globus

CloudyCluster™

Topic Modeling in the Cloud with Globus and CloudyCluster

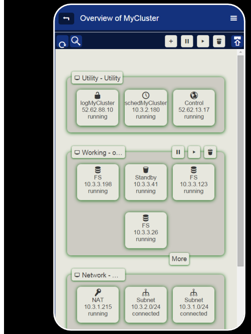
A DICE Lab Project Case Study

Topic Modeling PLDA+

- Machine Learning Content/Theme Understanding
  - Topic modeling, which is extremely compute-intensive, is based on Latent Dirichlet Allocation (LDA), published by Blei, Ng, and Jordan in 2003. This method has been implemented many times.
  - One recent implementation by Google, PLDA+, uses message passing in a distributed cluster environment to speed up the calculation of topics in a very large corpus.
  - Work at the Dr. Apon's Data Intensive Computing Ecosystems (DICE) lab is working on ways to apply and extend PLDA+
  - understanding of the themes and change of themes in many kinds of documents such as scientific journals, patents, historical documents, and more.

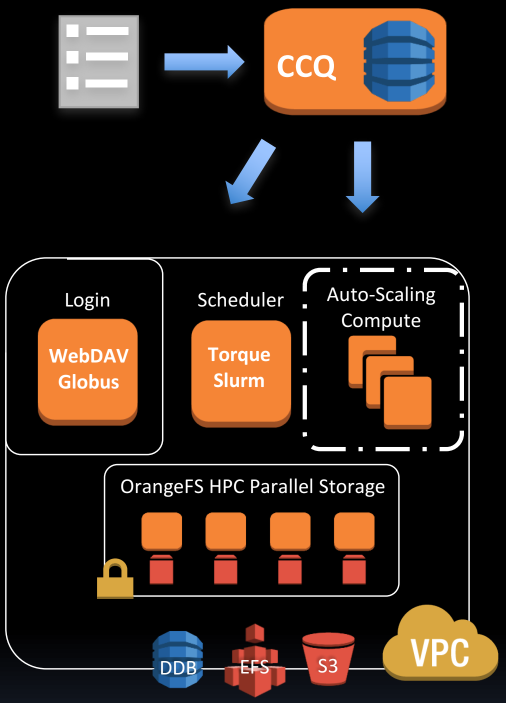
CloudyCluster™

Self-Service Elastic HPC


Overview of MyCluster
Utility - Utility
logMyCluster 52.62.88.10 running
schedMyCluster 10.3.2.180 running
Control 52.62.13.17 running
Working - o...
FS 10.3.3.198 running
Standby 10.3.3.41 running
FS 10.3.3.123 running
FS 10.3.3.26 running
More
Network - ...
NAT 10.3.1.215 running
Subnet 10.3.2.0/24 connected
Subnet 10.3.1.0/24 connected

Create a fully operational HPC Cluster in minutes, complete with:

- ***Storage***: OrangeFS on EBS, S3, EFS
- ***Compute***: Job Driven Elastic Compute through CCQ
- ***Scheduler***: with v.1.2 Torque/Maui & SLURM with CCQ MetaScheduler
- ***HPC Libraries***: Boost, Cuda Toolkit, Docker, FFTW, FLTK, GCC, Gengetopt, GRIB2, GSL, Hadoop, HDF5, ImageMagick, JasPer, NetCDF, NumPy, Octave, OpenCV, OpenMPI, PROJ, R, Rmpi, SciPy, SWIG, WGRIB, UDUNITS, .NET Core, Singularity, Queue, Picard and xrootd
- ***HPC Software***: Ambertools, ANN, ATLAS, BLAS, Blast, Blender, Burrows-Wheeler Aligner, CESM, GROMACS, LAMMPS, NCAR, NCL, NCO, nwchem, OpenFoam, papi, paraview, Quantum Espresso, SAMtools, WRF, Galaxy, Vtk, Su2, Dakota, Gatk and Jupyter Notebook



- You can also Install your own software in a custom AMI or in EFS
- All from an easy to use Web UI from mobile, tablet or desktop
- XDMoD are targeted for Future release.

