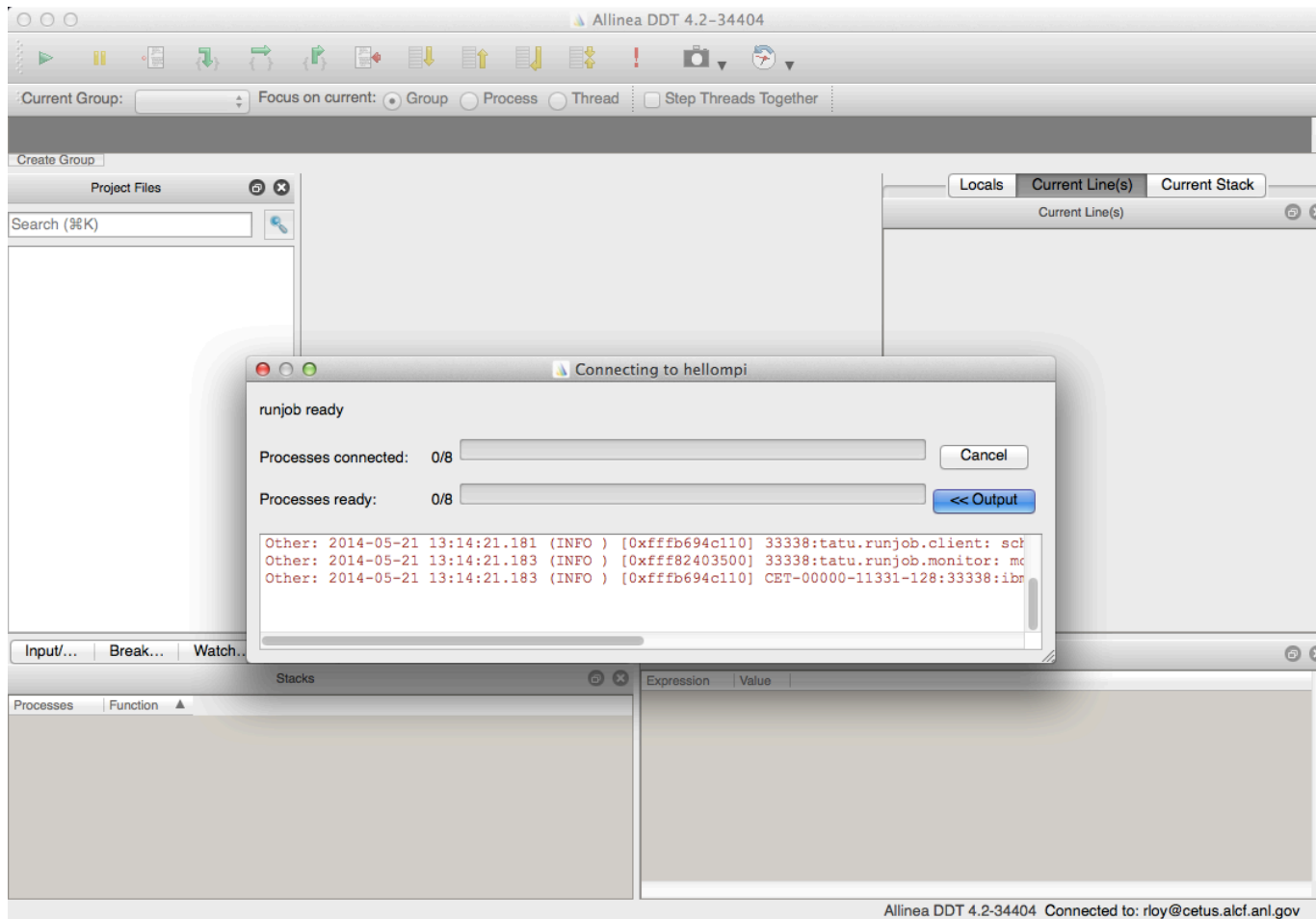
The screenshot displays the Allinea DDT 4.2-34404 application window. A modal dialog box is open in the center, titled "Allinea DDT 4.2-34404". It contains a message: "Your job has been submitted to the queue. Allinea DDT will continue automatically once the job has been started." Below this message is a table showing the status of various jobs. The table has columns: JobID, User, WallTime, Nodes, State, and Location. The job with ID 267163, user rloy, is highlighted in blue and is in the "starting" state. Below the table, it says "Waiting for job '267163' to start..." with "Help" and "Cancel" buttons. The background interface shows a sidebar with "Current Group:", "Create Group", "Project Files", and a "Search (%K)" field. The main area has tabs for "Locals", "Current Line(s)", and "Current Stack". At the bottom, there are tabs for "Input/...", "Break...", "Watch...", "Stacks", "Trace...", "Tracepoint...", and "Logbook". The "Stacks" tab is active, showing a "Processes" list and a "Function" list. To the right of the "Stacks" tab is an "Evaluate" panel with "Expression" and "Value" fields. The status bar at the bottom right indicates "Allinea DDT 4.2-34404 Connected to: rloy@cetus.alcf.anl.gov".

