Adaptive COMPUTING

Intelligent Workload Management Software for HPC and Cloud Environments

# Broadening Moab/TORQUE for Expanding User Needs

Gary D. Brown
HPC Product Manager

CUG 2016

# Agenda

- DataWarp
- Intel MIC “KNL”
- Viewpoint Web Portal
  - User Portal
  - Administrator Portal
  - Remote Visualization
  - Nitro HTC


© 2016 Adaptive Computing Enterprises, Inc.

# Agenda

- **DataWarp (fast I/O)**
- Intel MIC “KNL”
- Viewpoint Web Portal
  - User Portal
  - Administrator Portal
  - Remote Visualization
  - Nitro HTC


© 2016 Adaptive Computing Enterprises, Inc.

# Use Cases

- Stage job data with fast SSD-based storage
  - Input data before job starts
  - Output data after job ends
- Fast persistent storage for job workflows
- Fast job checkpoint/restore


© 2016 Adaptive Computing Enterprises, Inc.

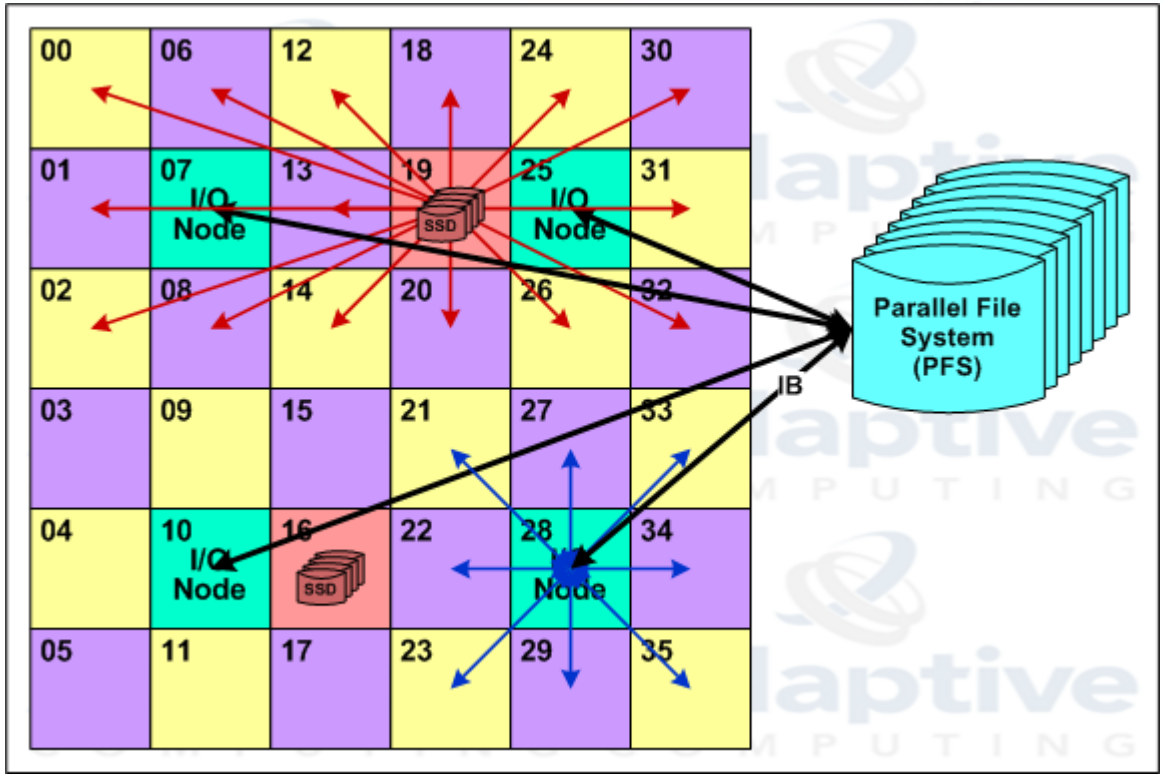
# Cray DataWarp

- SSD-based storage system independent of parallel file system (PFS)
- Hosted on DataWarp (DW) service nodes
  - Contain actual SSDs
  - Nodes connect to Aries high-speed network
- DataWarp Service (DWS)
  - Organizes DW storage for jobs and persistent storage
  - Manages SSD storage across DW service nodes


© 2016 Adaptive Computing Enterprises, Inc.
Adaptive COMPUTING

# Cray DataWarp Physical Architecture

- DW storage
  - 2-hop roundtrip
    - Aries HSN
  - Electronic access times
- PFS storage
  - 4-hop roundtrip
    - 2-hop Aries HSN
    - 2-hop InfiniBand
  - Electro-mechanical access times
- *DWS not shown*






© 2016 Adaptive Computing Enterprises, Inc.

# Cray DataWarp Support

- DataWarp Storage Types
  - Job-based
  - Persistent
- Script-based integration with Moab/TORQUE
- DW-enabled jobs processed by Moab “job submit filter”
- Moab automatically creates DW job workflow


© 2016 Adaptive Computing Enterprises, Inc.

# DataWarp Integration Project

- All DW storage scheduled by Moab
  - Job DW storage by job scheduling
  - Persistent DW storage by reservation scheduling
- Out-of-the-box functionality
  - Some site-specific information configuration required


© 2016 Adaptive Computing Enterprises, Inc.
Adaptive COMPUTING

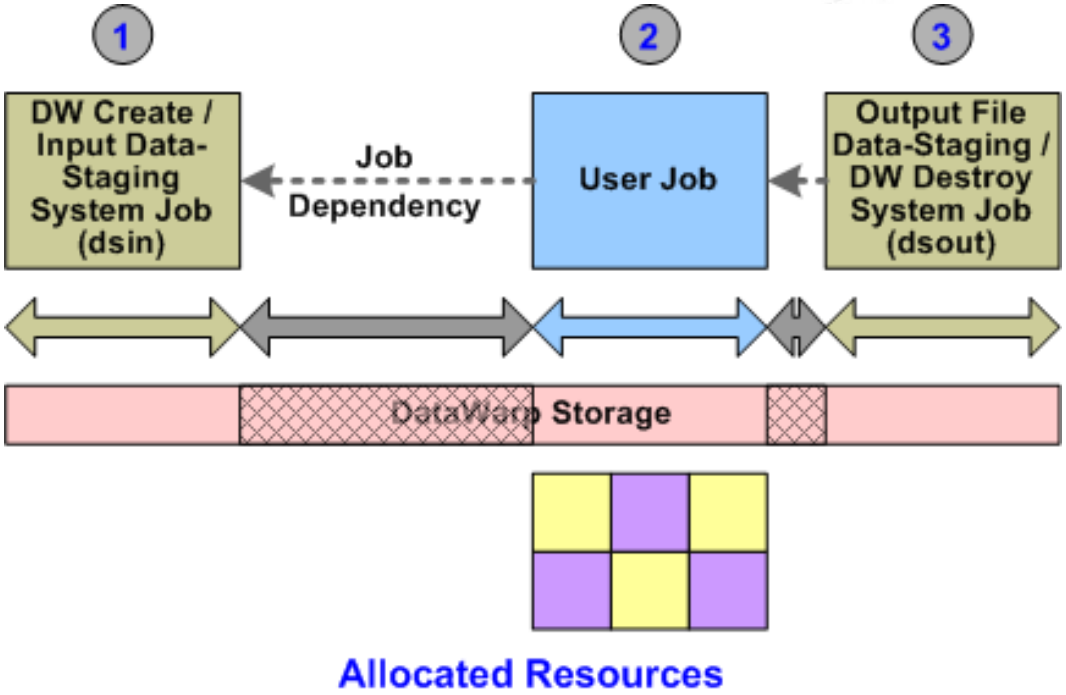
# Moab 3-Job DataWarp Model

- Follows Moab Data-staging Model
  - Uses Moab "system" jobs to create/destroy DW storage and stage data in/out
- Automatically creates 3-job workflow
  - DW creation/input data-staging system job
  - User job
  - Output data-staging/DW destruction system job


© 2016 Adaptive Computing Enterprises, Inc.

# DataWarp-enabled Job Workflow

- User submits user job
- Job Submit Filter detects #DW directives
- Job Submit Filter creates 3-job workflow
  1. DW Creation / Input Data-staging System job
  2. User Job
  3. Output Data-staging / DW Destruction System Job






© 2016 Adaptive Computing Enterprises, Inc.

Adaptive COMPUTING

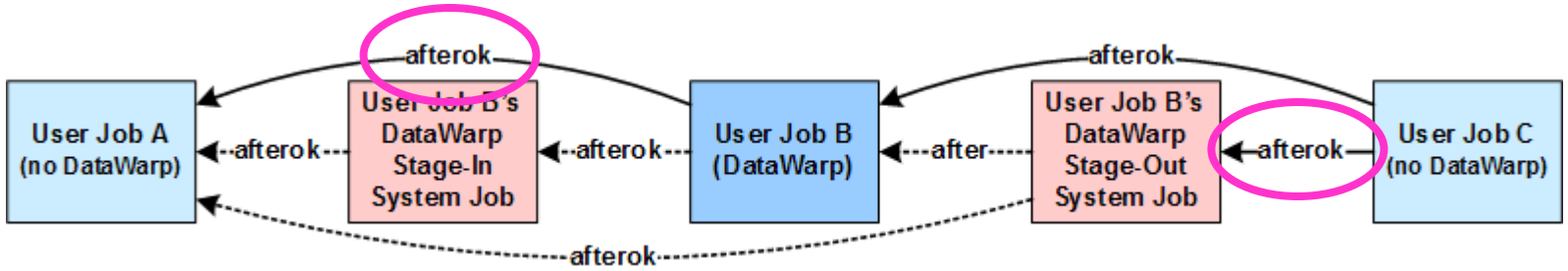
# Automatic Job/DW Workflow Generation

- What if user submits multiple-job workflow
  - And some user jobs in job workflow use DataWarp storage?
- Job Submit Filter creates proper DW job dependencies, but not for entire user job workflow
- Example
  - Data validation — *User Job A*
  - Compute (uses DataWarp) — *User Job B (DW)*
  - Visualization — *User Job C*


© 2016 Adaptive Computing Enterprises, Inc.

# Automatic Job/DW Workflow Generation

- User Job Dependencies
  - Data validation  User Job A
  - Compute  User Job B *(DataWarp-enabled job)*
  - Visualization  User Job C
  - Job A (ok) ← Job B (ok) ← Job C
- Job workflow generation script must work with Job Submit Filter to create proper job dependencies *(solid lines below)*






© 2016 Adaptive Computing Enterprises, Inc.

Adaptive COMPUTING

# Multi-Job Workflow Generation Script

```
# Submit data validation job A
jobAid=`msub -l nodes=1,walltime=10:00 myCheck.sh`
# Submit DataWarp-enabled compute job B
#  (3 jobids returned (dwin, user, dwout); put into array)
jobBids=(`msub -l nodes=16,walltime=01:30:00,\
                  depend=afterok:${jobAid} ...\
              mySim.sh`)
# Submit visualization job C
jobCid=`msub -l nodes=1,walltime=30:00,\
                 depend=afterok:${jobBids[1]}:${jobBids[2]} \
              myViz.sh`
```

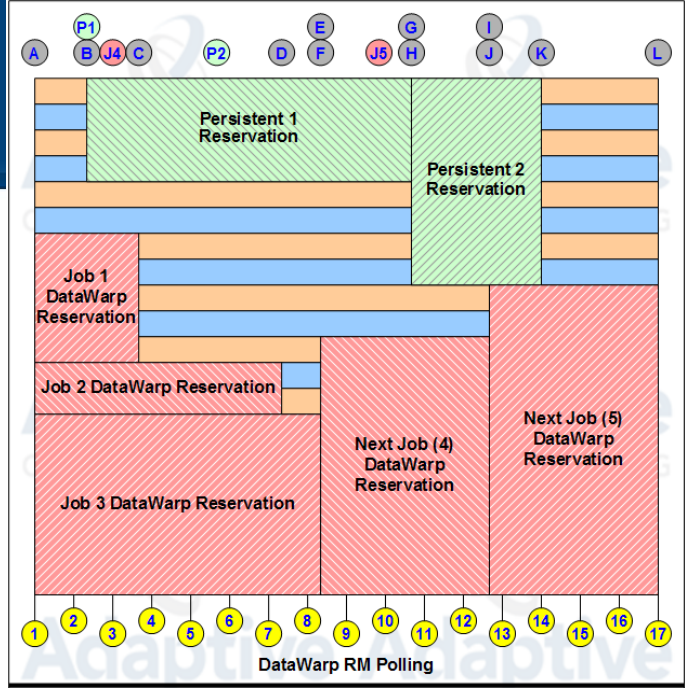


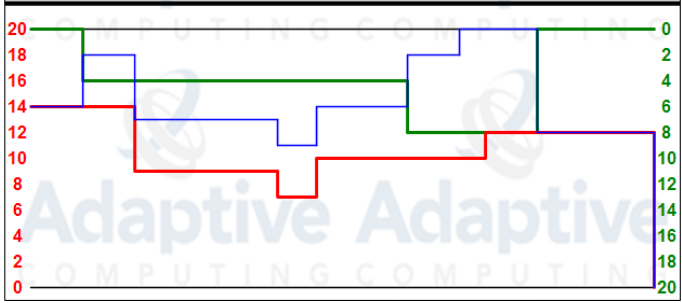
© 2016 Adaptive Computing Enterprises, Inc.

Adaptive COMPUTING

# Persistent DataWarp Storage

- Moab schedules and tracks all DW storage capacity
- Moab persistent DW storage uses DW storage reservation
- Persistent DataWarp storage management
  - Created by reservation start trigger
  - Destroyed by reservation end trigger
- Jobs get mount point






© 2016 Adaptive Computing Enterprises, Inc.

# Benefits

- Moab schedules and manages both job and persistent DataWarp storage
- Users have good user experience
  - *No race conditions cause cancelled jobs!*
- Jobs have fast I/O
- Jobs execute in less time


© 2016 Adaptive Computing Enterprises, Inc.

# Agenda

- DataWarp
- **Intel MIC "KNL" (re-provisioning)**
- Viewpoint Web Portal
  - User Portal
  - Administrator Portal
  - Remote Visualization
  - Nitro HTC


© 2016 Adaptive Computing Enterprises, Inc.

# Use Case

- Intel introduces 2nd-gen Xeon Phi processor
- KNL supports
  - 3 different NUMA configurations with 5 modes
  - 3 different MCDRAM configurations with 4 modes
  - 20 possible NUMA/MCDRAM configurations
- Some applications execute fastest with specific NUMA/MCDRAM configuration


© 2016 Adaptive Computing Enterprises, Inc.

# Intel MIC "Knights Landing" (KNL)

- 5 NUMA configurations for cores and DDR4 and MCDRAM memories
  - No NUMA (all2all, hemisphere, quadrant) and sub-NUMA cluster 2 and 4
  - *a2a, hemi, quad, snc2, snc4*
- 4 MCDRAM "mode" configurations
  - Cache, flat, 50/50 hybrid, 25/75 hybrid
  - *cache, flat, equal, split*


© 2016 Adaptive Computing Enterprises, Inc.

# Moab Cray KNL Support

- 20 pre-defined combinations
  - Select combination as "OS" at job submission

    ```
    msub nodes=20,os=CLE_snc4_flat myJobScript
    ```
- Moab re-provisions Cray compute nodes when necessary
  - Uses Moab's built-in re-provisioning capability
  - Quickly re-provisions Cray KNL compute nodes for jobs requesting specific NUMA/MCDRAM modes

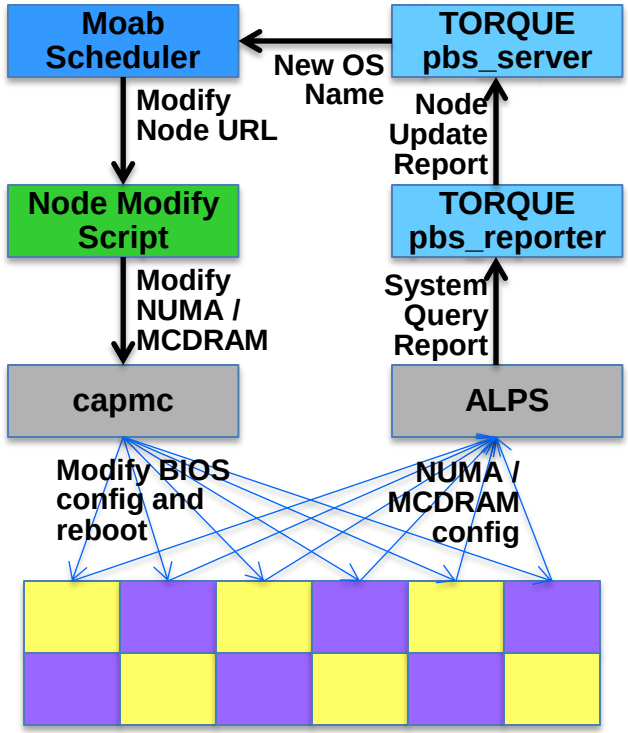

© 2016 Adaptive Computing Enterprises, Inc.

# Cray KNL Re-provisioning Process

- User submits job requiring specific KNL NUMA / MCDRAM configuration

```
msub -l nodes=20,os=CLE_snc4_flat myJob
```

- Moab schedules KNL job to nodes
- Moab runs NodeModifyURL script to modify compute node configurations and reboot them
- Script executes Cray capmc to set KNL NUMA node and MCDRAM configurations and to reboot nodes
- Cray's capmc modifies nodes' BIOS settings and reboots nodes
- Cray reports new NUMA node / MCDRAM configurations
- pbs_reporter maps NUMA node/MCDRAM configuration to OS name and reports OS name to pbs_server
- TORQUE pbs_server passes new OS name to Moab, which starts the KNL job






© 2016 Adaptive Computing Enterprises, Inc.

Adaptive COMPUTING

# Benefits

- Users can run applications with KNL NUMA and MCDRAM configuration for optimal job execution.
- Easy to request specific KNL configuration with job submission


© 2016 Adaptive Computing Enterprises, Inc.

# Agenda

- DataWarp
- Intel MIC “KNL”
- **Viewpoint Web Portal (non-technical users)**
  - User Portal
  - Administrator Portal
  - Remote Visualization
  - Nitro HTC


© 2016 Adaptive Computing Enterprises, Inc.

# Use Case

- Many HPC sites asked to expand user base
- Non-technical users
  - “What is a command line?”
- Non-traditional users
  - Non-research Engineers and Scientists
  - Humanities
  - Local or regional industries with no HPC experience


© 2016 Adaptive Computing Enterprises, Inc.

# Agenda

- DataWarp
- Intel MIC “KNL”
- Viewpoint Web Portal (non-technical users)
  - **User Portal (job submission)**
  - Administrator Portal
  - Remote Visualization
  - Nitro HTC

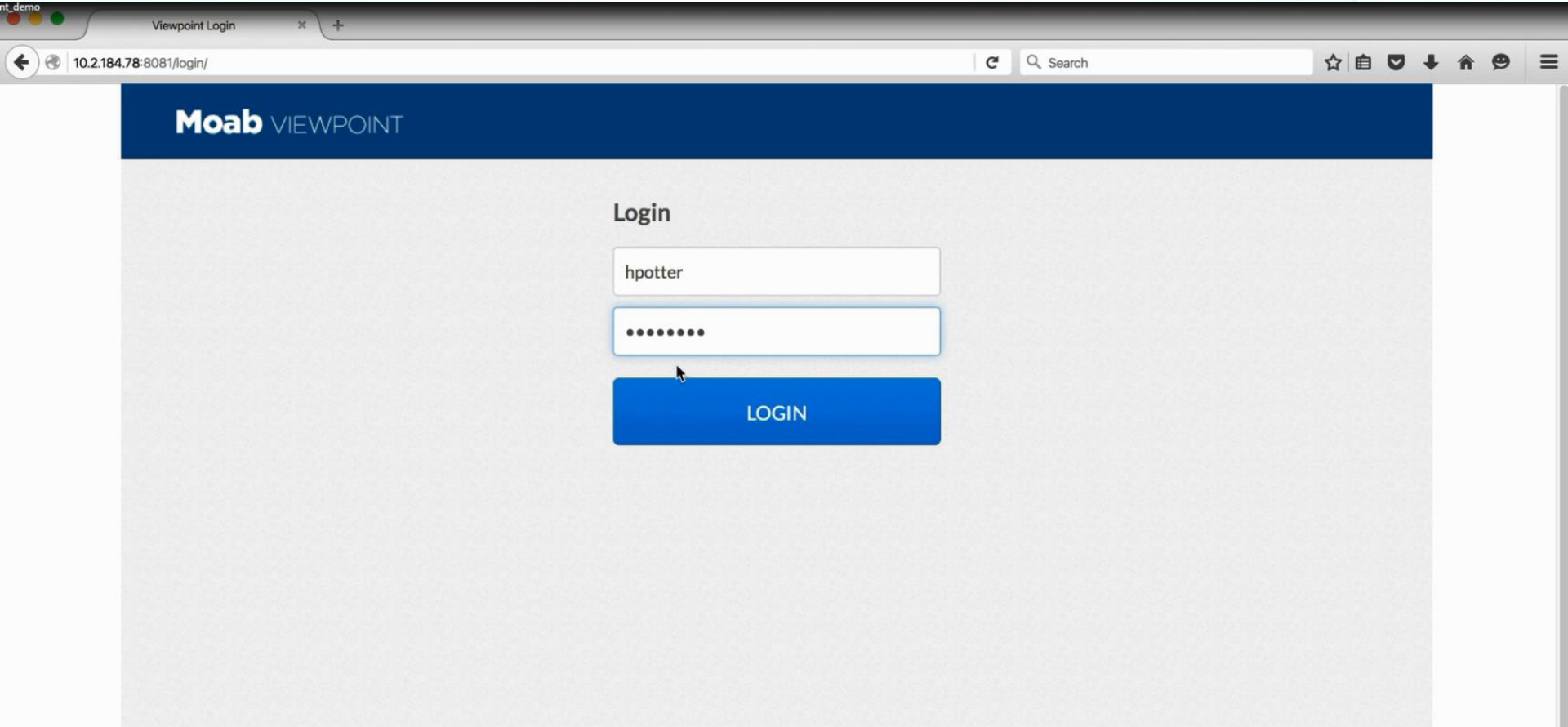

© 2016 Adaptive Computing Enterprises, Inc.

