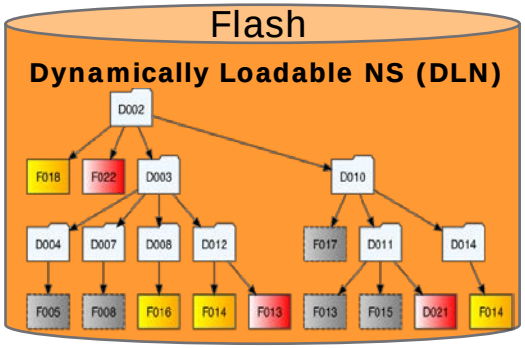
2 TIERS™ Architecture

POSIX-like namespace layered above an object store

Percy Tzelnic – SVP, EMC Fellow
Sorin Faibish – DE, Architect
Fast Data Group
Office of the CTO
EMC

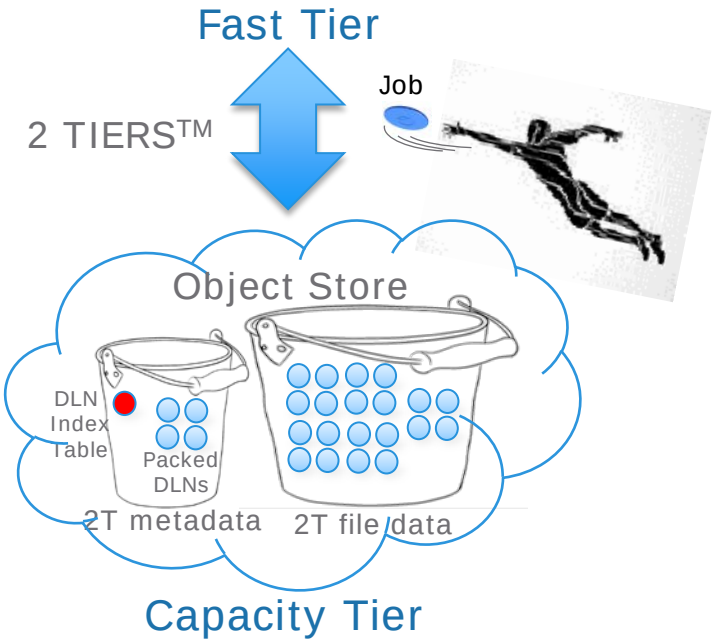
MODERNIZE

EMC WORLD 2016



2 TIERS™ –

Roots In Exascale Research



- EMC collaboration with DoE and Industry Consortia in FastForward Exascale
  - Fast acceleration tier → Performance of flash
  - Large capacity tier → Retention and capacity of object store
  - Global POSIX namespace over one trillion objects

EMC²

Fundamental Change In Storage Architecture

- Contemporary Storage Architecture is being disrupted:
  - Flash replaces disk for +100x performance (Flash Array)
  - Cloud replaces disk for +100x capacity (Object Store)
  - Capacity disks from arrays move to the cloud, leaving a Flash only Fast Tier on-premise, and an Object Store only Capacity Tier, in the Cloud or sometimes on premise
- We can no longer package Performance and Capacity in one box at an attractive price/value point
  - *Split the two, hence 2 TIERS™ (Fast and Capacity Tier)*

Instantiation Of 2 TIERS™ In EMC

- *In EMC, Fast Tier can be instantiated in the network by DSSD, or in the servers by ScaleIO local flash*
- *2 TIERS™ is Software Defined Storage with two key components:*
  - *PFS – Parallel File System (OrangeFS) and*
  - *Syncer (built by EMC for the DOE FastForward program, for Exascale I/O stacks)*

"The DSSD D5 Storage Appliance … this brings me to some future technologies that EMC was showing in their Innovation exhibit. There they were showing what they called Two-Tiers model …"

"The Two Tiers model provided consistent subsets of both data and metadata between the hot edge and cold core storage capacity (hence two tiers). The company said it was seeing performance and cost benefits from this solution prototype for IO and capacity intensive applications."

Tom Coughlin, Intersect 360

http://www.forbes.com/sites/tomcoughlin/2015/05/08/emc-cloud-storage-flash-memory-and-beyond/

- *In EMC, the Capacity Tier can be instantiated by ECS or Isilon*
- *EMC 2 TIERS™ software runs in the Fast Tier, presents the POSIX API and Namespace to apps and maps the apps data into objects on the Capacity Tier, with policy driven tiering between the two*

2 TIERS™ Was Designed For The 3rd Platform

But It Provides Equal Support To 2nd Platform Apps!

Disaggregate the monolithic memory / storage / IO Stack and recast it into loosely coupled “Fast Tier” and “Capacity Tier”, to enable

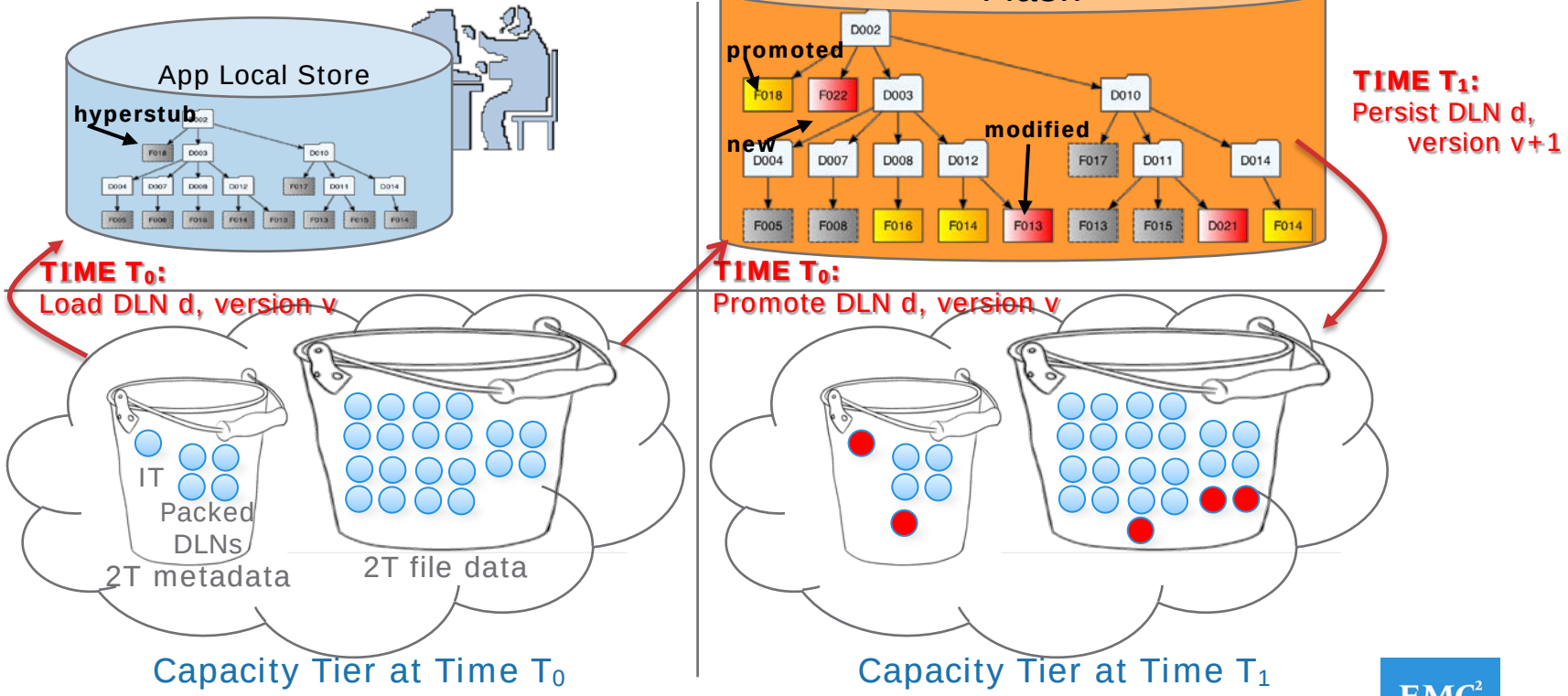
- Independent Scaling:
  - Scale-out for Fast Tier, O(100 to 1,000)
  - Hyperscale for Capacity Tier, O(100,000)

