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

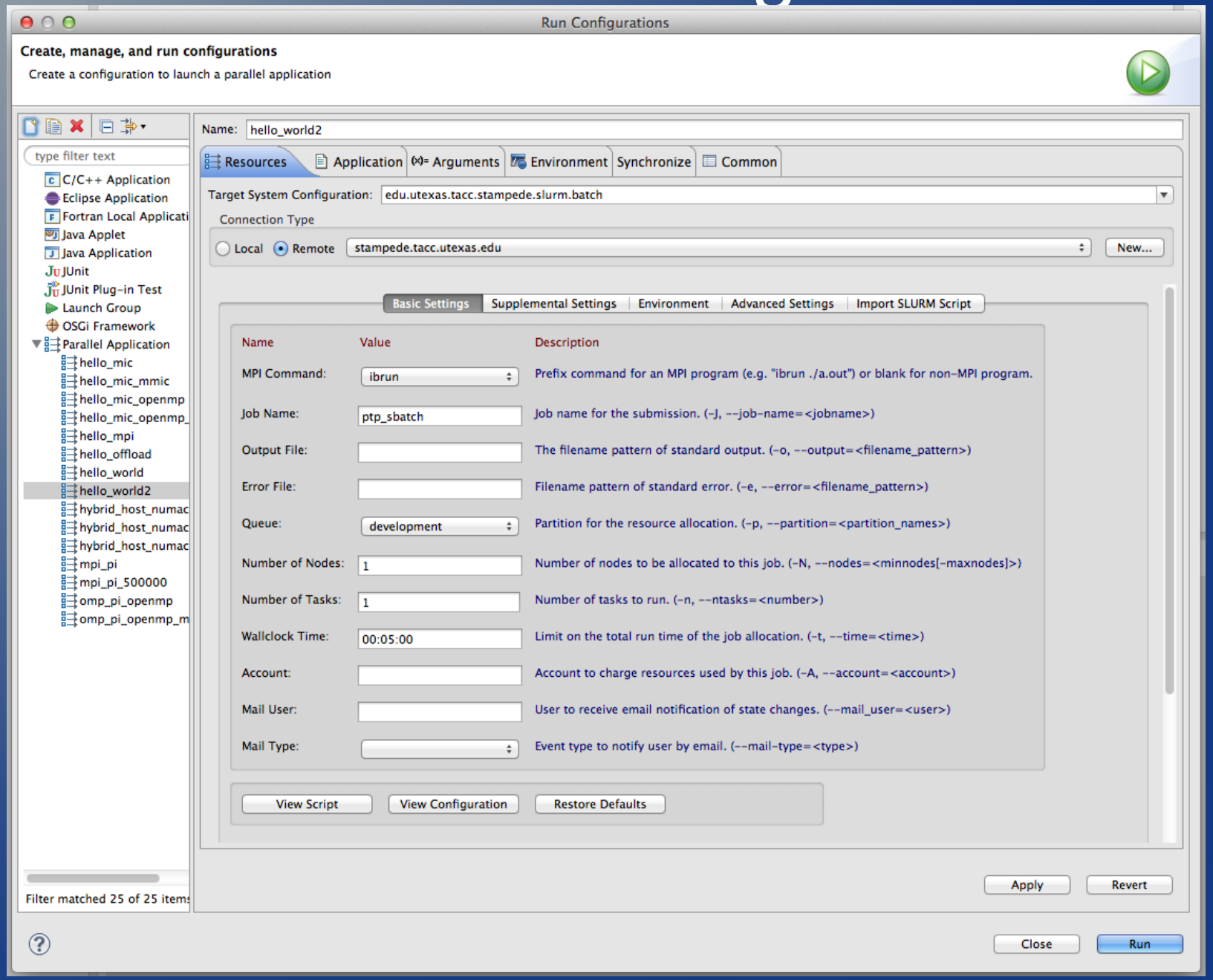
Stampede Environment

- 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