# Use Case

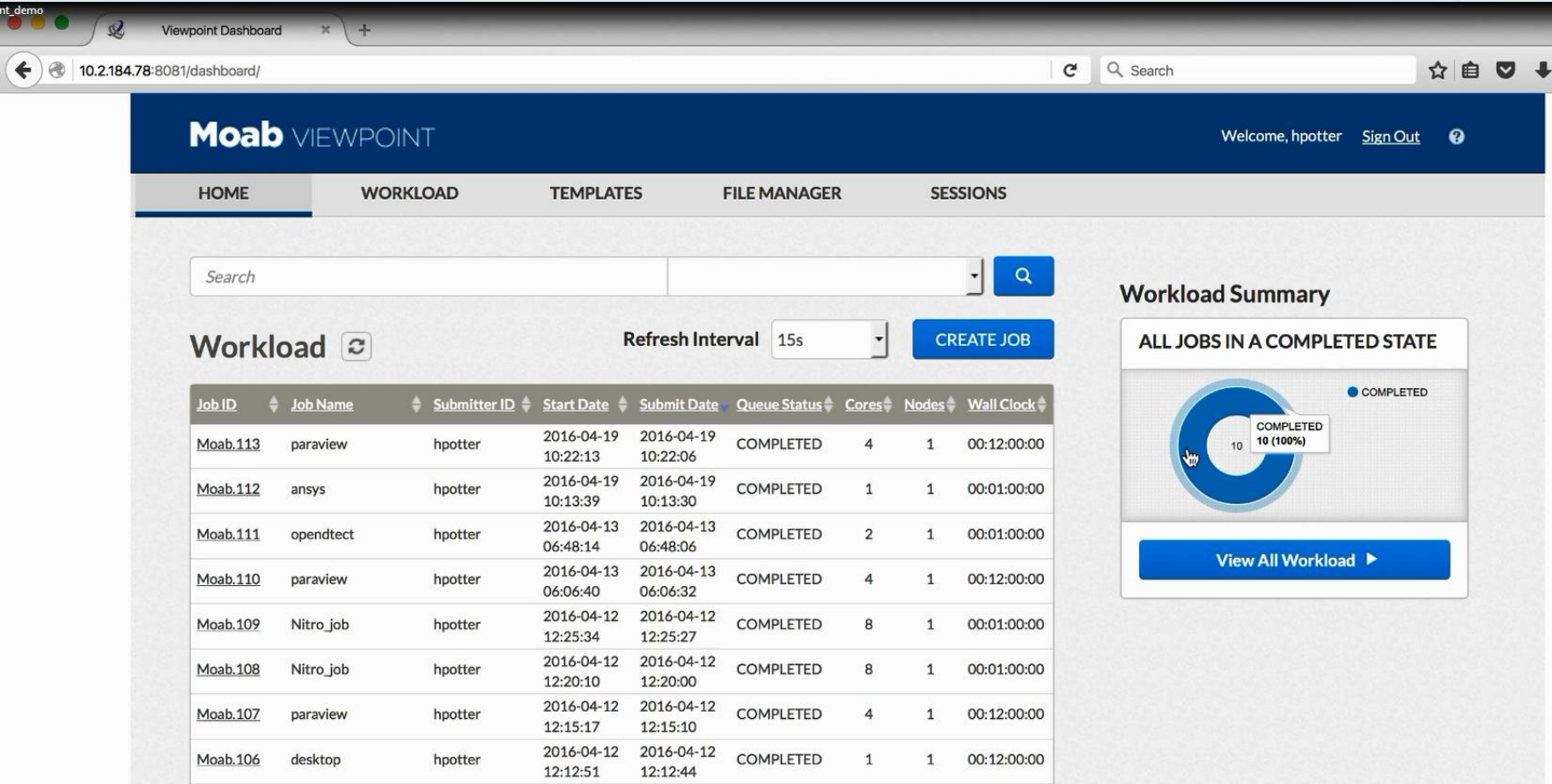
- Someone said I need to use “xyz” program
- I have data
  - Someone gave me my data
  - Someone mentored/trained me and I created my own data
  - I taught myself and created my own data
- How do I use “xyz” with my data?


© 2016 Adaptive Computing Enterprises, Inc.

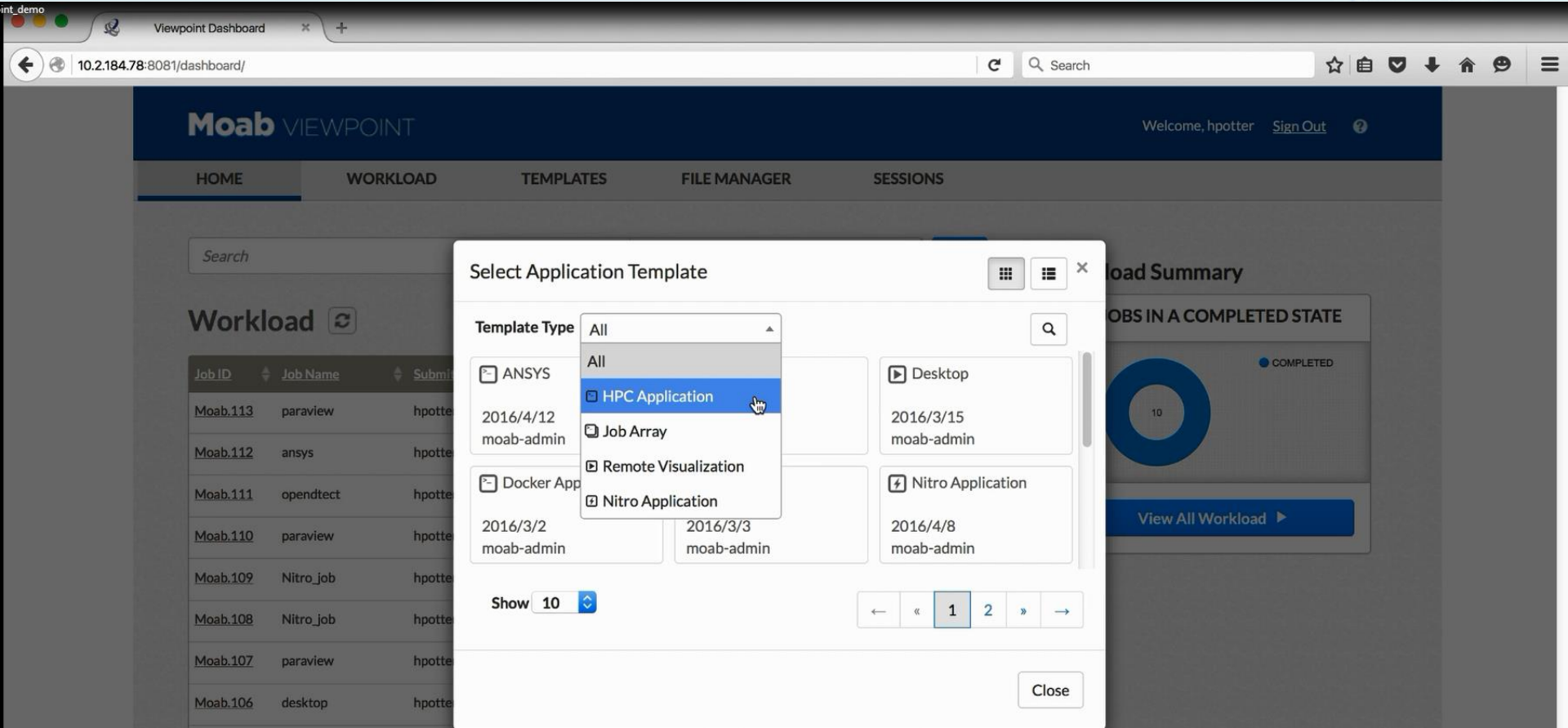
# Log in as user to Viewpoint Web Portal



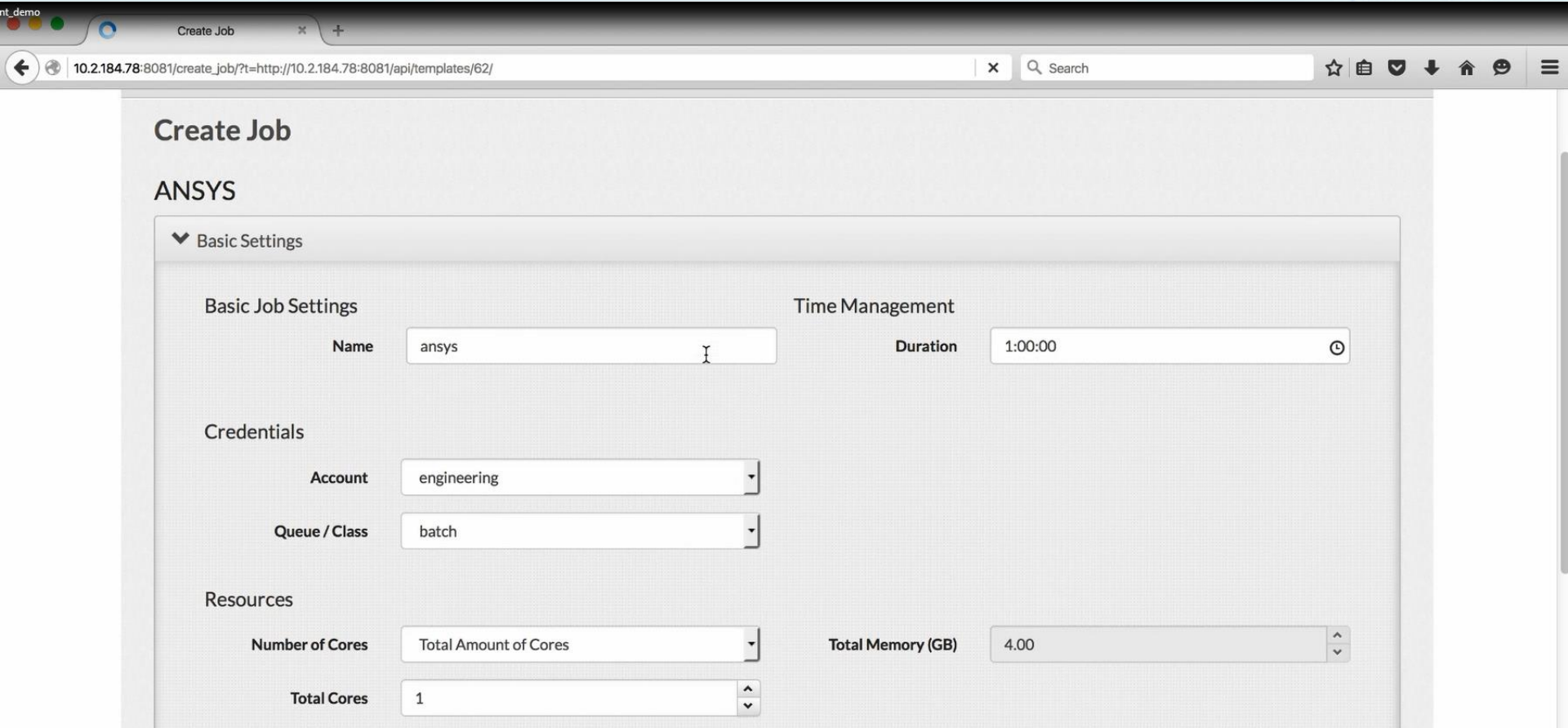
# Click Templates on your User Dashboard



