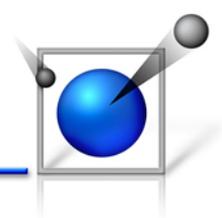


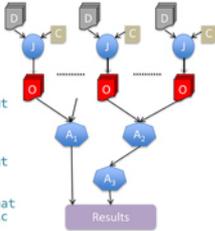
# On the Orchestration of the SNS Reduction Job Workflow



Sudharshan S. Vazhkudai<sup>1</sup>, Michael A. Reuter<sup>1</sup>, James A. Kohl<sup>1</sup>, Stephen D. Miller<sup>1</sup>, Shelly Ren<sup>1</sup>, Mark L. Green<sup>2</sup>, (<sup>1</sup>Oak Ridge National Laboratory and <sup>2</sup>TechX Corporation)

## Motivation

- SNS Data Reduction Job Workflow:
  - Converting raw data to scientifically analyzable data
  - Complex process involving:
    - Hundreds of jobs
    - GBs of datasets and configuration files as input per job
    - Numerous intermediate data products
  - Several aggregator jobs that collate jobs on different criteria
  - End result: set of files that users can perform scientific analyses on

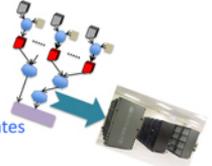


## Motivation (cont'd)

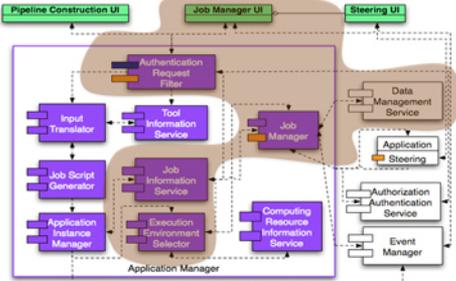
- Such a complex workflow can quickly get unmanageable
  - A single analysis computer cannot handle this workload in the long run
  - Need rich job management tools for: scalable job submission, job monitoring, fault tolerance (restarts) and job input/output data orchestration
  - Minimal user intervention (transparency)
  - Need SNS reduction job workflow management

## SNS Reduction Job Workflow Orchestration Infrastructure

- Seamlessly mapping SNS reduction workflow to a cluster:
  - Workflow definition
  - Composing PBS jobs
  - Community accounts
  - Data staging
  - Job monitoring and updates
  - Fault tolerance



## SNS Application Manager Helps Accomplish the Mapping



## Job Definition and Workflow Tags

- Individual Job command-line:
 

```
amrun -p oic -j Reduction_CNCS_1460_bank033-034 --batch --nostageout --remote-stagein-Files=vfs-201010051656-2282/in /home/jnu/software/bin/agg_reduction XXX_REMOTE_INPUT_DIR_XXX_CNCS_1460.nxx -l CNCS -v --corner=geom-XXX_REMOTE_INPUT_DIR_XXX_CNCS_1460.nxx --time-zero=offset=12.0,0,0 --no-norm --pc-norm --mon-int-range=3500 3500 --lambda-bin=0,10,0,0.1 --dump-ctof-comb --dump-wave-comb --dump-et-comb --tib-range=25000 28000 --mom-trans-bin=0,10,0,1 --split --make-spe --output-CNCS_bank033-034.txt --data-paths=33-34 --mask-file=/SNS/users/2/zr/mask/CNCS/CNCS_bank033-034_mask.dat /SNS/users/2/zr/results/CNCS/1742/CNCS_bank033-034.norm
```
- Tags tell the job manager to:
  - Read input from the appropriate locations, retain the output of the individual jobs on the target machine, etc.
- Aggregator Job command-line that processes many job outputs:
 

```
amrun -p oic -j testreductionaggagregator1 --batch --remote-stagein-Files=vfs-201010051737-1340/out/vfs-201010051737-1352/out/vfs-201010051737-1340/out /home/jnu/software/bin/agg_files CNCS_1460 XXX_REMOTE_OUTPUT_DIR_XXX_XXX_REMOTE_INPUT_DIR_XXX
```

