

Eclipse PTP Support for UT TACC Stampede Progress Report

Brian Watt, briandwatt@gmail.com

Doug James, UT TACC, djames@tacc.utexas.edu

September 13, 2013

Agenda

- UT TACC Stampede
- SLURM sbatch Support
- Stampede Target System Configuration
- Module Support
- Inject Commands into Batch Script
- System Monitoring View
- Acknowledgements

UT TACC Stampede - 1

- Dell PowerEdge C8220 Cluster with Intel Xeon Phi Many Integrated Core (MIC) CoProcessors
 - Cluster contains 6,400+ Dell Zeus C8220 Nodes
 - Typical node consists of
 - Two Xeon Intel 8-Core 64-bit E5-processors w/ 32GB
 - for a total of 2.2 PF
 - One or two 61-core Xeon Phi MIC coprocessor w/ 8GB – for an additional 7+ PF
 - Specialty nodes – 1TB large mem, GPU, etc.
- CentOS 6.3 OS
- SLURM 2.4 w/ mods
- Intel Compilers & Libs



UT TACC Stampede - 2

- Different ways to run on MIC coprocessor
 - Native – run on MIC (serial, MPI, OpenMP)
 - Offload – from host offload work to MIC
 - Symmetric – MPI across one or more hosts and MICs
- Initial Eclipse PTP support for MIC provides
 - Native and Offload
 - Future – Symmetric (dual executable launch)

SLURM sbatch Support

- Based upon slurm-generic-batch XML file in `org.eclipse.ptp.rm.jaxb.configs`
- Noticed that it did NOT contain all sbatch command line arguments. Why? So added
 - For example, `acctg_freq`, `clusters`, `comment`, `constraint`, `contiguous`, `cores_per_socket`, `cpu_bind`, `cpus_per_task`, `distribution`, `exclude`, `exclusive`, `export_file`, `extra_node_info`, `gres`, `hint`, `input`, `licenses`, `mem`, `mem_per_cpu`, `mem_bind`, `mincpus`, `network`, `nodefile`, `nodelist`, `no_kill`, `no_requeue`, `ntasks`, `ntasks_per_core`, `ntasks_per_node`, `ntasks_per_socket`, `overcommit`, `propagate`, `qos`, `requeue`, `share`, `sockets_per_node`, `switches`, `threads_per_core`, `time_min`, `tmp`, `uid`, `wait_all_nodes`, `wckey`
- Wrote Bug 416962 - Update slurm-generic-batch with all document sbatch command line arguments
 - There are multiple SLURM XML files for ALPS, BGP, BGQ, and generic - only change generic one
 - A future enhancement – UT TACC not interested in just generic SLURM

Stampede Environment

- Fast moving infrastructure
 - Intel software stack – Use Intel Eclipse Plugins
 - Support both mvapich2 and Intel MPI libraries
 - Propagating inherited environment to MIC
 - Symmetric MPI – `ibrun.symm -c <host exec> -m <MIC exec>`
- Custom SLURM batch XML file – `edu.utexas.tacc.stampede.batch` in `org.eclipse.ptp.rm.jaxb.configs`
 - Customize which sbatch arguments needed
 - No `srun` instead replaced with `ibrun`
 - Explicitly set environment variables -
`MIC_OMP_NUM_THREADS`, `MIC_PPN`, and
`OMP_NUM_THREADS`
 - Add Module support (LMOD) as GUI control

Stampede Target System Configuration

- Wrote Bug 412925 - UT Ranger decommissioned - remove `edu.utexas.tacc.ranger.sge.batch.xml` – DONE, checked into master
- Wrote Bug 412926 - Add UT Stampede - add `edu.utexas.tacc.stampede.slurm.batch.xml` – Work in progress, checked into master
- After adding all missing sbatch command line arguments, reviewed with Doug James and other TACC personnel
- Upon review needed to adjust the generic SLURM batch XML file
 - Some definitely worked and were primary - 10
 - Some definitely worked and were secondary - 7
 - Some might work and were secondary - 12
 - Some were not supported and should be removed! - 27
 - Some were discouraged/conflict and should be removed! - 5
 - Some were uncertain and should be removed! – 9
- Can NOT use generic SLURM batch, but has to reorganized and simplified it

Basic Settings

The screenshot shows the Eclipse IDE's Run Configurations dialog for a Parallel Application. The configuration is named 'hello_world2' and is set to run on a remote system 'stampede.tacc.utexas.edu'. The 'Basic Settings' tab is active, showing various parameters for the job submission.