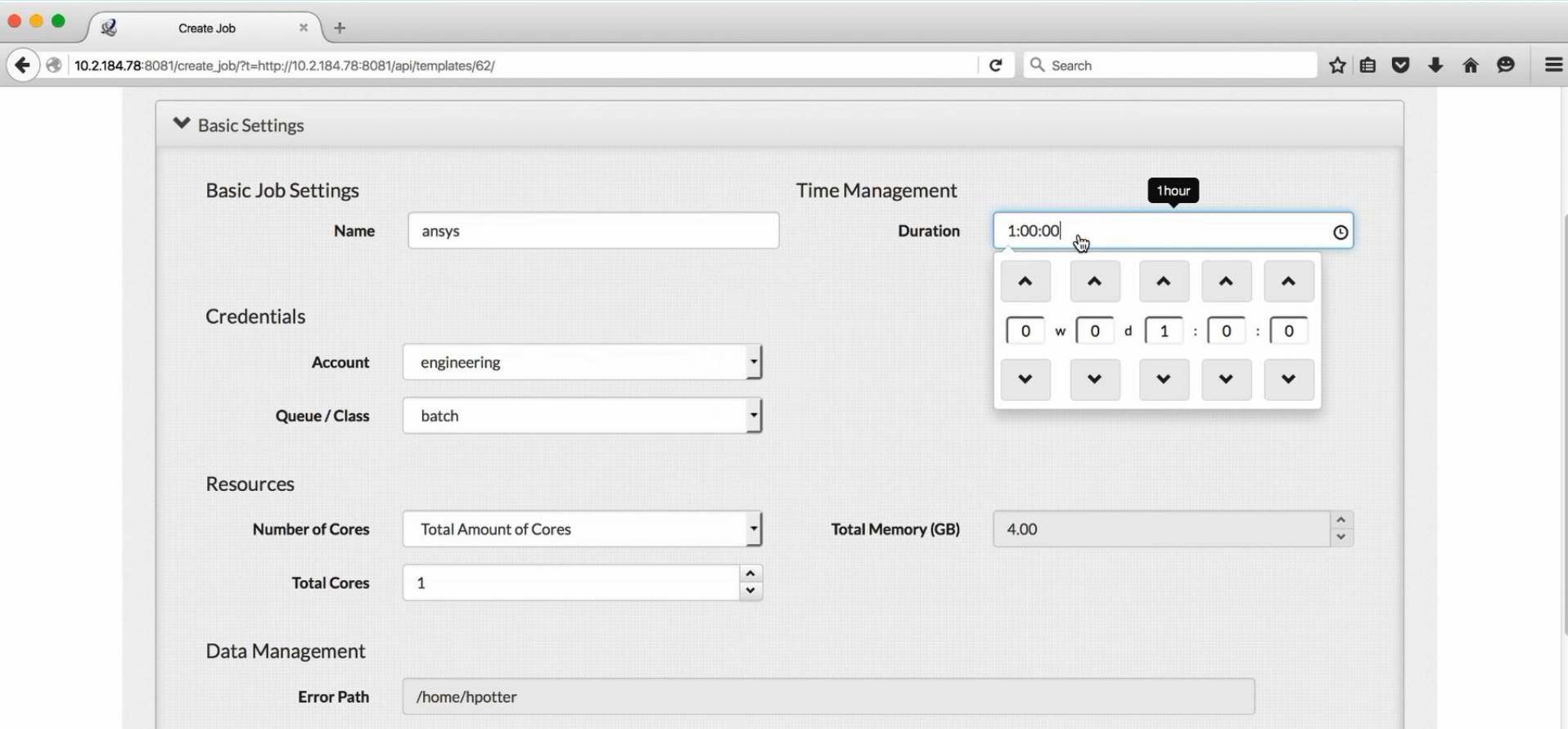
# Use “xyz” application template



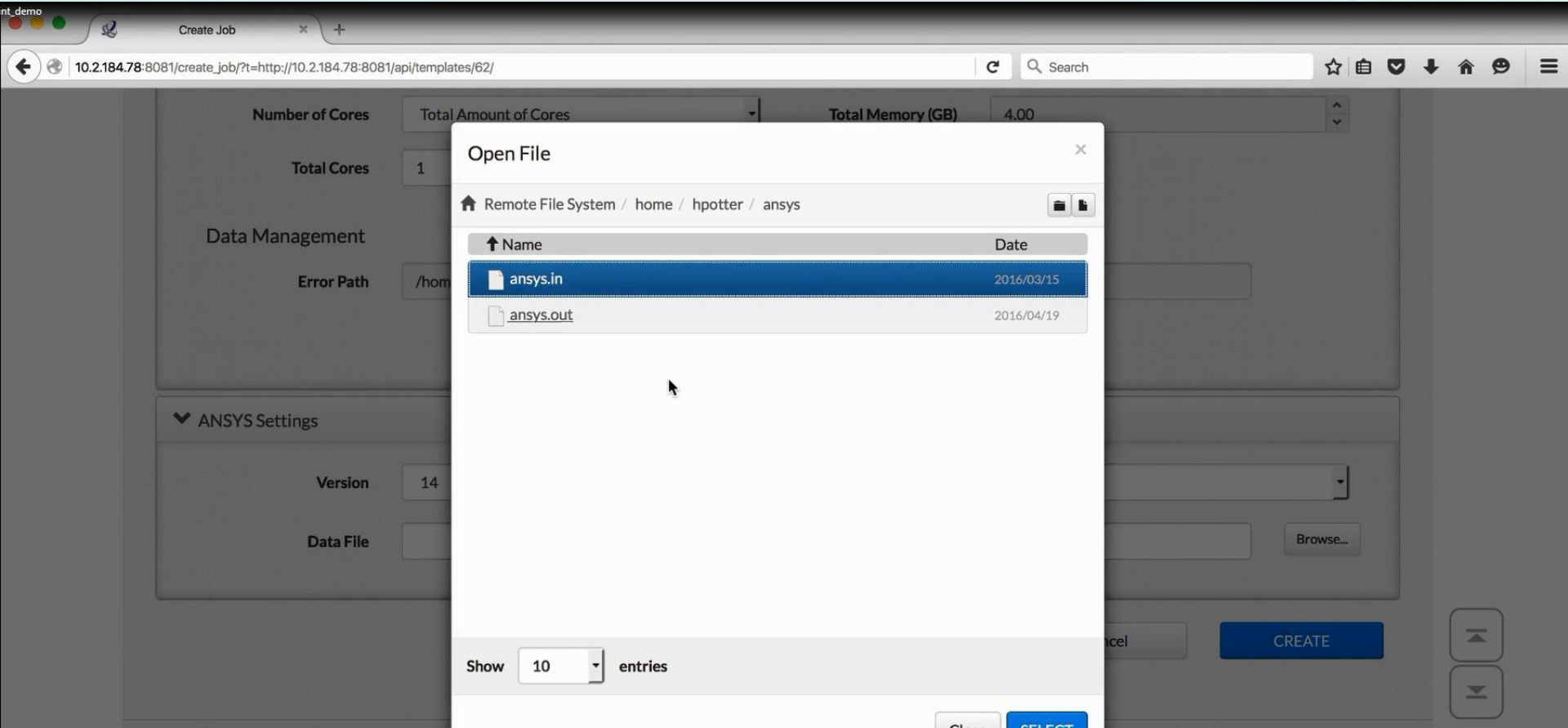
# Fill in application template



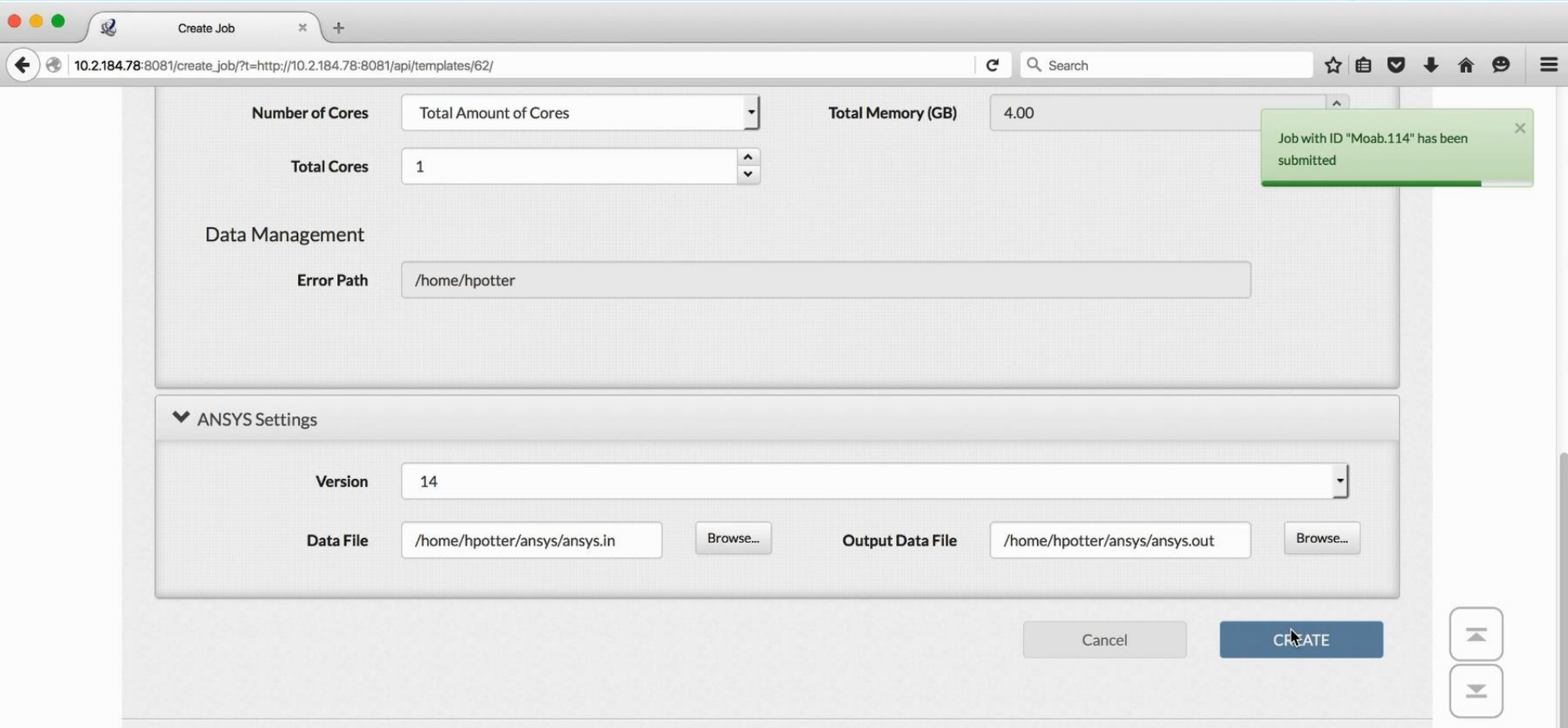
# Choose a time limit



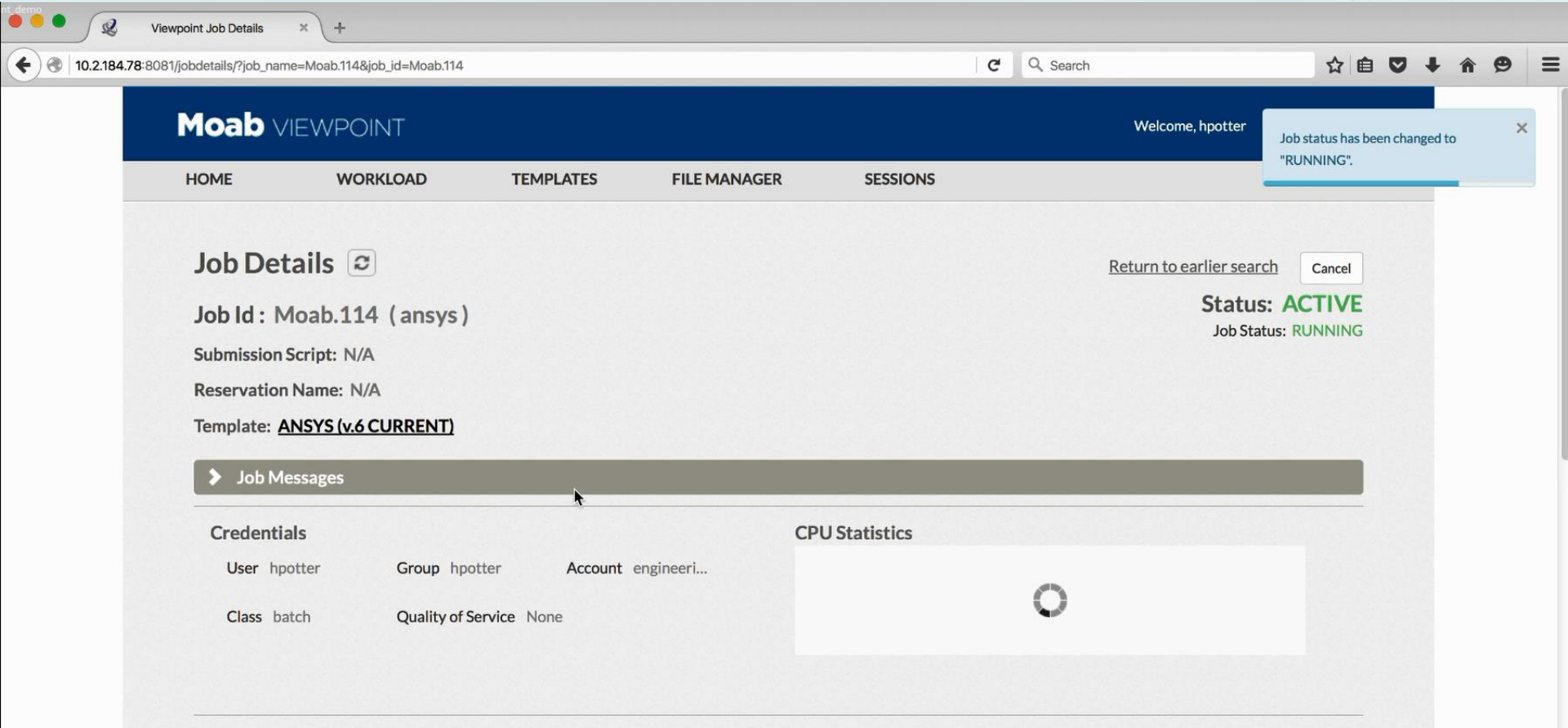
# Tell it where your data file is



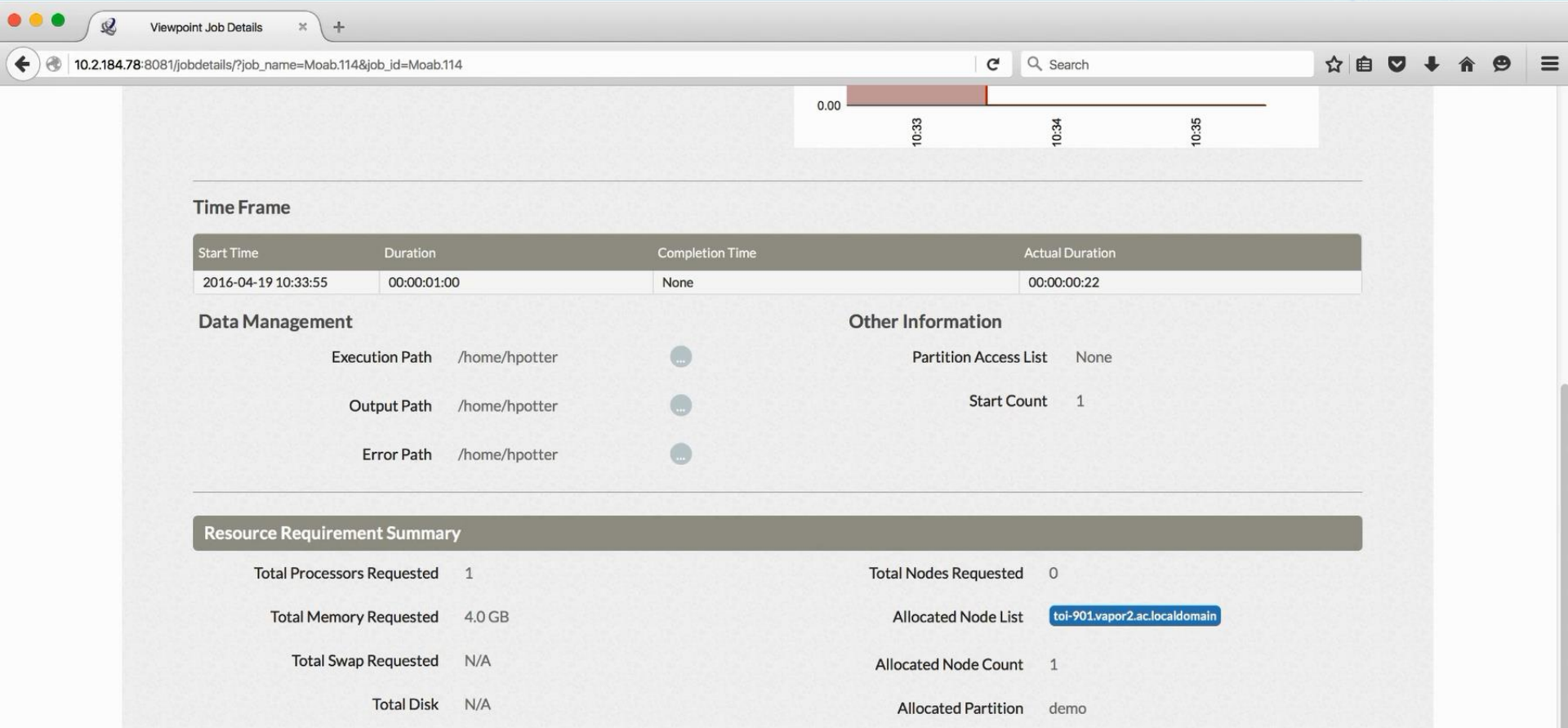
# Submit your “xyz” job



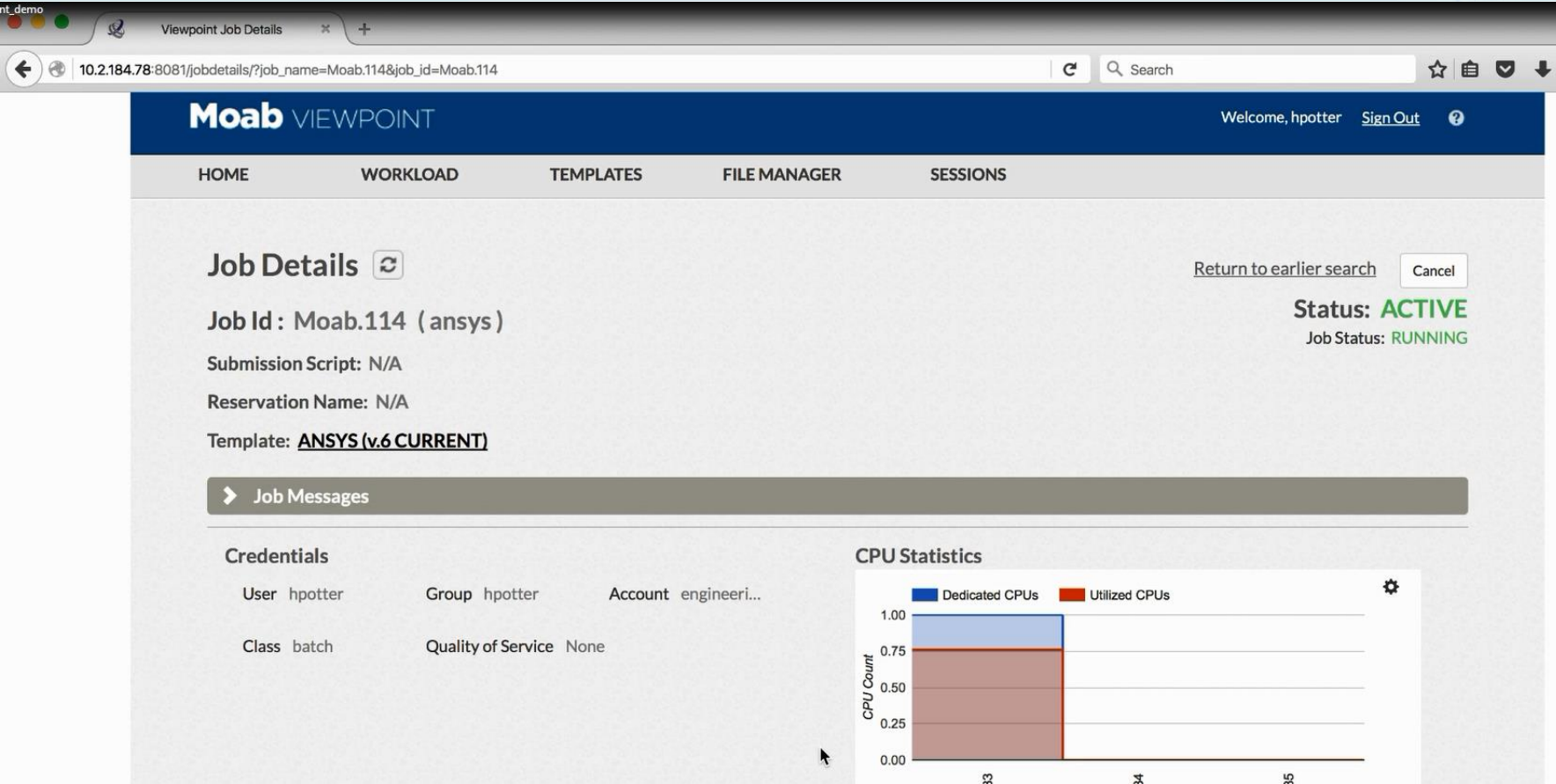
# Click Workload to check on your “xyz” job



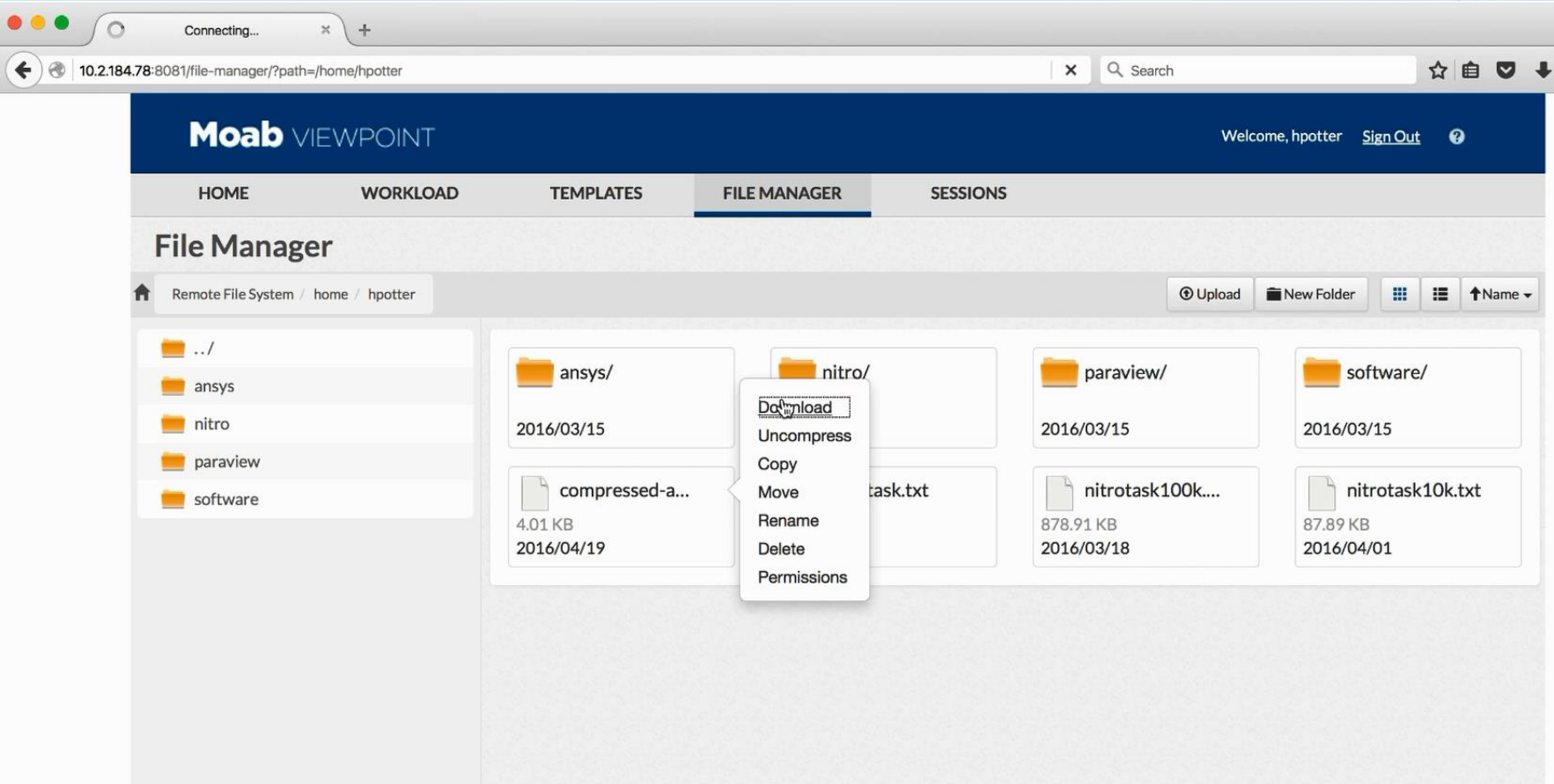
# Check out your “xyz” job’s details



# Look at your job's statistics



# Download your “xyz” job’s results



# Benefits

- Allows non-technical users to submit jobs using pre-defined application templates
- Application templates contain
  - Job script for running application
  - Web form design with pre-defined and/or custom widgets and their variables
  - Job submission replaces variables within job script with form's widget values


© 2016 Adaptive Computing Enterprises, Inc.

# Agenda

- DataWarp
- Intel MIC “KNL”
- Viewpoint Web Portal (non-technical users)
  - User Portal
  - **Administrator Portal (application templates)**
  - Remote Visualization
  - Nitro HTC

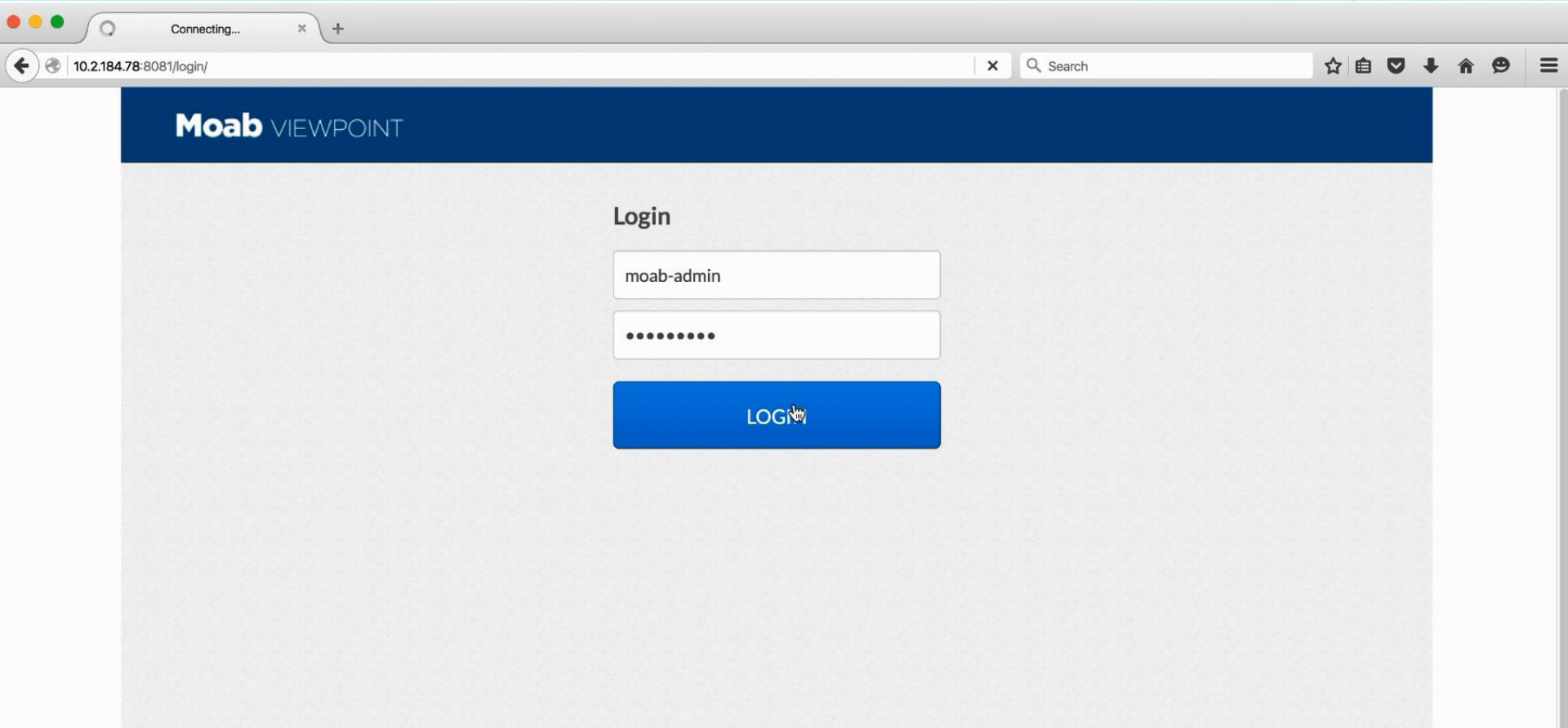

© 2016 Adaptive Computing Enterprises, Inc.

# Use Case

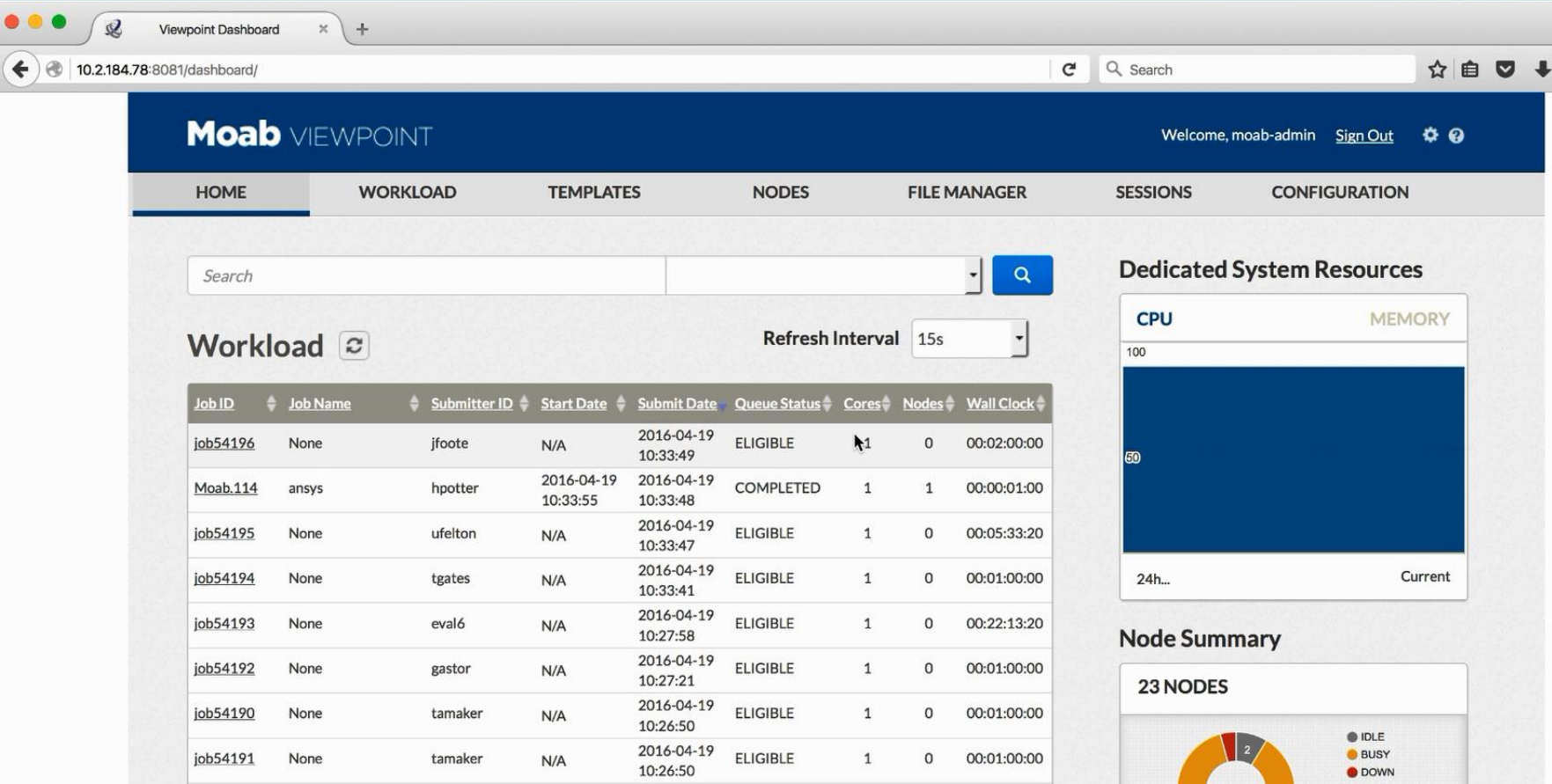
- Administrator needs to support many non-technical users and their applications
- Define application templates using editor
  - Define and customize application template form using pre-defined “widget” types
  - Define widgets’ variable name, default value, and other characteristics to build form
  - Create job script


© 2016 Adaptive Computing Enterprises, Inc.

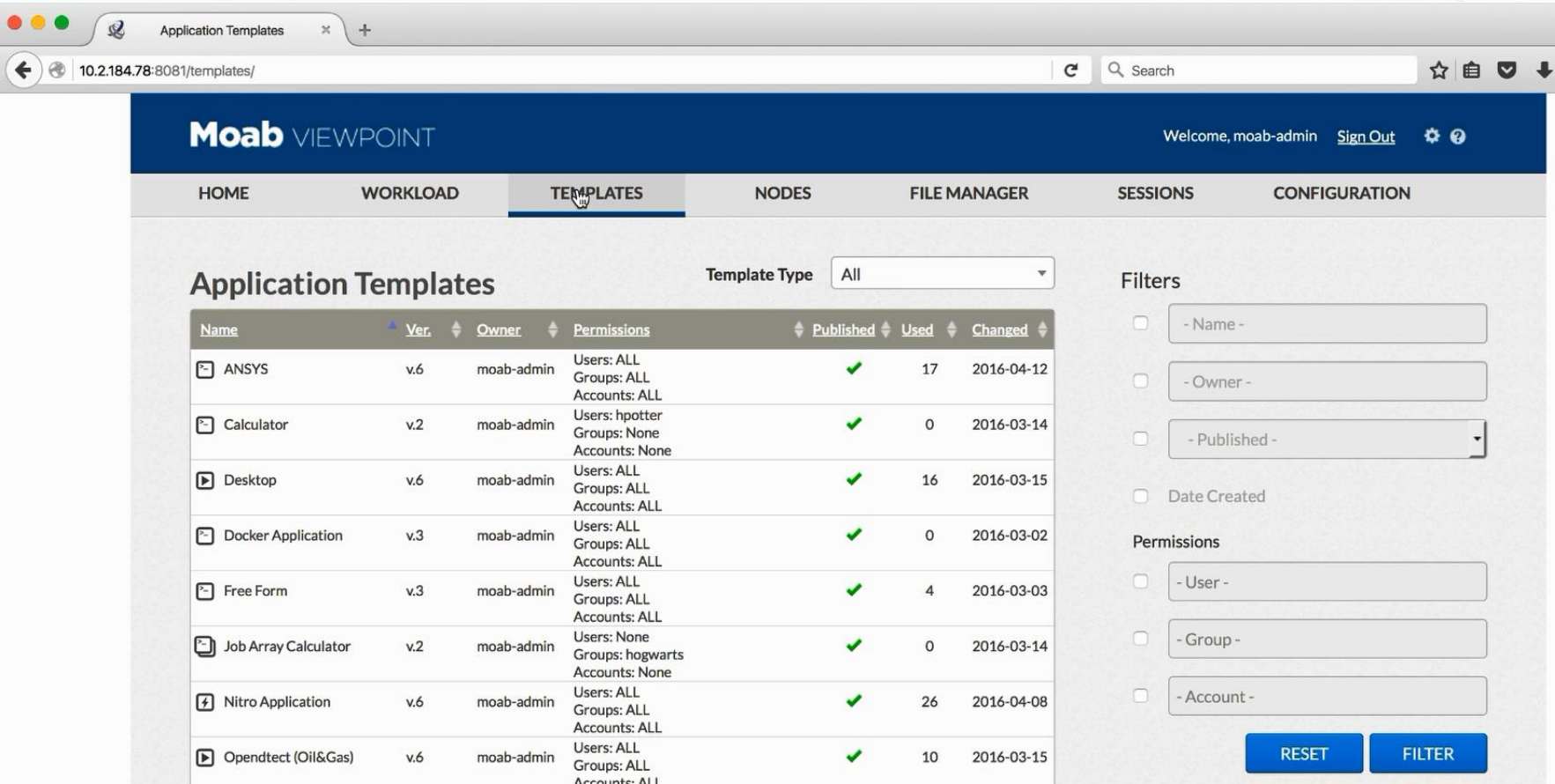
# Log in as admin to Viewpoint Web Portal