XDMoD METRICS ON DEMAND globus awsmarketplace

Quote

“We are using CloudyCluster to execute PLDA+ in the Amazon Web Services Cloud. With CloudyCluster and AWS we have access to a massive amount of resources that allow us to explore topics in a wide range of documents simultaneously”

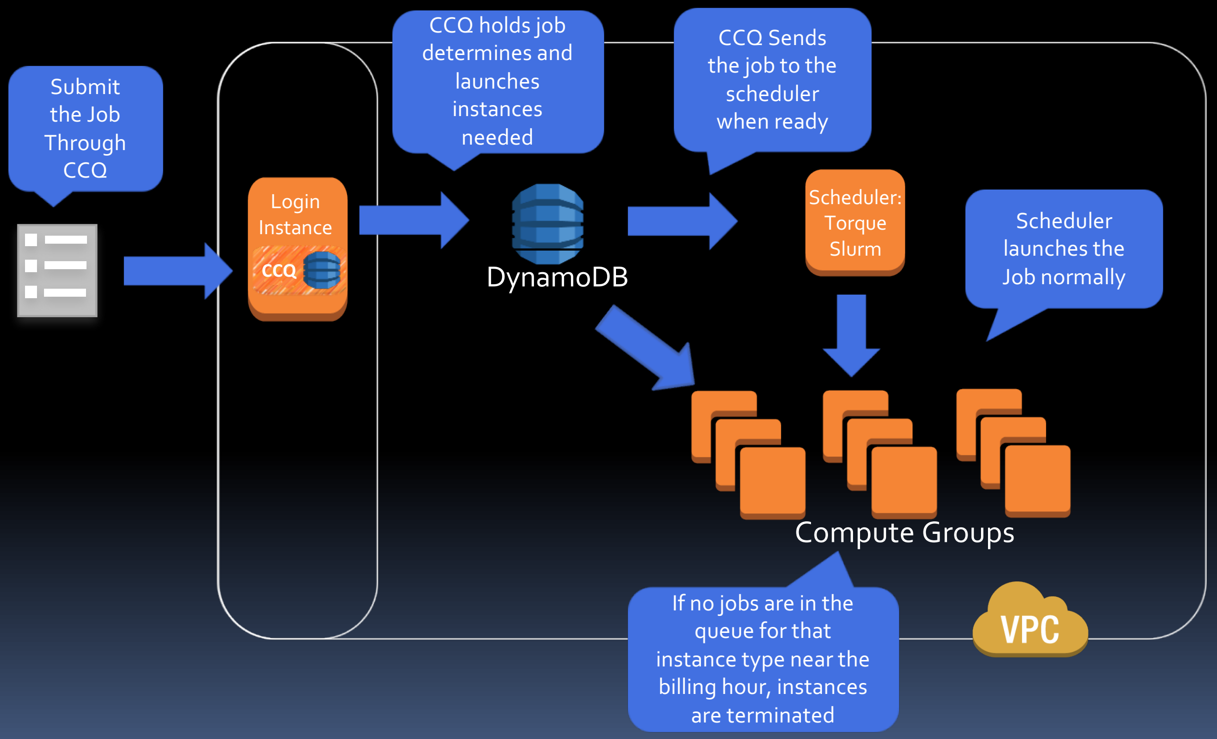
says Dr. Apon. *Professor and Chair, Division of Computer Science and DICE Lab Lead*

She continues, “We expect that our calculations will lead to a better understanding of how topic models work and how they can be applied to the study and understanding of the themes and change of themes in many kinds of documents such as scientific journals, patents, historical documents, and more.”

