Building a High Availability NFS Server

Mentors: Michael Gilbert, David Fox, Martin Baltezore, Jason Shortino

August 11, 2021

Arshita Sandhiparthi
Emily Ramirez-Serrano

LLNL-PRES-825698

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

Team Members

Arshita Sandhiparthi

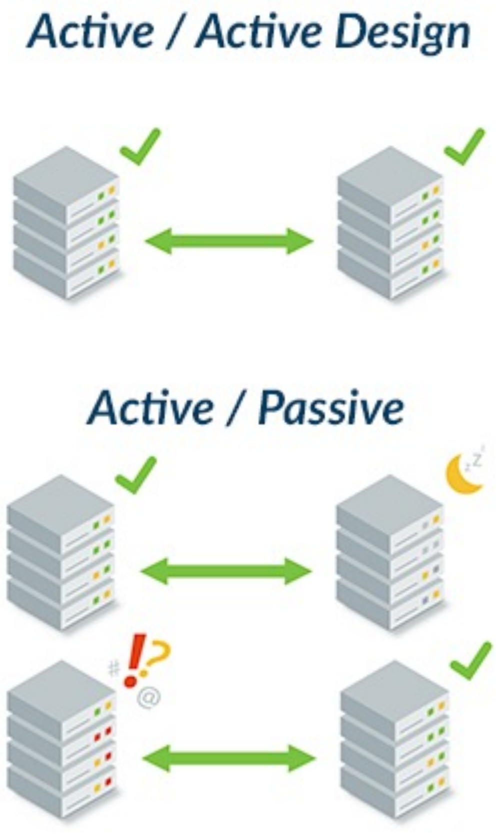
University of the Pacific
Political Science & Computer Science
Graduating Spring 2022

Emily Ramirez Serrano

Northern Arizona University
Computer Science
Graduating Spring 2022

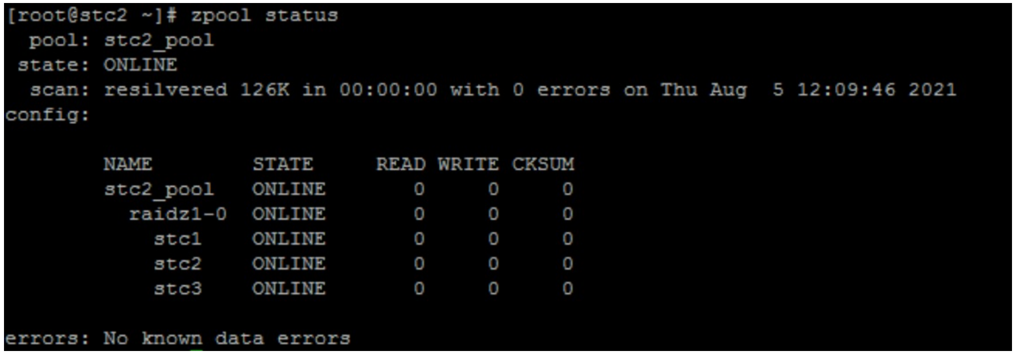
High Availability (HA)

- Why HA?
  - — Continuous operation
  - — Reliable protection
  - — Automatic failover procedures in outages or node failure
- The Biggest Use Case
  - — The Lustre file system
- Problem
  - — Don't have a system set up to failover NFS on mgmt nodes
  - — Need to explore CentOS

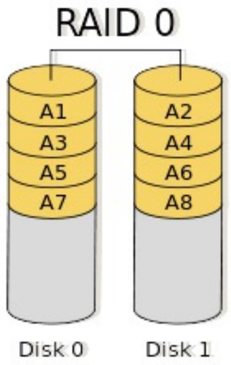


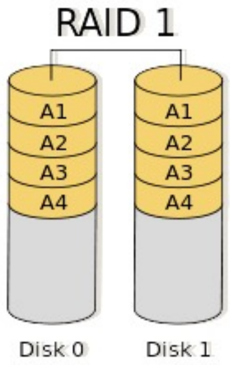
ZFS

- ZFS
  - zpools
  - RAIDz1
  - multihost
- SAN Arrays
  - Storage Area Network
  - Logical Unit Numbers (LUN)
  - Multipath



```
[root@stc2 ~]# zpool status
  pool: stc2_pool
 state: ONLINE
  scan: resilvered 126K in 00:00:00 with 0 errors on Thu Aug  5 12:09:46 2021
config:

        NAME          STATE     READ WRITE CKSUM
        stc2_pool     ONLINE       0     0     0
          raidz1-0    ONLINE       0     0     0
            stc1      ONLINE       0     0     0
            stc2      ONLINE       0     0     0
            stc3      ONLINE       0     0     0

errors: No known data errors
```