Name: hello_world2

Target System Configuration: edu.utexas.tacc.stampede.slurm.batch

Connection Type: Remote (stampede.tacc.utexas.edu)

Name	Value	Description
MPI Command:	ibrun	Prefix command for an MPI program (e.g. "ibrun ./a.out") or blank for non-MPI program.
Job Name:	ptp_sbatch	Job name for the submission. (-j, --job-name=<jobname>)
Output File:		The filename pattern of standard output. (-o, --output=<filename_pattern>)
Error File:		Filename pattern of standard error. (-e, --error=<filename_pattern>)
Queue:	development	Partition for the resource allocation. (-p, --partition=<partition_names>)
Number of Nodes:	1	Number of nodes to be allocated to this job. (-N, --nodes=<minnodes[-maxnodes]>)
Number of Tasks:	1	Number of tasks to run. (-n, --ntasks=<number>)
Wallclock Time:	00:05:00	Limit on the total run time of the job allocation. (-t, --time=<time>)
Account:		Account to charge resources used by this job. (-A, --account=<account>)
Mail User:		User to receive email notification of state changes. (--mail_user=<user>)
Mail Type:		Event type to notify user by email. (--mail-type=<type>)

Buttons: View Script, View Configuration, Restore Defaults, Apply, Revert, Close, Run

Supplemental Settings

Run Configurations

Create, manage, and run configurations
Create a configuration to launch a parallel application

Name: hello_world

Target System Configuration: edu.utexas.tacc.stampede.slurm.batch

Connection Type: Local Remote stampede.tacc.utexas.edu

Basic Settings Supplemental Settings OMP Environment MIC Environment Advanced Settings Import SLURM Script

Name	Value	Description
Begin:	<input type="text"/>	Defer the allocation of the job until the specified time. (--begin<time>)
Comment:	<input type="text"/>	An arbitrary comment. (--comment<string>)
Dependency:	<input type="text"/>	Defer the start of this job until the specified dependencies have been satisfied completed. (-d, --dependency=<dependency_list>)
Exclude Node List:	<input type="text"/>	Specify what nodes to exclude. (-x, --exclude=<node name list>)
Hold:	<input type="checkbox"/>	Specify the job is to be submitted in a held state (priority of zero). (-H, --hold)
Immediate:	<input type="checkbox"/>	The batch script is only submitted to the controller if the resources necessary to grant its job allocation are immediately available. (-I, --immediate)
Input:	<input type="text"/>	Specify filename pattern of batch script's standard input. (-i, --input=<filename_pattern>)
Node File:	<input type="text"/>	Specify filename containing nodes. (-F, --nodefile=<node file>)
Node Name List:	<input type="text"/>	Request a specific list of node names. (-w, --odelist=<node name list>)
MPI Ranks per Node:	<input type="text"/>	Request the maximum ntasks be invoked on each node. (--ntasks-per-node=<ntasks>)
Open Mode:	<input type="text"/>	Open the output and error files using append or truncate mode as specified. (--open_mode=append truncate)
Quiet:	<input type="checkbox"/>	Suppress informational messages from sbatch. (-Q, --quiet)
Reservation:	<input type="text"/>	Allocate resources for the job from the named reservation. (--reservation=<name>)
Time Minimum:	<input type="text"/>	Set a minimum time limit on the job allocation. (--time-min=<time>)
Verbose:	<input type="checkbox"/>	Increase the verbosity of sbatch's informational messages. (-v, --verbose)
Working Directory:	<input type="text" value="\$\${ptp_rm:directory#value}"/> <input type="button" value="Browse"/>	Set the working directory of the batch script to "directory" before it it executed. (-D, --workdir=<directory>)

View Script View Configuration Restore Defaults

Filter matched 24 of 24 items

Apply Revert

Close Run

Environment

The screenshot shows the Eclipse IDE's Run Configurations dialog for a Parallel Application named "hello_world2". The dialog is titled "Run Configurations" and has a subtitle "Create, manage, and run configurations". Below the subtitle is the instruction "Create a configuration to launch a parallel application".