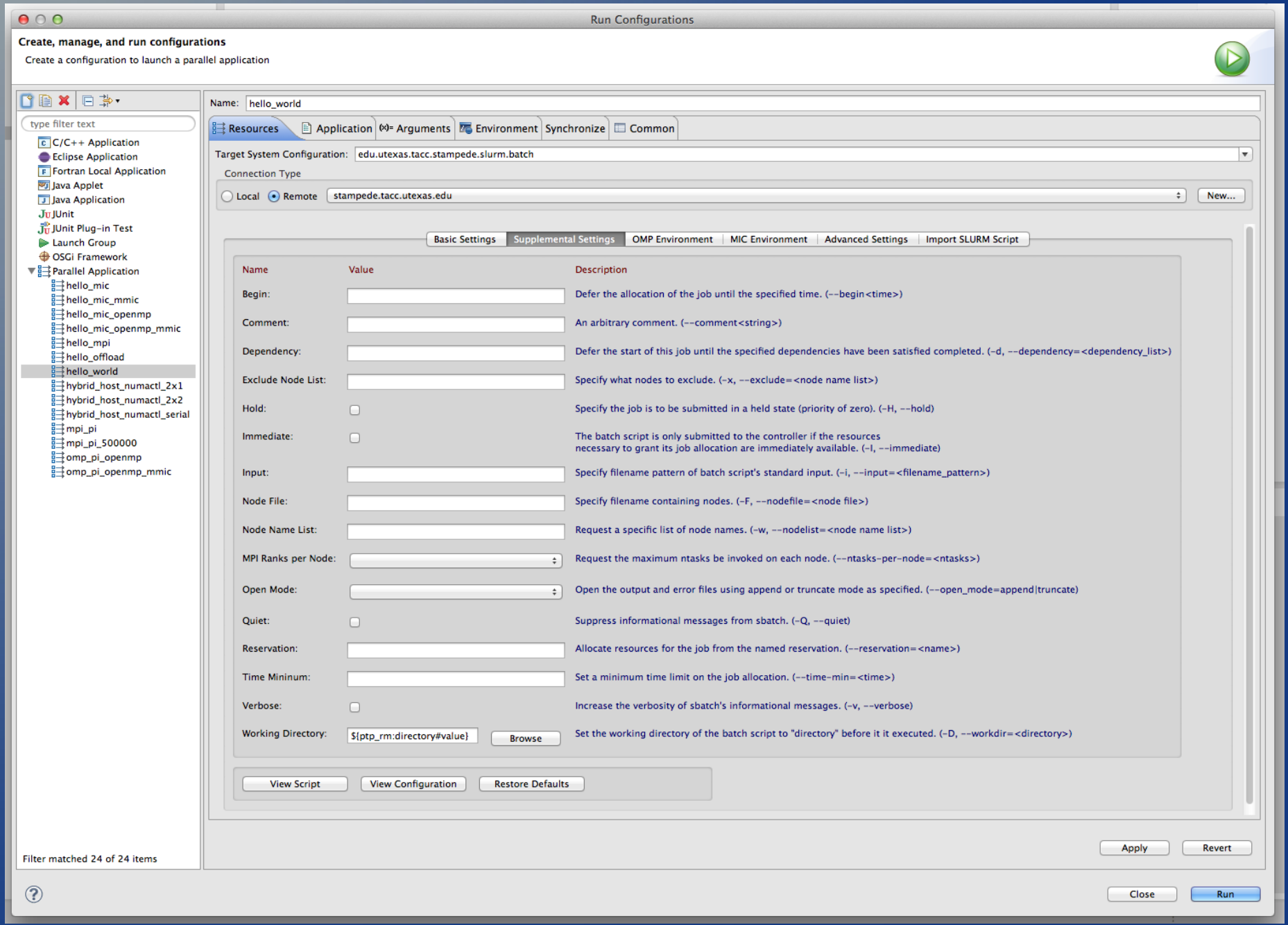
Stampede Target System Configuration

- 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

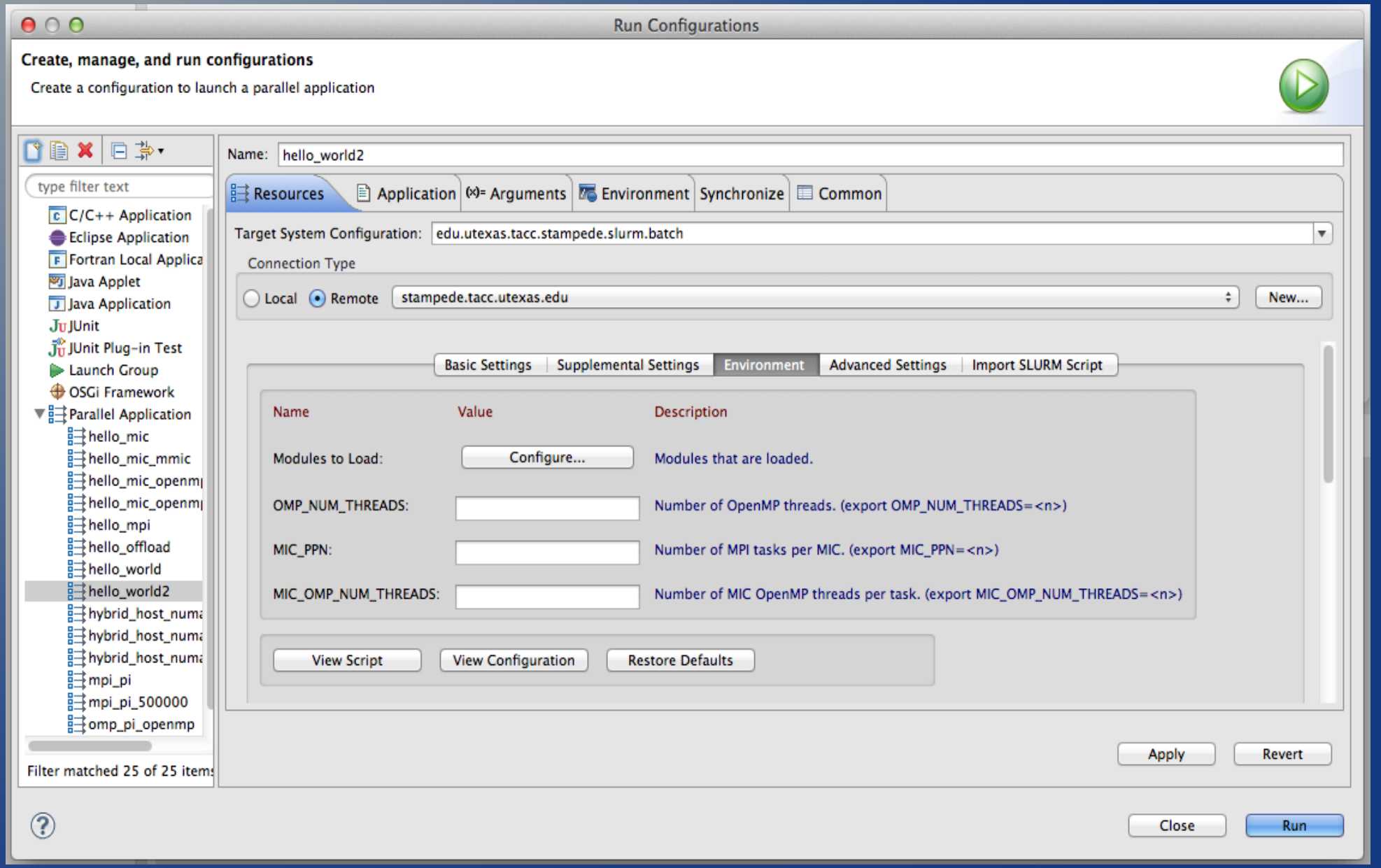