- POSIX API and Namespace
  - Required by the majority of 2nd platform apps (at a scale-out level lower than for a 3rd platform infrastructure)

How Does 2 TIERS™Work?

Read-only, Read-through Translation Service on Local FUSE File System

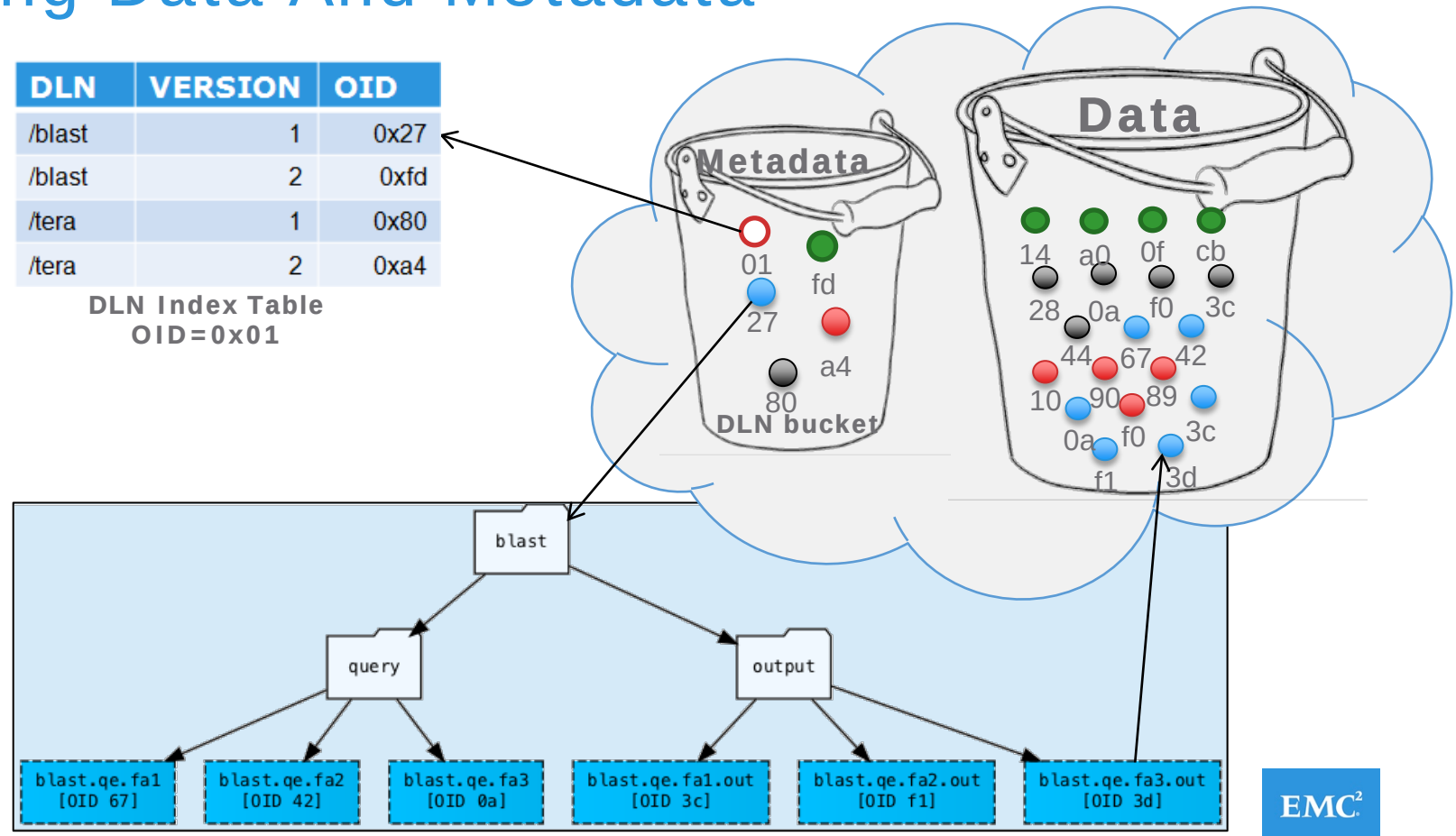
Fast Tier on Distributed OrangeFS



Tiering Data And Metadata

DLN	VERSION	OID
/blast	1	0x27
/blast	2	0xfd
/tera	1	0x80
/tera	2	0xa4

DLN Index Table
OID=0x01



How Does Metadata Tiering Work?

Similar To A UNIX FFS File System

- Expose a global namespace view to the app, similar to a File System
- Use a pre-defined Global Unique OID for the DLN Index Table (Cassandra KVS); similar to a SuperBlock of a File System
- Each Entry in the Index Table points to a DLN; similar to inodes for directories in a File System
- Each DLN points to a partial view of global namespace; similar to a subtree of a File System
- Each namespace view has a pointer to an object; similar to file inodes in a directory in a File System

EMC 2 TIERS™ – One Of Many Similar Approaches

- Since June 2014, while EMC develops 2 TIERS™, an ever increasing number of similar projects have emerged in the industry, both in EMC and outside (university research, new ventures, etc.)
  - MarFS (LANL) – the most similar to 2 TIERS™, in production for Campaign Archival
  - DeltaFS
  - BatchFS
  - IndexFS
  - TableFS
  - SlimFS
  - ShardFS
  - KVFS
  - BetrFS
  - GiraffaFS
- This is good confirmation of two widely resonating concepts:
  - Object Storage for Capacity, Flash for Performance!
  - Users think in folders, not objects! … They need a File System Namespace!