The left sidebar shows a tree view of configurations under "Parallel Application", with "hello_world2" selected. The main area has tabs for "Resources", "Application", "Arguments", "Environment", "Synchronize", and "Common". The "Environment" tab is active, showing the "Target System Configuration" set to "edu.utexas.tacc.stampede.slurm.batch" and the "Connection Type" set to "Remote" with the host "stampede.tacc.utexas.edu".

Below the connection type, there are tabs for "Basic Settings", "Supplemental Settings", "Environment", "Advanced Settings", and "Import SLURM Script". The "Environment" tab is active, displaying a table of environment variables:

Name	Value	Description
Modules to Load:	<input type="button" value="Configure..."/>	Modules that are loaded.
OMP_NUM_THREADS:	<input type="text"/>	Number of OpenMP threads. (export OMP_NUM_THREADS=<n>)
MIC_PPN:	<input type="text"/>	Number of MPI tasks per MIC. (export MIC_PPN=<n>)
MIC_OMP_NUM_THREADS:	<input type="text"/>	Number of MIC OpenMP threads per task. (export MIC_OMP_NUM_THREADS=<n>)

At the bottom of the "Environment" tab, there are three buttons: "View Script", "View Configuration", and "Restore Defaults".

At the bottom right of the dialog, there are buttons for "Apply", "Revert", "Close", and "Run".

Advanced Settings

Run Configurations

Create, manage, and run configurations
Create a configuration to launch a parallel application

Name: hello_world

Resources Application Arguments Environment Synchronize Common

Target System Configuration: edu.utexas.tacc.stampede.slurm.batch

Connection Type
 Local Remote stampede.tacc.utexas.edu

Basic Settings Supplemental Settings OMP Environment MIC Environment **Advanced Settings** Import SLURM Script

Show Only Checked Items

Name	Value	Description
<input checked="" type="checkbox"/> MIC_OMP_NUM_THREADS		Number of MIC OpenMP threads per task. (export MI
<input checked="" type="checkbox"/> MIC_PPN		Number of MPI tasks per MIC. (export MIC_PPN=<n>
<input checked="" type="checkbox"/> OMP_NUM_THREADS		Number of OpenMP threads. (export OMP_NUM_THR
<input checked="" type="checkbox"/> account		Charge resources used by this job to specified accou
<input checked="" type="checkbox"/> begin		Defer the allocation of the job until the specified tim
<input checked="" type="checkbox"/> comment		An arbitrary comment. (--comment<string>)
<input checked="" type="checkbox"/> dependency		Defer the start of this job until the specified depend
<input checked="" type="checkbox"/> error		Instruct SLURM to connect the batch script's standar
<input checked="" type="checkbox"/> exclude		Specify what nodes to exclude. (-x, --exclude=<no
<input checked="" type="checkbox"/> hold	false	Specify the job is to be submitted in a held state (pri
<input checked="" type="checkbox"/> immediate	false	The batch script is only submitted to the controller i
<input checked="" type="checkbox"/> input		Specify filename pattern of batch script's standard ir
<input checked="" type="checkbox"/> job_name	ptp_sbatch	Specify a name for the job allocation. (-j, --job-nam
<input checked="" type="checkbox"/> mail_type		Notify user by email when certain event types occur.
<input checked="" type="checkbox"/> mail_user		User to receive email notification of state changes. (-
<input checked="" type="checkbox"/> nodefile		Specify filename containing nodes. (-F, --nodefile=.
<input checked="" type="checkbox"/> nodelist		Request a specific list of node names. (-w, --nodeli
<input checked="" type="checkbox"/> nodes	1	Request that a minimum of "minnodes" nodes be all
<input checked="" type="checkbox"/> ntasks	1	Specify the number of tasks to run. (-n, --ntasks=<
<input checked="" type="checkbox"/> ntasks_per_node		Request the maximum ntasks be invoked on each ne
<input checked="" type="checkbox"/> open_mode		Open the output and error files using append or tru
<input checked="" type="checkbox"/> other		Other Options.
<input checked="" type="checkbox"/> output		Instruct SLURM to connect the batch script's standar
<input checked="" type="checkbox"/> partition	development	Request a specific partition for the resource allocat
<input checked="" type="checkbox"/> quiet	false	Suppress informational messages from sbatch. (-Q, .
<input checked="" type="checkbox"/> reservation		Allocate resources for the job from the named reser