Basic Settings



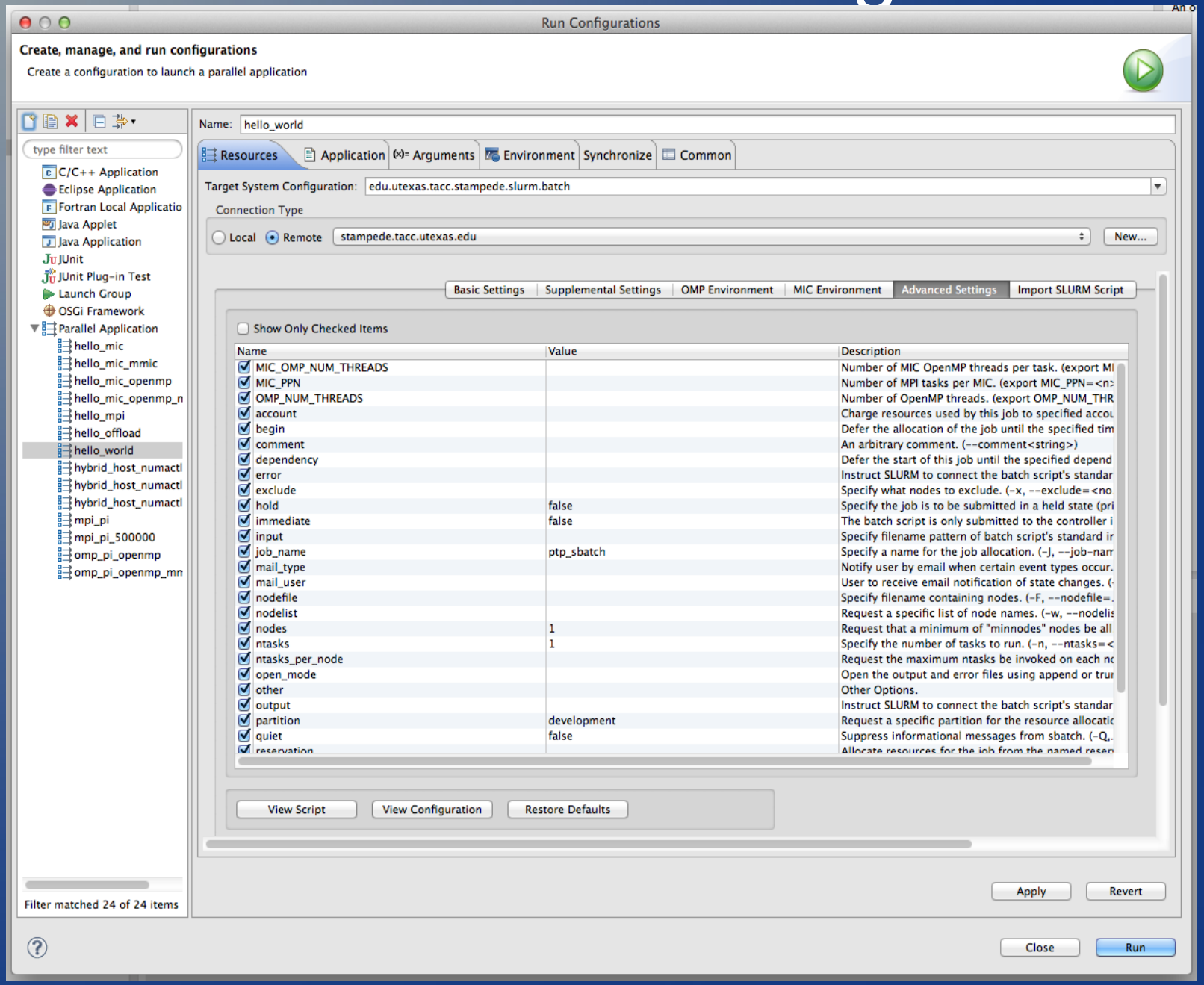
Supplemental Settings



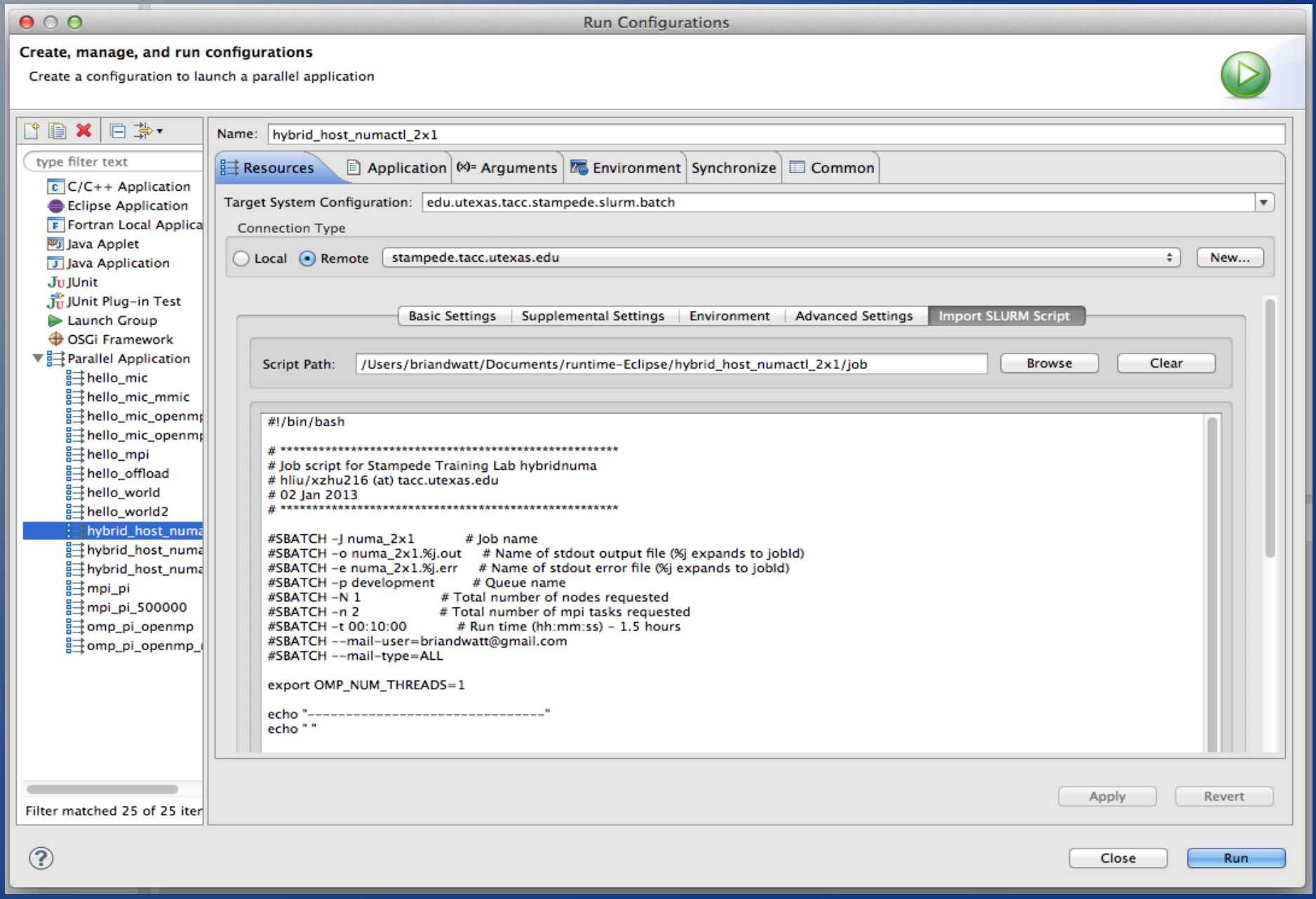
Environment



Advanced Settings



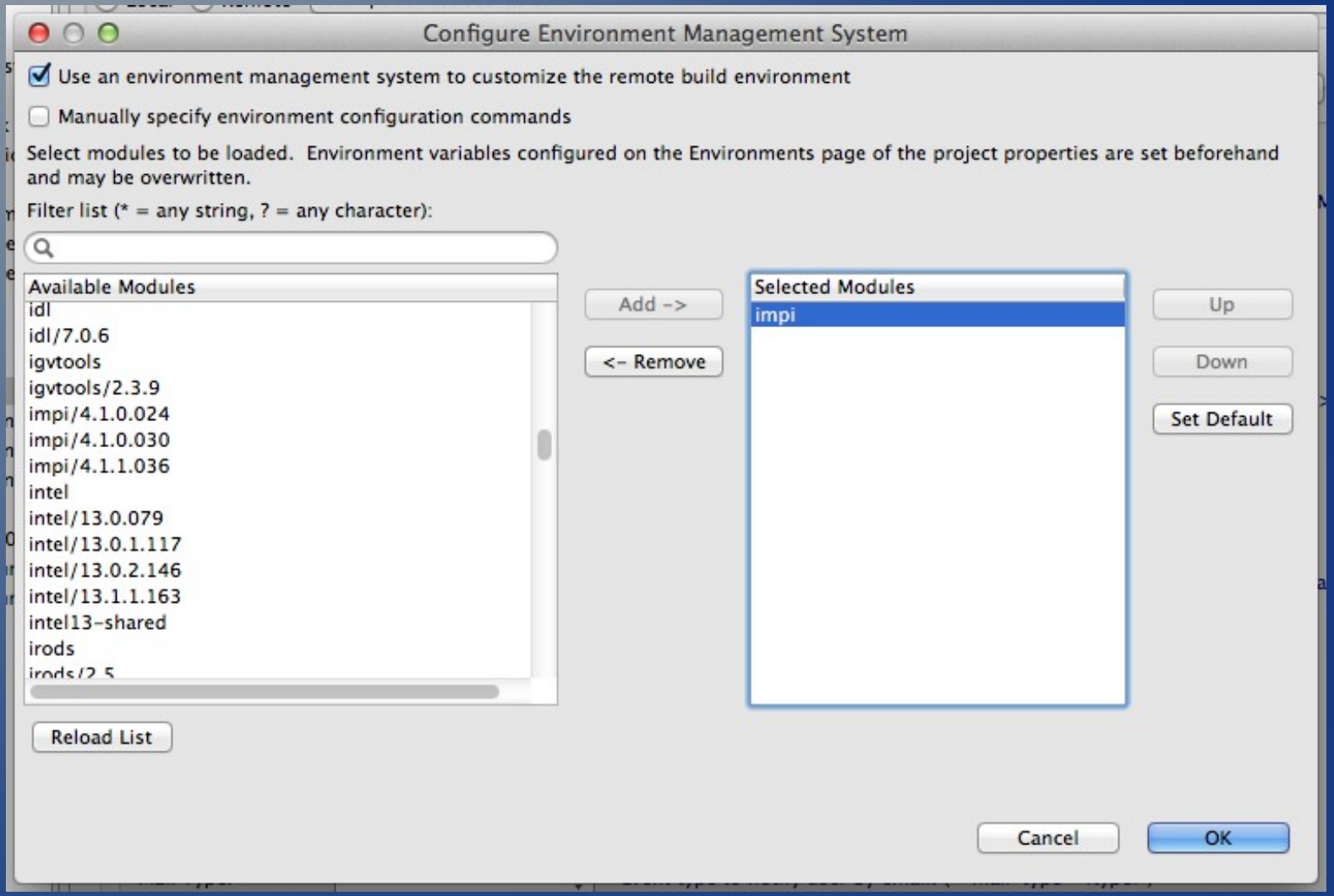