Differentiation Of EMC 2 TIERS™

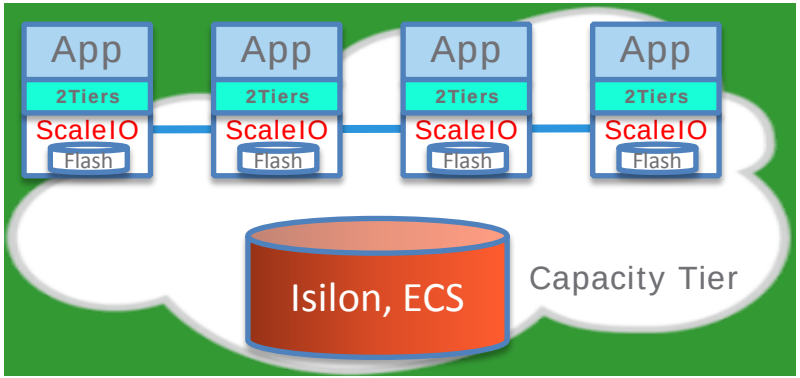
Unique Characteristics Of 2 Tiers™

1. Single Global Namespace with Dynamically Loadable Namespaces (DLNs)
2. Tiering of both Data and Metadata
3. Fast Tier Performance Target: greater than 10X Capacity Tier
4. Direct access (read-only) to the Capacity Tier, bypassing the Fast Tier
5. 2 TIERS™ provides Tiering and Non-Tiering modes
6. No client changes required
7. No changes to the EMC products required for EMC instantiation

2 TIERS™: Local Or Network Fast Tier Examples

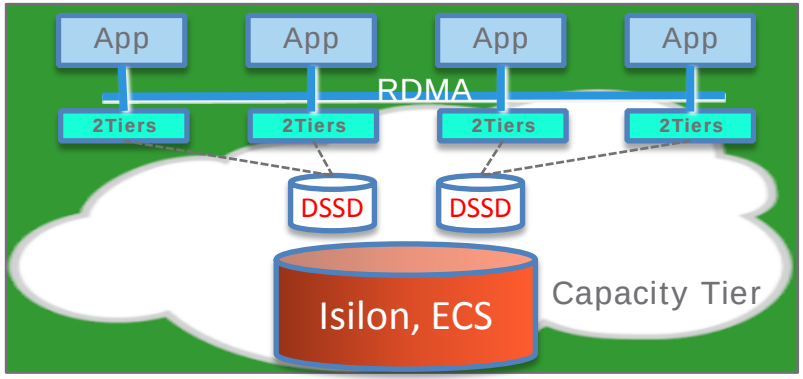
Local Fast Tier

App Cluster
(Compute Servers)
+ 2T Servers + Local Flash



Network Fast Tier

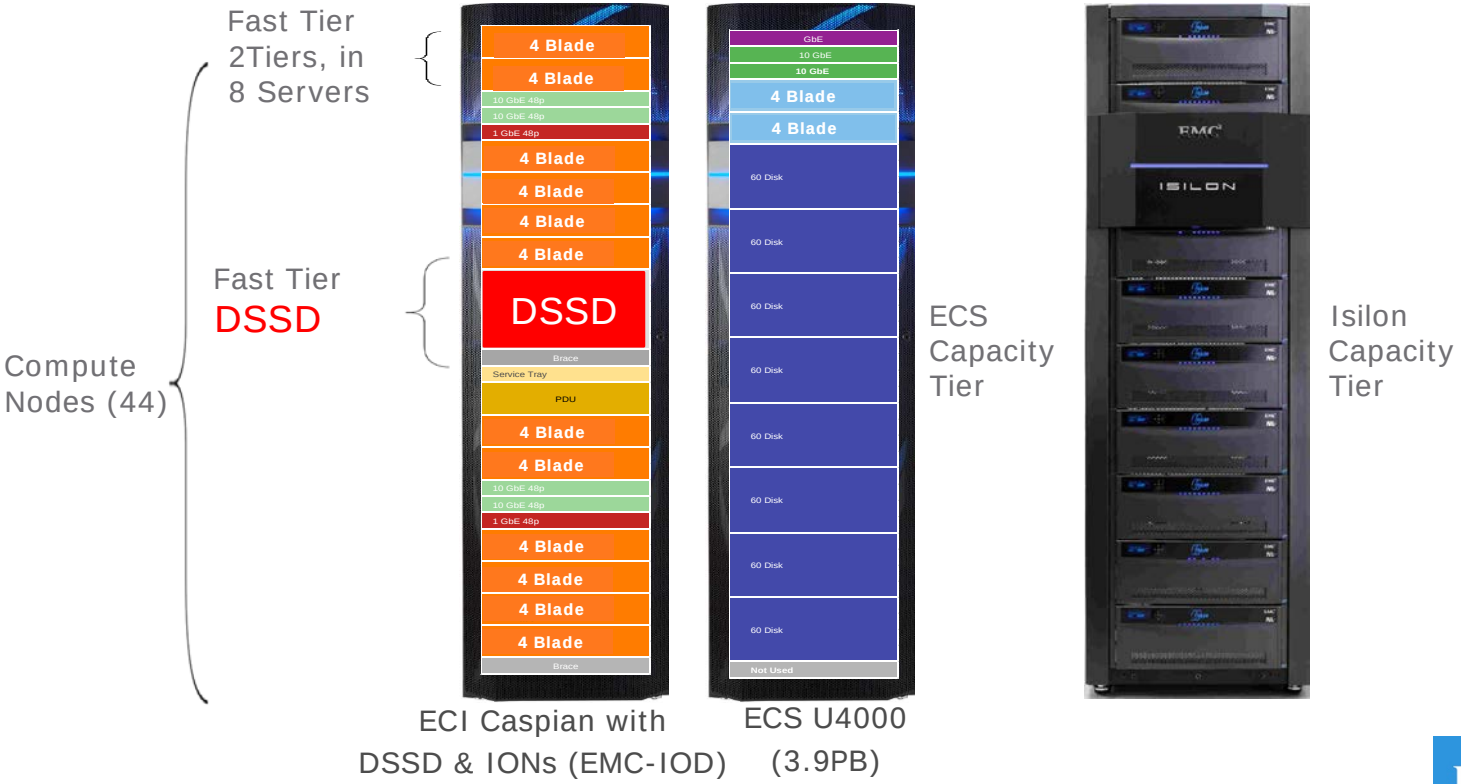
App Cluster
(Compute Servers)



