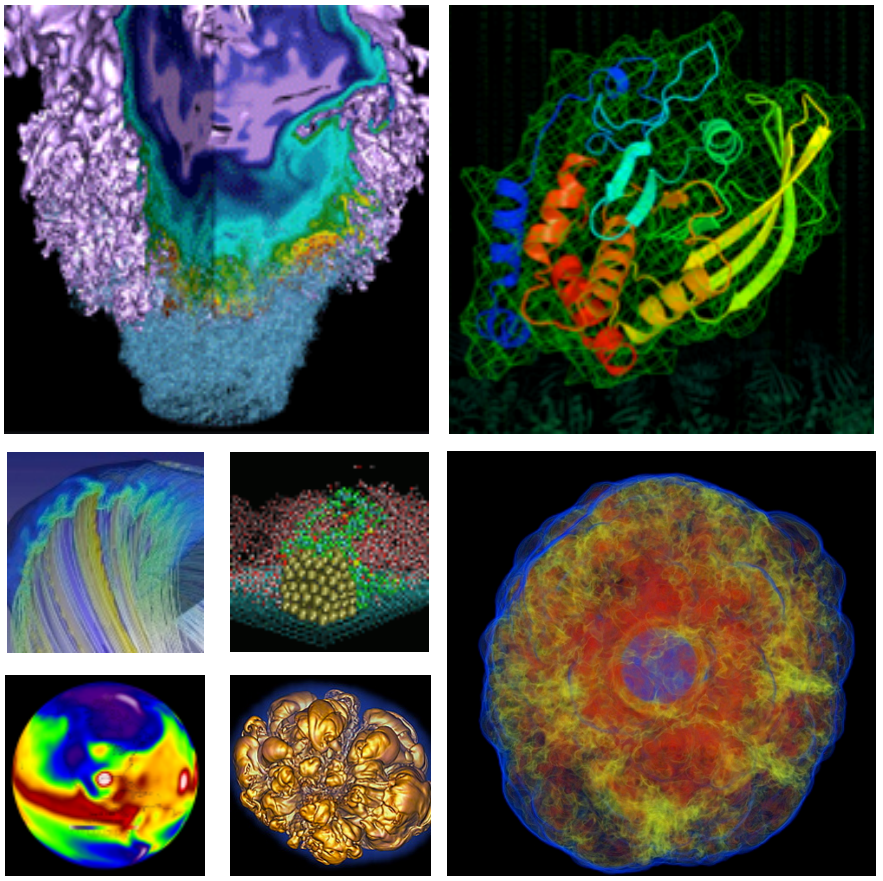
External Torque / Moab on an XC30 and Fairshare



Tina M. Declerck
Iwona Sakrejda

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Outline

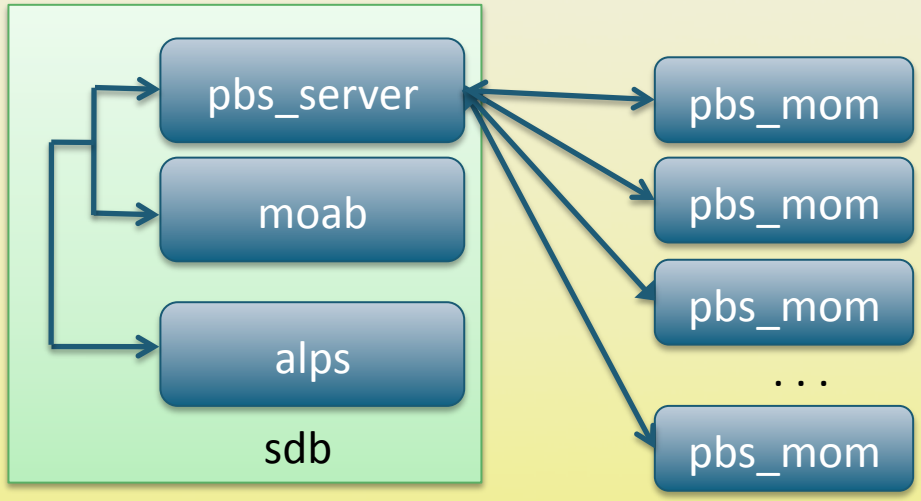
- **External Torque / Moab**
  - Description
  - Why use it?
  - Installation and Configuration
  - Issues
- **Fairshare**
  - Component description
  - Goals
  - Configuration
  - Next Steps