Import SLURM Script



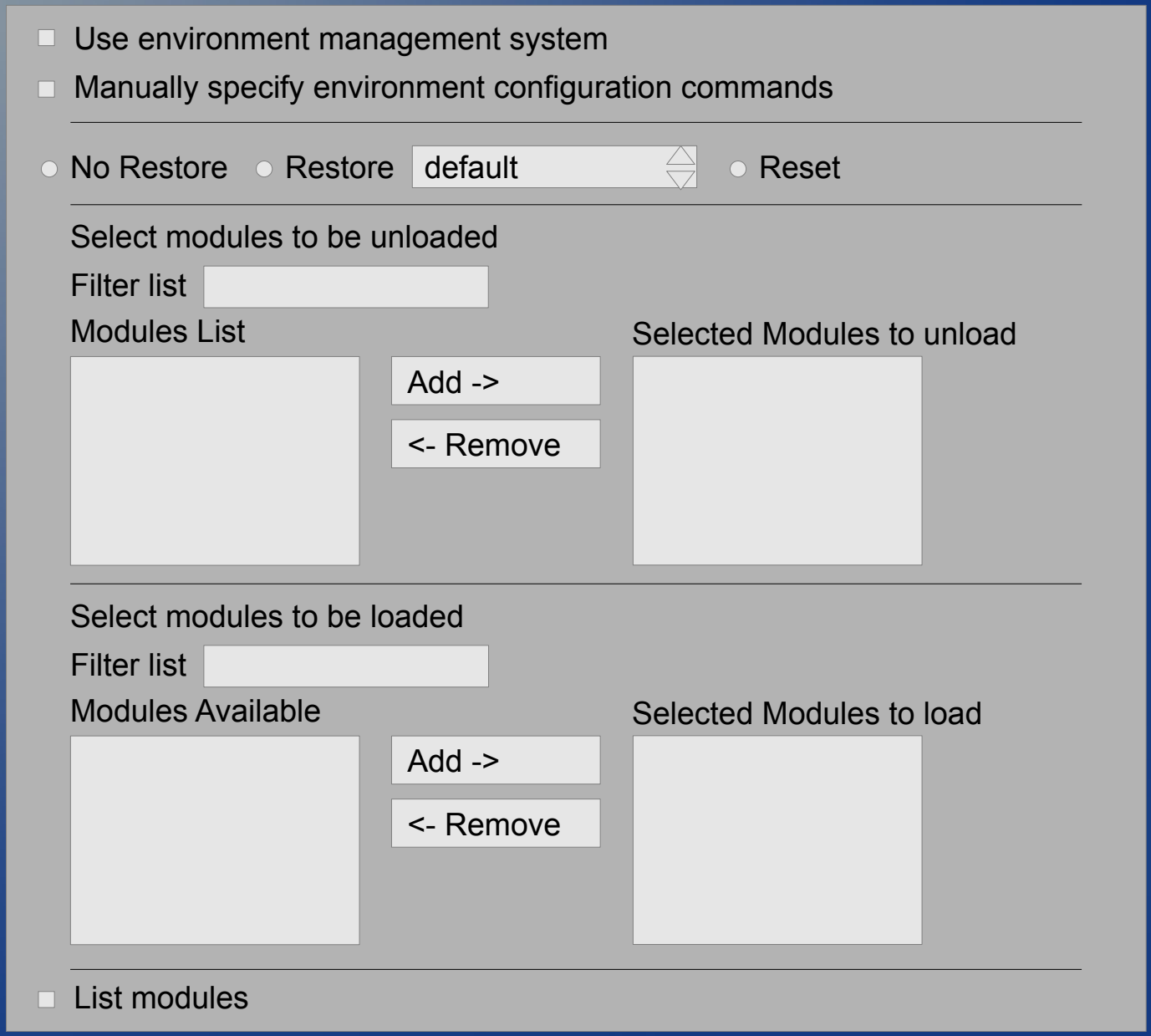
Module Support

- 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



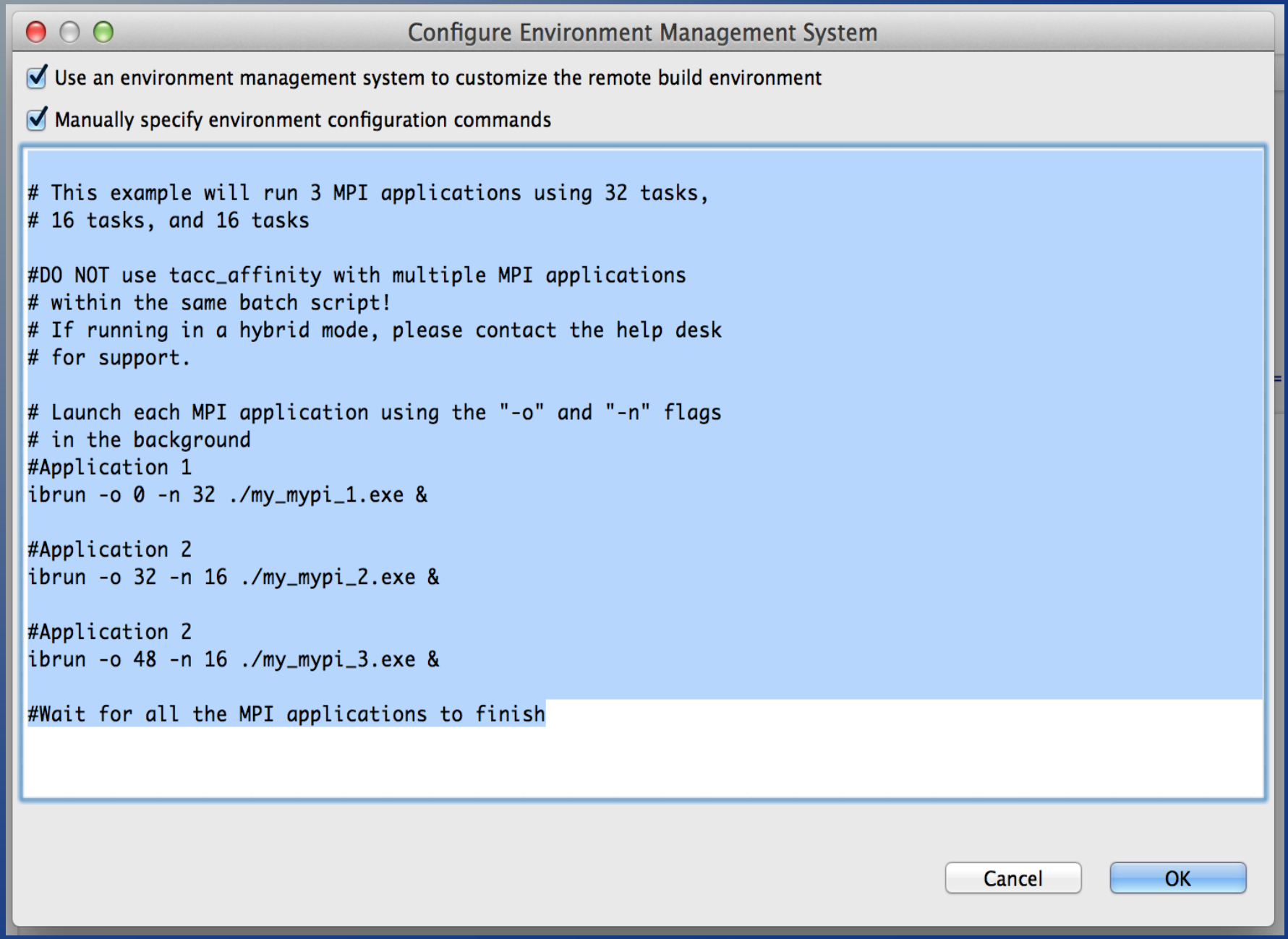