View Script View Configuration Restore Defaults

Apply Revert

Close Run

Filter matched 24 of 24 items

Import SLURM Script

The screenshot shows the Eclipse IDE's Run Configurations dialog. The configuration name is 'hybrid_host_numactl_2x1'. The 'Resources' tab is selected, showing the 'Target System Configuration' as 'edu.utexas.tacc.stampede.slurm.batch' and the 'Connection Type' as 'Remote' with the host 'stampede.tacc.utexas.edu'. The 'Import SLURM Script' tab is active, displaying the script path and the SLURM script content.

Name: hybrid_host_numactl_2x1

Target System Configuration: edu.utexas.tacc.stampede.slurm.batch

Connection Type: Local Remote stampede.tacc.utexas.edu

Script Path: /Users/briandwatt/Documents/runtime-Eclipse/hybrid_host_numactl_2x1/job

```
#!/bin/bash

# *****
# Job script for Stampede Training Lab hybridnuma
# hliu/xzhu216 (at) tacc.utexas.edu
# 02 Jan 2013
# *****

#SBATCH -J numa_2x1      # Job name
#SBATCH -o numa_2x1.%j.out  # Name of stdout output file (%j expands to jobid)
#SBATCH -e numa_2x1.%j.err  # Name of stdout error file (%j expands to jobid)
#SBATCH -p development    # Queue name
#SBATCH -N 1              # Total number of nodes requested
#SBATCH -n 2              # Total number of mpi tasks requested
#SBATCH -t 00:10:00      # Run time (hh:mm:ss) - 1.5 hours
#SBATCH --mail-user=briandwatt@gmail.com
#SBATCH --mail-type=ALL

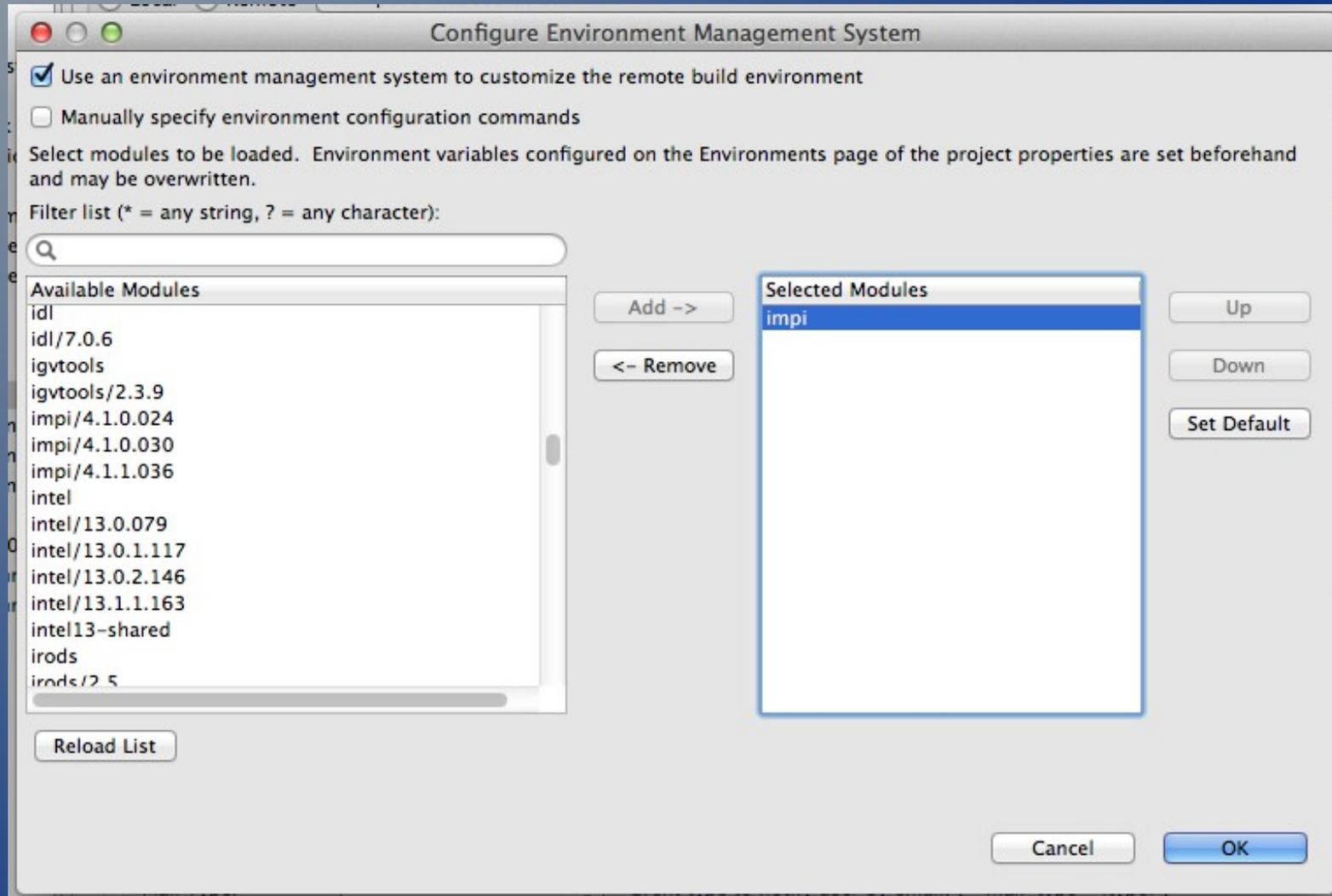
export OMP_NUM_THREADS=1

echo "-----"
echo " "
```