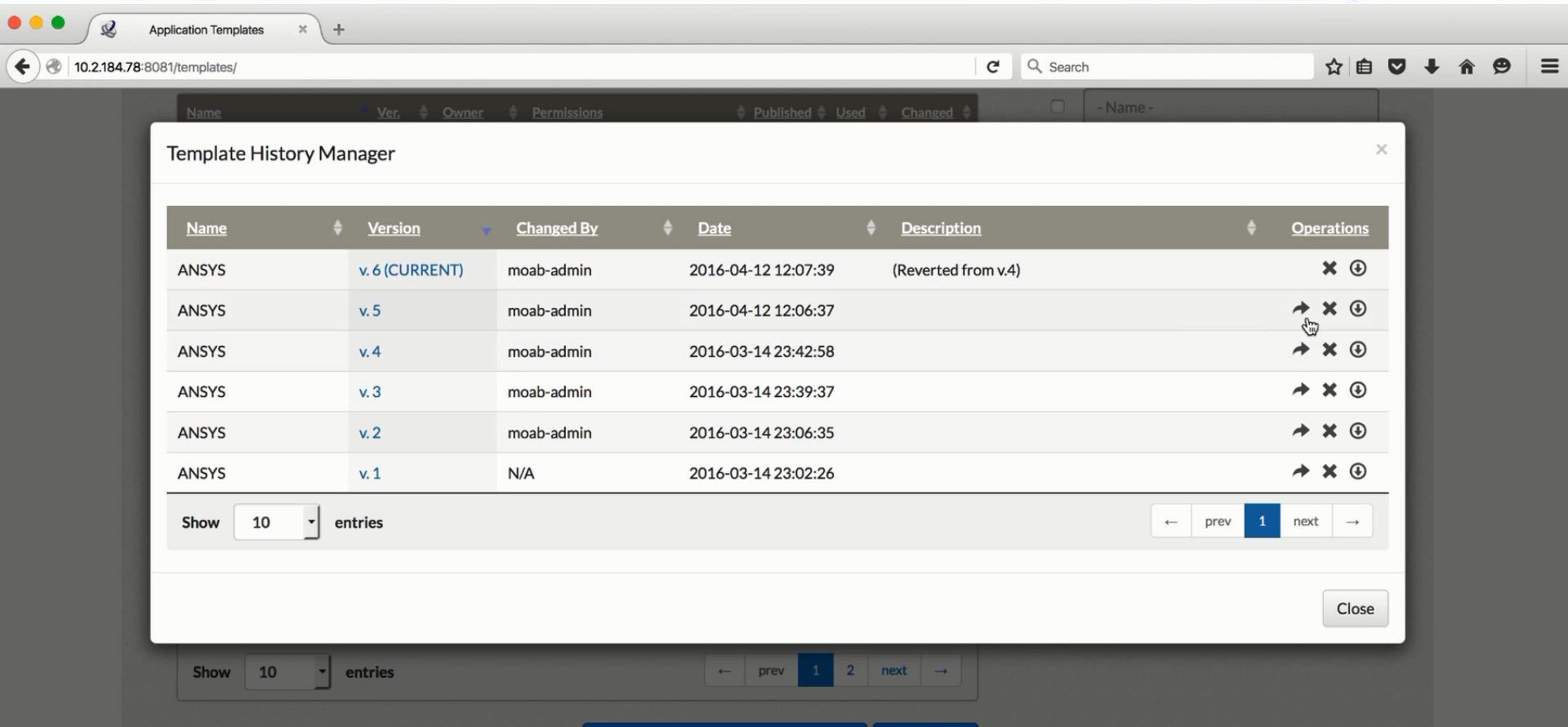
# Click on Templates in Admin Dashboard



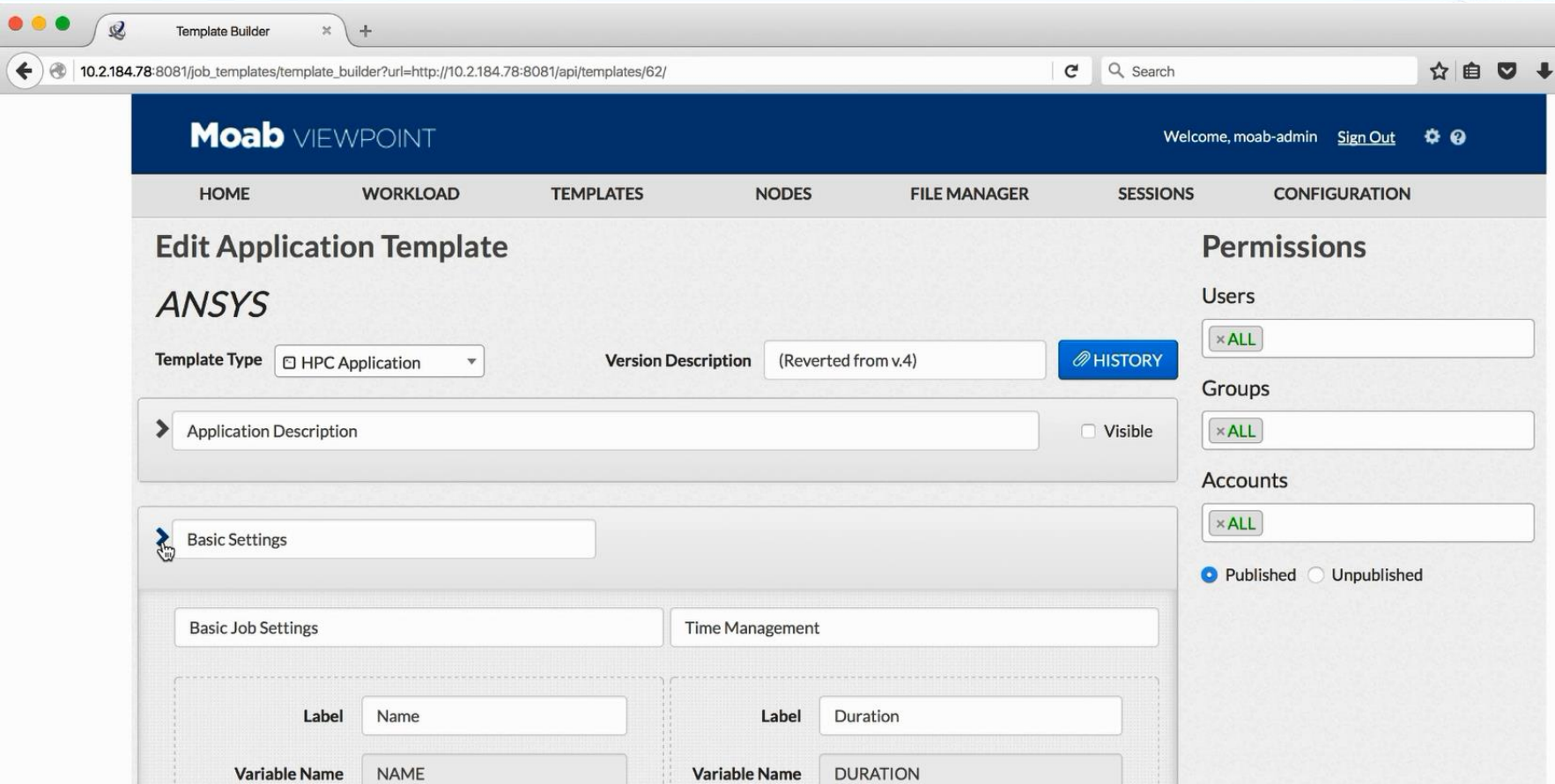
# Choose an application template



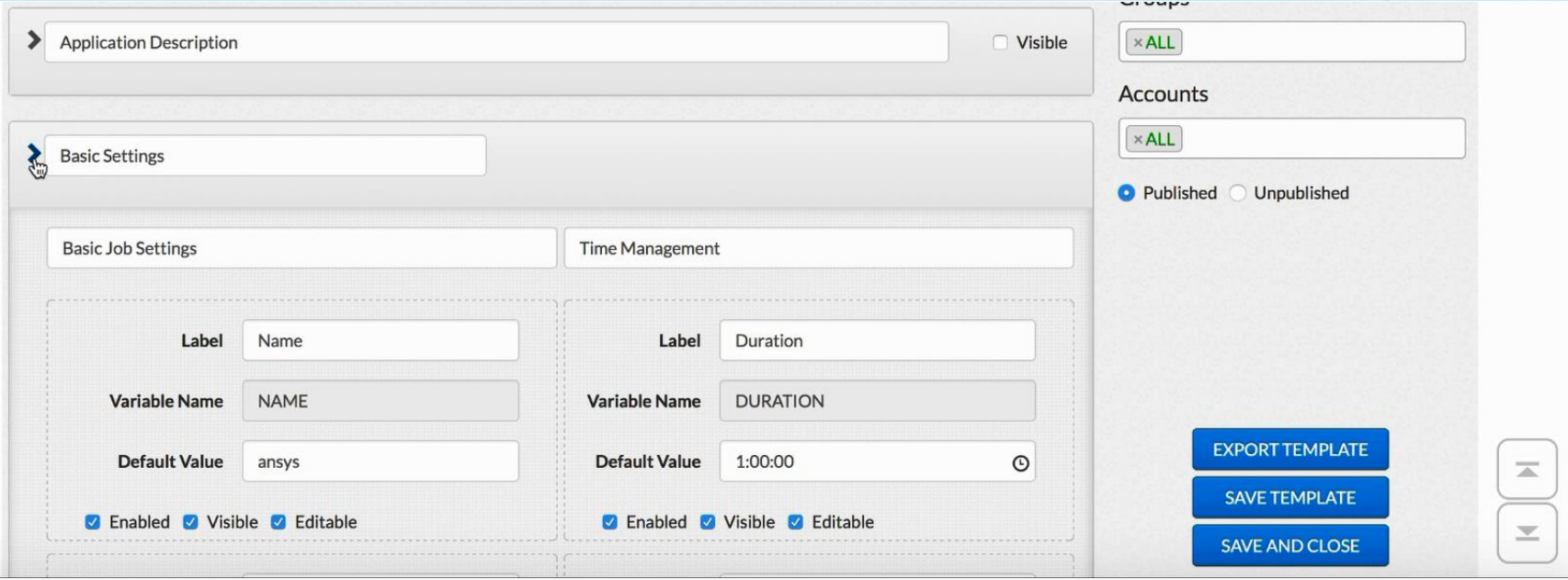
# View application template history



# Edit and design application template



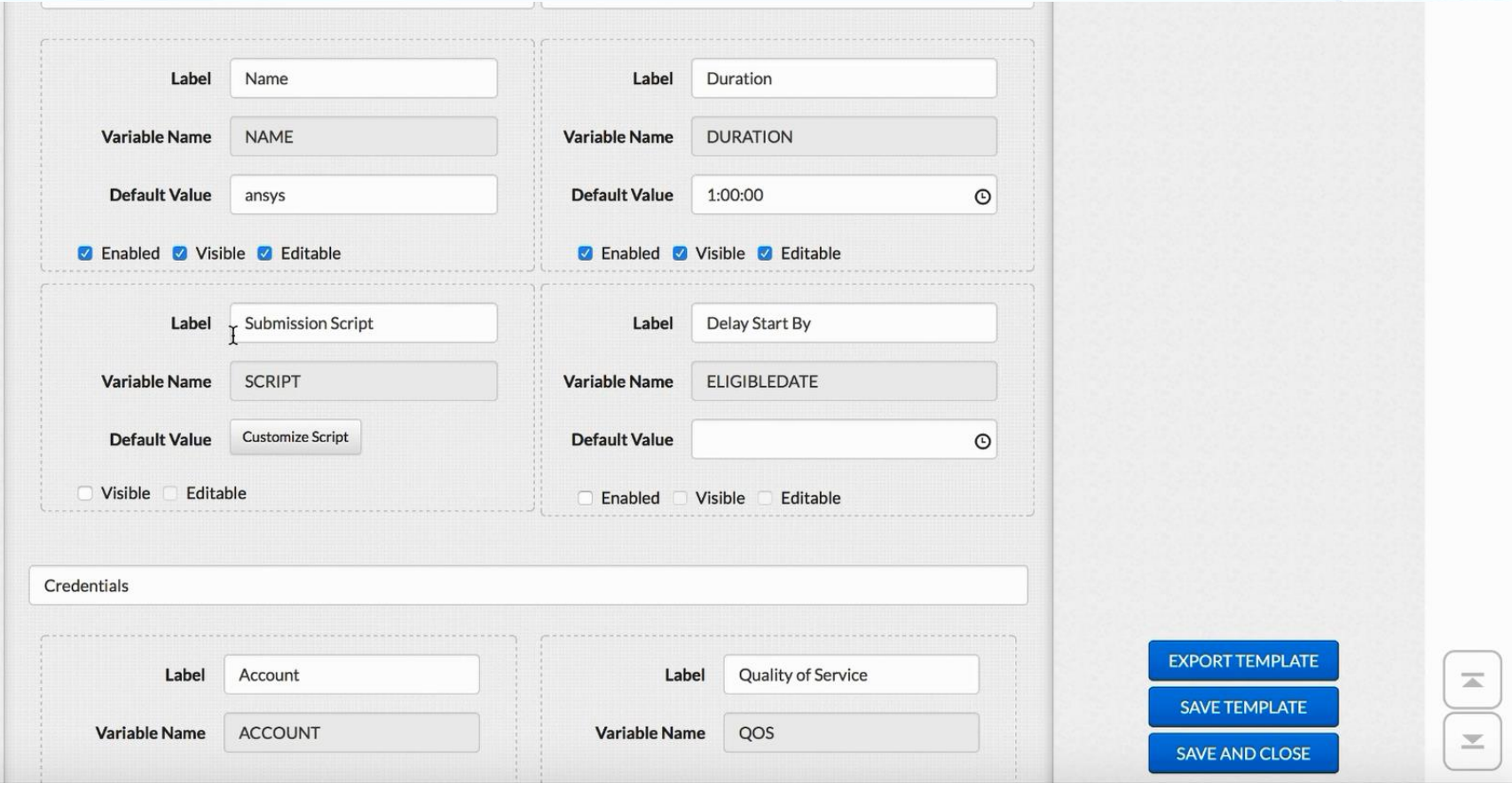
# Edit application template (lower part)



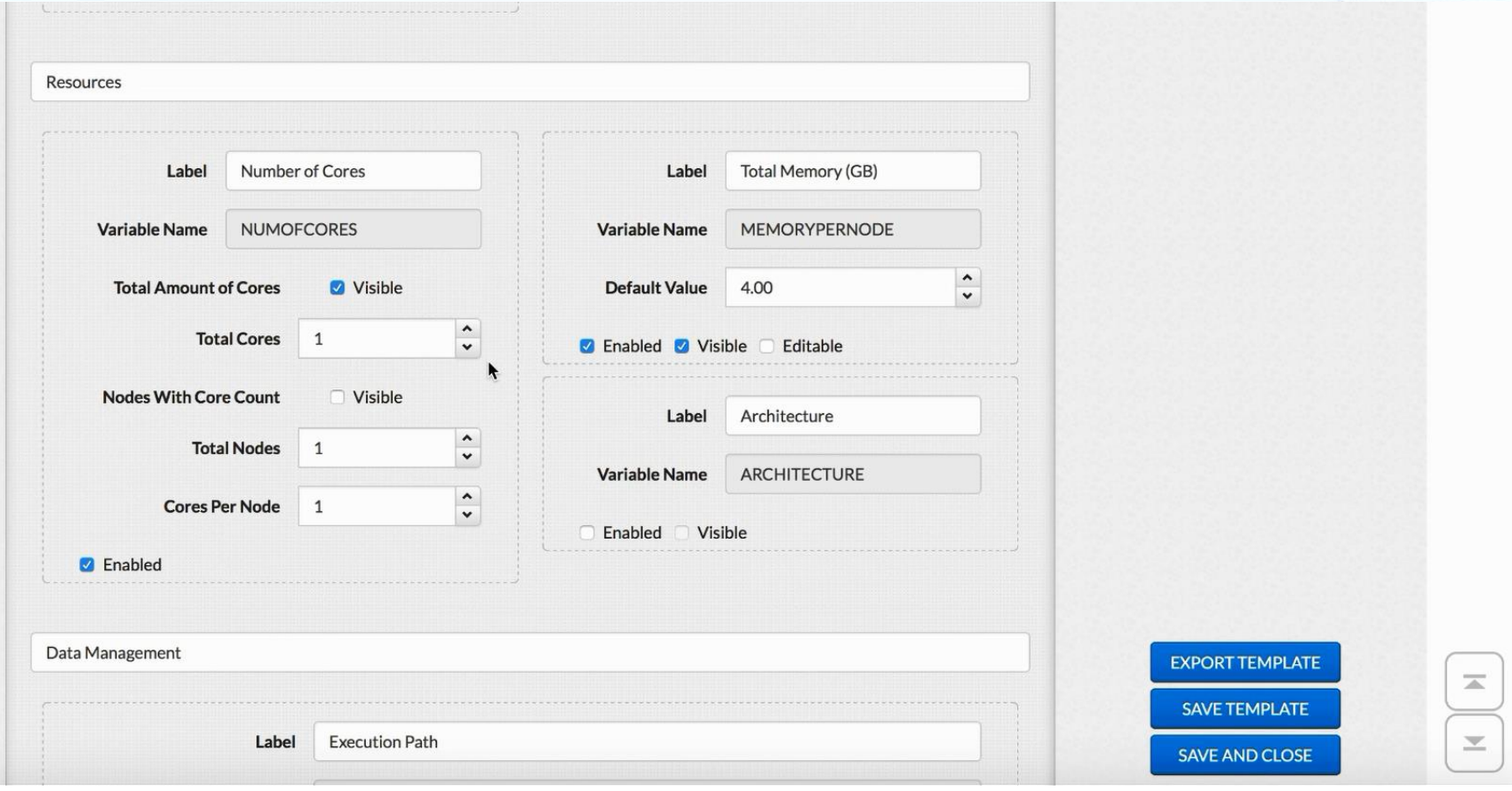



© 2016 Adaptive Computing Enterprises, Inc.