2T Servers

Note: Compute Server interconnect should be RDMA, for best performance

A Possible EMC Product Packaging



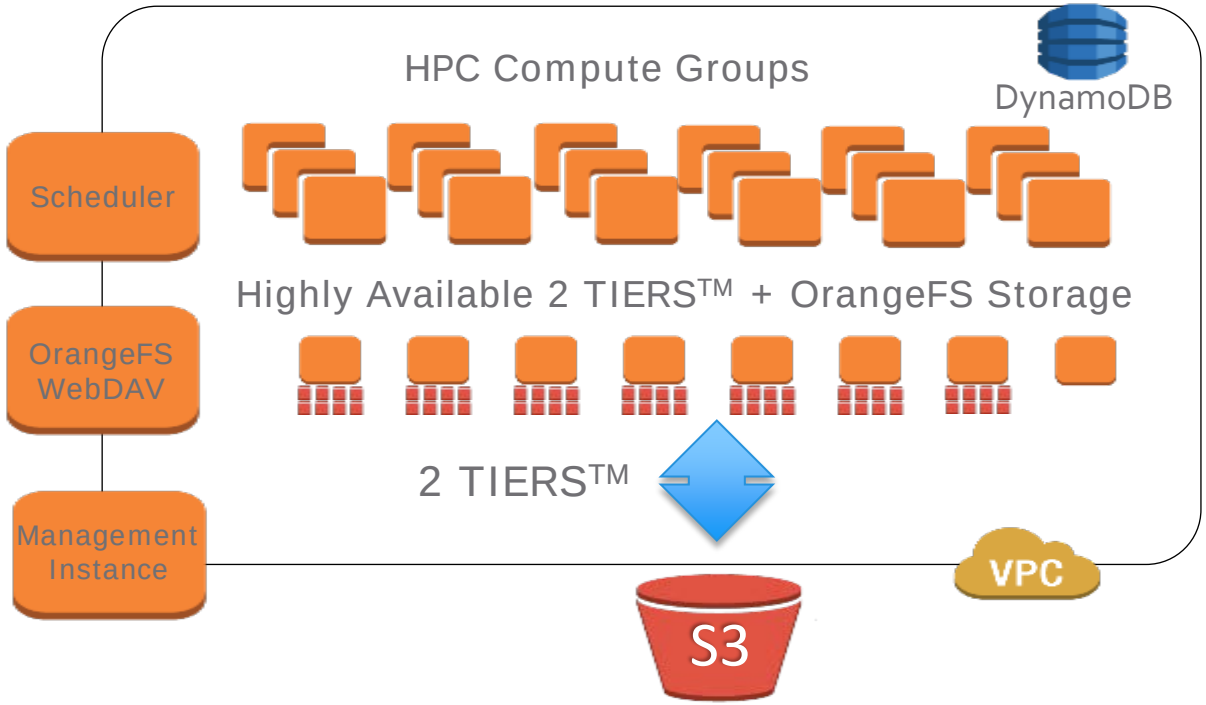
OrangeFS: EMC Choice For 2 TIERS™

- Stateless design of underlying PVFS2
  - Light weight Linux kernel module, multi-threaded client
  - Performance comparable to other PFS
  - HDFS with JNI client; support for Windows, Mac
- Modular design
  - Abstract key-value interface for metadata
  - Abstract storage interface for data
  - Abstract networking allows RDMA, IP
- Client changes NOT required

- Future roadmap: OFS V3 – changes for CloudyCluster™ (Cloud PaaS); already deployed on AWS
- OrangeFS is maintained and developed by Omnibond, Clemson, SC
  - Agile and responsive open source community
  - Committed to open source community development
  - History of 4-5 years in production, at major customers

EMC 2 TIERS™ On Omnibond CloudyCluster On AWS

- 2 TIERS™ Customer Demo built on CloudyCluster to host POSIX apps as POC on AWS
- Several Large Customers of EMC have asked for access waiting to 2 TIERS™ on AWS
- After for POC ready, significant dev work remains
- Initial Custom AMI built by EMC with 2 TIERS™ and CloudyCluster

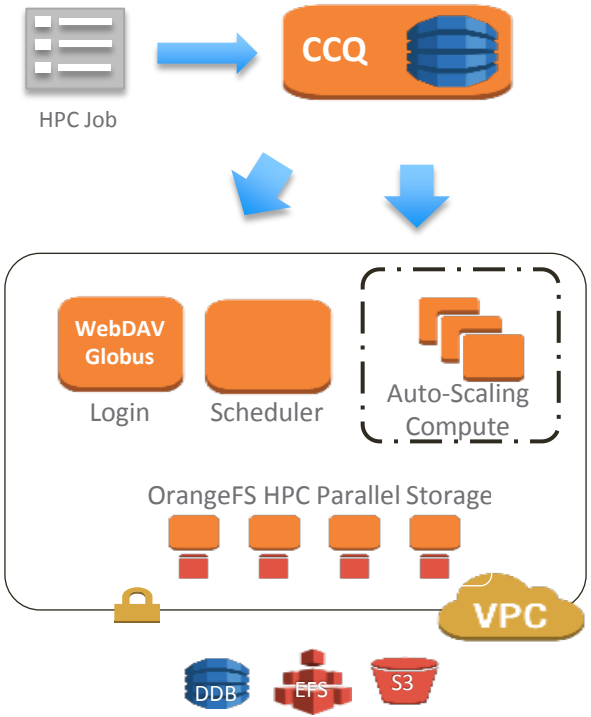


awsmarketplace Omnibond EMC²

CloudyCluster

Self Service Elastic HPC

OrangeFS



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

- ***Storage***: OrangeFS on EBS, S3, EFS
- ***Compute***: Job Driven Elastic Compute through CCQ
- ***Scheduler***: Initially Torque 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