Module Support

- Presently generates in script – fetches once at start
 - `module purge > /dev/null 2>&1`
 - `module load <module> - repeated`
- Stampede uses LMOD environmental module system
 - Supports hierarchical modules
 - Supports named saved environments
- Future (for LMOD on Stampede only) – refetches for each change because restore affects unloads affects loads
 - `Nothing or module restore <name> or module reset`
 - `module unload <module> - repeated`
 - `module load <module> - repeated`
 - `module list - optional`

Current Module Support GUI



Future Stampede Module Support GUI

- Use environment management system
 - Manually specify environment configuration commands
-

No Restore Restore  Reset

Select modules to be unloaded

Filter list

Modules List

Add ->

<- Remove

Selected Modules to unload

Select modules to be loaded

Filter list

Modules Available

Add ->

<- Remove

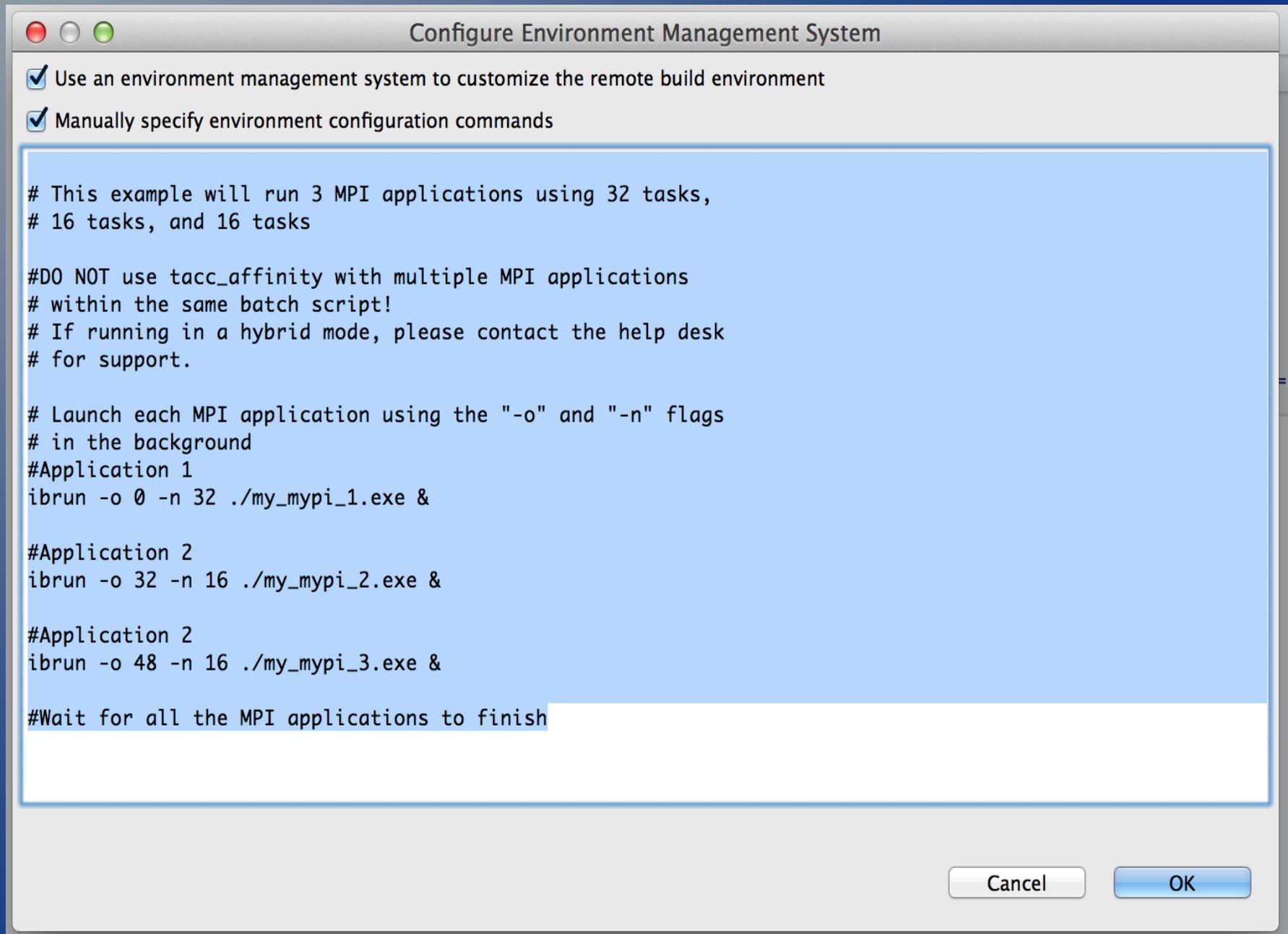
Selected Modules to load

-
- List modules

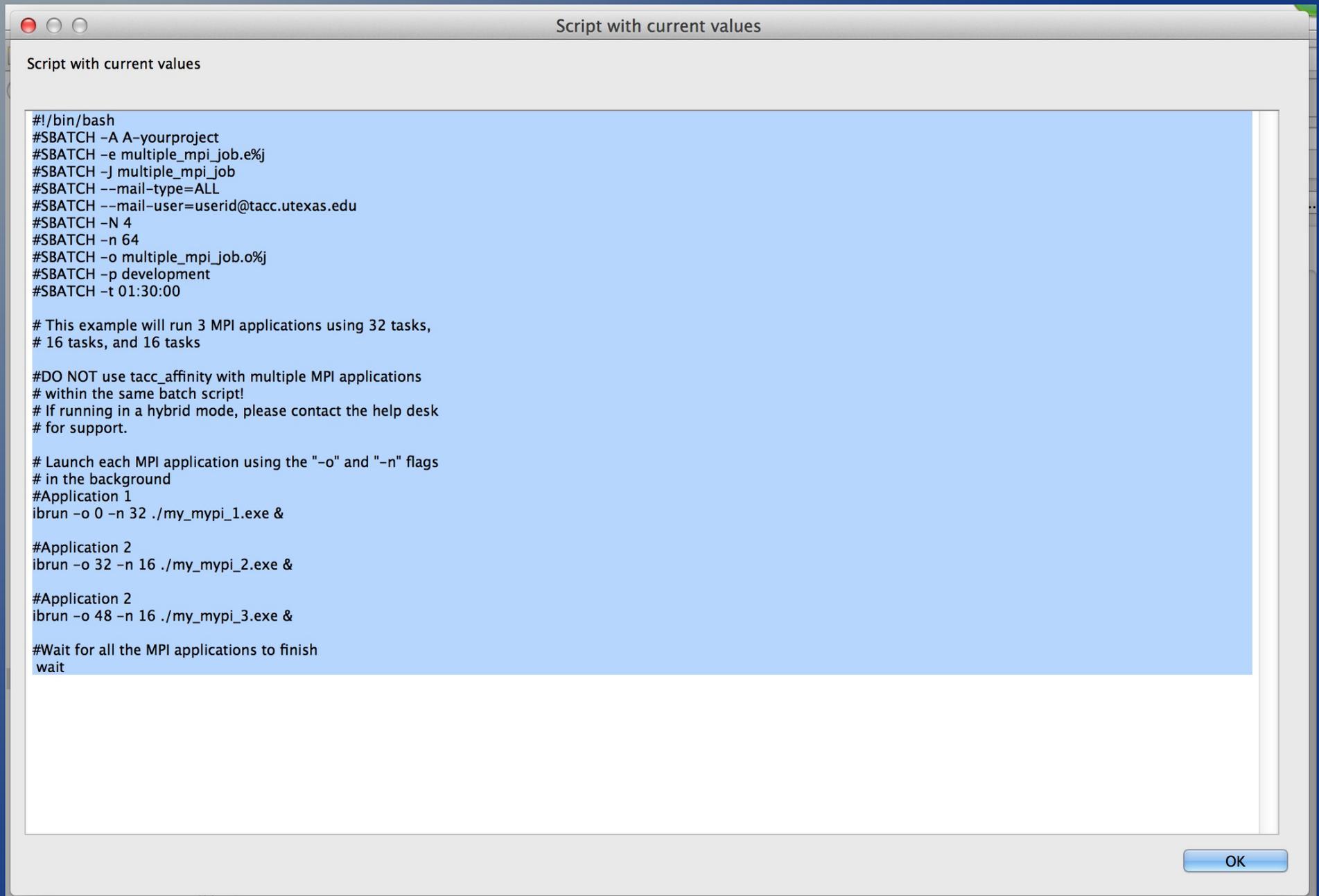
Inject Commands into Batch Script