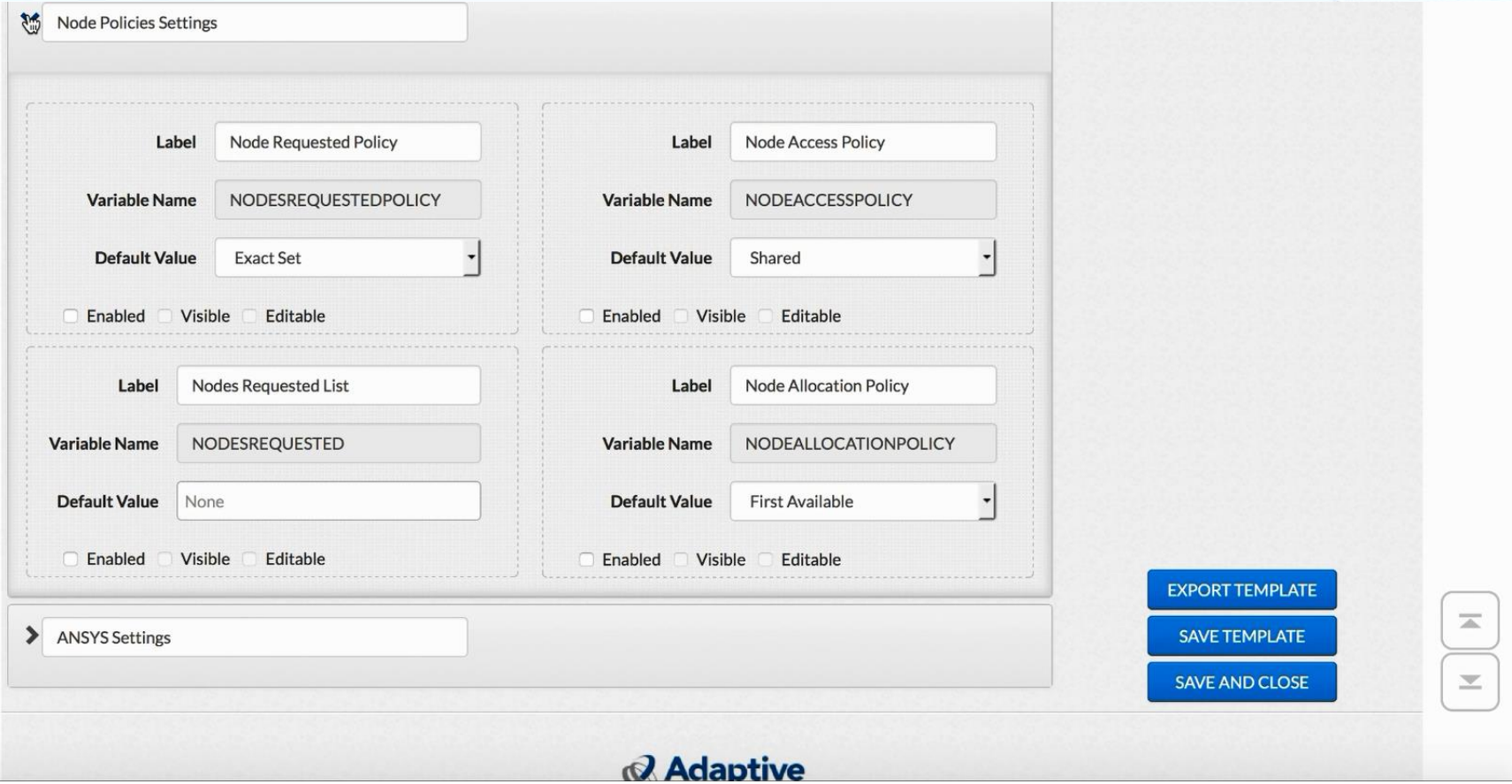
# Choose Basic Job Settings widgets



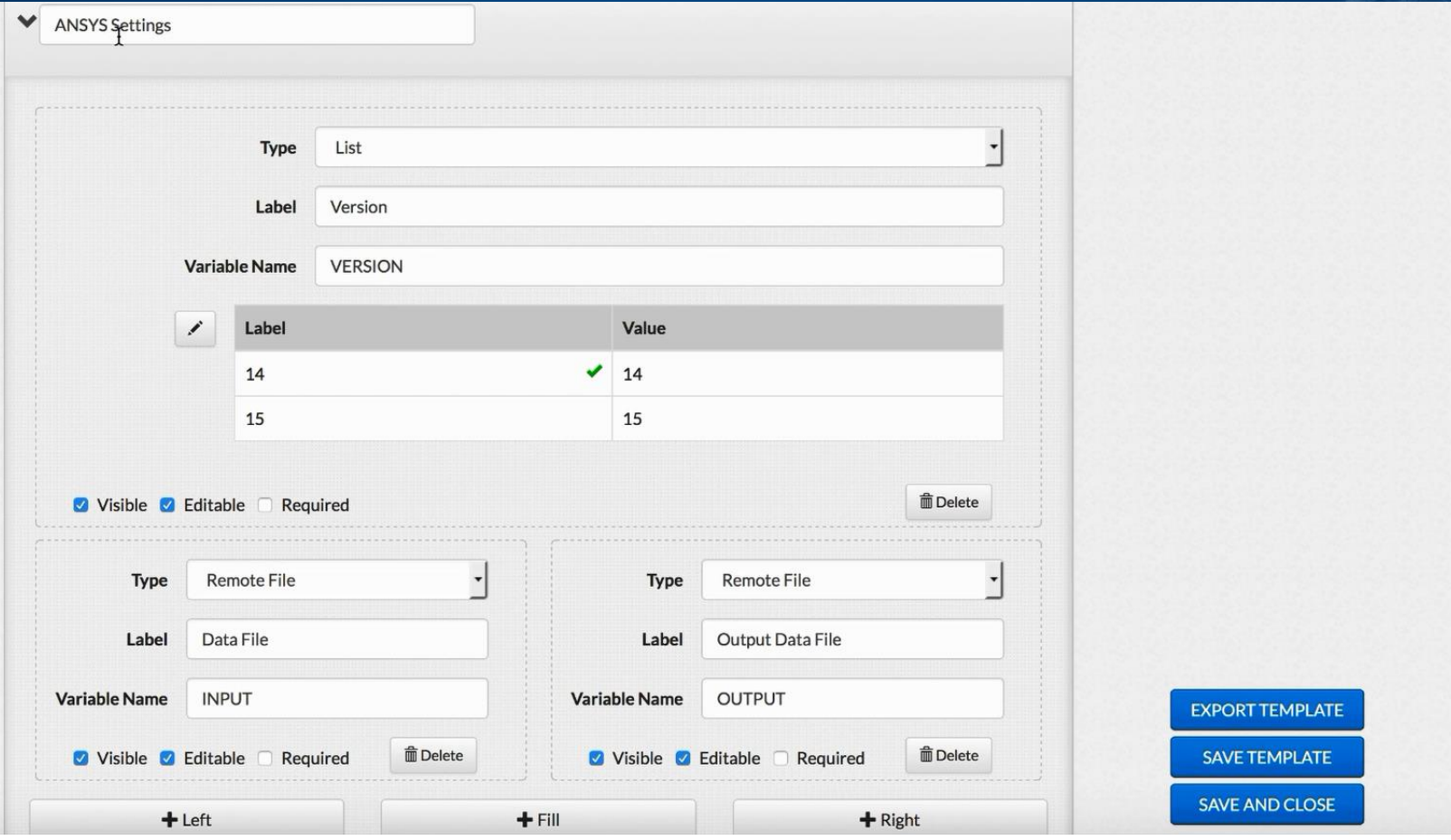
# Choose Resources widgets



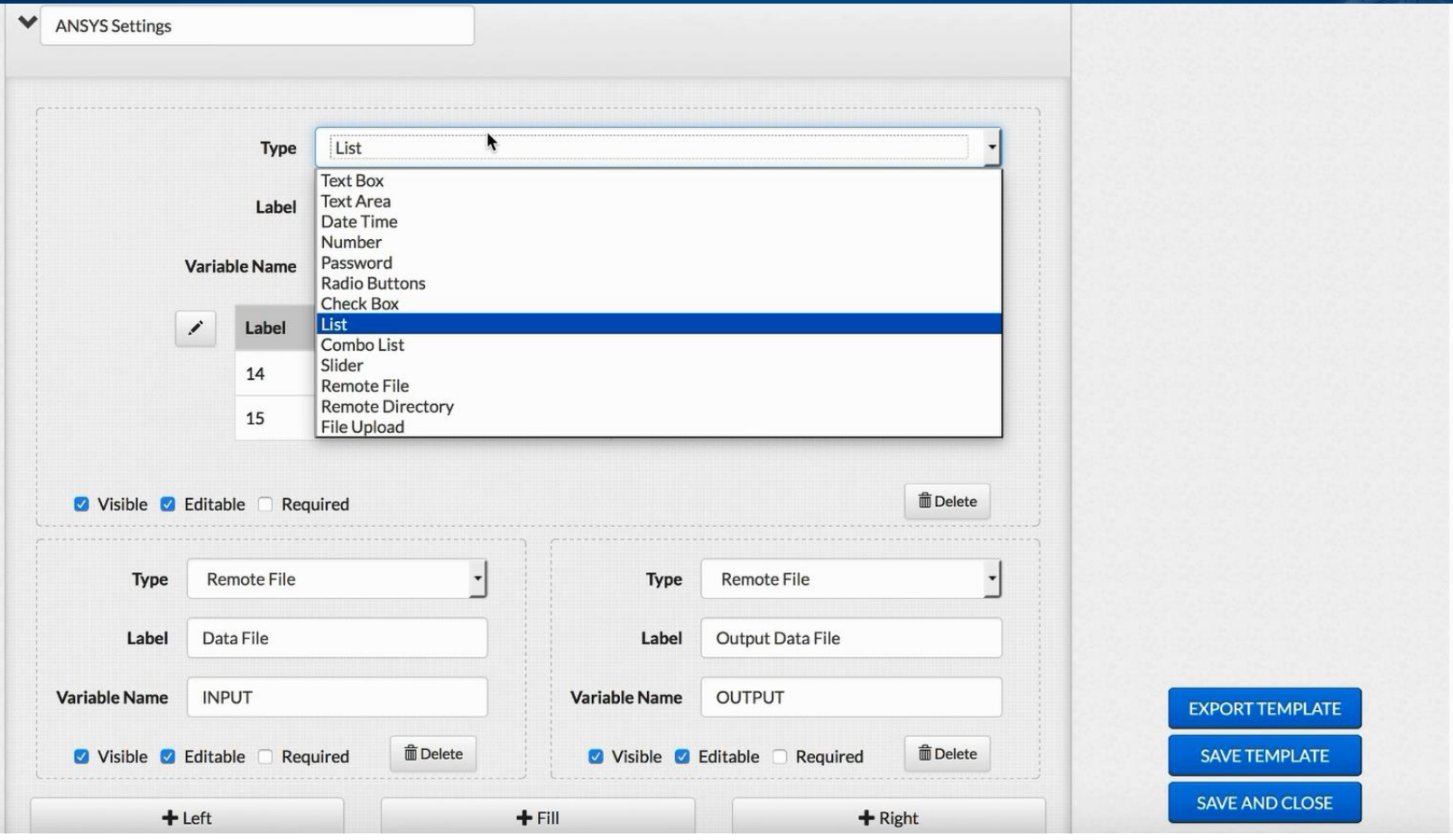
# Choose Node Policies Settings widgets



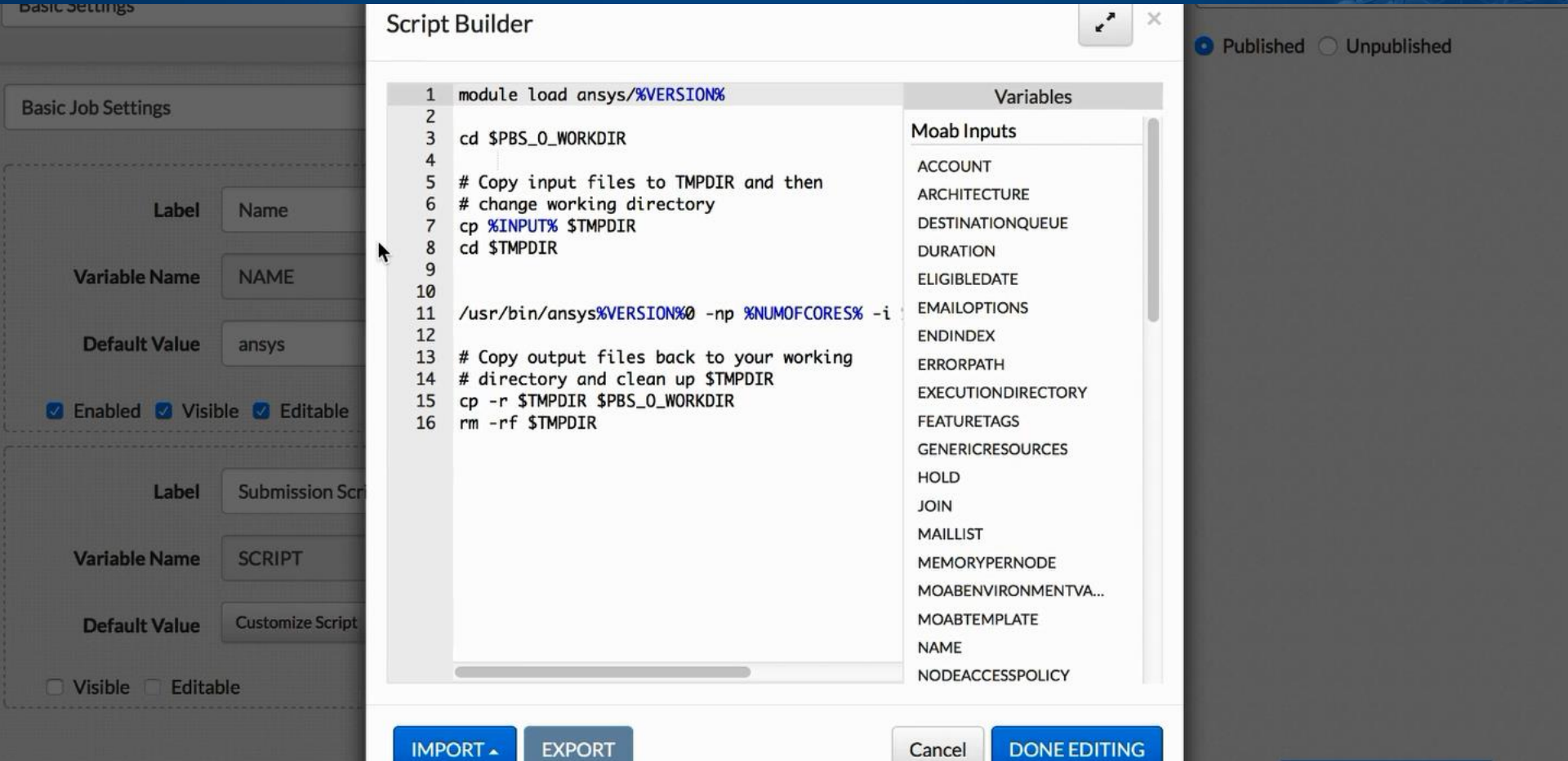
# Design custom “xyz” Settings widgets



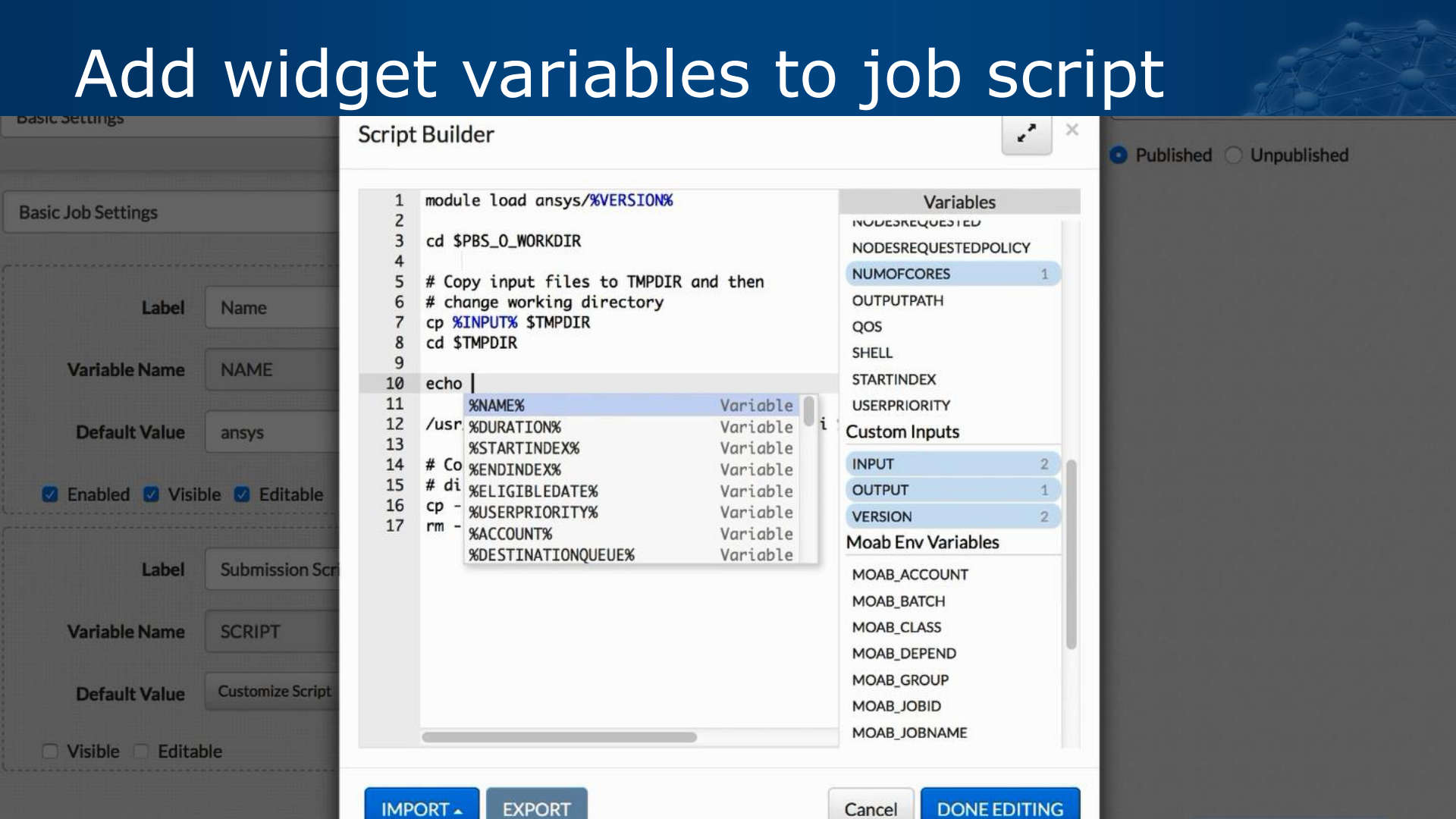
# Choose custom widget from widget type list



# Create/import job script with Script Builder



# Add widget variables to job script



# Benefits

- Application templates let an administrator create application templates for non-technical users.
- Non-technical users have a simple way to submit jobs for specific applications.


© 2016 Adaptive Computing Enterprises, Inc.

# Agenda

- DataWarp
- Intel MIC “KNL”
- Viewpoint Web Portal (non-technical users)
  - User Portal
  - Administrator Portal
  - Remote Visualization (graphical interfaces)
  - Nitro HTC

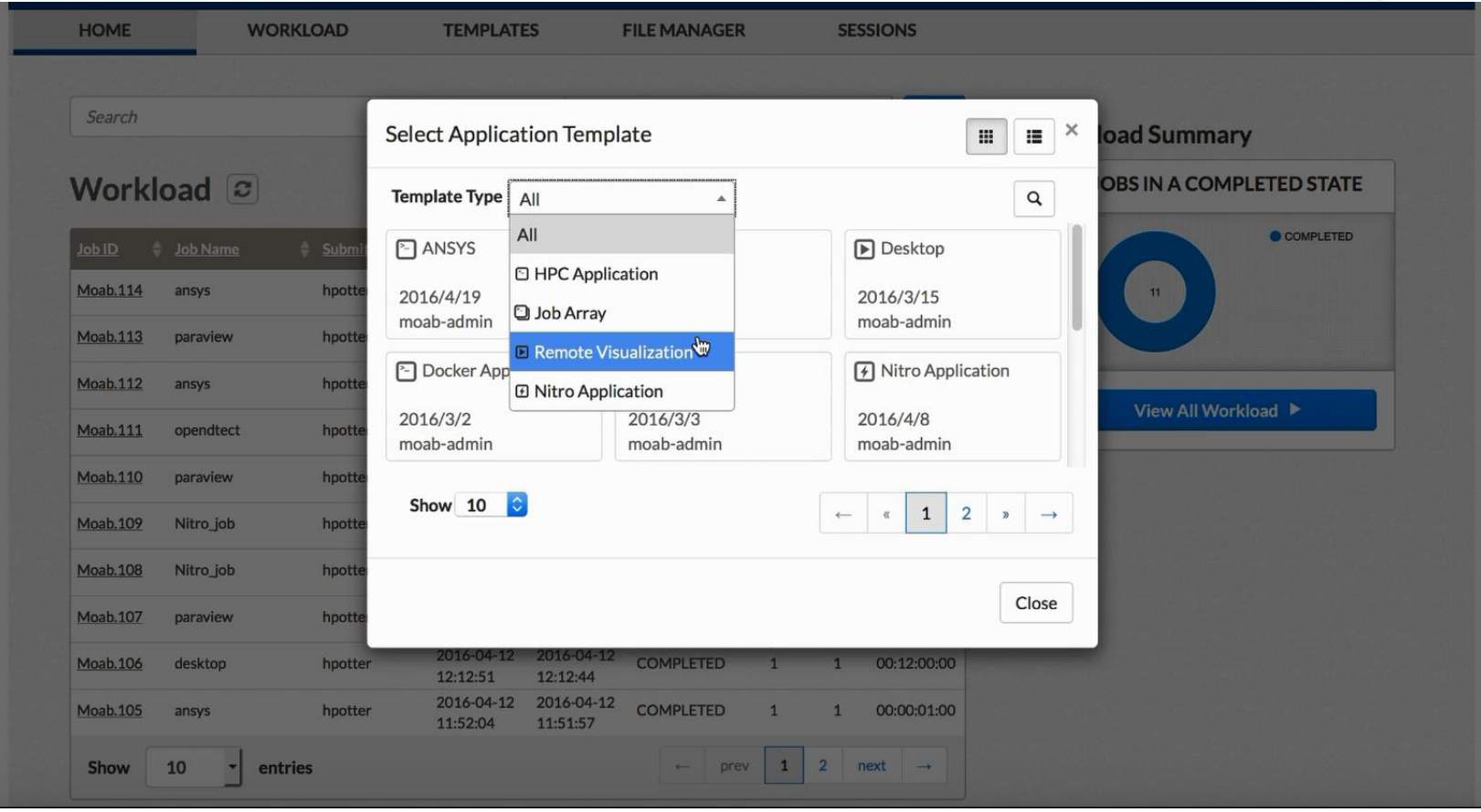

© 2016 Adaptive Computing Enterprises, Inc.

# Use Cases

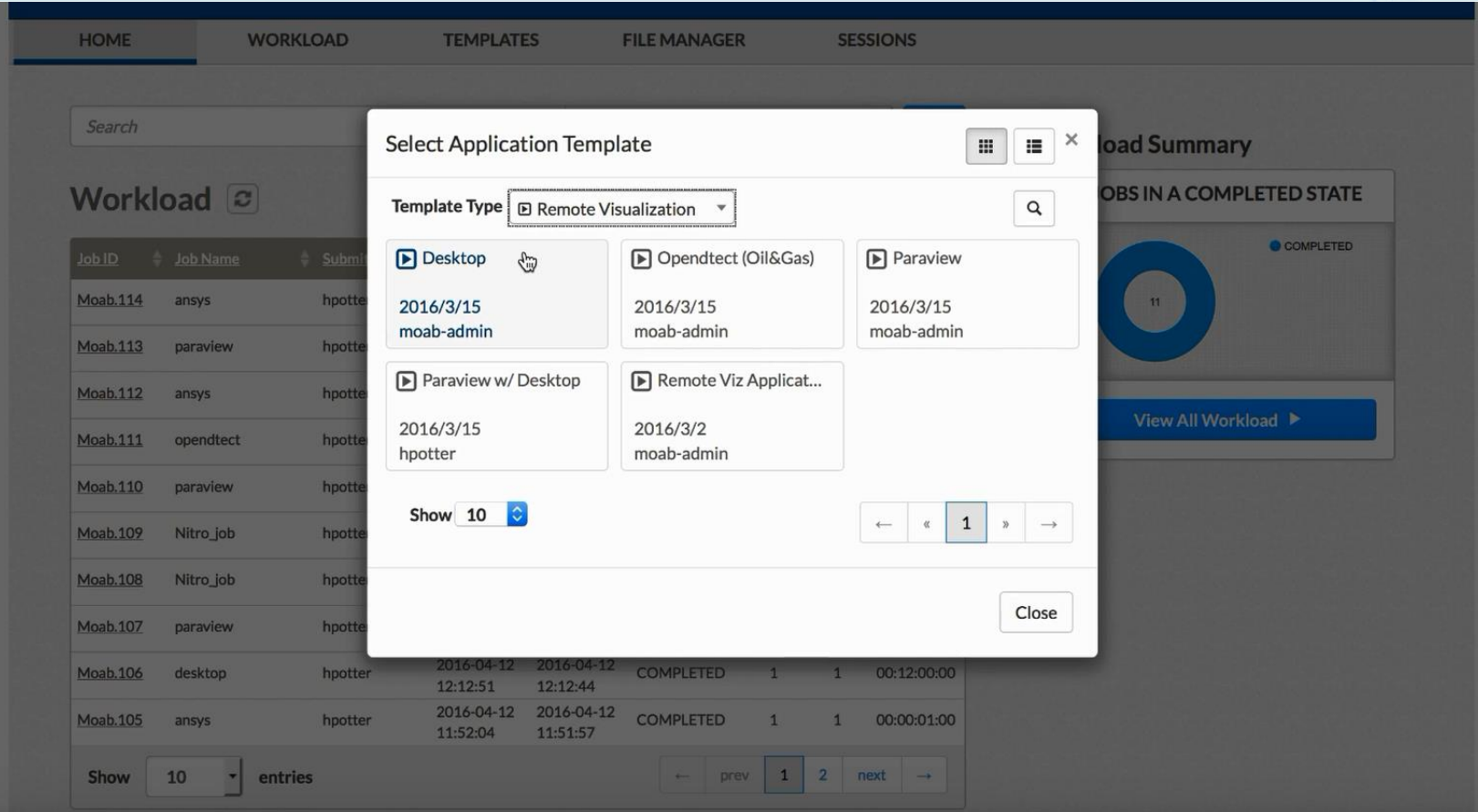
- Users need to visualize their data to gain insights from it.
- Desktops or workstations with decent visualization hardware can be expensive.
- Data can be too large to move to another location.
- Non-technical users need a way to easily set up a remote visualization session.


© 2016 Adaptive Computing Enterprises, Inc.