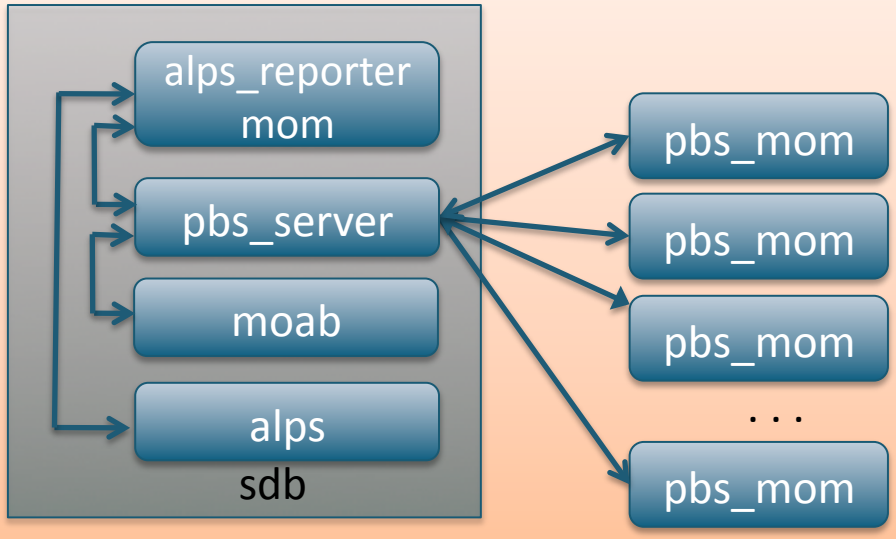
Description

- **Torque / Moab Components**
  - pbs_server – cluster information and status
    - Serverdb keeps information about torque operation and queues
  - qmgr – command line interface to configure torque
  - pbs_mom – tracks jobs from start through completion
  - trqauthd – used for authentication between torque components
  - moab.cfg – holds moab configuration information
  - Alps – Cray's job launcher
- **What changes with external Torque / Moab?**

Changes to Torque / Moab

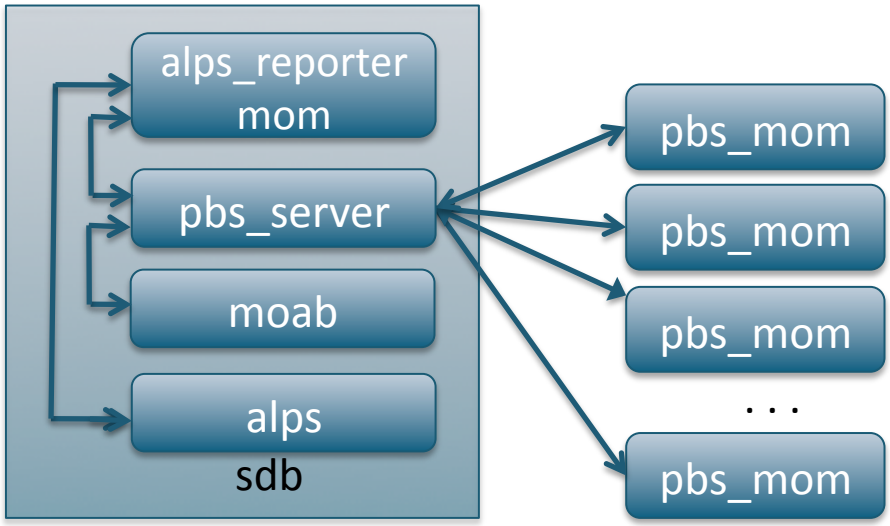
Torque / Moab configuration prior to version 4.1 and 7.0



