Los Alamos
NATIONAL LABORATORY

Automated AI-Analysis of the Lustre-Development Mailing List
(and TASSI)

John Bent, Dominic Manno

May 2024

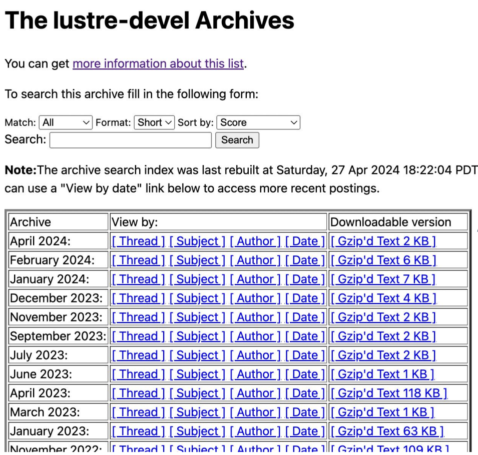
LA-UR-23-33506

Suspected Motivation: Current Challenges with HPC Storage Systems

- Very long release cycles
- Extremely long resolution periods
- Lack of diagnostic tools in deployed systems
- Onerous requirements of arcane expertise

Biggest challenges in distributed storage systems?

Biggest challenges in distributed storage systems?



The lustre-devel Archives

You can get more information about this list.

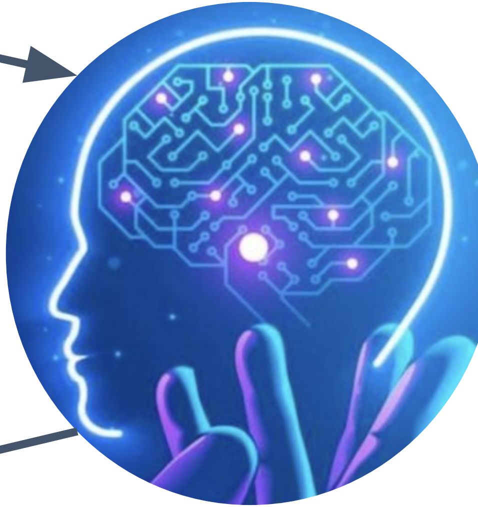
To search this archive fill in the following form:

Match: All Format: Short Sort by: Score
Search: [ ] Search

Note: The archive search index was last rebuilt at Saturday, 27 Apr 2024 18:22:04 PDT
can use a "View by date" link below to access more recent postings.

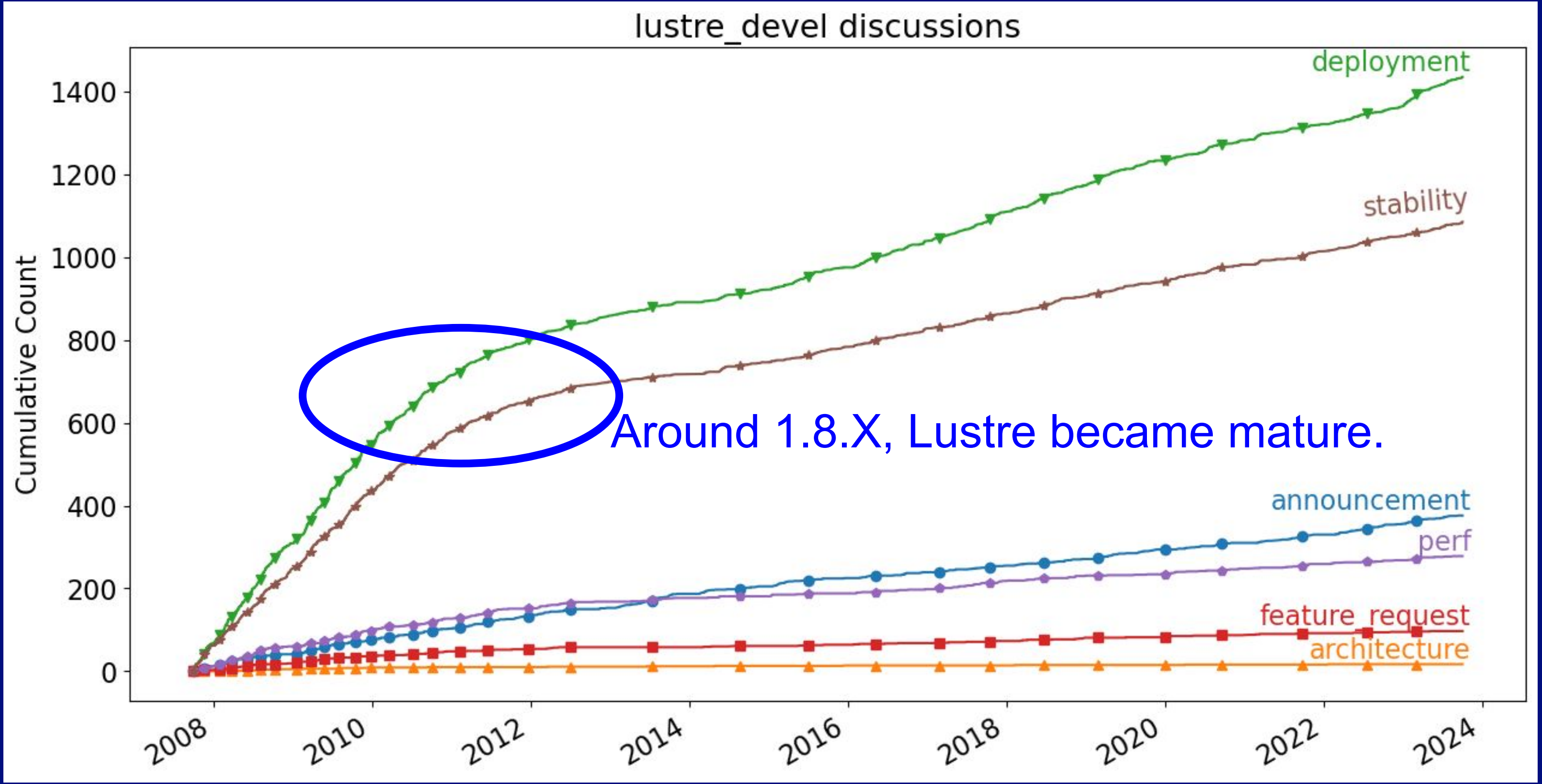
Archive	View by:	Downloadable version
April 2024:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 2 KB]
February 2024:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 6 KB]
January 2024:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 7 KB]
December 2023:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 4 KB]
November 2023:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 2 KB]
September 2023:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 2 KB]
July 2023:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 2 KB]
June 2023:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 1 KB]
April 2023:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 118 KB]
March 2023:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 1 KB]
January 2023:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 63 KB]
November 2022:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 109 KB]