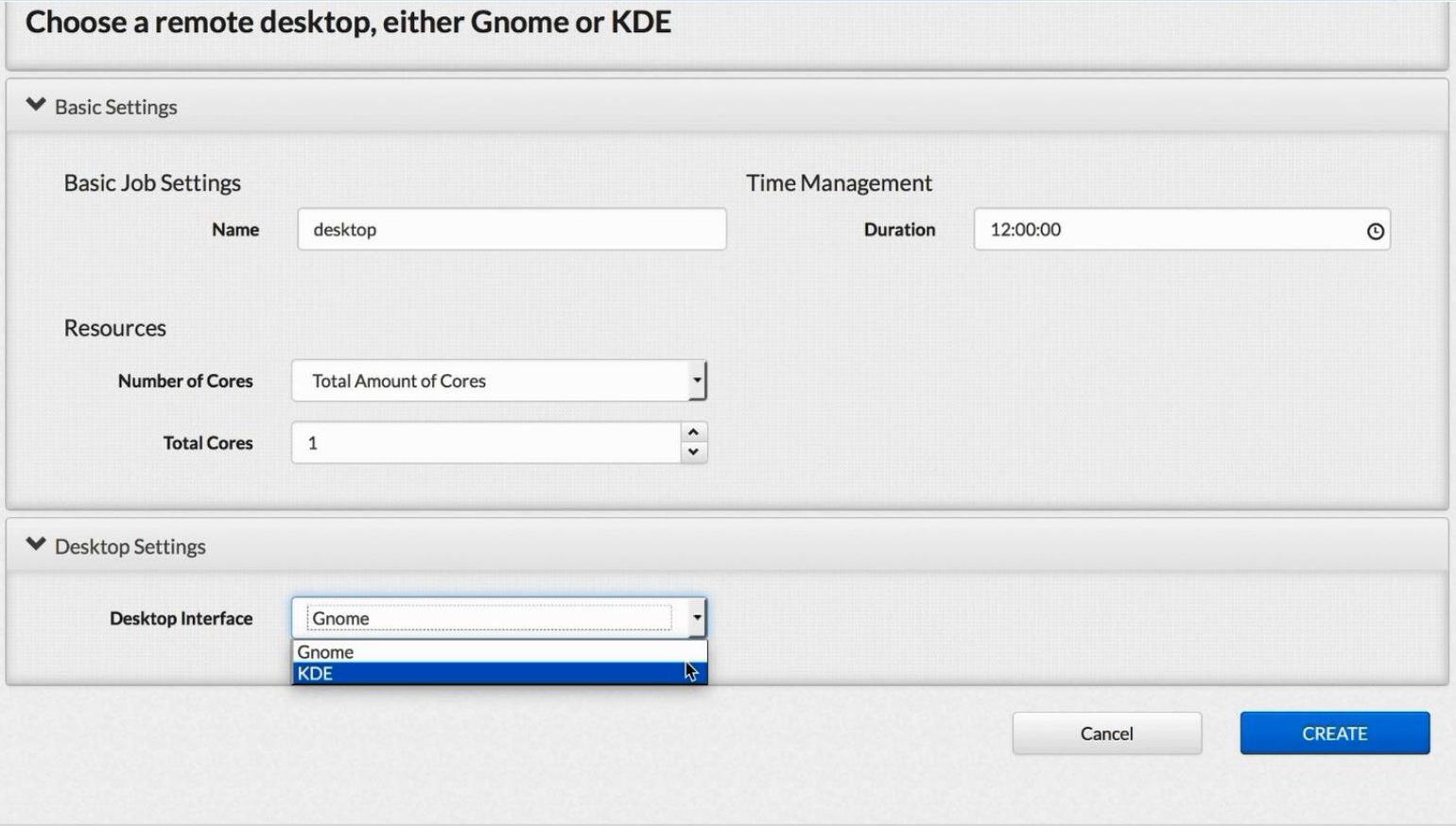
# Choose Remote Visualization Templates



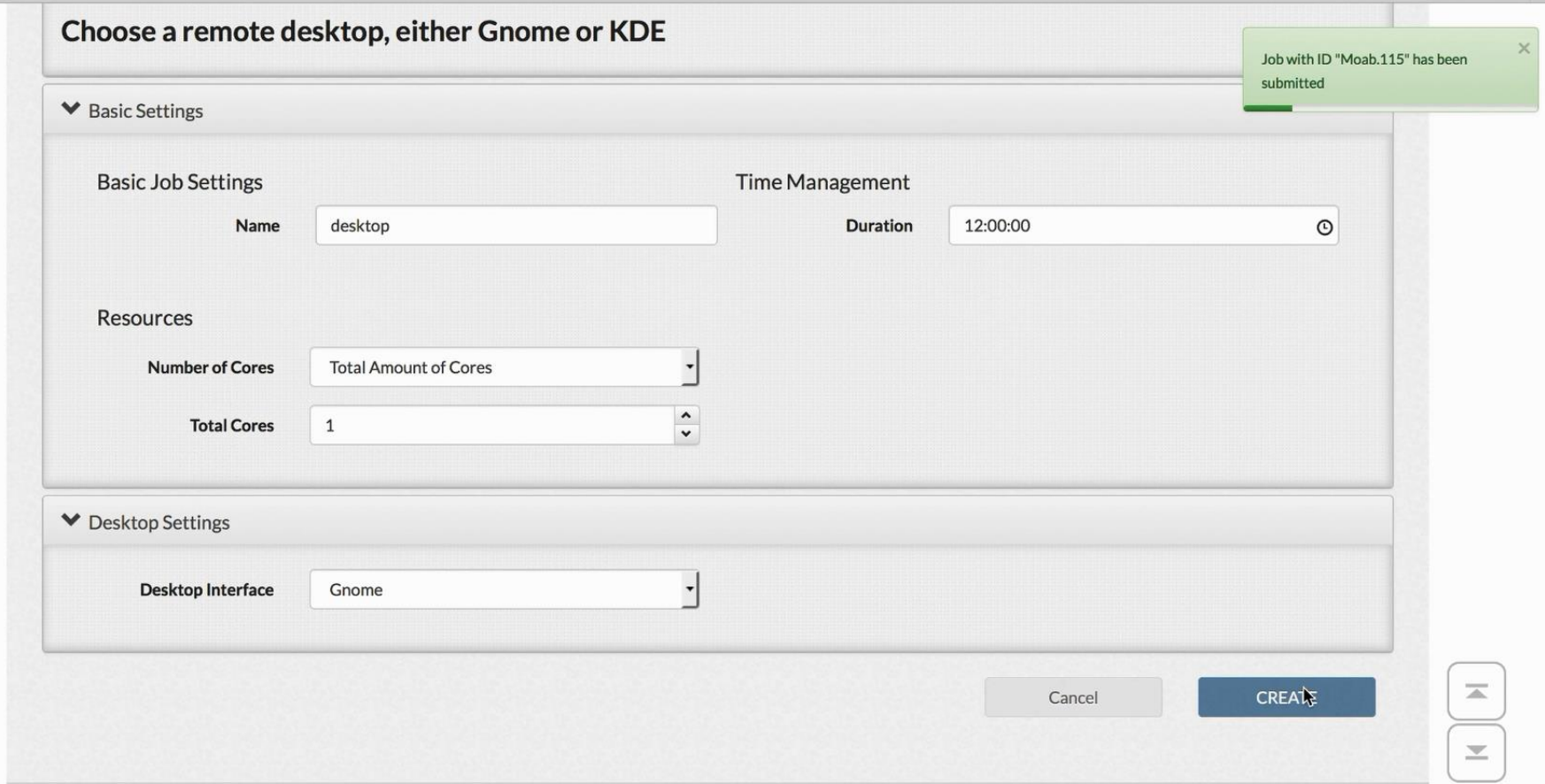
# Choose Remote Desktop App Template



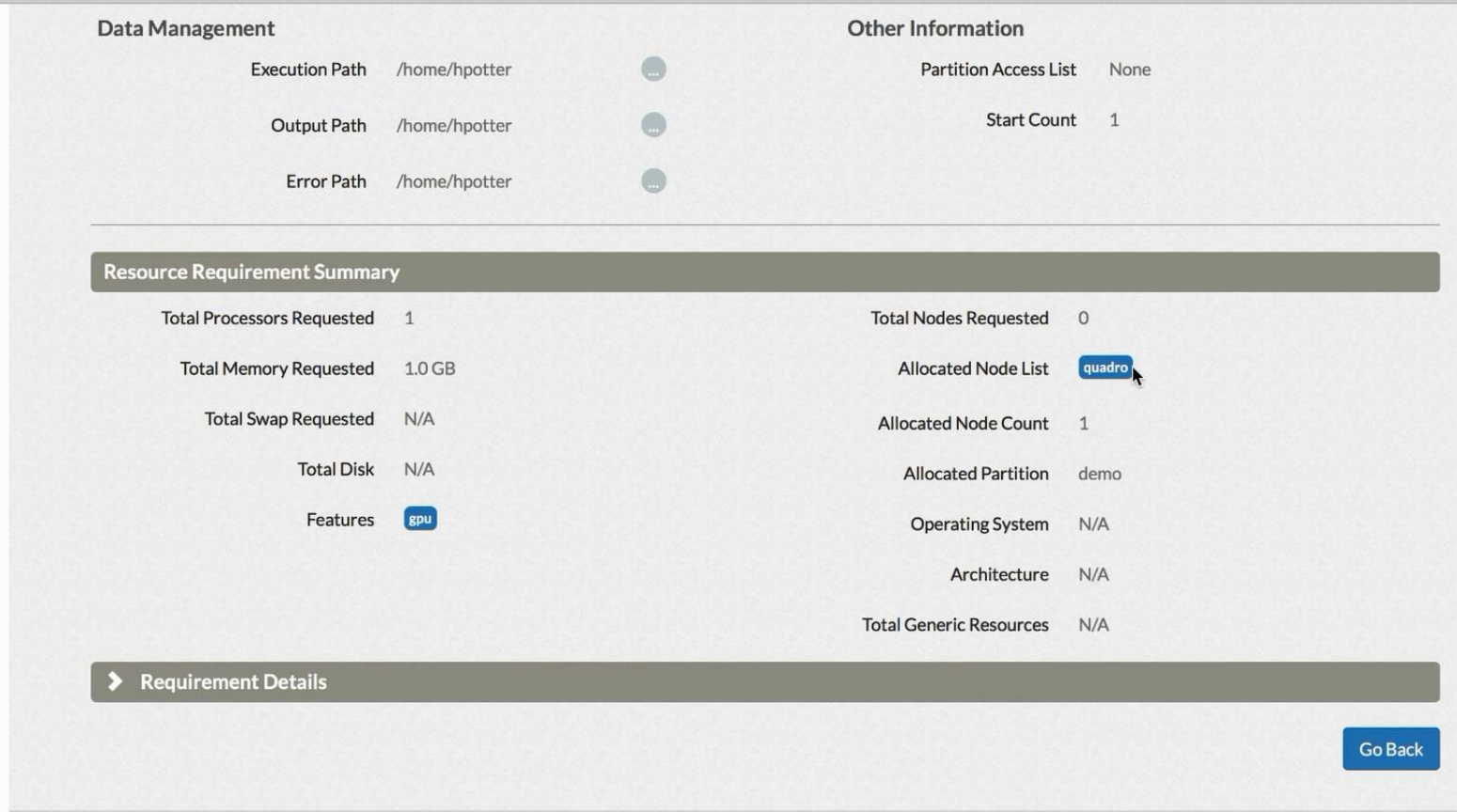
# Choose Graphical Desktop Type



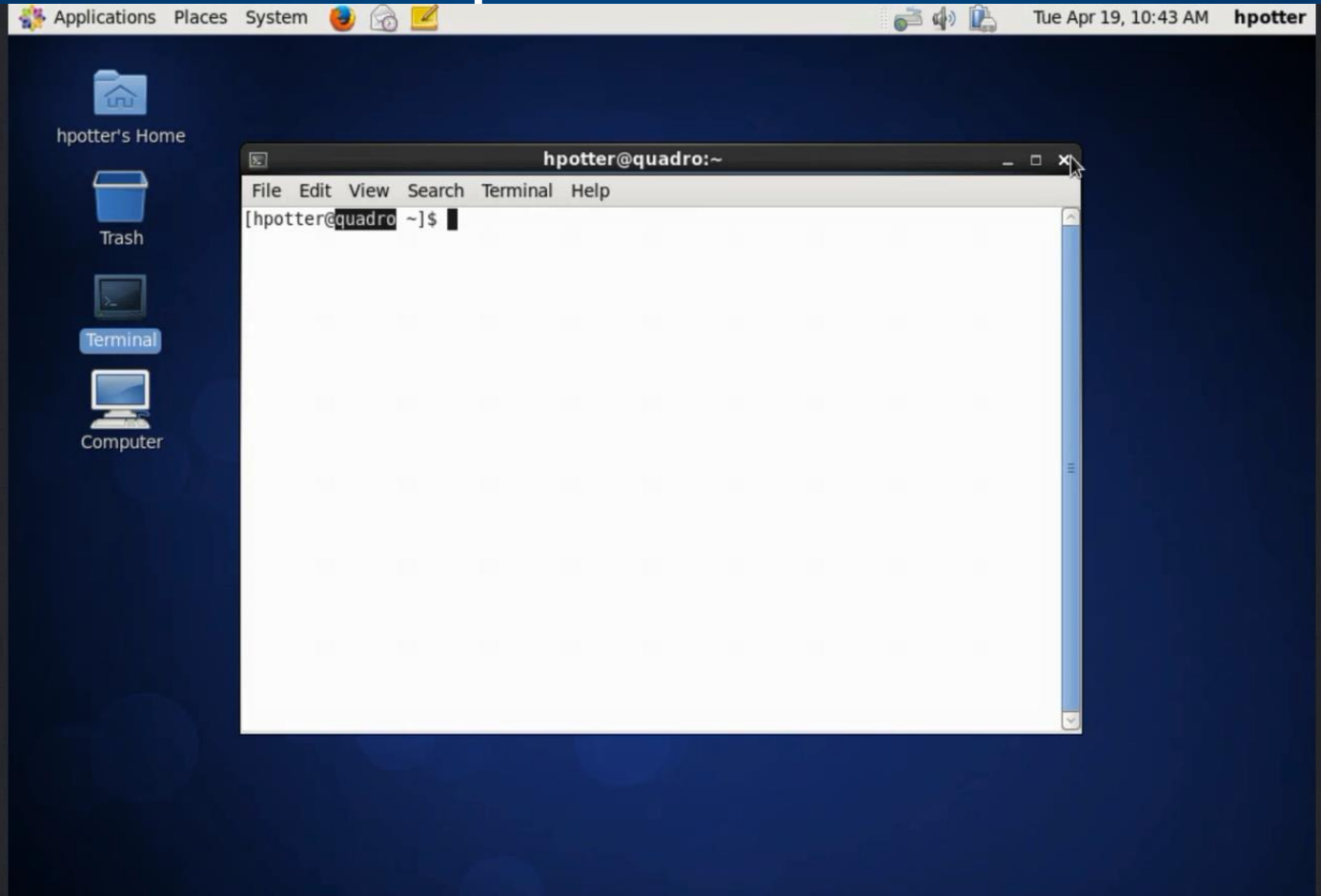
# Submit Remote Desktop Job



# Remote Desktop Job's Compute Node



# Remote Desktop Session Screen

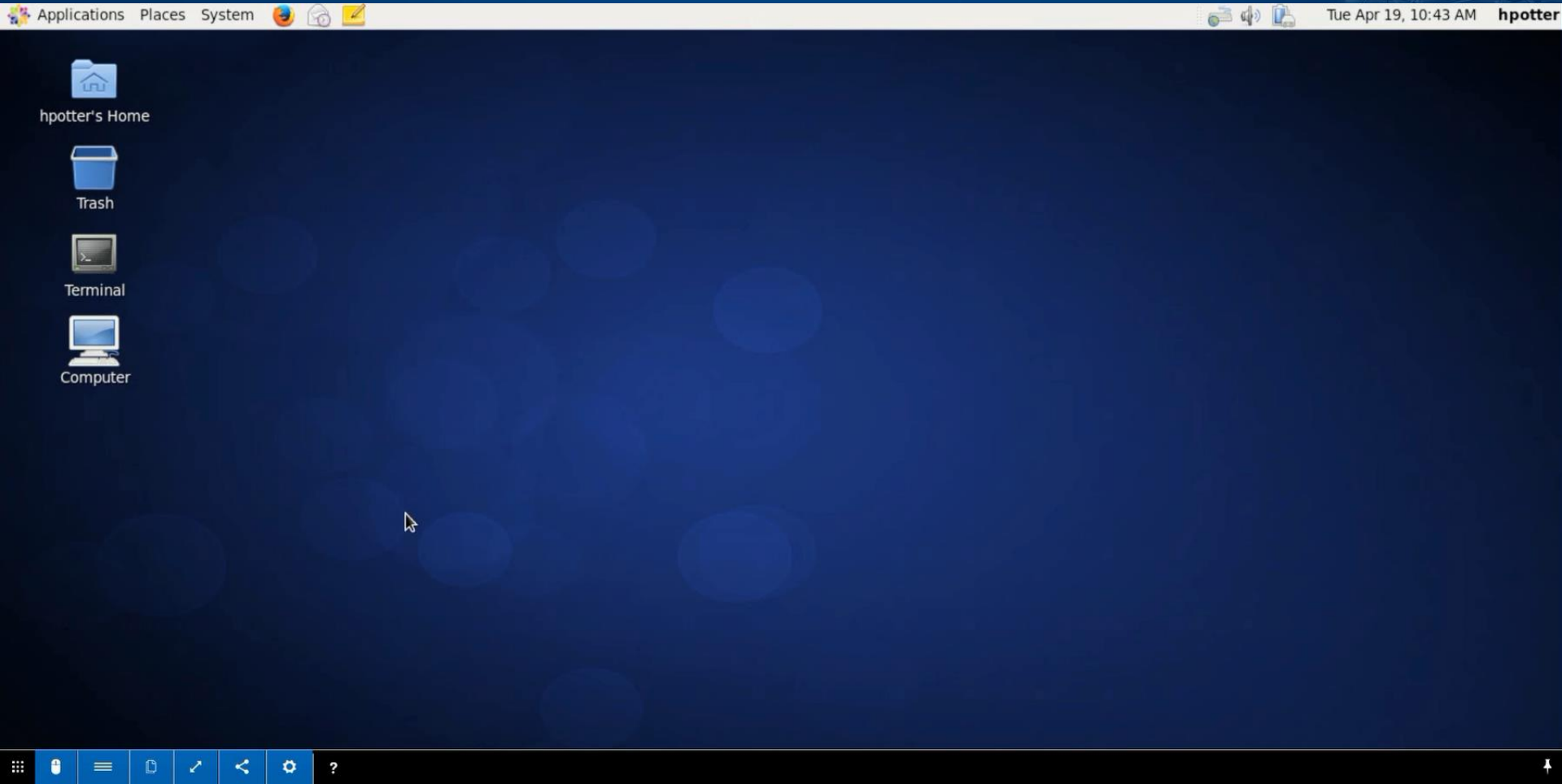


# Remote Visualization Session Controls

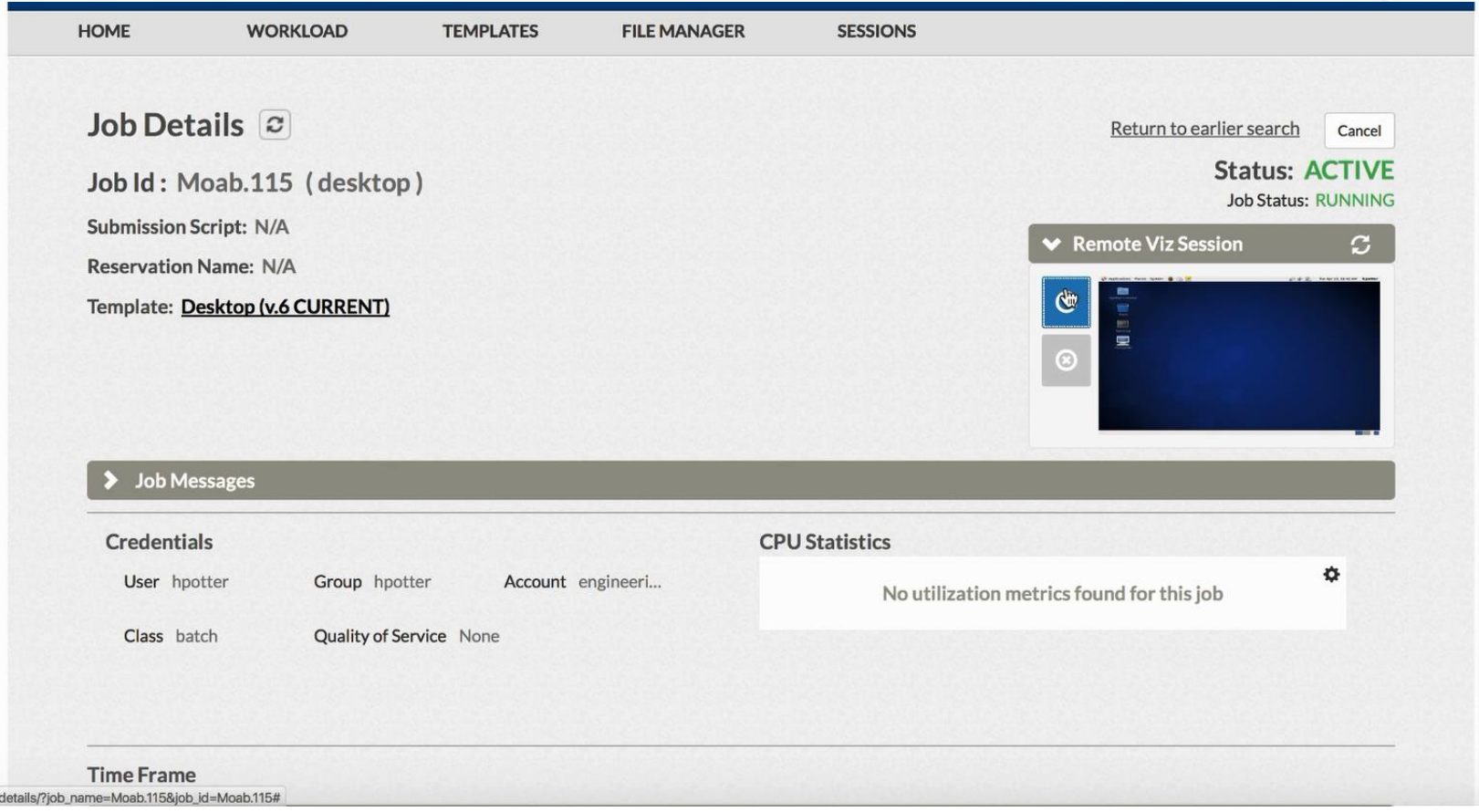
© 2016 Adaptive Computing Enterprises, Inc.