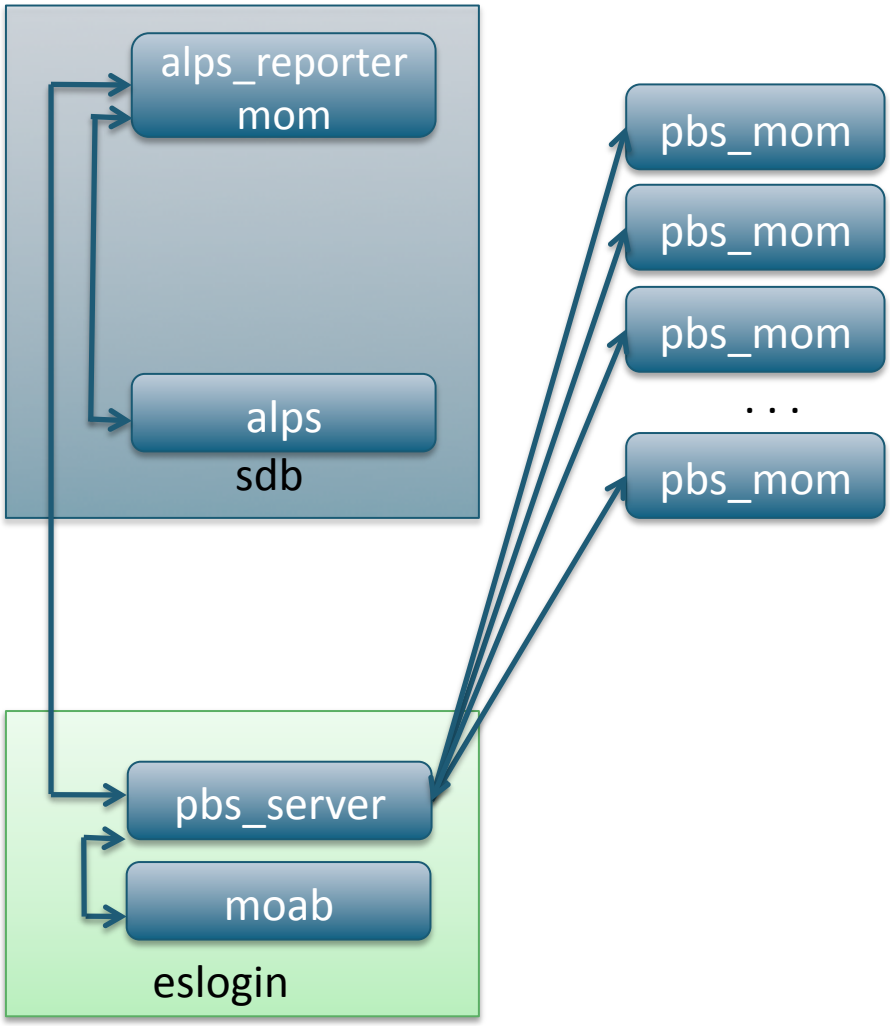


Torque / Moab configuration with version 4.1 and 7.0

External Torque / Moab



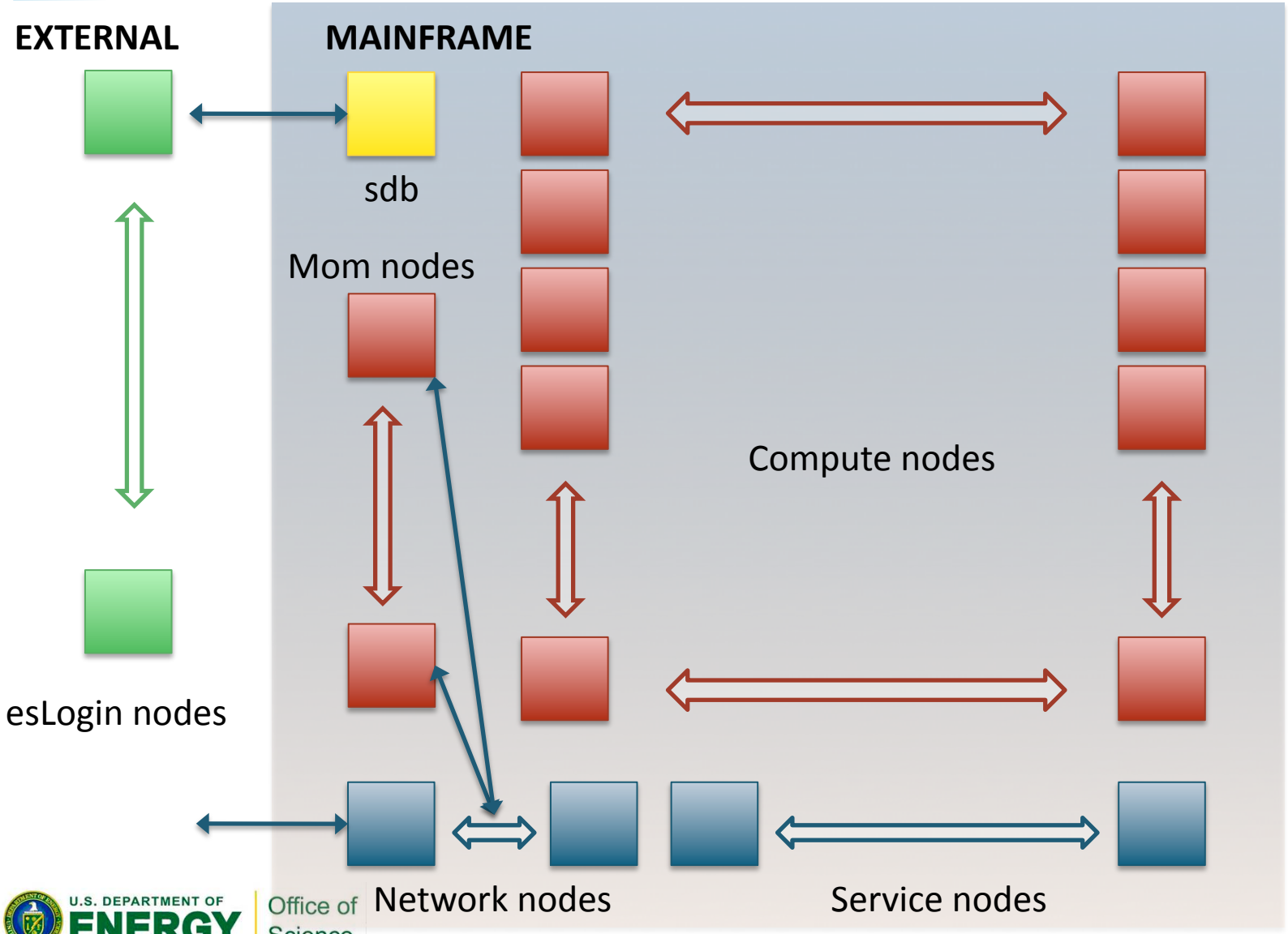
External Torque / Moab



Why use External Torque / Moab?

- **Allows use of a more powerful server**
- **Allows HA configuration**
- **Allows job submission and access when the system is unavailable**
- **Eliminates** (in our case)
  - Running multiple instances of torque
  - Wrappers for torque commands

NERSC Configuration



Torque Installation

- **Preparation**
  - If moving from internal to external
  - Verify network connections
  - NERSC uses a persistent /var; changes needed for other configurations
- **Building the software**
  - Appendix G of the Moab 7.2.x manual
  - Configure & make (on external server and boot node)

Torque Install (cont)

- **Edit the <server_home>/server_priv/nodes file**

```
sdb cray_compute alps_reporter
nid00264 np=16 alps_login
nid00265 np=16 alps_login
. . .
nid00704 np=16 alps_login
nid00705 np=16 alps_login
```

- **Allow access and submit privileges from mom nodes**

```
set server acl_hosts += nid00264
set server acl_hosts += nid00265
. . .
set server acl_hosts += nid00704
set server acl_hosts += nid00705

set server submit_hosts += nid00264