- Need for 'escape' capability to support more intermediate to advanced users
- Injection of commands into batch script
 - Provides custom processing and/or setup prior to application launch
- Option 1 - User Specified Module Commands (a fudge)
- Option 2 - Propose a new RM 'custom' tab
 - For example, `<inject title="Additional Lines">`
 - Provides editor text area where the user enters one or more commands
 - Injects commands after module commands, and before application launch

User Specified Module Commands



Resulting Batch Script



```
Script with current values

#!/bin/bash
#SBATCH -A A-yourproject
#SBATCH -e multiple_mpi_job.e%j
#SBATCH -J multiple_mpi_job
#SBATCH --mail-type=ALL
#SBATCH --mail-user=userid@tacc.utexas.edu
#SBATCH -N 4
#SBATCH -n 64
#SBATCH -o multiple_mpi_job.o%j
#SBATCH -p development
#SBATCH -t 01:30:00

# This example will run 3 MPI applications using 32 tasks,
# 16 tasks, and 16 tasks

#DO NOT use tacc_affinity with multiple MPI applications
# within the same batch script!
# If running in a hybrid mode, please contact the help desk
# for support.

# Launch each MPI application using the "-o" and "-n" flags
# in the background
#Application 1
ibrun -o 0 -n 32 ./my_mypi_1.exe &

#Application 2
ibrun -o 32 -n 16 ./my_mypi_2.exe &

#Application 2
ibrun -o 48 -n 16 ./my_mypi_3.exe &

#Wait for all the MPI applications to finish
wait
```

OK

System Monitor Display

- Based upon recent presentation: Customizing the PTP Monitoring Layout by Carsten Karbach
 - Custom LML-Layout
 - Define Machine Topology – TBD
 - Setup/Usage
- Update for UT TACC Stampede specifics
- Details/specifics working with Doug James, and consultation with Carsten Karbach
- Review refresh/update performance

Current System Monitor Display

The screenshot displays the Eclipse Platform System Monitoring interface. The main window is titled "System Monitoring - Eclipse Platform" and shows a grid of 16x16 job progress bars. Each bar is color-coded to represent the status of a job: green for completed, yellow for submitted, and red for other states. The grid is titled "system: login3.stampede.tacc.utexas.edu".