Adaptive COMPUTING

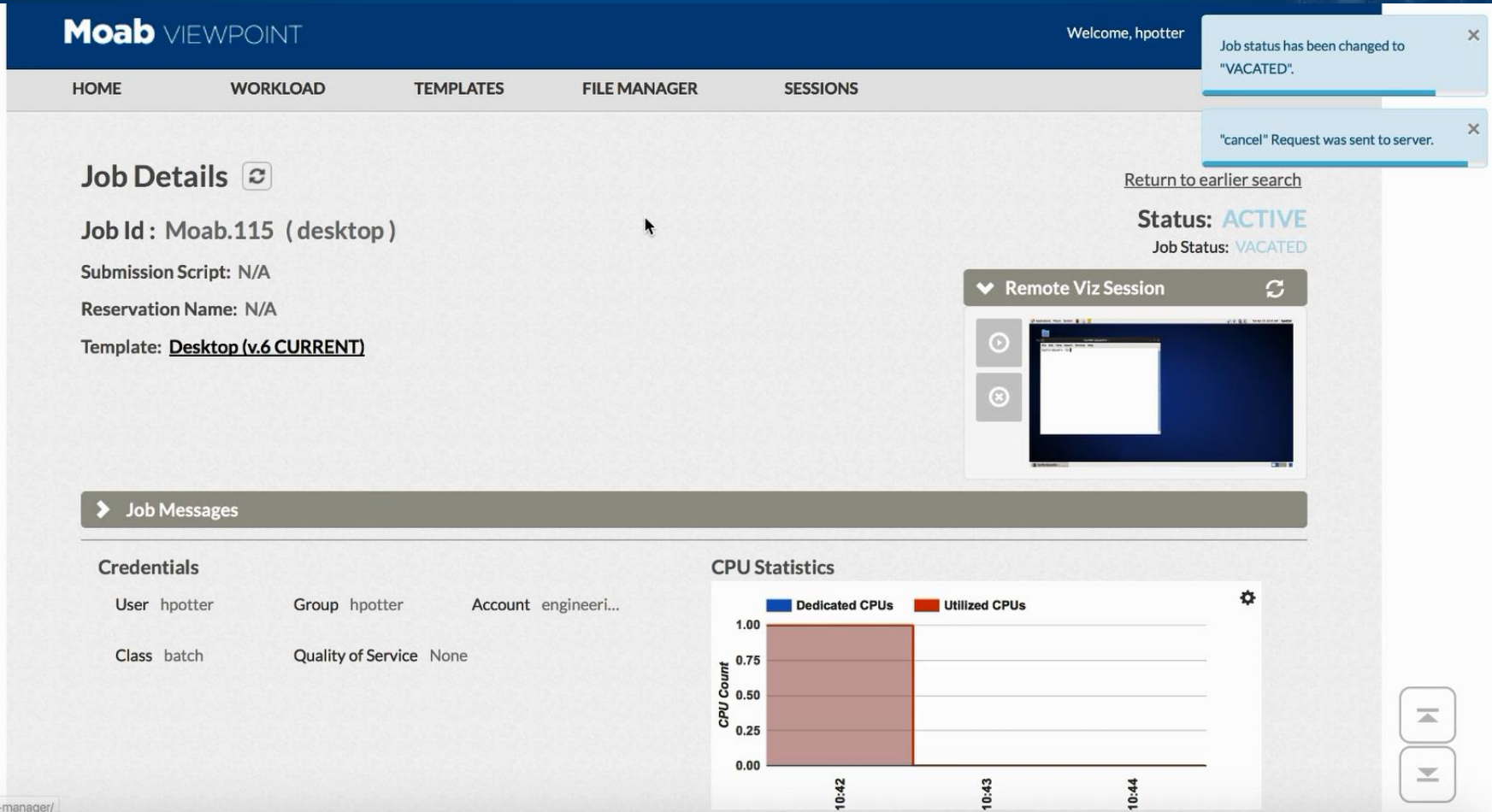
# Resized Remote Viz Session Desktop



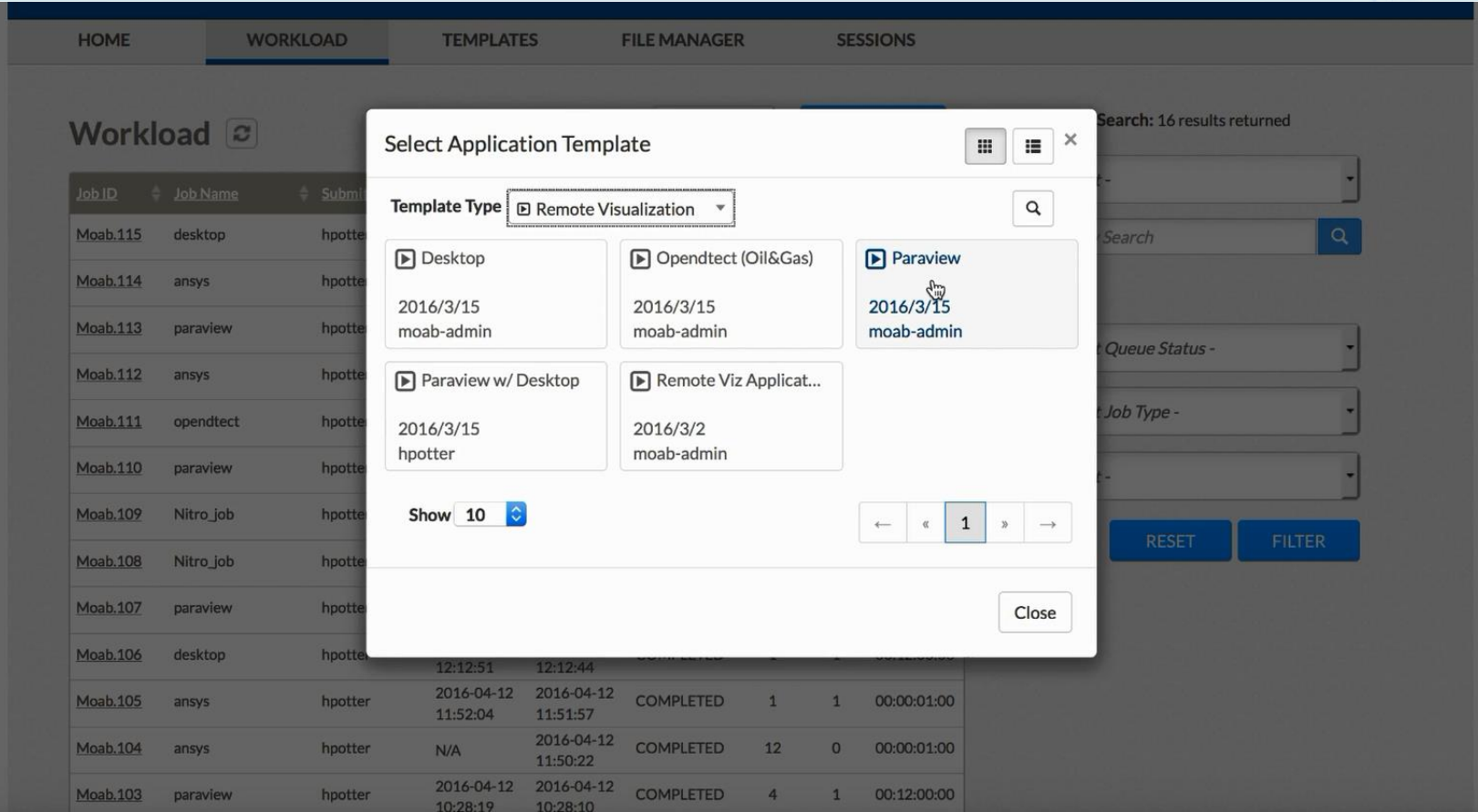
# Job Detail with Remote Viz Session



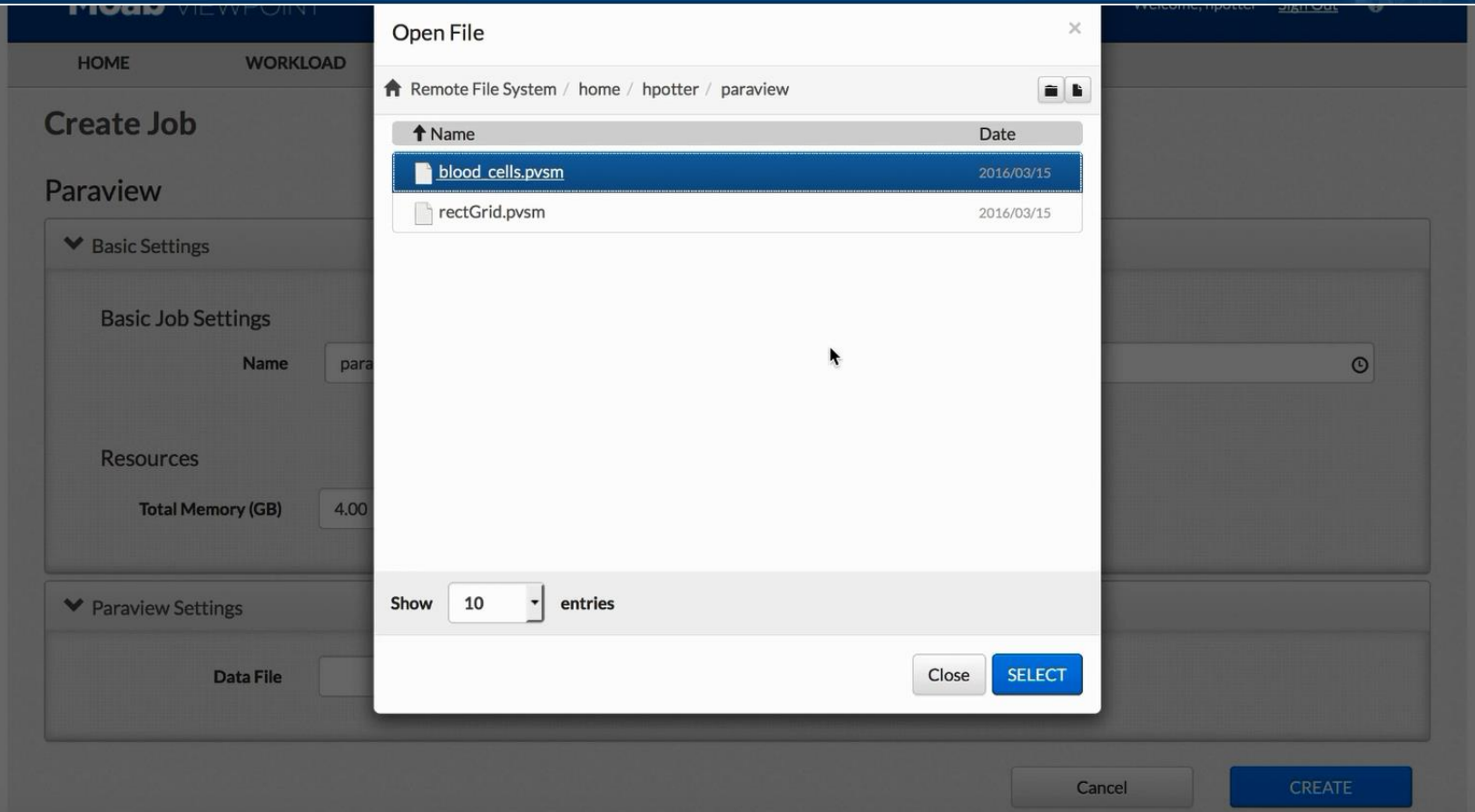
# Canceling the Remote Desktop Session



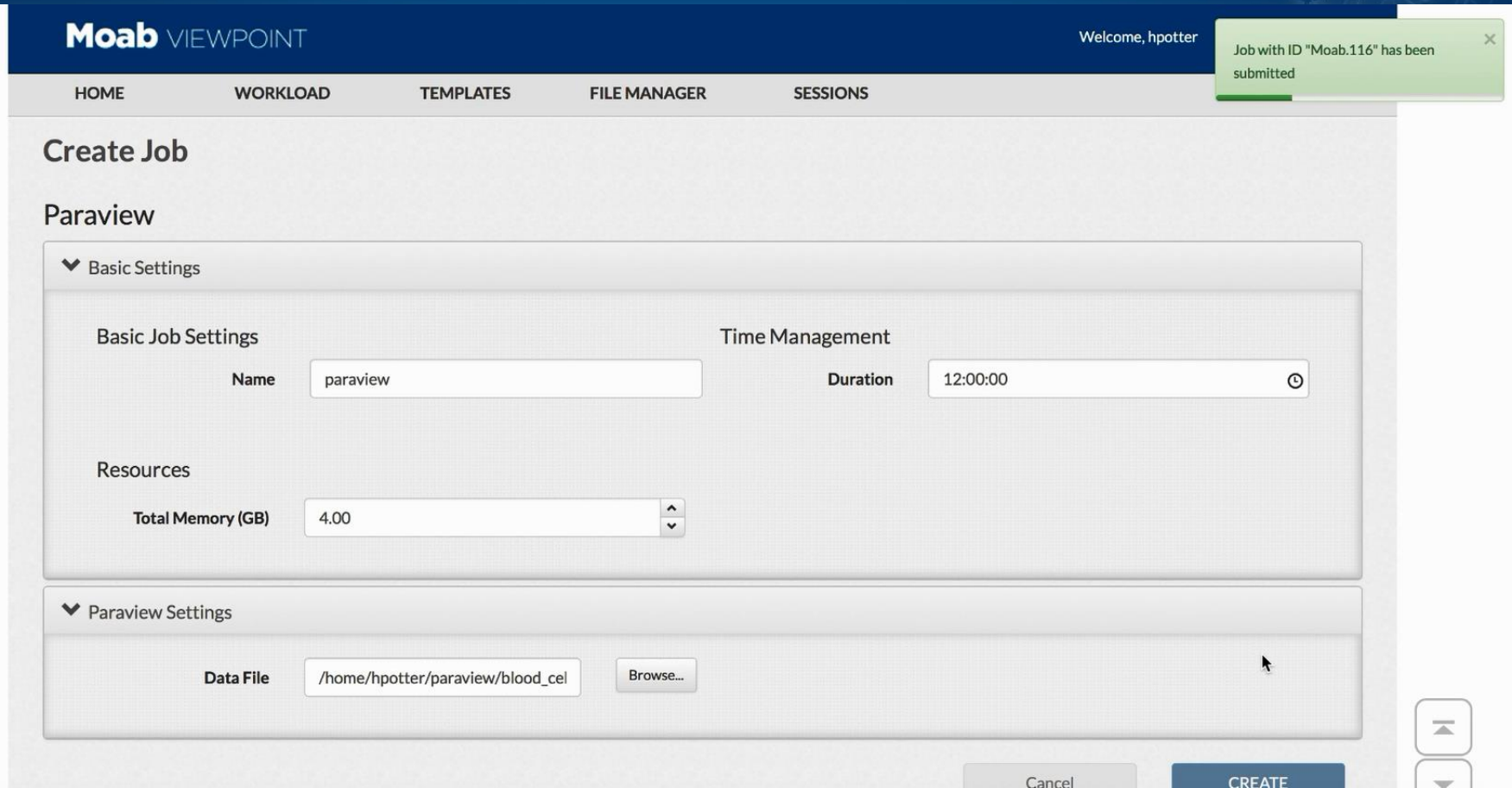
# Choose ParaView Application Template



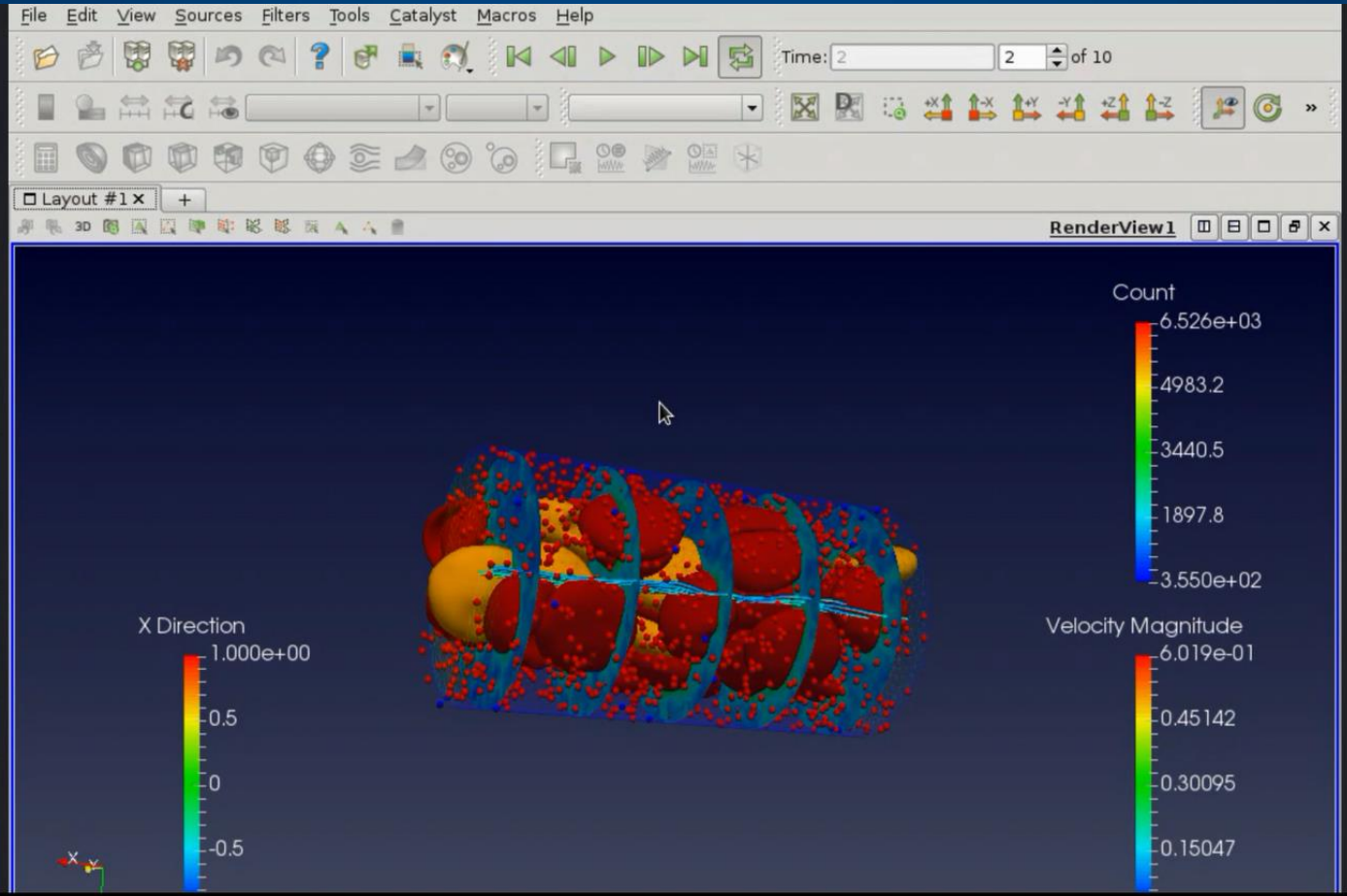
# ParaView Input Data file selection



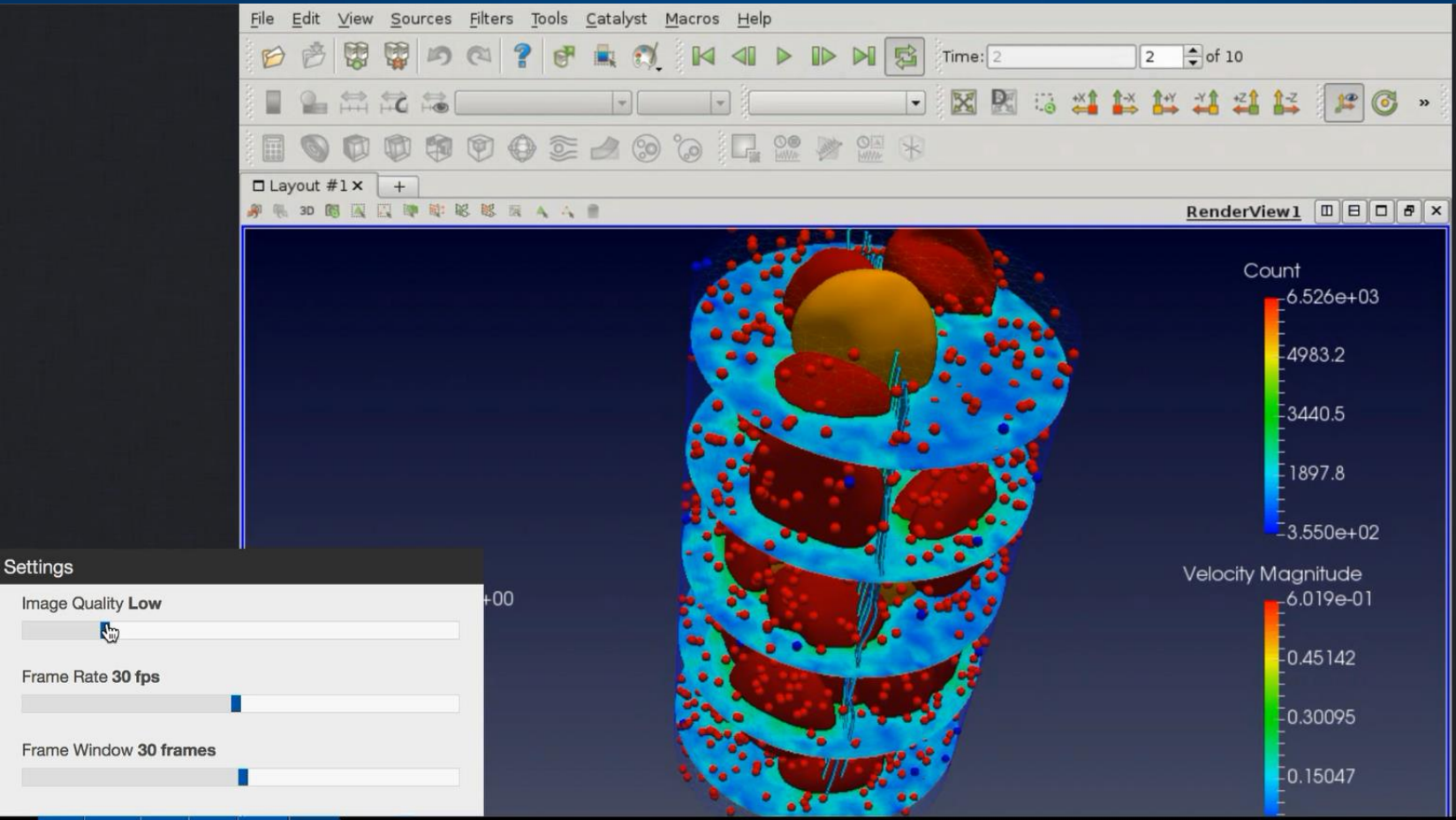
# Submitting ParaView Remote Viz Job



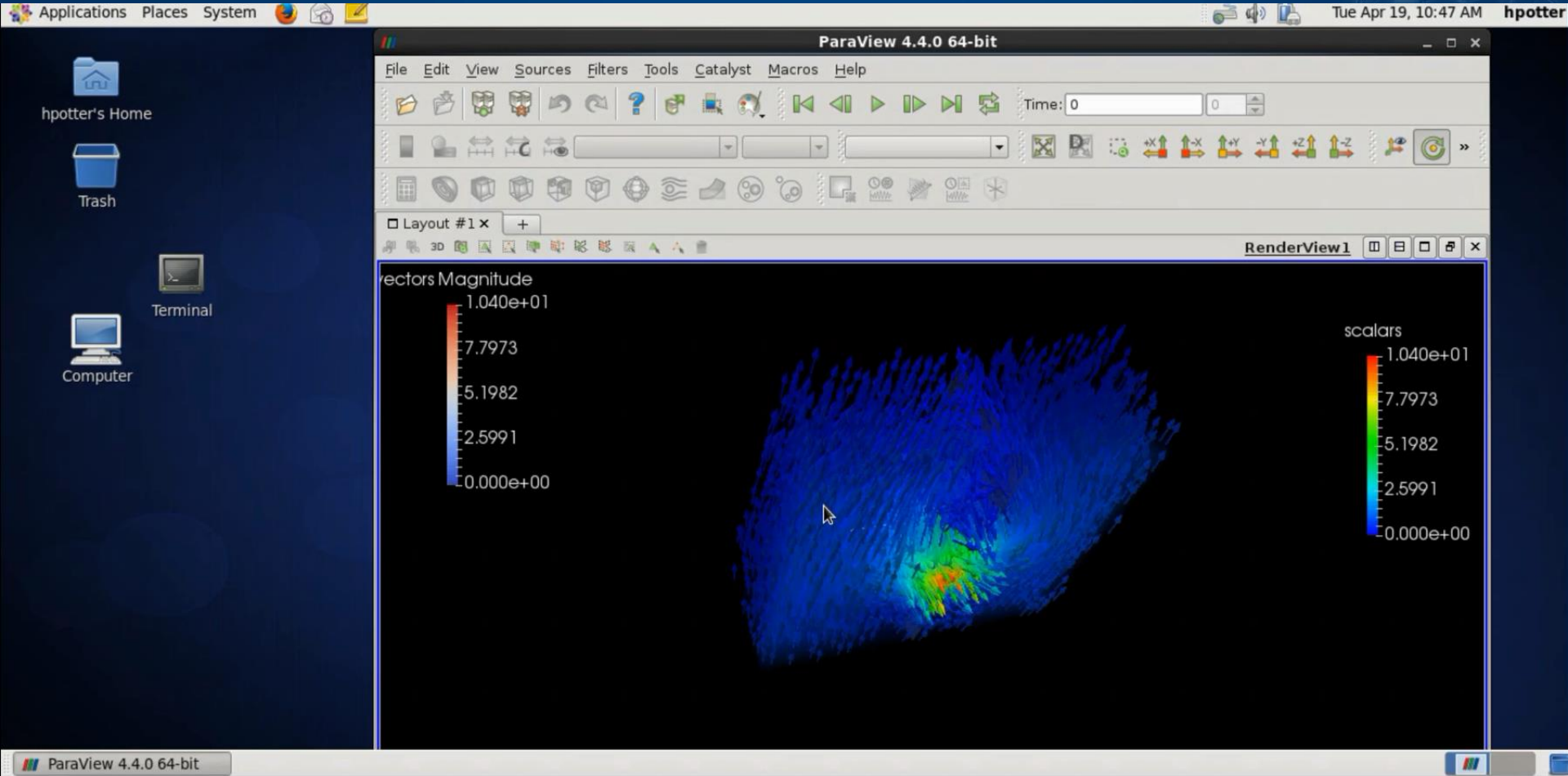
# ParaView Remote Viz Session



# Adjusting Session Image Quality



# Remote Desktop / ParaView Session



# Benefits

- Technical and non-technical users can easily create remote visualization sessions with their applications and data.


© 2016 Adaptive Computing Enterprises, Inc.