set server submit_hosts += nid00265
. . .
set server submit_hosts += nid00704
set server submit_hosts += nid00705
```

- **Other qmgr parameters**

```
set server cray_enabled = true
set server resources_default.partition=<clustername>
```

Torque Install (cont)

- **Copy installed /var/spool/torque to mom nodes and sdb persistent var**
- **echo [server_name] > /snv/[nid]/var/spool/torque/server_name**
- **mom: add to /var/spool/torque/mom_priv/config:**

  $login_node true

  $apbasil_path <path-to-apbasil> *default /usr/bin/apbasil

  $apbasil_protocol 1.2

- **sdb: add to /var/spool/torque/mom_priv/config:**

  $reporter_mom true

  $apbasil_path <path-to-apbasil> *default /usr/bin/apbasil

  $apbasil_protocol 1.2

Moab Configuration

- **Preparation**
- **Building**
- **Configuration – edit moab.cfg**

```
NODECFG[nid00264] Partition=login
NODECFG[nid00265] Partition=login
. . .
NODECFG[nid00704] Partition=login
NODECFG[nid00705] Partition=login
CLIENTCFG[DEFAULT]  DEFAULTSUBMITPARTITION=<clustername>
```

Other Info

- **Network**
  - add sdb and mom node nid names to /etc/hosts

```
118.5.2.12      edisdb     sdb
118.5.1.2   edimom01    nid00264
118.5.1.3   edimom02    nid00265
. . .
118.5.1.7   edimom07    nid00704
118.5.1.8   edimom08    nid00705
```

