Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

Progress with supporting CMS workflows on ALCF Theta

Dirk Hufnagel (FNAL)

HEPCloud Stakeholder Meeting

12th August 2020

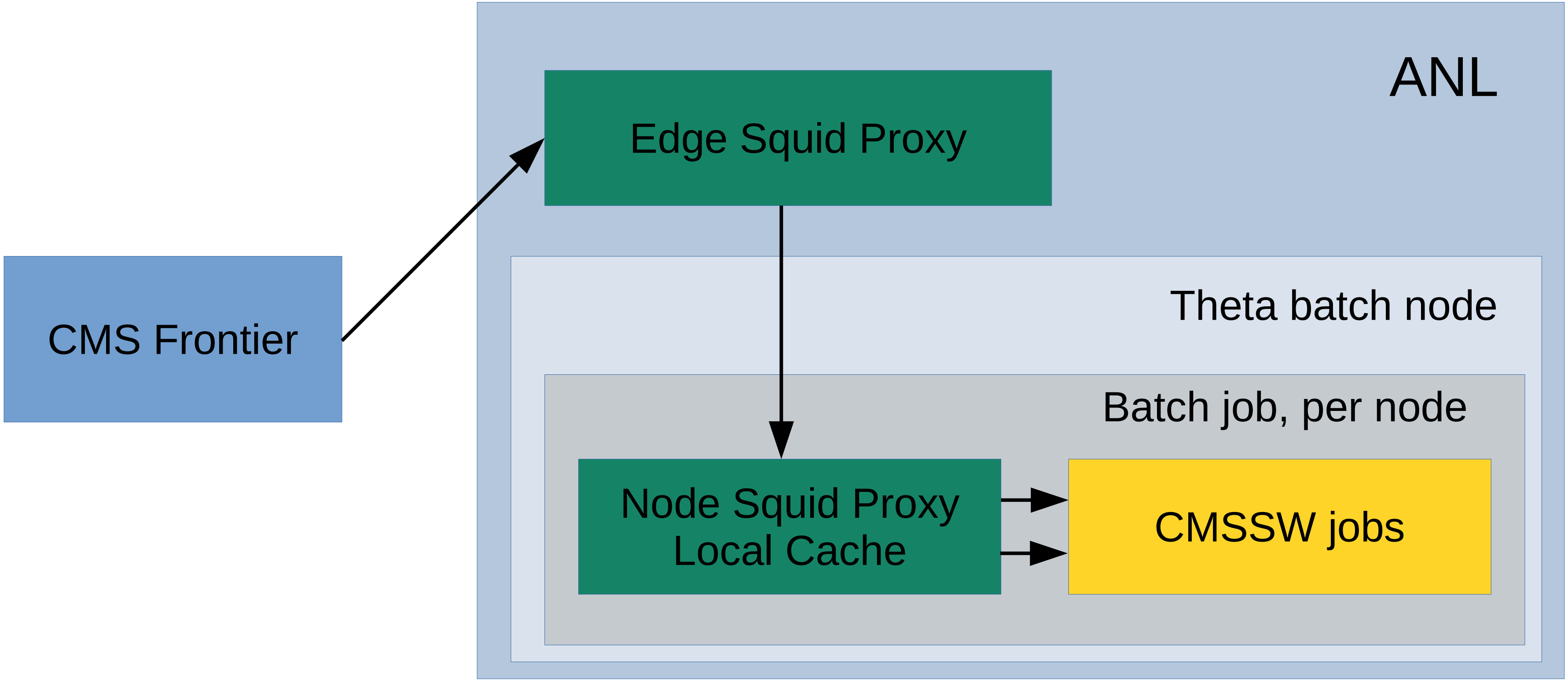
CMS runtime requirements

- Over the last two weeks I have taken a fresh look at the CMS runtime requirements, namely frontier and cvmfs and how they apply to Theta.

- The proposed backup solutions in our planning (reading conditions from sqlite files and building workflow specific containers) are “not great” (to say it mildly)

- I think I found solutions for both frontier and cvmfs access that will avoid having to go to our backups. More testing is needed, but initial results are very promising.