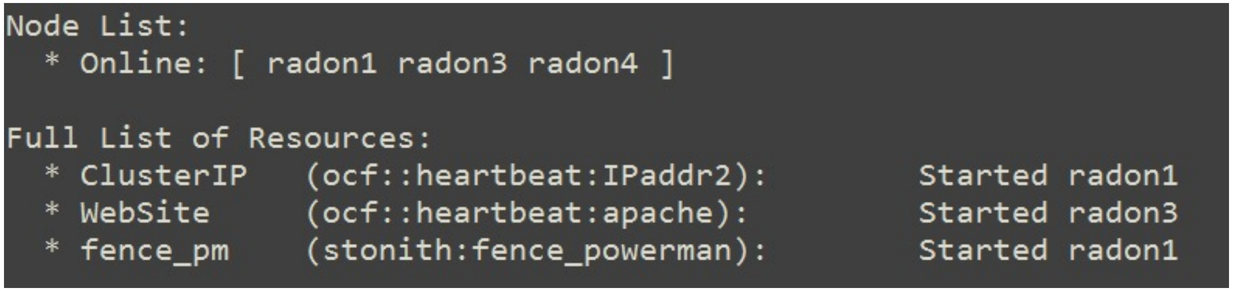
openzfs.github.io/openzfs-docs

Pacemaker

- Pacemaker
  — HA Resource Manager software

- Fencing and Shoot The Other Node In The Head (STONITH)
  — Powerman
  — Small Computer System Interface (SCSI)

- Safely manage resources across the system



```
Node List:
  * Online: [ radon1 radon3 radon4 ]

Full List of Resources:
  * ClusterIP   (ocf::heartbeat:IPaddr2):        Started radon1
  * WebSite     (ocf::heartbeat:apache):         Started radon3
  * fence_pm    (stonith:fence_powerman):        Started radon1
```