- **mppnppn is now a required field**
  - Need to set a default in qmgr

Fairshare Components

- **What is fairshare?**
  - Allows historical usage to be used for priority calculations
- **Two levels**
  - Data collection
  - Applying Fairshare criteria
- **Priority based Fairshare**
  - Fairshare weights
    - Applied to subcomponents
      - Account, class, group, user, and QoS
  - Time
    - Interval, depth, and decay rate

NERSC's Goals

- **Primary goal of providing DARPA with 25% of compute time**
- **Other considerations:**
  - Queue wait time should make a difference
  - A single user shouldn't be able to take over most of the system
  - A single user should not use all shares for a repo
  - We should be able to favor large jobs (or at least not discriminate against them)
  - Maintain the ability to favor individuals or repos when necessary

NERSC Process

- **Ran with fairshare without weights to get a baseline**
- **Tested fairshare on a couple of repos**
  - Set fairshare baseline with DARPA at 25% and all other accounts (repos) to 10%
  - Set a couple of users shares higher than others to see impact (15%)
- **Created a fairshare tree for each set of repos in each DOE office of science**
  - Initially broke fairshare – needed an additional parameter

Configuration

```
FSINTERVAL  12:00:00
FSDEPTH     8
FSDECAY     0.8
FSPOLICY    DEDICATEDPS
FSWEIGHT    1
FSACCOUNTWEIGHT 100

# darpa repo should have 25%

FSTREE[root] SHARES=100 MEMBERLIST=darpa,nersc
FSTREE[darpa] SHARES=25 MEMBERLIST=acct:darpa
FSTREE[nersc] SHARES=75 MEMBERLIST=ascr,ber,bes,ccc,fes,hep,np

FSTREE[ascr] SHARES=5 MEMBERLIST=acct:m888,acct:m945, +85 more
FSTREE[ber] SHARES=17 MEMBERLIST=acct:m917,acct:m950, +126 more
FSTREE[bes] SHARES=30 MEMBERLIST=acct:m881,acct:m894, +303 more
FSTREE[ccc] SHARES=7 MEMBERLIST=acct:mpesnet,acct:mpccc, +10 more
FSTREE[fes] SHARES=17 MEMBERLIST=acct:m908,acct:m916, +63 more
FSTREE[hep] SHARES=13 MEMBERLIST=acct:m981,acct:m1067, +63 more
FSTREE[np] SHARES=11 MEMBERLIST=acct:m1401,acct:m327, +44 more
```

Monitoring Fairshare

- **mdiag –p**

edison06:~ # mdiag -p
diagnosing job priority information (partition: ALL)

```
Job             PRIORITY*   Cred( User:Group:Accnt:Class)    FS(Accnt)  Serv(QTime)
        Weights  --------      1( 1440: 1440: 1440: 1440)     1(  100)     1(    1)

37999               -4161    17.1(  0.0:  0.0:  0.0:  1.0)  74.7(-62.8)   8.1(683.0)
39588               -1746     0.0(  0.0:  0.0:  0.0:  0.0)  70.8(-29.7)  29.2(1223.)
39592               -1745     0.0(  0.0:  0.0:  0.0:  0.0)  70.8(-29.7)  29.2(1224.)
45878               -9715     0.0(  0.0:  0.0:  0.0:  0.0)  70.5(-167.)  29.5(7012.)
. . .
99575              -15051     7.8(  0.0:  0.0:  0.0:  1.0)  90.9(-167.)   1.3(236.0)
99631               -4367     0.0(  0.0:  0.0:  0.0:  0.0)  99.1(-44.1)   0.9( 39.0)
99643               -6267     0.0(  0.0:  0.0:  0.0:  0.0)  99.7(-62.8)   0.3( 17.0)
99652              -16716     0.0(  0.0:  0.0:  0.0:  0.0)  99.9(-167.)   0.1( 11.0)
99661               -6283     0.0(  0.0:  0.0:  0.0:  0.0) 100.0(-62.8)   0.0(  1.0)

Percent Contribution  --------    1.9(  0.0:  0.0:  0.0:  1.9)  80.8( 80.8)  17.3( 17.3)

* indicates absolute/relative system prio set on job
```