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

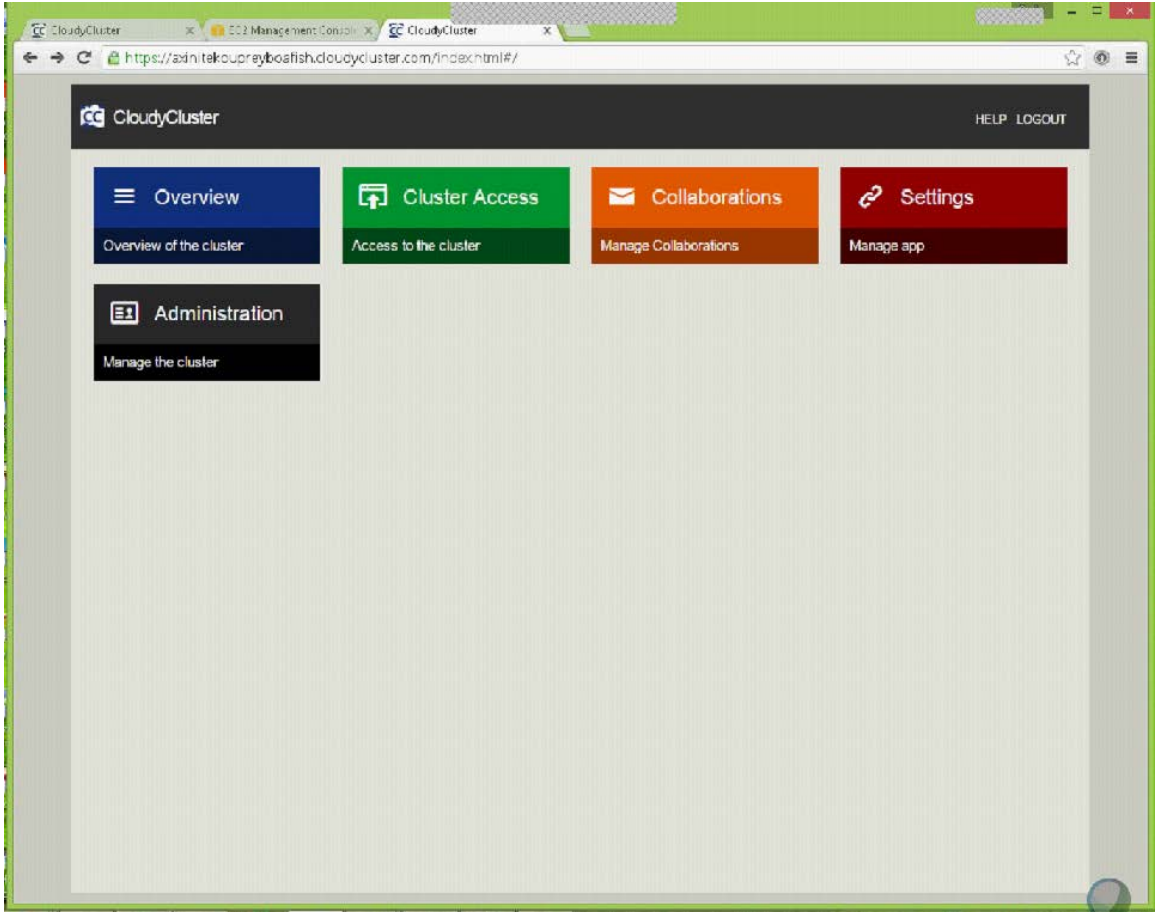
- All from an easy to use Web UI from mobile, tablet or desktop