JobID	User	WallTime	Nodes	State	Location
261856	aksenova	00:15:00	4	user_hold	None
264656	mati	01:00:00	512	user_hold	None
264657	mati	01:00:00	512	user_hold	None
266872	rrahaman	02:00:00	4	queued	None
266899	zhong	01:00:00	256	queued	None
266906	bshipman	01:00:00	2	queued	None
267144	baip	01:00:00	512	running	CET-00040-33371-512
267159	sameer	01:00:00	4	queued	None
267160	sameer	01:00:00	4	queued	None
267163	rloy	00:10:00	1	starting	CET-00000-11331-128



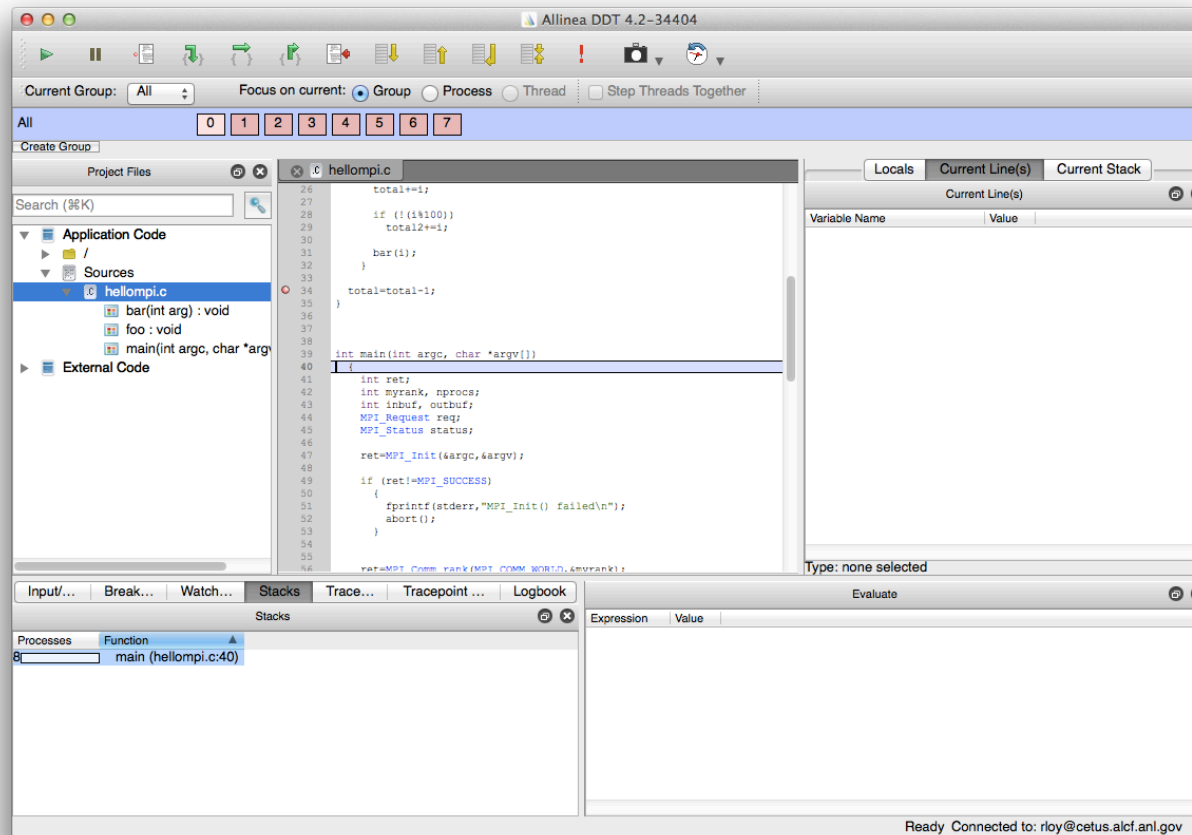
DDT (8)

When job starts running, connection status will show



DDT (9)

Ready to debug!



Questions

- See also
 - <http://www.alcf.anl.gov/user-guides/mira-cetus-vesta>
 - support@alcf.anl.gov