Monitoring Fairshare

- **mdiag –fv**

edison06:~ # mdiag -fv
FairShare Information

Depth: 8 intervals Interval Length: 12:00:00 Decay Rate: 0.80

FS Policy: DEDICATEDPS
System FS Settings: Target Usage: 0.00

FSInterval	%	Target	0	1	2	3	4	5	6	7
FSWeight			1.0000	0.8000	0.6400	0.5120	0.4096	0.3277	0.2621	0.2097
TotalUsage	100.00		14298.9	119618.7	128607.6	131376.5	120110.3	135019.9	123260.6	119206.9
USER										
wangw	0.00					0.01				
nsai	0.00							0.00		
vartyukh	1.97				2.02	2.60	3.06	5.09	2.82	
yong	10.77			6.87	21.67	18.69	14.63			
jptrinas	1.47		1.70	1.04	1.27	3.95	1.28	0.24	0.82	0.32
lujian	0.28		0.03	0.12	0.45	0.09	0.40	0.43	0.26	0.43

Monitoring Fairshare

GROUP

---

wangw 0.00 ------- ------- ------- 0.01 ------- ------- ------- ------- -------

nsai 0.00 ------- ------- ------- ------- ------- ------- 0.00 ------- -------

vartyukh 1.97 ------- ------- ------- 2.02 2.60 3.06 5.09 2.82 -------

yong 10.77 ------- ------- 6.87 21.67 18.69 14.63 ------- ------- -------

jptrinas 1.47 ------- 1.70 1.04 1.27 3.95 1.28 0.24 0.82 0.32

lujian 0.28 ------- 0.03 0.12 0.45 0.09 0.40 0.43 0.26 0.43

sellner 0.00 ------- ------- ------- 0.00 ------- ------- ------- 0.00 0.00