Available now in the awsmarketplace

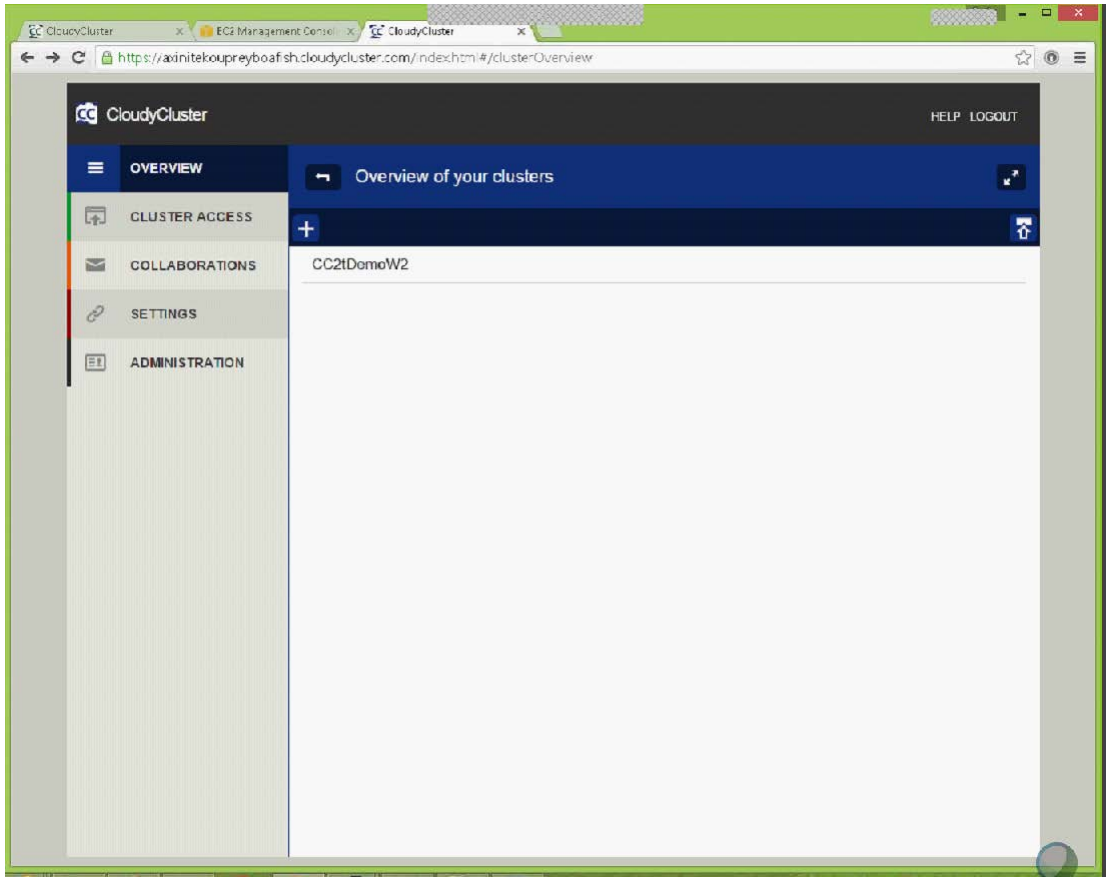
Omnibond EMC²

Accessing 2 TIERS™ On AWS

View available
2 Tiers clusters



List Of 2 TIERS™ Clusters



Single instance
2 Tiers cluster
On AWS West

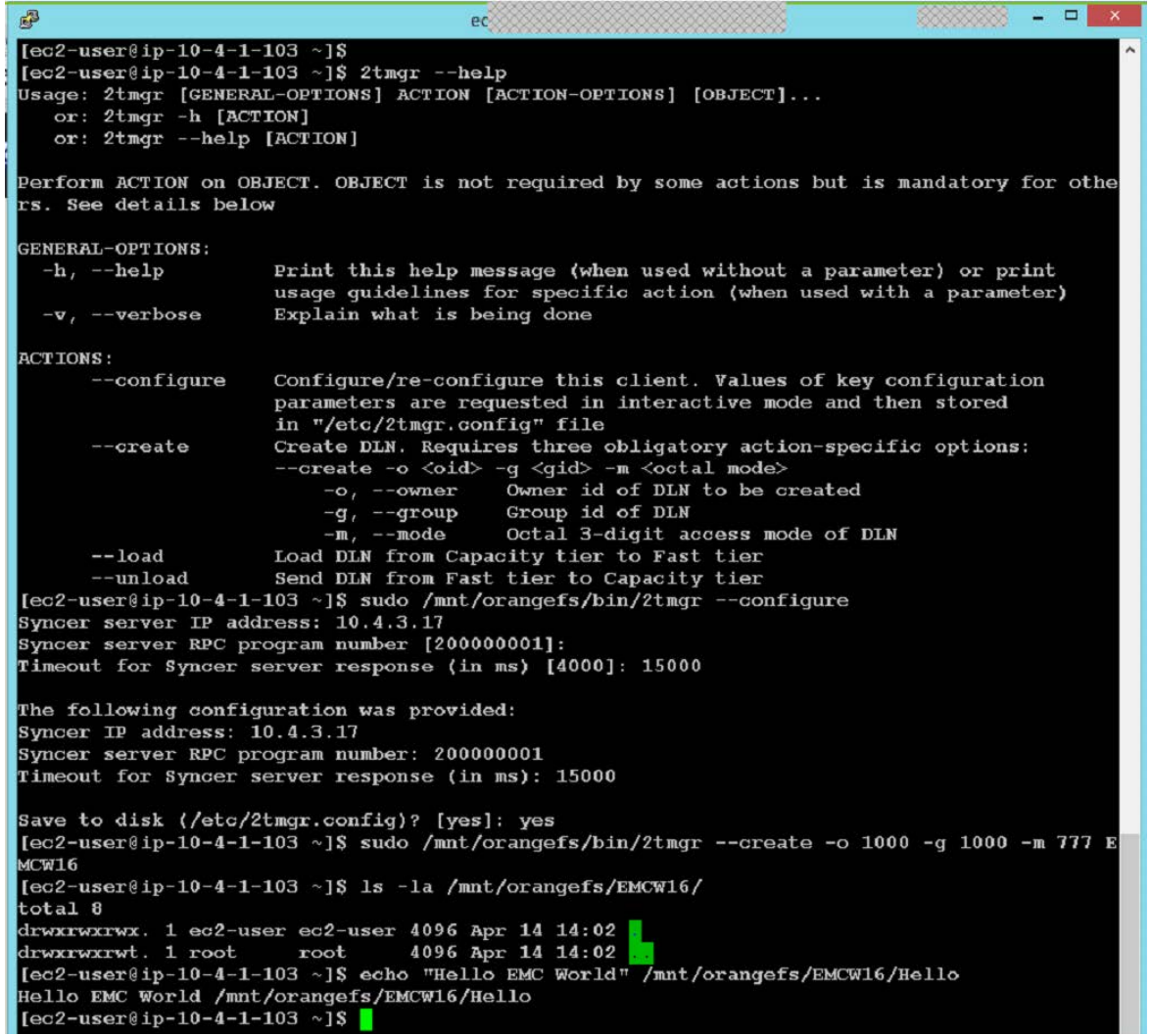
View Of 2 TIERS™Cluster On AWS

Control Nodes

Compute Nodes

OrangeFS Nodes

Management Nodes

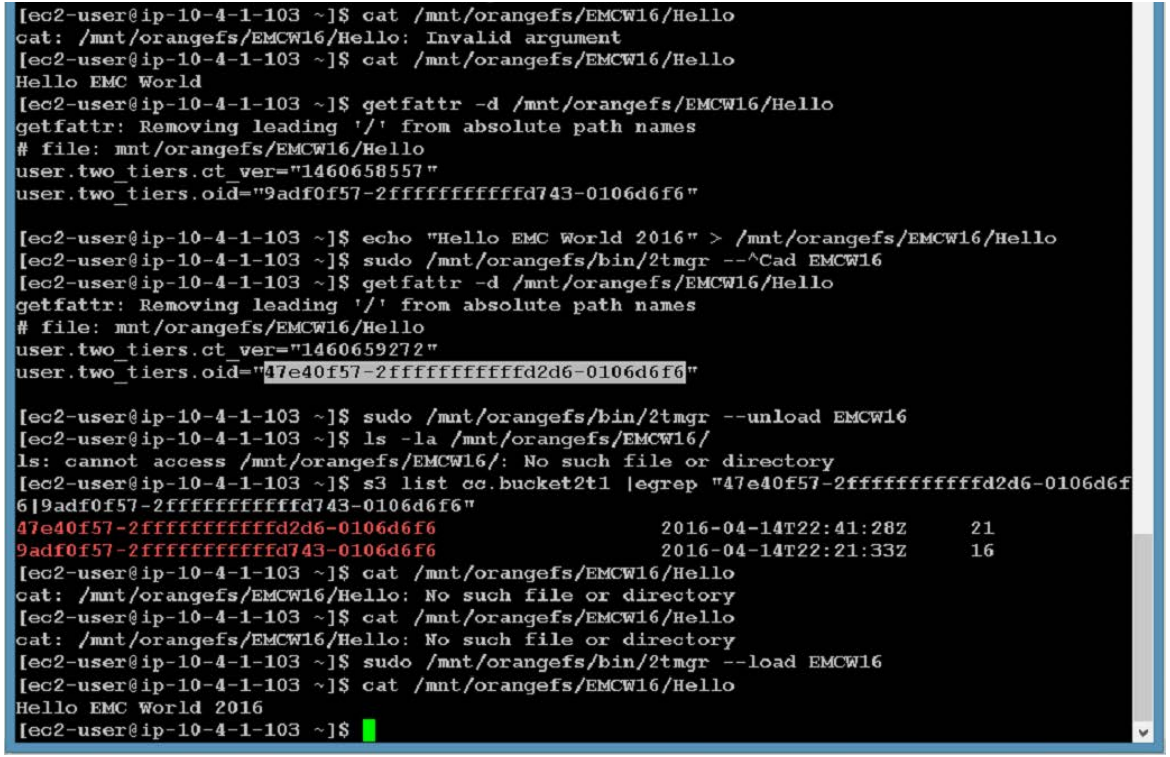

[ec2-user@ip-10-4-1-103 ~]$
[ec2-user@ip-10-4-1-103 ~]$ 2tmgr --help
Usage: 2tmgr [GENERAL-OPTIONS] ACTION [ACTION-OPTIONS] [OBJECT]...
or: 2tmgr -h [ACTION]
or: 2tmgr --help [ACTION]
Perform ACTION on OBJECT. OBJECT is not required by some actions but is mandatory for othe
rs. See details below
GENERAL-OPTIONS:
-h, --help Print this help message (when used without a parameter) or print
usage guidelines for specific action (when used with a parameter)
-v, --verbose Explain what is being done
ACTIONS:
--configure Configure/re-configure this client. Values of key configuration
parameters are requested in interactive mode and then stored
in "/etc/2tmgr.config" file
--create Create DLN. Requires three obligatory action-specific options:
--create -o <oid> -g <gid> -m <octal mode>
-o, --owner Owner id of DLN to be created
-g, --group Group id of DLN
-m, --mode Octal 3-digit access mode of DLN
--load Load DLN from Capacity tier to Fast tier
--unload Send DLN from Fast tier to Capacity tier
[ec2-user@ip-10-4-1-103 ~]$ sudo /mnt/orangefs/bin/2tmgr --configure
Syncer server IP address: 10.4.3.17
Syncer server RPC program number [200000001]:
Timeout for Syncer server response (in ms) [4000]: 15000
The following configuration was provided:
Syncer IP address: 10.4.3.17
Syncer server RPC program number: 200000001
Timeout for Syncer server response (in ms): 15000
Save to disk (/etc/2tmgr.config)? [yes]: yes
[ec2-user@ip-10-4-1-103 ~]$ sudo /mnt/orangefs/bin/2tmgr --create -o 1000 -g 1000 -m 777 E
MCW16
[ec2-user@ip-10-4-1-103 ~]$ ls -la /mnt/orangefs/EMCW16/
total 8
drwxrwxrwx. 1 ec2-user ec2-user 4096 Apr 14 14:02 .
drwxrwxrwt. 1 root root 4096 Apr 14 14:02 ..
[ec2-user@ip-10-4-1-103 ~]$ echo "Hello EMC World" /mnt/orangefs/EMCW16/Hello
Hello EMC World /mnt/orangefs/EMCW16/Hello
[ec2-user@ip-10-4-1-103 ~]$