Frontier



CMS Frontier
ANL
Edge Squid Proxy
Theta batch node
Batch job, per node
Node Squid Proxy Local Cache
CMSSW jobs

Frontier

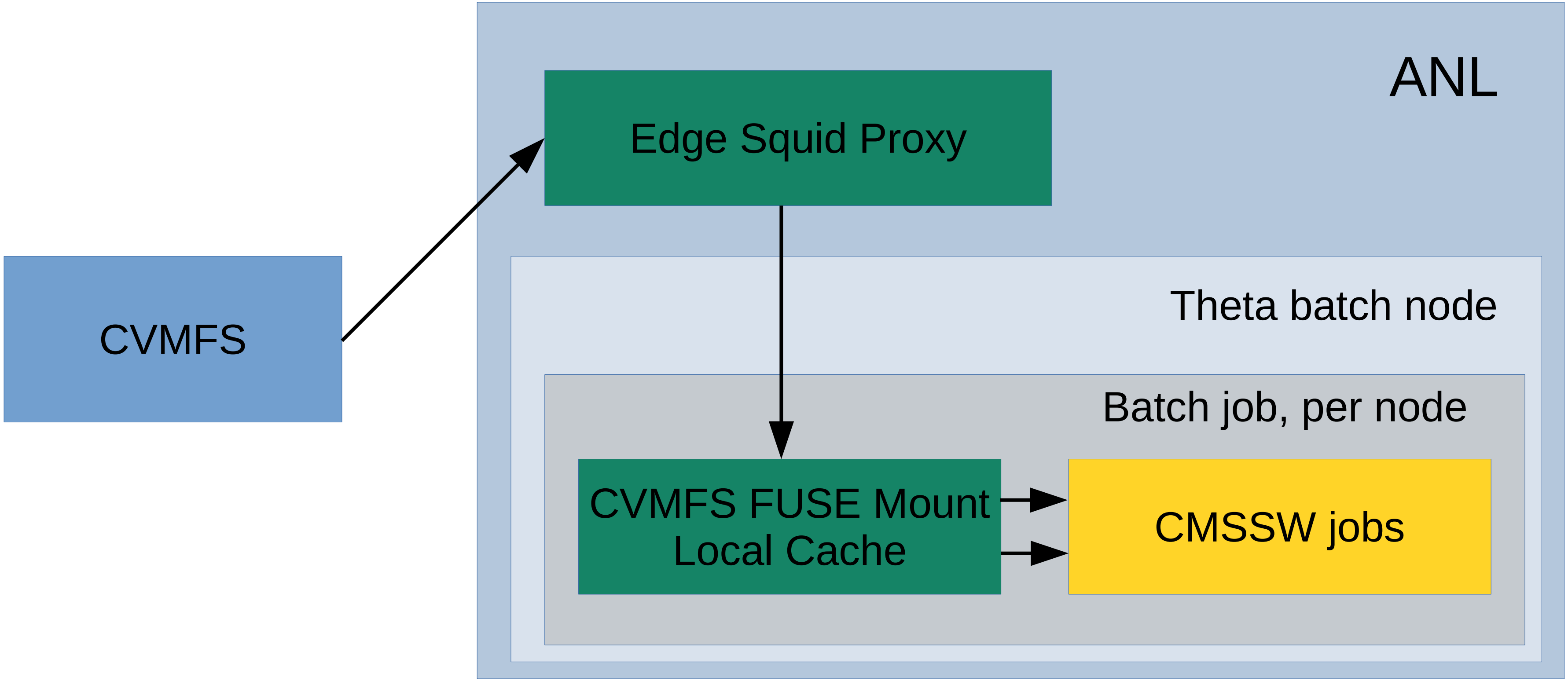
- Had to compile frontier-squid on Theta batch node (Cray Compute Node Linux, based on Suse Enterprise)

- Worked out node squid configuration with Dave Dykstra

- Tested with script from Dave Dykstra that simulates the Frontier load of a CMS job (basically a long list of wget calls)

- Without node squid proxy: 802 nodes, 8 instances/node => failure (due to known edge/node network limitations)
- With node squid proxy: 802 nodes, 64 instances/node => success

CVMFS


ANL
Edge Squid Proxy
CVMFS
Theta batch node
Batch job, per node
CVMFS FUSE Mount
Local Cache
CMSSW jobs

CVMFS

- Using the cvmfsexec package from Dave Dykstra to fuse mount cvmfs (we already do this at TACC Stampede2 in production)

- Package didn’t support Cray/Suse, had to go looking for correct cvmfs rpm and some additional libraries it needed

- After my own (successful) super-hacky version I worked with Dave Dykstra to add Suse support to the cvmfsexec package (mostly providing Theta system info and testing changes)
  - Currently using his latest git branch

Both together, complete picture (sans HTCondor)

- Many-Node Theta Batch job
  - Launcher to run our own script on every node
    - On the node initialize local squid with local cache
    - On the node fuse mount cvmfs (in /local/cvmfs)
    - Singularity (RHEL container) runs our payload(s)
      - Bind mounts /local/cvmfs to /cvmfs

- For now local storage is /dev/shm (will move to node SSD soon)

- Tested on 8 nodes with 64 real CMS Generator jobs per node
  (802 nodes x 64 jobs/node test is still queuing)