Simplified with Globus

- Large Corpus of Data, up to 100’s of GB
- Initially scp data to CloudyCluster Login Instance
- Introduced Globus and CloudyCluster seamless Integration
- Ease of use for Data Transfers was improved
- Data Transfer Performance
  - **50% Increase with Globus** for single node transfers
  - **Over Internet2 to AWS**

CCQ - Elastic HPC Dispatching



Submit the Job Through CCQ

Login Instance

CCQ

CCQ holds job determines and launches instances needed

DynamoDB

CCQ Sends the job to the scheduler when ready

Scheduler: Torque Slurm

Scheduler launches the Job normally

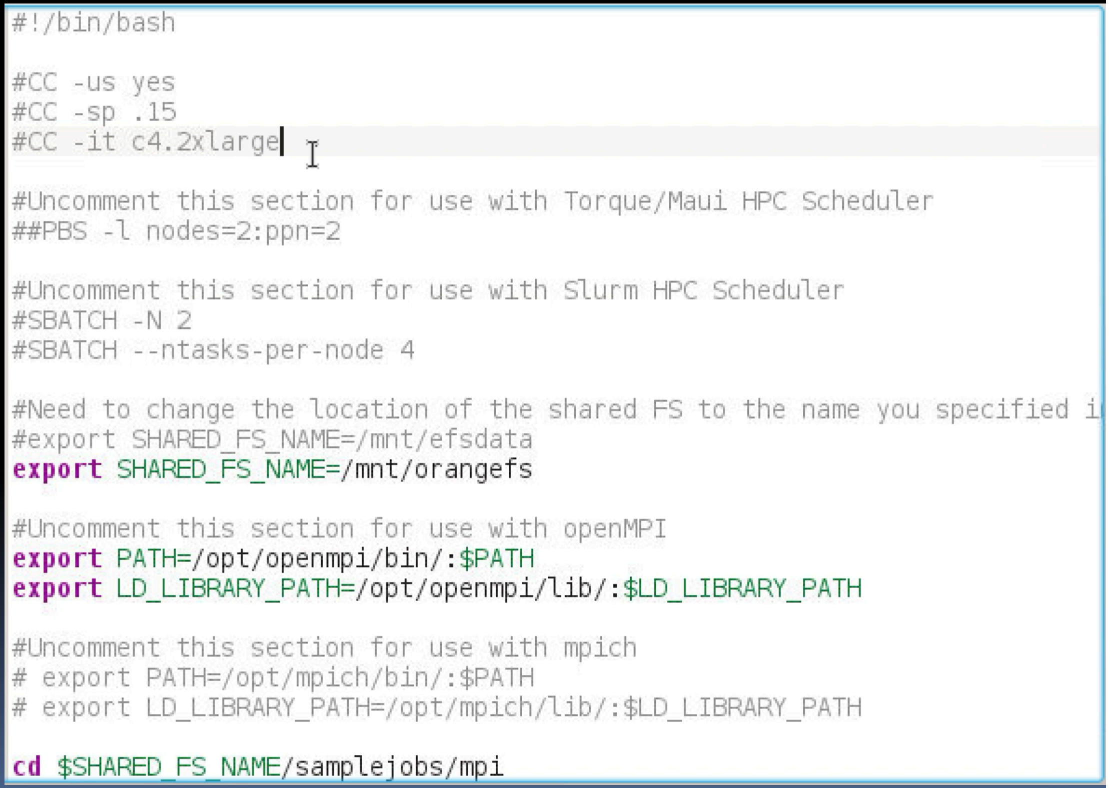
Compute Groups

If no jobs are in the queue for that instance type near the billing hour, instances are terminated

VPC

Cost Savings with Spot

- With the tight integration with the Spot Market, CCQ allowed jobs for the DICE lab to save up to 90% off of the AWS On-Demand price.
- Allowing them to perform more computation for less with minimal effort.
- CCQ Job Directives enable Spot Instances Seamlessly
- Jobs should run quickly or be preemptable