On the left side, there are several panels:

- Monitors:** A table showing connection and configuration names.
- Active Jobs:** A table listing job details such as step, owner, queue, wall time, queue date, dispatch date, total cores, and status.
- Remote Environments:** A list of target environments and their status.

Status	Connection Name	Configuration Name
	lonestar.tacc.utexas.edu	edu.utexas.tacc.lonestar.sge.batch
	stampede.tacc.utexas.edu	edu.utexas.tacc.stampede.slurm.batch

step	owner	queue	wall	queue date	dispatch date	total cores	status
1150903	jcmingta...	normal	9000	2013-07-16...	2013-07-16...	16	COMPLETED
1150909	us3(8091...	normal	50640	2013-07-16...	2013-07-16...	32	COMPLETED
1150928	tg45923...	normal	3540	2013-07-16...	?	256	SUBMITTED
1150930	tg45923...	normal	3540	2013-07-16...	?	256	SUBMITTED
1150933	tg45923...	normal	2940	2013-07-16...	?	96	SUBMITTED
1150934	tg45923...	normal	2940	2013-07-16...	?	96	SUBMITTED
1150935	tg45923...	normal	2940	2013-07-16...	?	96	SUBMITTED
1150937	tg45923...	normal	2940	2013-07-16...	?	96	SUBMITTED
1150938	tg45923...	normal	1740	2013-07-16...	?	64	SUBMITTED
1150958	beckver...	large	36000	2013-07-16...	2013-07-16...	4992	COMPLETED
1150960	yingyue...	normal	900	2013-07-16...	2013-07-16...	32	COMPLETED
1150963	yingyue...	normal	86400	2013-07-16...	2013-07-16...	32	COMPLETED
1150964	yingyue...	serial	7200	2013-07-16...	2013-07-16...	16	COMPLETED
1150968	yingyue...	normal	900	2013-07-16...	2013-07-16...	32	COMPLETED
1150970	us3(8091...	normal	86400	2013-07-16...	2013-07-16...	256	COMPLETED
1150973	tg45923...	normal	2340	2013-07-16...	2013-07-16...	256	COMPLETED
1150974	tg45923...	normal	2940	2013-07-16...	?	48	SUBMITTED
1150975	tg45923...	normal	2940	2013-07-16...	?	48	SUBMITTED
1150976	tg45923...	normal	2940	2013-07-16...	?	48	SUBMITTED
1150977	yingyue...	normal	900	2013-07-16...	2013-07-16...	32	COMPLETED
1150981	us3(8091...	normal	86400	2013-07-16...	2013-07-16...	256	COMPLETED
1150984	jflanaga...	vis	21600	2013-07-16...	2013-07-17...	16	SUBMITTED
1150986	us3(8091...	normal	86400	2013-07-16...	2013-07-16...	256	COMPLETED
1150987	yingyue...	normal	900	2013-07-16...	2013-07-16...	32	COMPLETED
1150990	minjung...	normal	28800	2013-07-16...	2013-07-16...	16	COMPLETED

Target Environment	Status
Remote Host	
lonestar.tacc.utexas.edu	Started
stampede.tacc.utexas.edu	Started

Acknowledgements

- Doug James for design, consultation and direction
- Jay Alameda for providing travel and userid/sign-on through XSEDE
 - NSF SI2 grant, NSF OCI-1047956 "SI2-SSI: A Productive and Accessible Development Workbench for HPC Applications Using the Eclipse Parallel Tools Platform"
 - This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575.
 - XSEDE Allocation TG-ASC110002, "A Productive and Accessible Development Workbench for HPC Applications Using the Eclipse Parallel Tools Platform"
- Greg Watson for Eclipse PTP consulting