arago 0.03 ------- ------- ------- ------- ------- 0.09 ------- 0.11 0.17

ankitb 9.36 ------- 53.04 32.22 ------- ------- ------- ------- ------- -------

ACCT

---

mp173 0.47 ------- ------- 0.85 0.16 ------- 0.13 0.21 0.80 2.19

mp19 0.00 ------- ------- ------- 0.01 ------- ------- ------- ------- -------

mp13 3.23 ------- ------- 0.43 10.19 ------- ------- ------- ------- 17.88

m747 1.01 ------- ------- ------- 3.62 0.10 ------- 0.09 3.11 0.17

m906 0.00 ------- ------- ------- ------- 0.01 ------- ------- ------- -------

mp7 0.20 ------- 3.99 0.27 ------- ------- ------- ------- ------- -------

m172 0.59 ------- ------- 2.55 ------- ------- ------- ------- ------- -------

gc8 0.39 ------- 0.36 0.10 0.66 0.29 0.45 0.23 0.16 1.32

mp48 0.00 ------- ------- ------- ------- ------- ------- 0.00 ------- -------

mp261 1.54 ------- 1.70 1.26 1.34 3.98 1.30 0.24 0.82 0.32

Monitoring Fairshare

QOS

---

reserve 39.19 ------- 53.04 39.09 38.69 36.54 31.57 41.85 51.19 35.16

CLASS

---

system 0.85 ------- ------- ------- ------- ------- 0.92 6.87 ------- -------

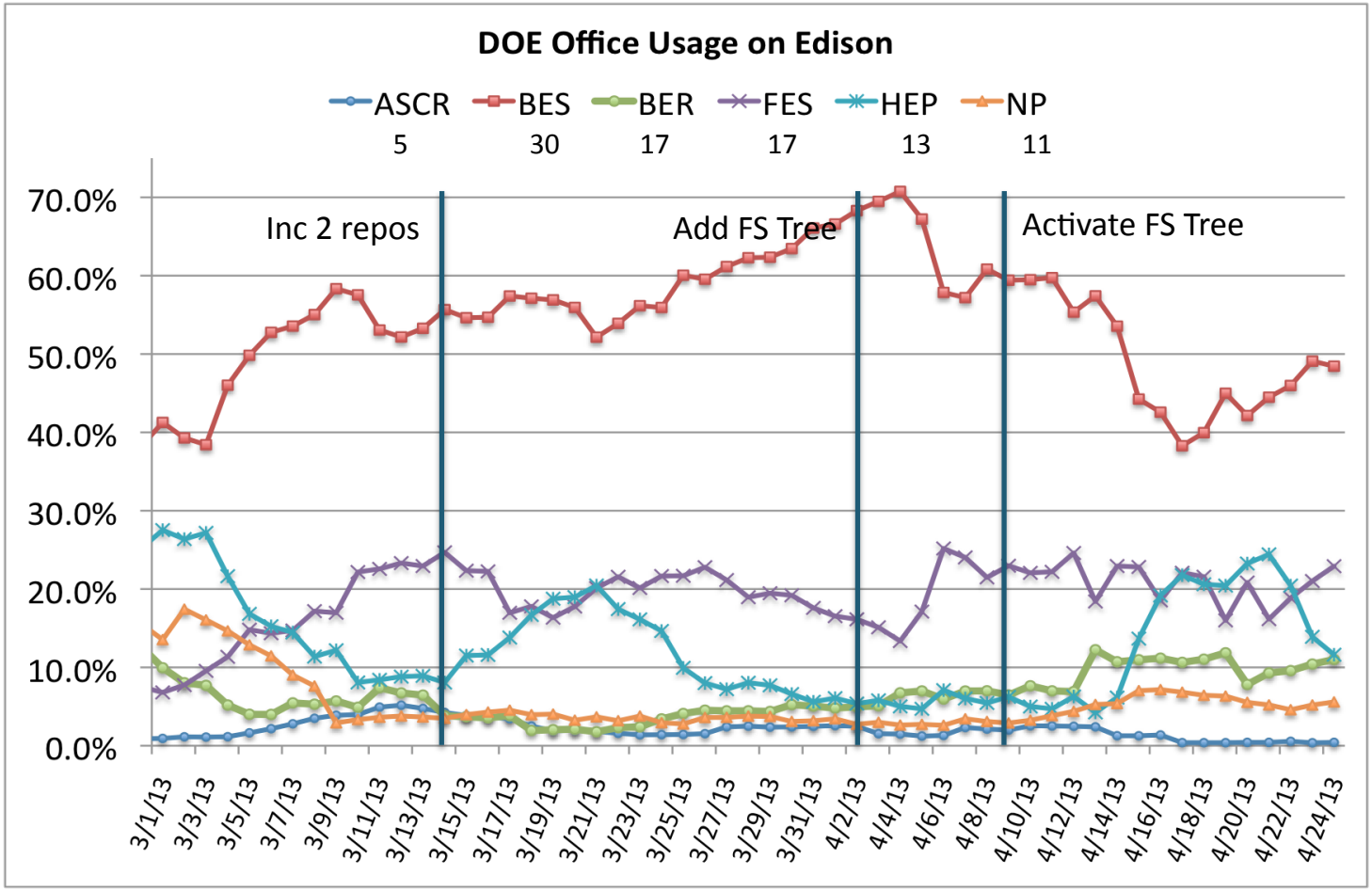
debug 3.52 ------- 5.39 2.68 6.23 1.93 3.03 0.79 5.61 4.20