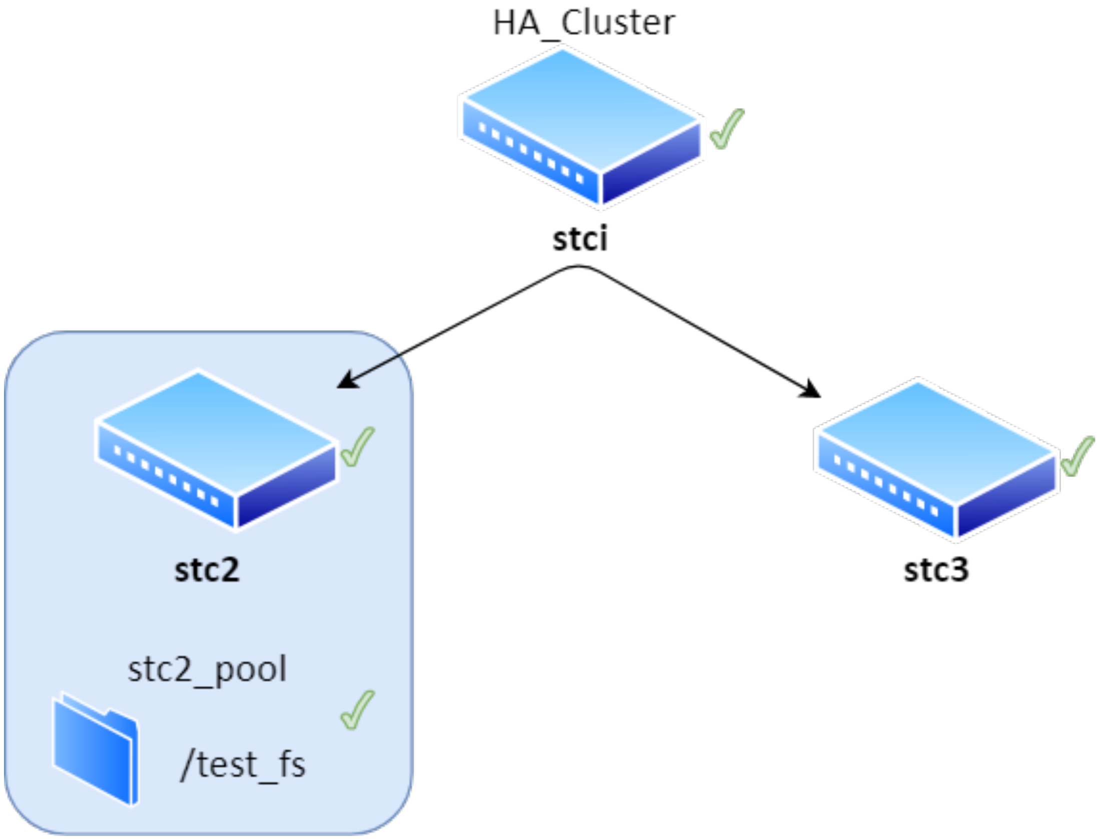
clusterlabs.org/pacemaker

Project Accomplishments

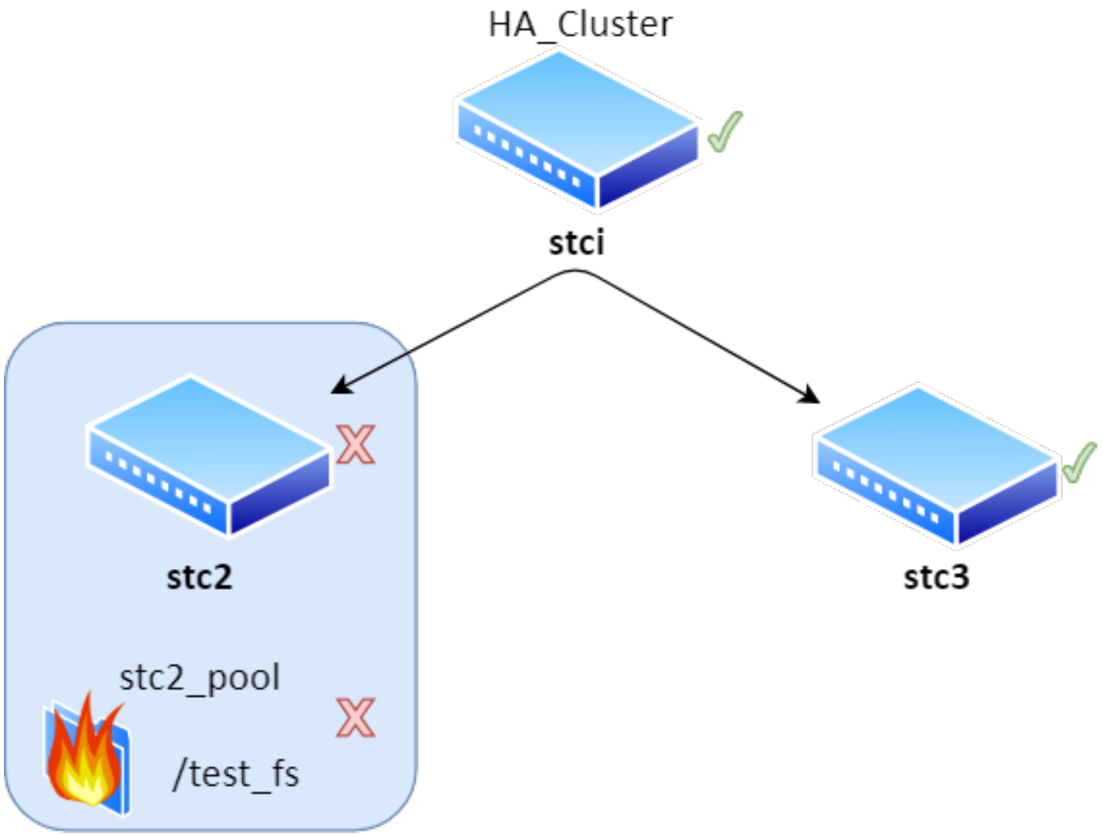
Integrating NFS With ZFS

- Goal: Setup pacemaker to support a HA setup and manage ZFS and NFS resource migration.
- Configuring Pacemaker and ZFS
  - Migrating resources
    - Importing/Exporting ZFS pools
    - Floating IP
  - Using multipath devices
- NFS on top of ZFS
  - ZFS pools are already widely used at the lab but not with NFS