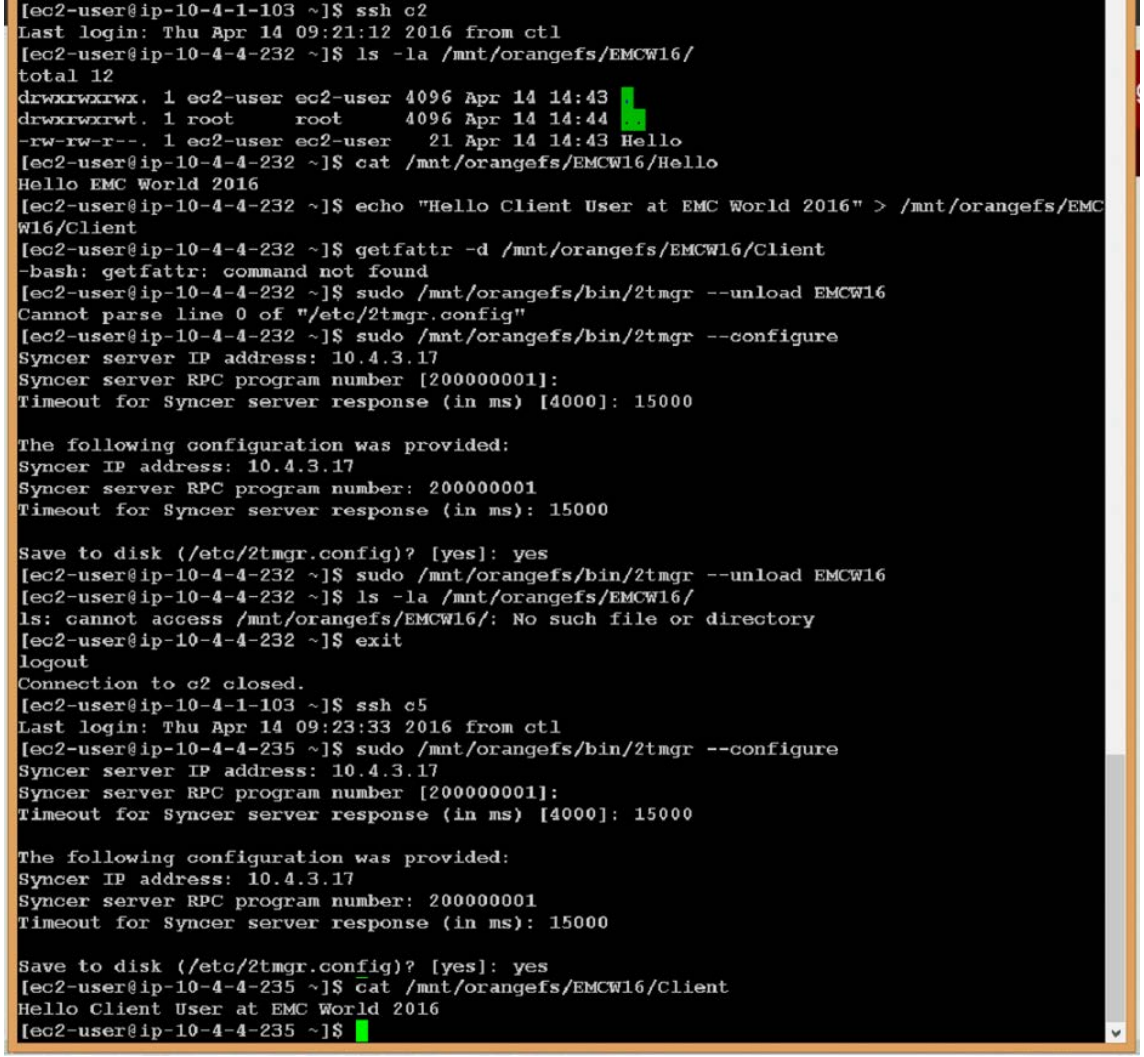
```
[ec2-user@ip-10-4-1-103 ~]$ echo "Hello EMC World" > /mnt/orangefs/EMCW16/Hello
[ec2-user@ip-10-4-1-103 ~]$ cat /mnt/orangefs/EMCW16/Hello
Hello EMC World
[ec2-user@ip-10-4-1-103 ~]$ getfattr -d /mnt/orangefs/EMCW16/Hello
getfattr: Removing leading '/' from absolute path names
# file: mnt/orangefs/EMCW16/Hello
user.two_tiers.ct_ver="1460658075"
user.two_tiers.oid="9adf0f57-2fffffffffffd743-0106d6f6"

[ec2-user@ip-10-4-1-103 ~]$ ls -la /mnt/orangefs/EMCW16/Hello
-rw-rw-r--. 1 ec2-user ec2-user 16 Apr 14 14:21 /mnt/orangefs/EMCW16/Hello
[ec2-user@ip-10-4-1-103 ~]$ s3 list
                          Bucket                                    Created
------------------------------------------------------------  --------------------
cc.2t.testbucket4.1                                         2016-04-14T07:46:01Z
cc.bucket2t1                                                2016-03-18T22:45:51Z
cc.bucket2twest11                                           2016-04-08T02:40:04Z
cc.bucket2twest2                                            2016-04-08T21:44:30Z
cc.dln.store                                                2016-02-09T00:21:35Z
cc.testbucket9.1                                            2016-04-12T19:56:05Z
cf-templates-1s9guled17jiw-ap-northeast-1                   2016-04-12T06:03:34Z
cf-templates-1s9guled17jiw-eu-central-1                     2016-04-13T08:46:54Z
cf-templates-1s9guled17jiw-eu-west-1                        2016-04-13T22:44:03Z
cf-templates-1s9guled17jiw-us-east-1                        2016-01-13T21:44:16Z
cf-templates-1s9guled17jiw-us-west-1                        2016-01-14T12:49:52Z
cf-templates-1s9guled17jiw-us-west-2                        2016-01-22T17:46:42Z
nevolinbucket                                               2016-03-17T16:56:28Z
[ec2-user@ip-10-4-1-103 ~]$ s3 list cc.bucket2t1 |grep 9adf0f57-2fffffffffffd743-0106d6f6
9adf0f57-2fffffffffffd743-0106d6f6                    2016-04-14T22:21:33Z     16
[ec2-user@ip-10-4-1-103 ~]$ sudo /mnt/orangefs/bin/2tmgr --unload EMCW16
[ec2-user@ip-10-4-1-103 ~]$ ls -la /mnt/orangefs/EMCW16/
ls: cannot access /mnt/orangefs/EMCW16/: No such file or directory
[ec2-user@ip-10-4-1-103 ~]$ sudo /mnt/orangefs/bin/2tmgr --load EMCW16
[ec2-user@ip-10-4-1-103 ~]$ getfattr -d /mnt/orangefs/EMCW16
getfattr: Removing leading '/' from absolute path names
# file: mnt/orangefs/EMCW16
user.two_tiers.dln_ref_count="1"

[ec2-user@ip-10-4-1-103 ~]$ ls -la /mnt/orangefs/EMCW16/
total 12
drwxrwxrwx. 1 ec2-user ec2-user 4096 Apr 14 14:29 .
drwxrwxrwt. 1 root     root     4096 Apr 14 14:29 ..
-rw-rw-r--. 1 ec2-user ec2-user   16 Apr 14 14:29 Hello
[ec2-user@ip-10-4-1-103 ~]$
```