reg_small 33.23 ------- 34.03 35.71 29.84 45.23 30.23 29.66 23.37 27.20

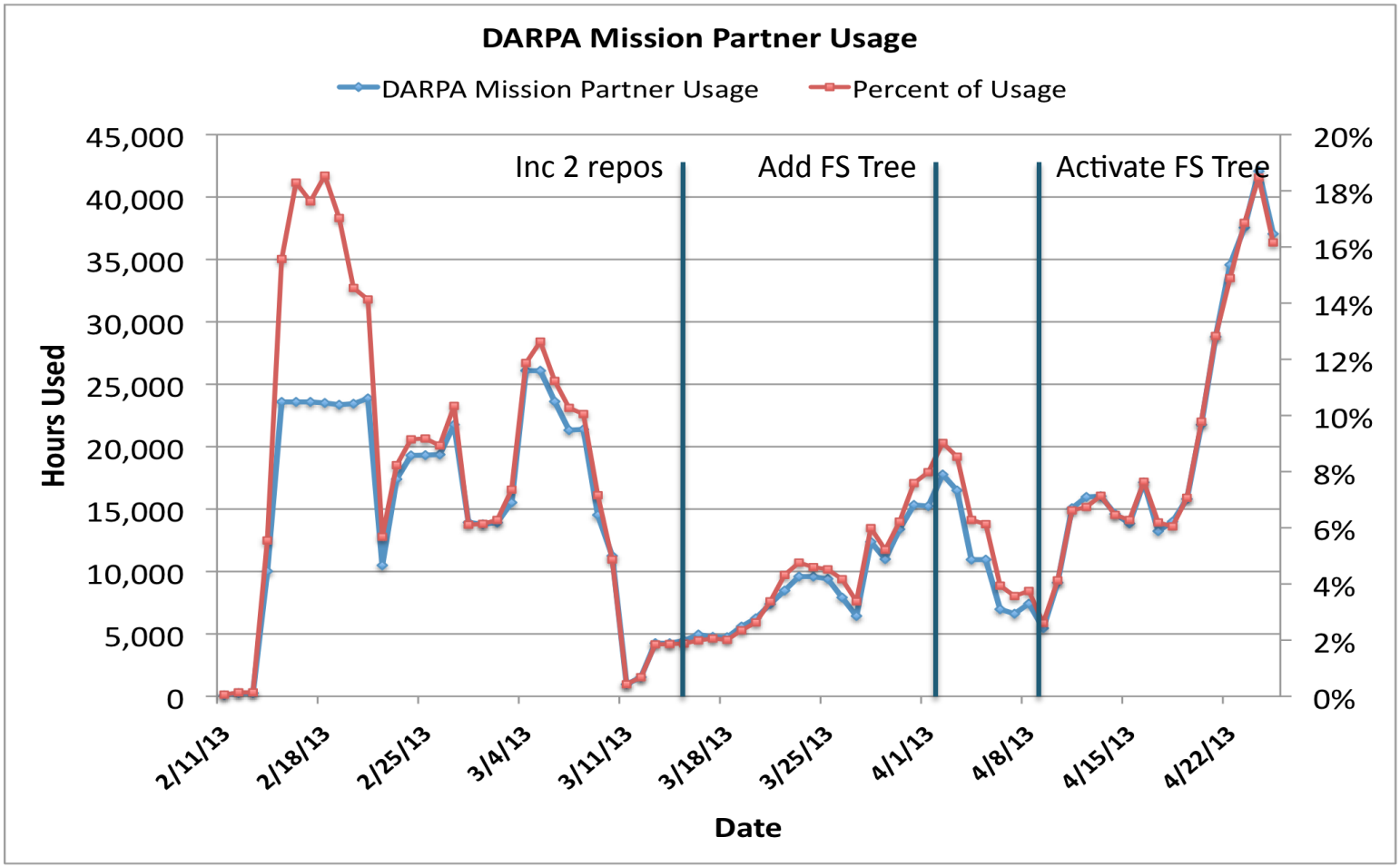
Share Tree Overview for partition 'ALL'

Name	Usage	Target	(FSFACTOR)
root	100.00	100.00 of	100.00 (node: 1327390960.06) (0.00)
- darpa	10.19	25.00 of	100.00 (node: 135299969.11) (0.00)
- darpa	25.00	0.00 of	25.00 (acct: 135299969.11) (65.10)
- nersc	89.81	75.00 of	100.00 (node: 1192090990.94) (0.00)
- ascr	8.45	5.00 of	100.00 (node: 100787159.45) (0.00)
- m888	0.01	0.00 of	5.00 (acct: 241715.69) (-78.05)
- m945	0.00	0.00 of	5.00 (acct: 0.00) (-78.05)
- m1142	0.00	0.00 of	5.00 (acct: 0.00) (-78.05)
- m1209	0.00	0.00 of	5.00 (acct: 0.00) (-78.05)
- m1175	0.00	0.00 of	5.00 (acct: 0.00) (-78.05)
- m1222	0.00	0.00 of	5.00 (acct: 0.00) (-78.05)

Results


DOE Office Usage on Edison
ASCR
BES
BER
FES
HEP
NP
5
30
17
17
13
11
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
Inc 2 repos
Add FS Tree
Activate FS Tree
3/1/13
3/3/13
3/5/13
3/7/13
3/9/13
3/11/13
3/13/13
3/15/13
3/17/13
3/19/13
3/21/13
3/23/13
3/25/13
3/27/13
3/29/13
3/31/13
4/2/13
4/4/13
4/6/13
4/8/13
4/10/13
4/12/13
4/14/13
4/16/13
4/18/13
4/20/13
4/22/13
4/24/13

Results


DARPA Mission Partner Usage
DARPA Mission Partner Usage
Percent of Usage
Inc 2 repos
Add FS Tree
Activate FS Tree
Hours Used
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
2/11/13
2/18/13
2/25/13
3/4/13
3/11/13
3/18/13
3/25/13
4/1/13
4/8/13
4/15/13
4/22/13
Date

Next Steps

- **Test additional parameters**
  - Limit usage for a single user
  - Apply fairshare to each repo rather than a group
  - Set QoS to provide more priority for larger jobs