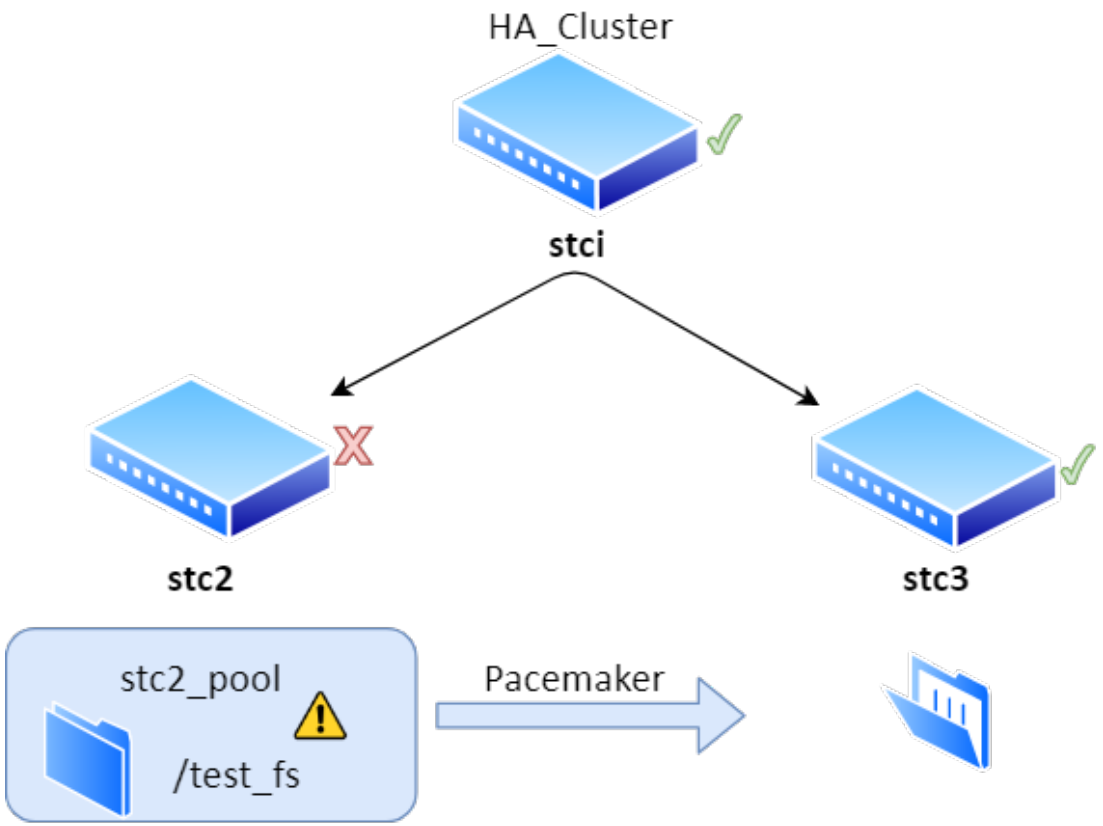
Integrating NFS With ZFS



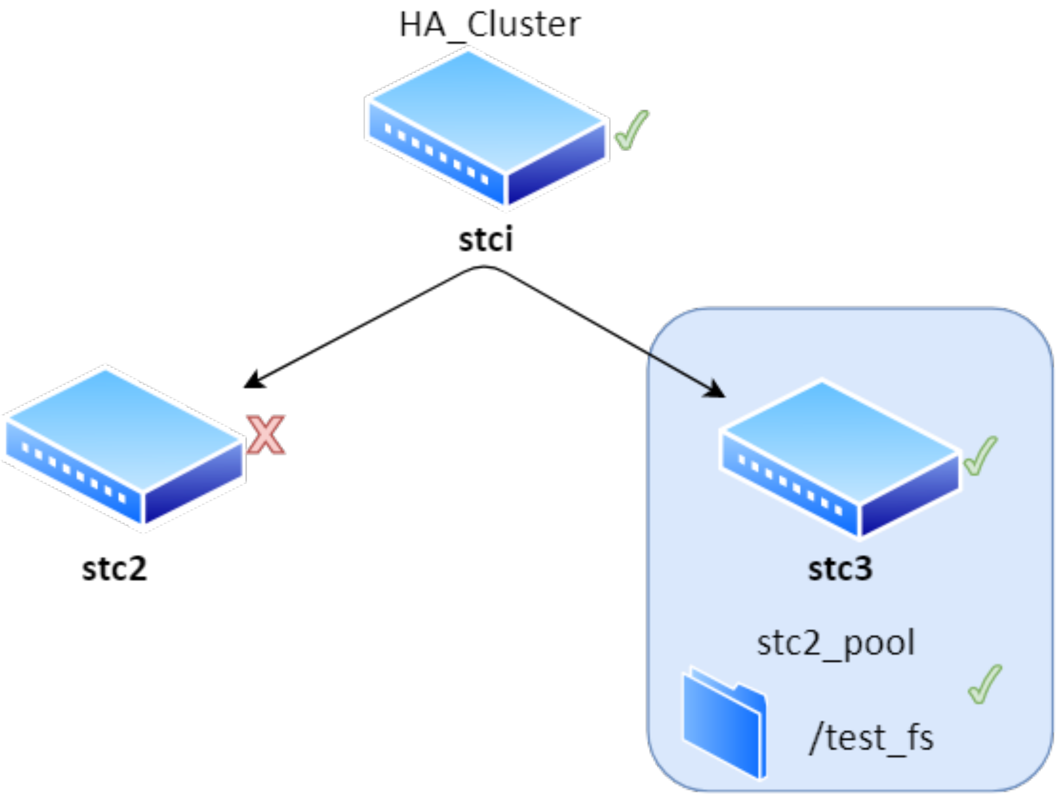
Integrating NFS With ZFS



Integrating NFS With ZFS

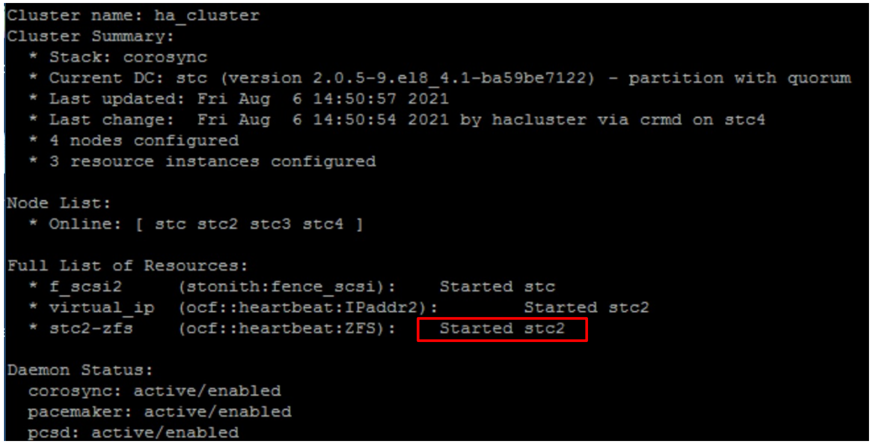


Integrating NFS With ZFS

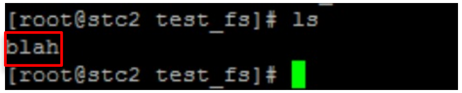


Integrating NFS With ZFS

Before fencing stc2

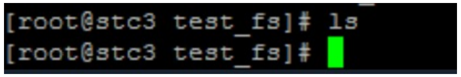


```
Cluster name: ha_cluster
Cluster Summary:
  * Stack: corosync
  * Current DC: stc (version 2.0.5-9.el8_4.1-ba59be7122) - partition with quorum
  * Last updated: Fri Aug  6 14:50:57 2021
  * Last change:  Fri Aug  6 14:50:54 2021 by hacluster via crmd on stc4
  * 4 nodes configured
  * 3 resource instances configured

Node List:
  * Online: [ stc stc2 stc3 stc4 ]

Full List of Resources:
  * f_scsi2     (stonith:fence_scsi):    Started stc
  * virtual_ip  (ocf::heartbeat:IPaddr2):        Started stc2
  * stc2-zfs    (ocf::heartbeat:ZFS):    Started stc2

Daemon Status:
  corosync: active/enabled
  pacemaker: active/enabled
  pcsd: active/enabled
```



```
[root@stc2 test_fs]# ls
blah
[root@stc2 test_fs]#
```