Future Stampede Module Support GUI



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

Script with current values

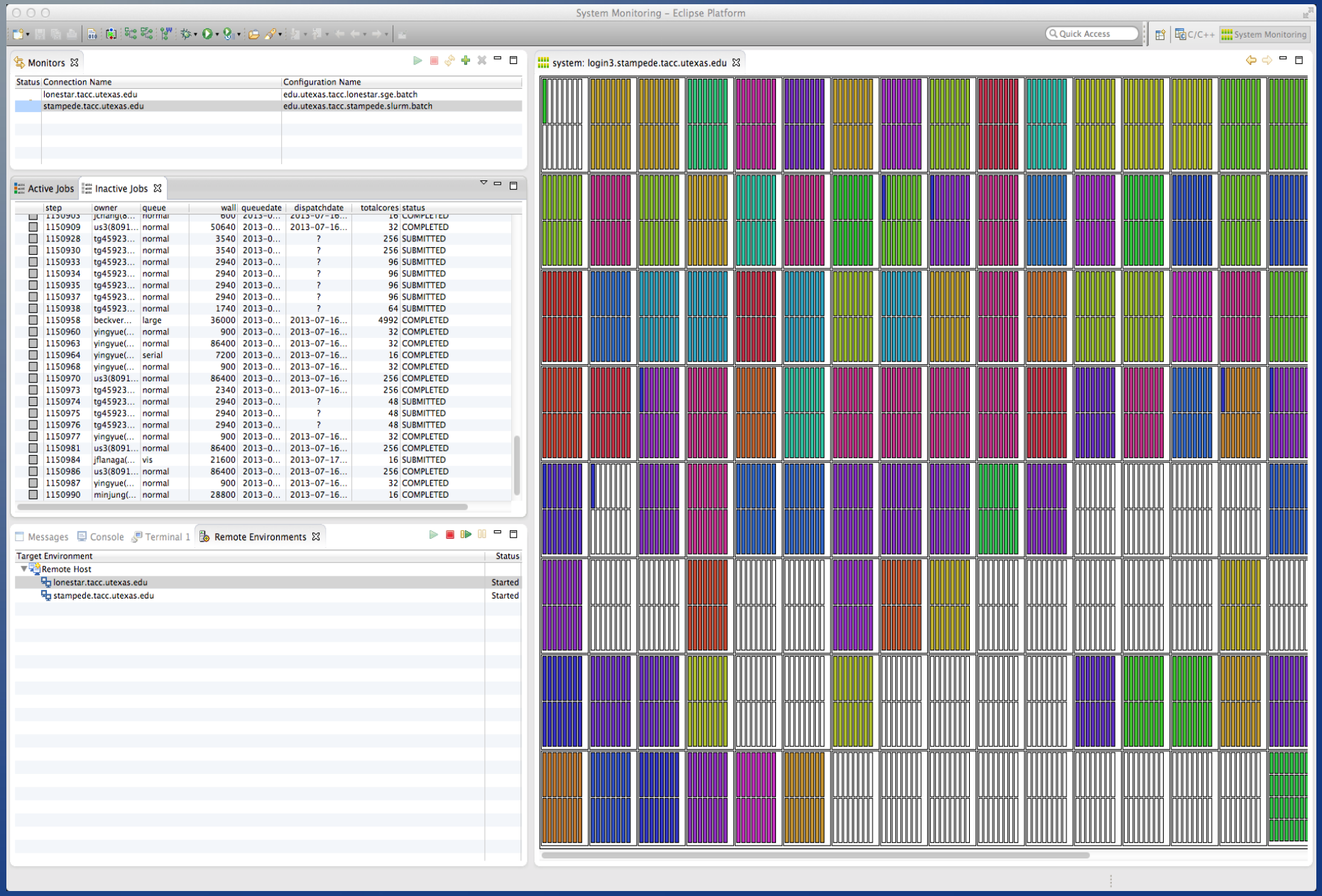