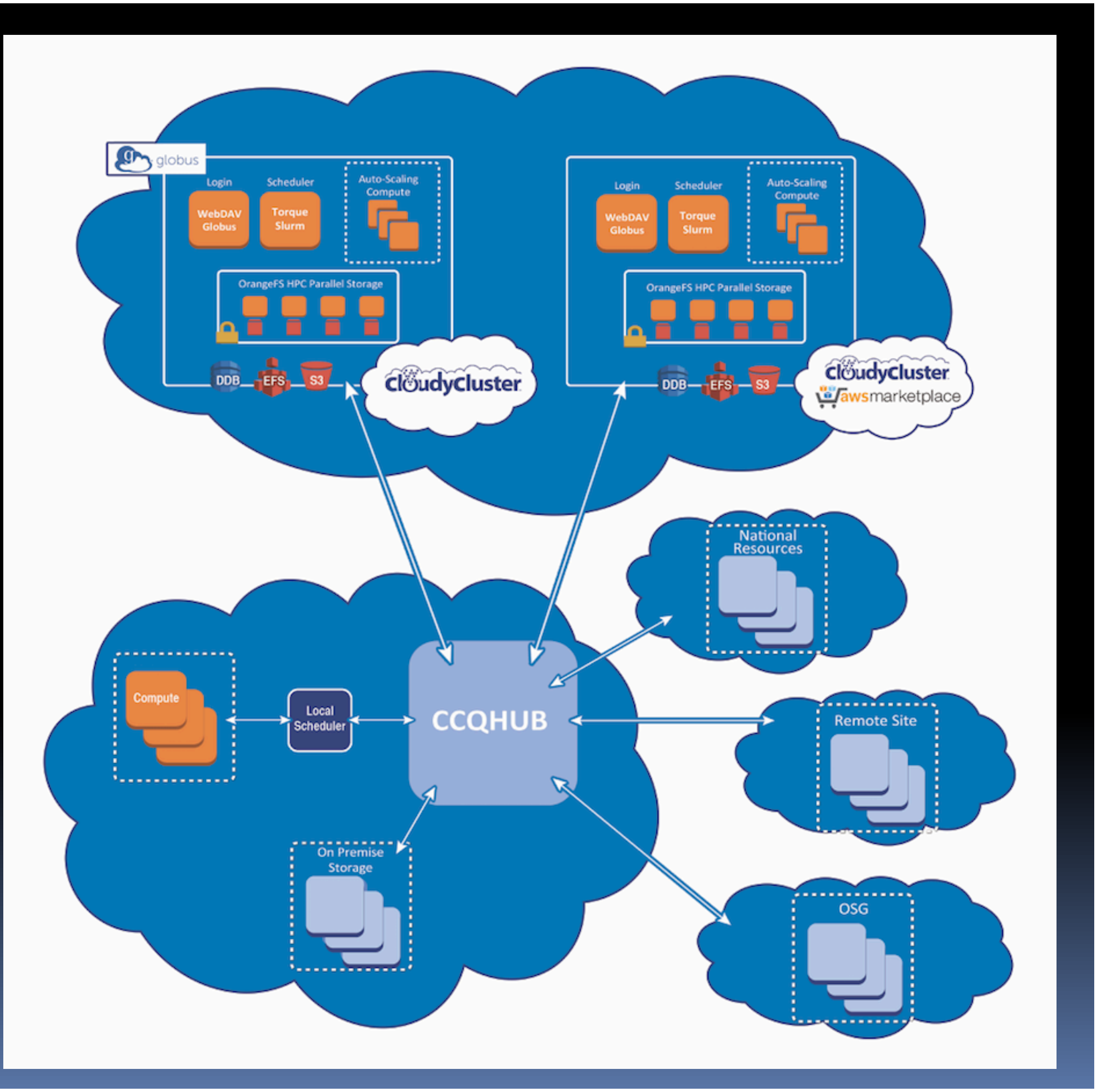
```
#!/bin/bash

#CC -us yes
#CC -sp .15
#CC -it c4.2xlarge

#Uncomment this section for use with Torque/Maui HPC Scheduler
##PBS -l nodes=2:ppn=2

#Uncomment this section for use with Slurm HPC Scheduler
#SBATCH -N 2
#SBATCH --ntasks-per-node 4

#Need to change the location of the shared FS to the name you specified i
#export SHARED_FS_NAME=/mnt/efsdata
export SHARED_FS_NAME=/mnt/orangefs

#Uncomment this section for use with openMPI
export PATH=/opt/openmpi/bin/:$PATH
export LD_LIBRARY_PATH=/opt/openmpi/lib/:$LD_LIBRARY_PATH

#Uncomment this section for use with mpich
# export PATH=/opt/mpich/bin/:$PATH
# export LD_LIBRARY_PATH=/opt/mpich/lib/:$LD_LIBRARY_PATH

cd $SHARED_FS_NAME/samplejobs/mpi
```