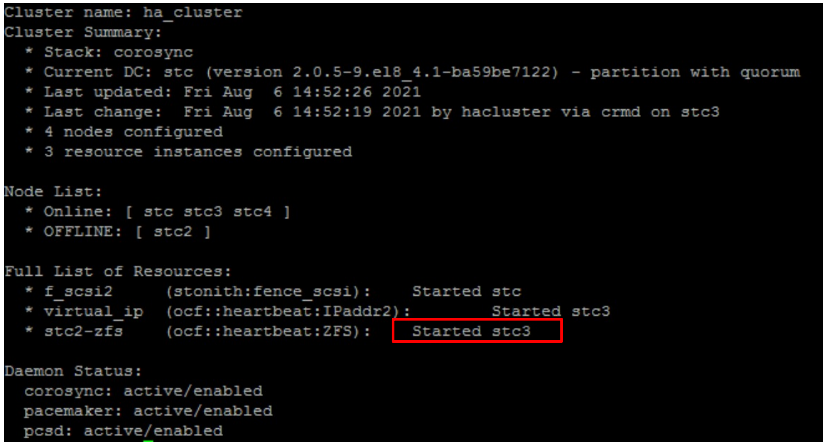
On stc2



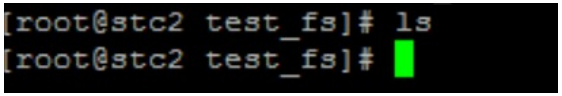
```
[root@stc3 test_fs]# ls
[root@stc3 test_fs]#
```

On stc3

After fencing stc2

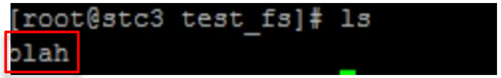


```
Cluster name: ha_cluster
Cluster Summary:
  * Stack: corosync
  * Current DC: stc (version 2.0.5-9.el8_4.1-ba59be7122) - partition with quorum
  * Last updated: Fri Aug  6 14:52:26 2021
  * Last change:  Fri Aug  6 14:52:19 2021 by hacluster via crmd on stc3
  * 4 nodes configured
  * 3 resource instances configured

Node List:
  * Online: [ stc stc3 stc4 ]
  * OFFLINE: [ stc2 ]

Full List of Resources:
  * f_scsi2     (stonith:fence_scsi):    Started stc
  * virtual_ip  (ocf::heartbeat:IPaddr2):        Started stc3
  * stc2-zfs    (ocf::heartbeat:ZFS):    Started stc3

Daemon Status:
  corosync: active/enabled
  pacemaker: active/enabled
  pcsd: active/enabled
```



```
[root@stc2 test_fs]# ls
[root@stc2 test_fs]#
```

On stc2



```
[root@stc3 test_fs]# ls
blah
```

On stc3

Challenges

- CentOs8 Compatibility
  - Fencing agents (powerman)
    - Custom fencing resource
    - Too simplistic for ZFS management
- Pacemaker and ZFS
  - Importing and Exporting ZFS pools
  - SCSI Fencing
  - ZFS set up took a lot of time
- Lack of Documentation
  - Had to dig around for a lot of information

Future Work and High End Goals

- Migrate ZFS pool and NFS servers across management nodes
- High availability between multiple management nodes

References

- https://github.com/ewwhite/zfs-ha/wiki
- https://openzfs.github.io/openzfs-docs/Project%20and%20Community/index.html
- https://www.clusterlabs.org/pacemaker/doc/2.1/Clusters_from_Scratch/singlehtml/
- https://books.clusterapps.com/books/deployments/page/nfs-on-zfs-ha-cluster
- https://docs.oracle.com/cd/E19253-01/819-5461/gayog/index.html
- https://wiki.lustre.org/Creating_Pacemaker_Resources_for_Lustre_Storage_Services

Lawrence Livermore
National Laboratory

Disclaimer
This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.