# Agenda

- DataWarp
- Intel MIC “KNL”
- Viewpoint Web Portal (non-technical users)
  - User Portal
  - Administrator Portal
  - Remote Visualization
  - **Nitro (HTC job scheduling)**


© 2016 Adaptive Computing Enterprises, Inc.

# Use Cases

- A user can have thousands or millions of single-core to single-node jobs.
  - Parameter sweeps
  - Monte Carlo simulations
  - Regression tests
  - Other
- Large job queues
- Long scheduling cycles


© 2016 Adaptive Computing Enterprises, Inc.

# Nitro HTC Scheduler Application

- Scheduler-agnostic HTC scheduler application
  - TORQUE
  - Slurm
  - Platform LSF
  - Cray ALPS
- Submitted as HPC job to scheduler


© 2016 Adaptive Computing Enterprises, Inc.

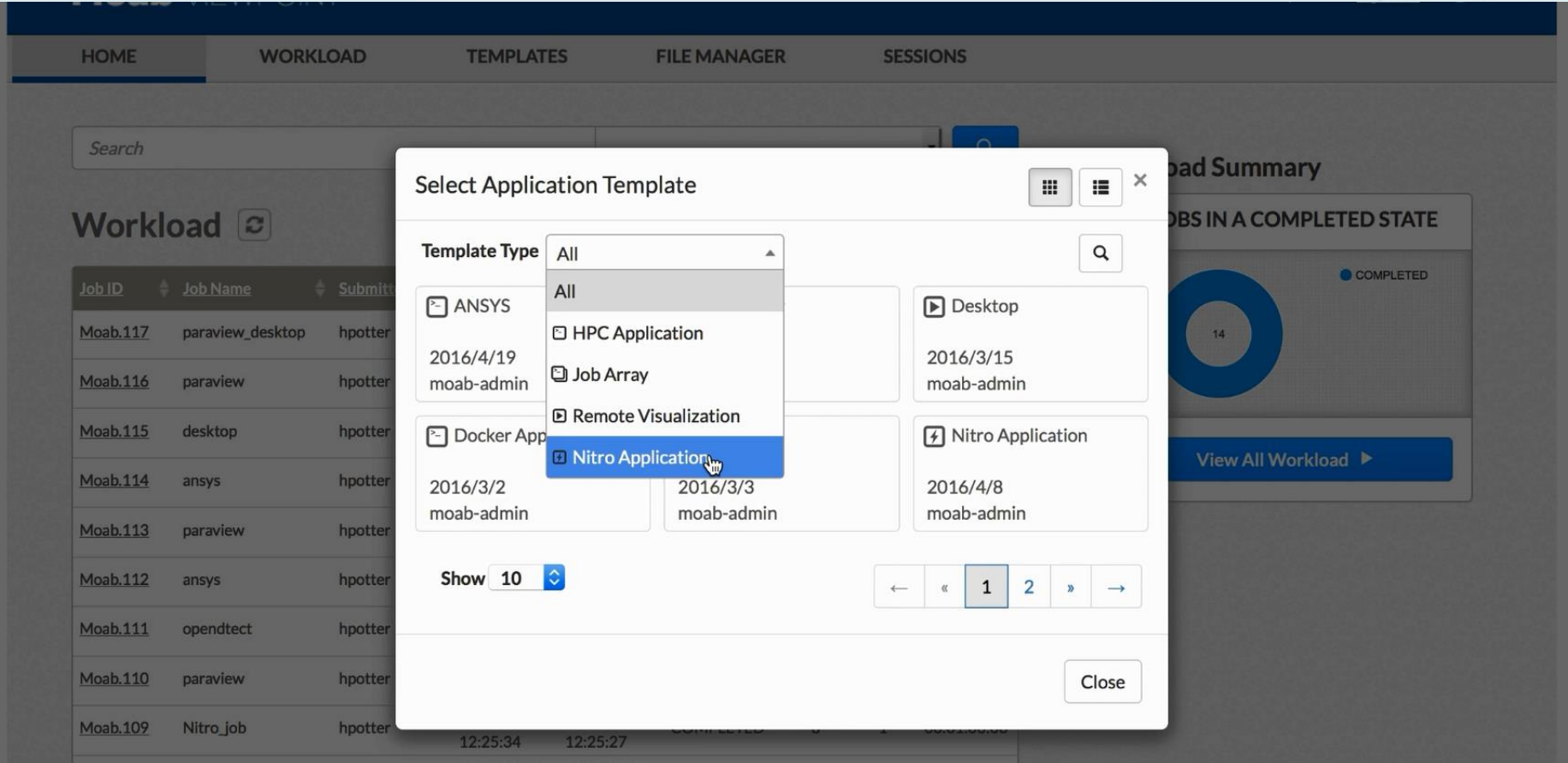
# Nitro Job

- User converts HTC jobs to Nitro “tasks”
  - Tasks defined in Nitro Task file
- Submit 1 Nitro job with Task file
  - Schedule 1 parallel HPC job
  - Execute millions of HTC tasks

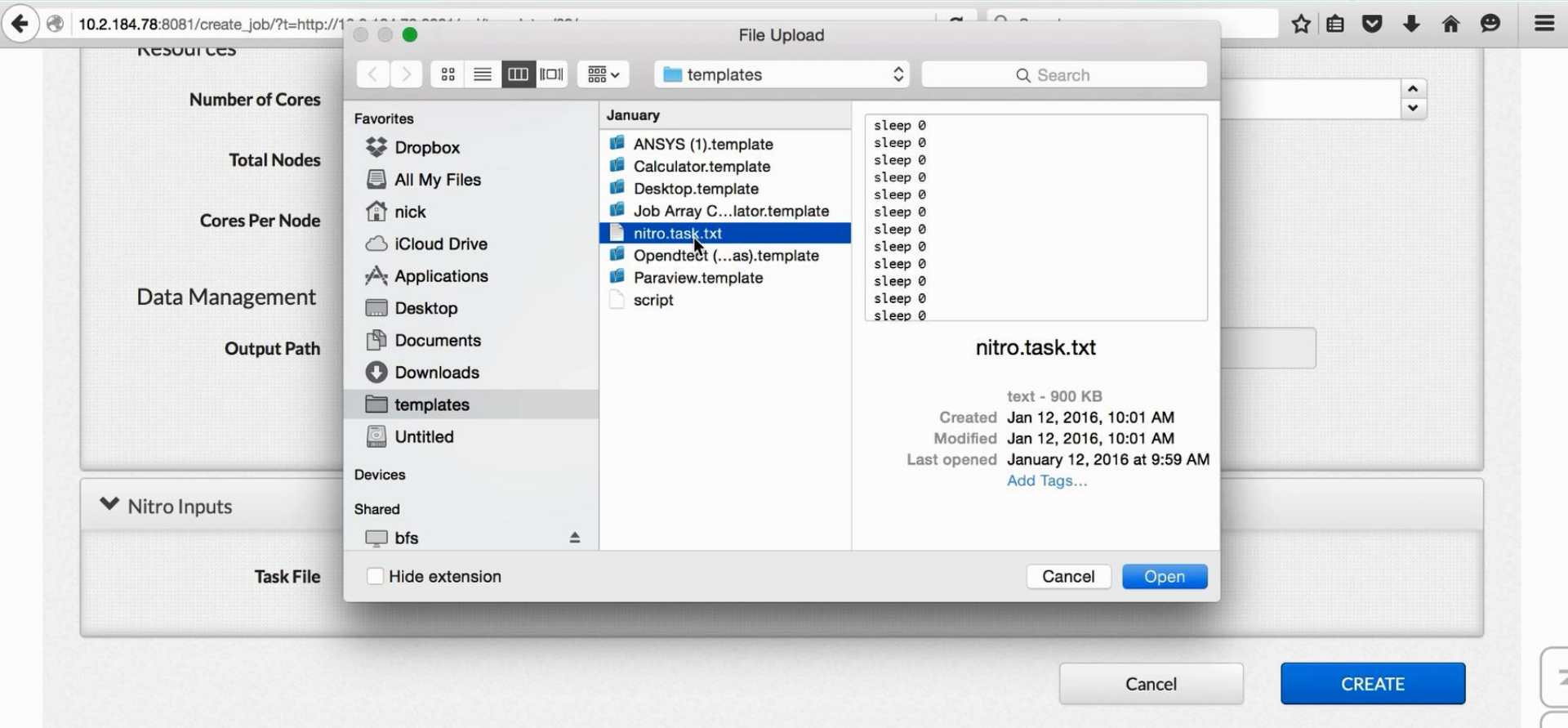
- What about HTC for non-technical users?


© 2016 Adaptive Computing Enterprises, Inc.

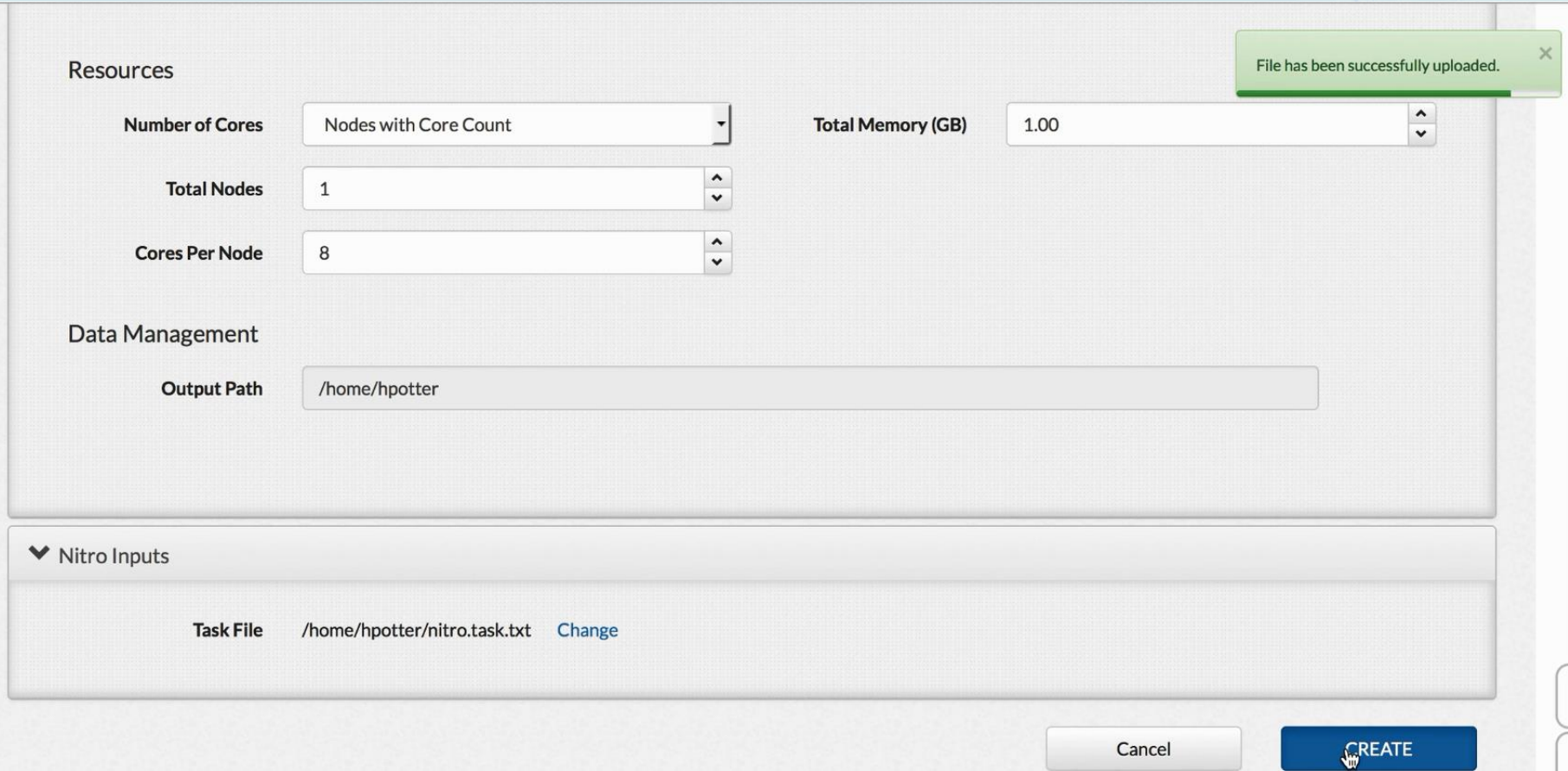
# Choose Nitro Application Template



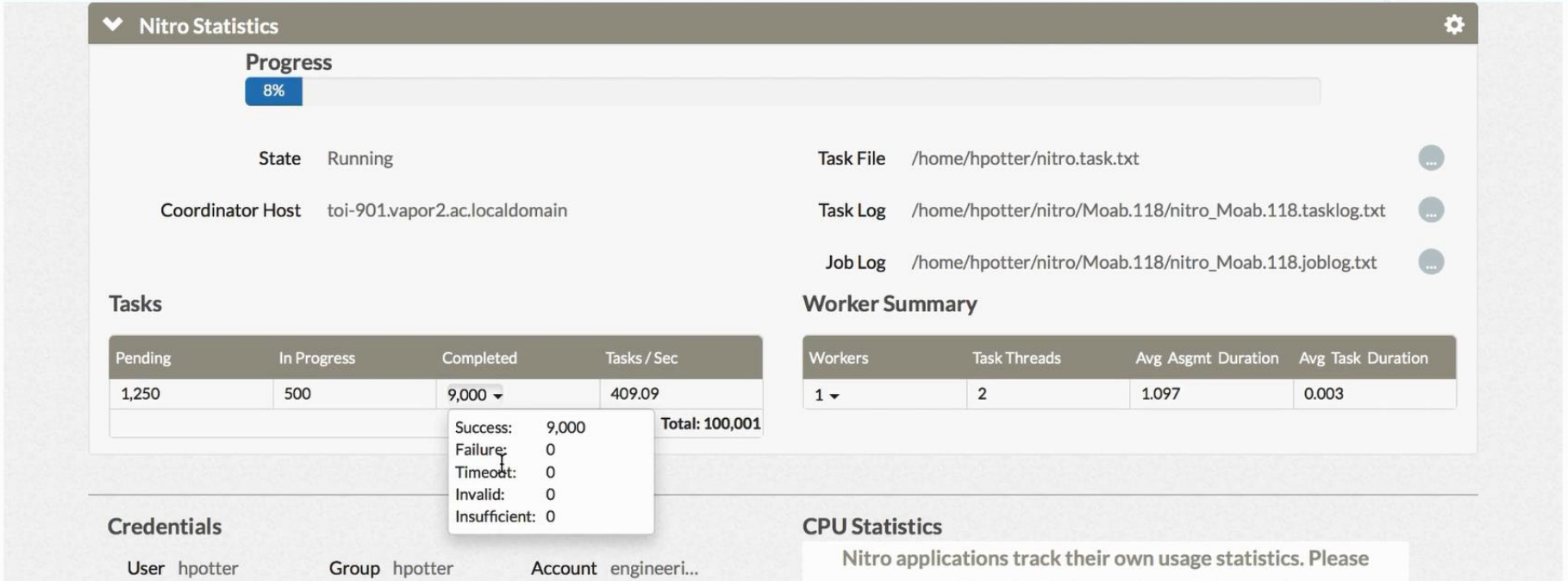
# Nitro Task File selection



# Nitro Task file upload and job submission



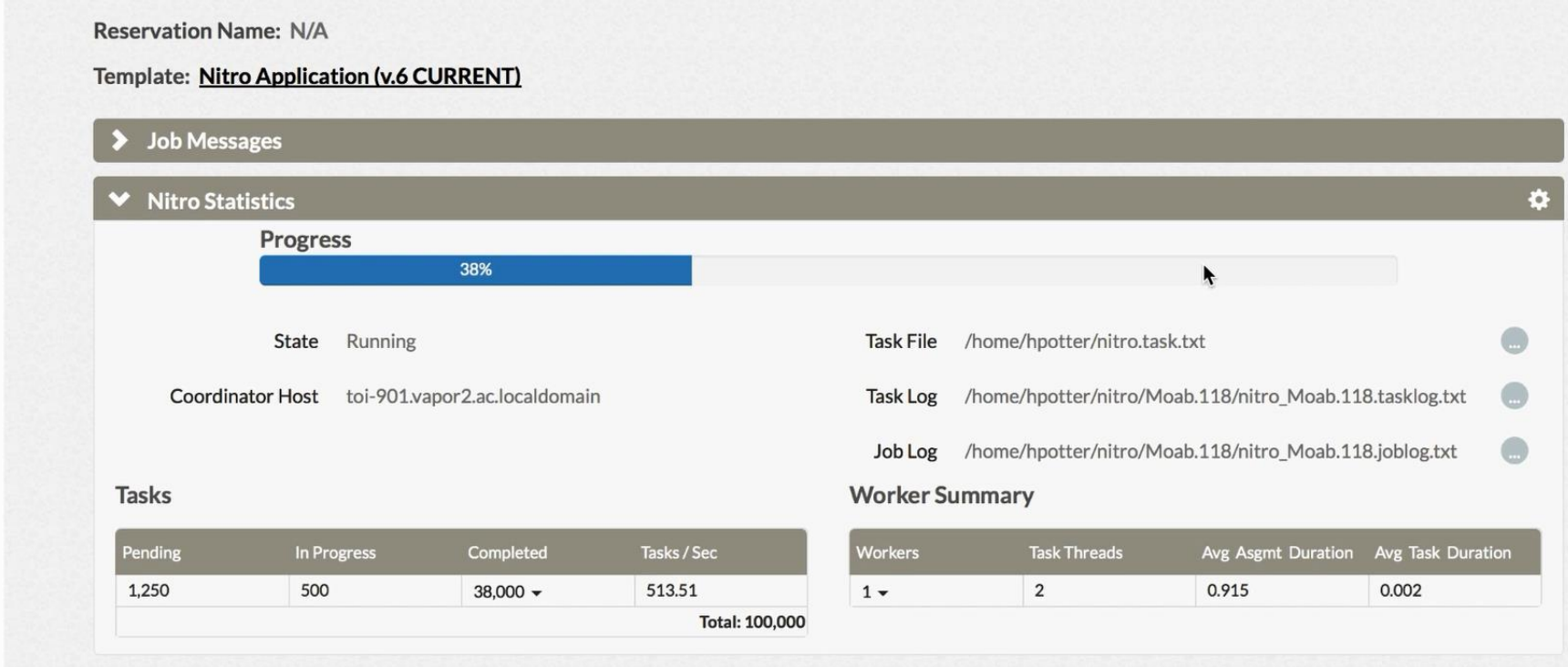
# Nitro job progress and statistics



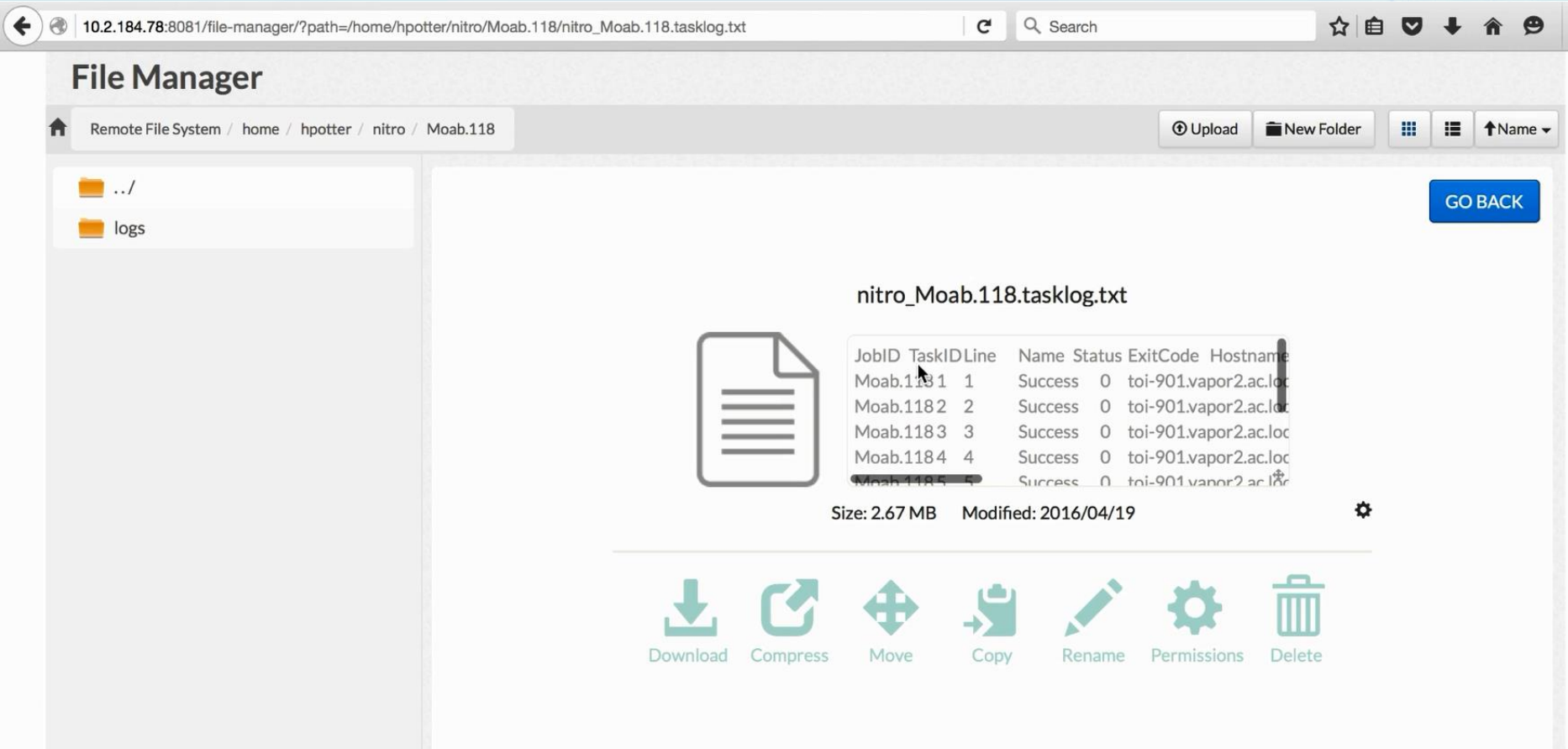



© 2016 Adaptive Computing Enterprises, Inc.

# Nitro job progress



# Nitro Task Log file display during job



# Benefits

- Remove thousands-to-millions of HTC jobs from HPC scheduler queue(s)
- Execute HTC jobs faster
- Non-technical users can
  - Submit Nitro jobs
  - Monitor Nitro job progress
  - Examine Nitro Task results


© 2016 Adaptive Computing Enterprises, Inc.

# Questions?

# PMIx Support

- Process Management Interface for eXascale
  - Open-source Project on github
  - Intel, Mellanox, IBM, Adaptive Computing, SchedMD
- Benefits
  - "Instant On" very fast job start (10 minutes → 2 seconds)
  - Asynchronous, non-blocking operations
  - Blob-based data transfers
  - Fast broadcasts/collectives
  - Interfaces with RMs (e.g., TORQUE)
  - *Will work with malleable/evolving jobs in future*


© 2016 Adaptive Computing Enterprises, Inc. | CONFIDENTIAL AND PROPRIETARY

# Malleable and Evolving Jobs

- Next-generation application/RTE support
  - Malleable jobs (scheduler → application/RTE)
  - Evolvable jobs (application/RTE → scheduler)
  - Need standard API to negotiate resources
    - Like MPI for parallel communications
  - *Gathering use cases for PMIx API definition*
- Benefits
  - Standard application API for scheduler/app dialog
    - Scheduler and Malleable/Evolving Application Dialog (SMEAD) API definition
    - Allows applications to ask for and return resources
    - Allows scheduler to give and take resources to and from a job


© 2016 Adaptive Computing Enterprises, Inc. | CONFIDENTIAL AND PROPRIETARY

# Backup Slides