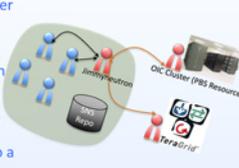
## Job XML Definition

- Example job XML definition (automatically generated):
 

```
<?xml version="1.0" encoding="UTF-8"?>
<SNSJOB NAME="Reduction_CNCS_1460_bank033-034"
  USER="nu"
  REMOTEADDR="portal.ans.gov"
  RESOURCES="beel.ornl.gov">
  <OPERATION NAME="stagein">
    <FILESOURCES>
      <FILE FILE_PREFIX_XXX/SNS/users/2/zr/mask/CNCS/CNCS_bank033-034_mask.dat XXX_FILE_ABB_PREFIX_XXX/SNS/users/2/zr/results/CNCS/1742/CNCS_bank033-034.nxx />
    </FILESOURCES>
  </OPERATION NAME="stagein">
  <OPERATION NAME="nostageout">
    <OPERATION NAME="compute">
      REMOTEADDRESSFILES=XXX_REMOTE_FILE_PREFIX_XXX/VFS-201010051656-2282/in COMPUTEDIR=/home/jnu/software/bin/agg_reduction
      XXX_REMOTE_INPUT_DIR_XXX_CNCS_1460.nxx XXX_REMOTE_INPUT_DIR_XXX_CNCS_1460.nxx -l CNCS -v --corner=geom-XXX_REMOTE_INPUT_DIR_XXX_CNCS_1460.nxx --time-zero=offset=12.0,0,0 --no-norm --pc-norm --mon-int-range=3500 3500 --lambda-bin=0,10,0,0.1 --dump-ctof-comb --dump-wave-comb --dump-et-comb --tib-range=25000 28000 --mom-trans-bin=0,10,0,1 --split --make-spe --output-CNCS_bank033-034.txt --data-paths=33-34 --mask-file=XXX_FILE_ABB_PREFIX_XXX/SNS/users/2/zr/mask/CNCS/CNCS_bank033-034_mask.dat XXX_FILE_ABB_PREFIX_XXX/SNS/users/2/zr/results/CNCS/1742/CNCS_bank033-034.norm </>
    </OPERATION NAME="compute">
  </OPERATION NAME="nostageout">
  </SNSJOB>
```
- Operations
  - Stagein, compute and stageout are orchestrated by the job manager based on the XML definition

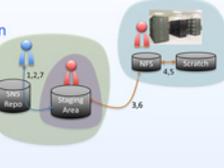
## Community Account: Jimmy Neutron

- User jobs are run through a single community account on specialized resources
  - Oblivates the need for each user to have an account on every compute resource
  - Both local machine and target cluster have the jimmyneutron account with the proper credentials
  - Accounting maintained locally (user jobs to jimmyneutron mapping) to trace jobs back to a user
  - Data movement accomplished as jimmyneutron



## Data Staging

- Sequence of Data Staging Operations:
  - SNS repo to jimmyneutron staging area on local machine as "user"
  - Grant access to jimmyneutron
  - Local staging area to cluster NFS as "jimmyneutron"
  - Cluster NFS to scratch file system of the running job as "jimmyneutron"
  - scratch file system to cluster NFS as "jimmyneutron"
  - Cluster NFS to local staging area
  - Local staging area to user home as "user"



## Composing a PBS Script per Workflow

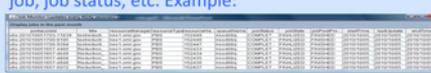
```
#!/bin/bash
PBS_M=be
PBS_O=AmisbOic
PBS_L=be
PBS_Q=mas08q
PBS_W=walltime=00:30:00
PBS_M=mem=1000mb

Staging orchestration from jimmyneutron on
cluster to scratch space of running job
scp -r /home/jnu/stagein/out/vfs-201010051737-13521/nxx /in
scp -r /home/jnu/stagein/out/vfs-201010051736-13145/nxx /in
cd $PBS_SCRATCH/out

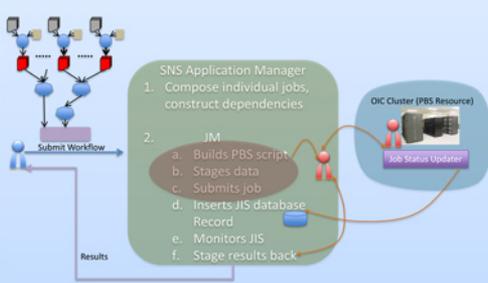
Compute operation
/home/jnu/software/bin/agg_reduction -l /in/CNCS_1460.nxx -l /in/CNCS_1460.nxx -l /in/CNCS_1460.nxx --time-zero=offset=12.0,0,0 --no-norm --pc-norm --mon-int-range=3500 3500 --lambda-bin=0,10,0,0.1 --dump-ctof-comb --dump-wave-comb --dump-et-comb --tib-range=25000 28000 --mom-trans-bin=0,10,0,1 --split --make-spe --output-CNCS_bank033-034_mask.dat /in/CNCS_bank033-034_mask.dat --norm=/in/CNCS_bank033-034.norm /in/CNCS_bank033-034.norm

Staging orchestration from scratch space of running job
to jimmyneutron on cluster
scp -r /home/jnu/stagein/out/vfs-201010051737-13521/nxx /out
```

## Monitoring and Fault Tolerance

- Individual jobs in the workflow monitored:
  - A record per job in the workflow
  - Job handle, start and finish times, target host, type of job, job status, etc. Example:
 
  - Ability to periodically update and query the status of jobs
  - Ability to treat data transfers as an integral part of the workflow
  - Any one component fails, appropriate error code returned for debugging and tracing for future restarts

## Putting it all together



## Conclusion

- SNS reduction workflow management:
  - Transparent to the user
  - Scalable and seamless induction of many target execution platforms
  - Allows us to farm out remote portal users to clusters
  - Orchestrate expensive data operations elegantly by performing global optimizations for the entire workflow
  - Elegant job monitoring and status updates
- Future Work: Application Instance Manager to maintain workflow state, Webservice enabling workflow orchestration for use by tools such as Mantid, ISAW, Commander, etc.