Future Work:
CCQHub Project

HPC Job Routing

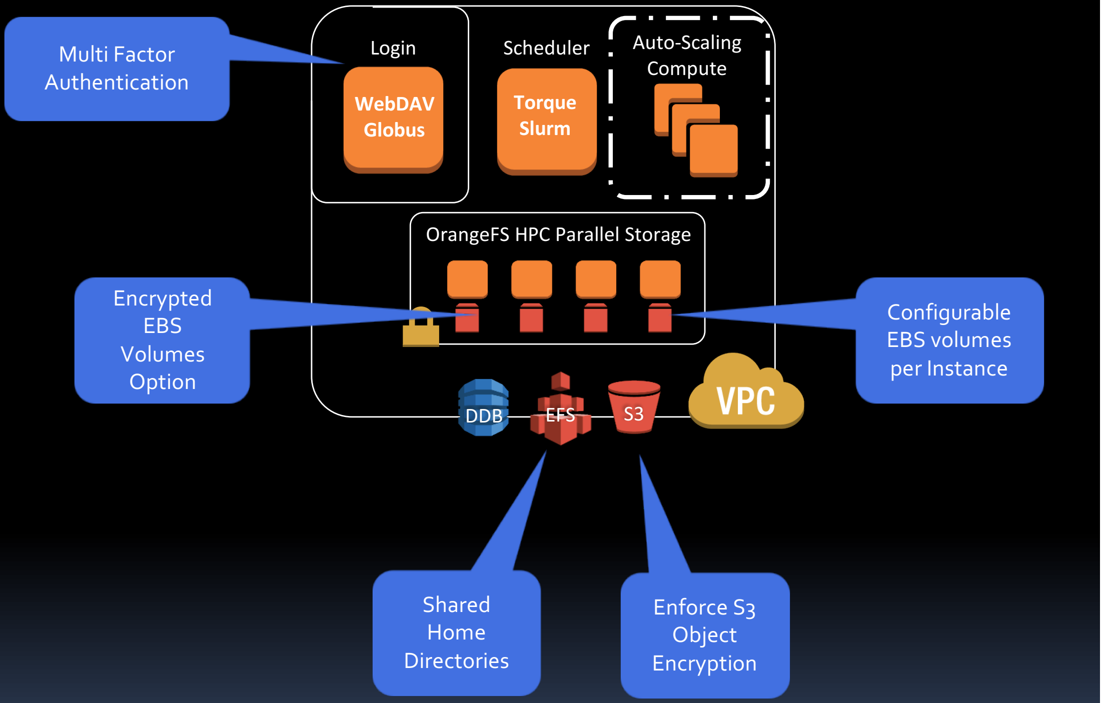
Project goals:

- Route HPC Jobs on Premise or to the Cloud
- Stage Data prior to launching the job. Including **Globus** Transfers
- Return job results when complete.
- Scale cloud resources with CloudyCluster



New Features in V1.3

- Shared Home Directories in EFS
- Configurable EBS volumes per instance for OrangeFS
- Encrypted EBS volume options
- Enforce S3 object encryption
- MFA support
- Support for CCQHub
- New Libraries including Machine Learning Codes
  - Mlpack, .Net Core, NuPIC, Octave, OpenCV, PICARD, Queue, Scikit-learn, Tensor Flow and Theano.


Multi Factor Authentication
Login
WebDAV Globus
Scheduler
Torque Slurm
Auto-Scaling Compute
OrangeFS HPC Parallel Storage
Encrypted EBS Volumes Option
Configurable EBS volumes per Instance
DDB
EFS
S3
VPC
Shared Home Directories
Enforce S3 Object Encryption

globus

CloudyCluster

Thank You