```
#!/bin/bash
#SBATCH -A A-yourproject
#SBATCH -e multiple_mpi_job.e%j
#SBATCH -J multiple_mpi_job
#SBATCH --mail-type=ALL
#SBATCH --mail-user=userid@tacc.utexas.edu
#SBATCH -N 4
#SBATCH -n 64
#SBATCH -o multiple_mpi_job.o%j
#SBATCH -p development
#SBATCH -t 01:30:00

# 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

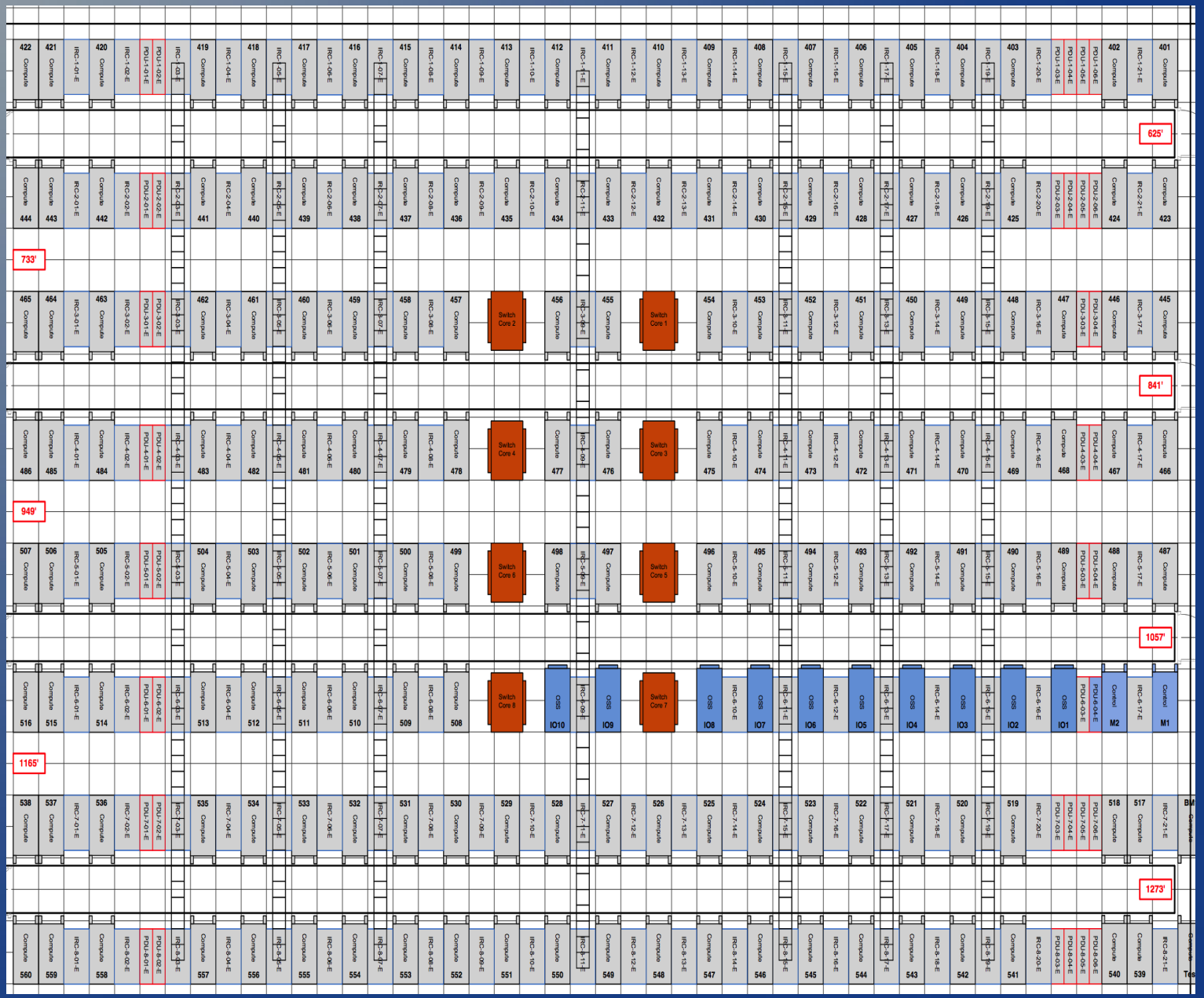
System Monitor Display

- 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



Future System Monitor Display



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