EMC²


[ec2-user@ip-10-4-1-103 ~]$ cat /mnt/orangefs/EMCW16/Hello
cat: /mnt/orangefs/EMCW16/Hello: Invalid argument
[ec2-user@ip-10-4-1-103 ~]$ cat /mnt/orangefs/EMCW16/Hello
Hello EMC World
[ec2-user@ip-10-4-1-103 ~]$ getfattr -d /mnt/orangefs/EMCW16/Hello
getfattr: Removing leading '/' from absolute path names
# file: mnt/orangefs/EMCW16/Hello
user.two_tiers.ct_ver="1460658557"
user.two_tiers.oid="9adf0f57-2fffffffffffd743-0106d6f6"
[ec2-user@ip-10-4-1-103 ~]$ echo "Hello EMC World 2016" > /mnt/orangefs/EMCW16/Hello
[ec2-user@ip-10-4-1-103 ~]$ sudo /mnt/orangefs/bin/2tmgr --^Cad EMCW16
[ec2-user@ip-10-4-1-103 ~]$ getfattr -d /mnt/orangefs/EMCW16/Hello
getfattr: Removing leading '/' from absolute path names
# file: mnt/orangefs/EMCW16/Hello
user.two_tiers.ct_ver="1460659272"
user.two_tiers.oid="47e40f57-2fffffffffffd2d6-0106d6f6"
[ec2-user@ip-10-4-1-103 ~]$ sudo /mnt/orangefs/bin/2tmgr --unload EMCW16
[ec2-user@ip-10-4-1-103 ~]$ ls -la /mnt/orangefs/EMCW16/
ls: cannot access /mnt/orangefs/EMCW16/: No such file or directory
[ec2-user@ip-10-4-1-103 ~]$ s3 list cc.bucket2t1 |egrep "47e40f57-2fffffffffffd2d6-0106d6f
6|9adf0f57-2fffffffffffd743-0106d6f6"
47e40f57-2fffffffffffd2d6-0106d6f6 2016-04-14T22:41:28Z 21
9adf0f57-2fffffffffffd743-0106d6f6 2016-04-14T22:21:33Z 16
[ec2-user@ip-10-4-1-103 ~]$ cat /mnt/orangefs/EMCW16/Hello
cat: /mnt/orangefs/EMCW16/Hello: No such file or directory
[ec2-user@ip-10-4-1-103 ~]$ cat /mnt/orangefs/EMCW16/Hello
cat: /mnt/orangefs/EMCW16/Hello: No such file or directory
[ec2-user@ip-10-4-1-103 ~]$ sudo /mnt/orangefs/bin/2tmgr --load EMCW16
[ec2-user@ip-10-4-1-103 ~]$ cat /mnt/orangefs/EMCW16/Hello
Hello EMC World 2016
[ec2-user@ip-10-4-1-103 ~]$


[ec2-user@ip-10-4-1-103 ~]$ ssh c2
Last login: Thu Apr 14 09:21:12 2016 from ctl
[ec2-user@ip-10-4-4-232 ~]$ ls -la /mnt/orangefs/EMCW16/
total 12
drwxrwxrwx. 1 ec2-user ec2-user 4096 Apr 14 14:43 .
drwxrwxrwt. 1 root root 4096 Apr 14 14:44 ..
-rw-rw-r--. 1 ec2-user ec2-user 21 Apr 14 14:43 Hello
[ec2-user@ip-10-4-4-232 ~]$ cat /mnt/orangefs/EMCW16/Hello
Hello EMC World 2016
[ec2-user@ip-10-4-4-232 ~]$ echo "Hello Client User at EMC World 2016" > /mnt/orangefs/EMC
W16/Client
[ec2-user@ip-10-4-4-232 ~]$ getfattr -d /mnt/orangefs/EMCW16/Client
-bash: getfattr: command not found
[ec2-user@ip-10-4-4-232 ~]$ sudo /mnt/orangefs/bin/2tmgr --unload EMCW16
Cannot parse line 0 of "/etc/2tmgr.config"
[ec2-user@ip-10-4-4-232 ~]$ sudo /mnt/orangefs/bin/2tmgr --configure
Syncer server IP address: 10.4.3.17
Syncer server RPC program number [200000001]:
Timeout for Syncer server response (in ms) [4000]: 15000
The following configuration was provided:
Syncer IP address: 10.4.3.17
Syncer server RPC program number: 200000001
Timeout for Syncer server response (in ms): 15000
Save to disk (/etc/2tmgr.config)? [yes]: yes
[ec2-user@ip-10-4-4-232 ~]$ sudo /mnt/orangefs/bin/2tmgr --unload EMCW16
[ec2-user@ip-10-4-4-232 ~]$ ls -la /mnt/orangefs/EMCW16/
ls: cannot access /mnt/orangefs/EMCW16/: No such file or directory
[ec2-user@ip-10-4-4-232 ~]$ exit
logout
Connection to c2 closed.
[ec2-user@ip-10-4-1-103 ~]$ ssh c5
Last login: Thu Apr 14 09:23:33 2016 from ctl
[ec2-user@ip-10-4-4-235 ~]$ sudo /mnt/orangefs/bin/2tmgr --configure
Syncer server IP address: 10.4.3.17
Syncer server RPC program number [200000001]:
Timeout for Syncer server response (in ms) [4000]: 15000
The following configuration was provided:
Syncer IP address: 10.4.3.17
Syncer server RPC program number: 200000001
Timeout for Syncer server response (in ms): 15000
Save to disk (/etc/2tmgr.config)? [yes]: yes
[ec2-user@ip-10-4-4-235 ~]$ cat /mnt/orangefs/EMCW16/Client
Hello Client User at EMC World 2016
[ec2-user@ip-10-4-4-235 ~]$


EMC²
®