foreach thread:
  what is topic?
  was it answered?



[analysis]

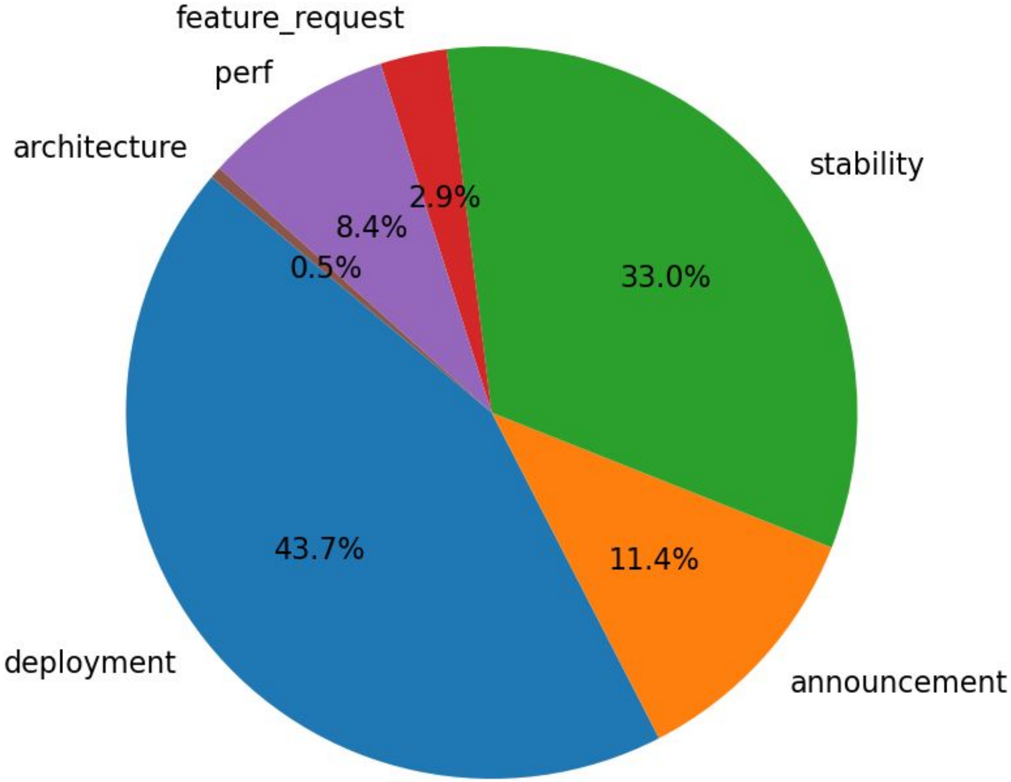
Automated AI-Analysis of Lustre-Devel



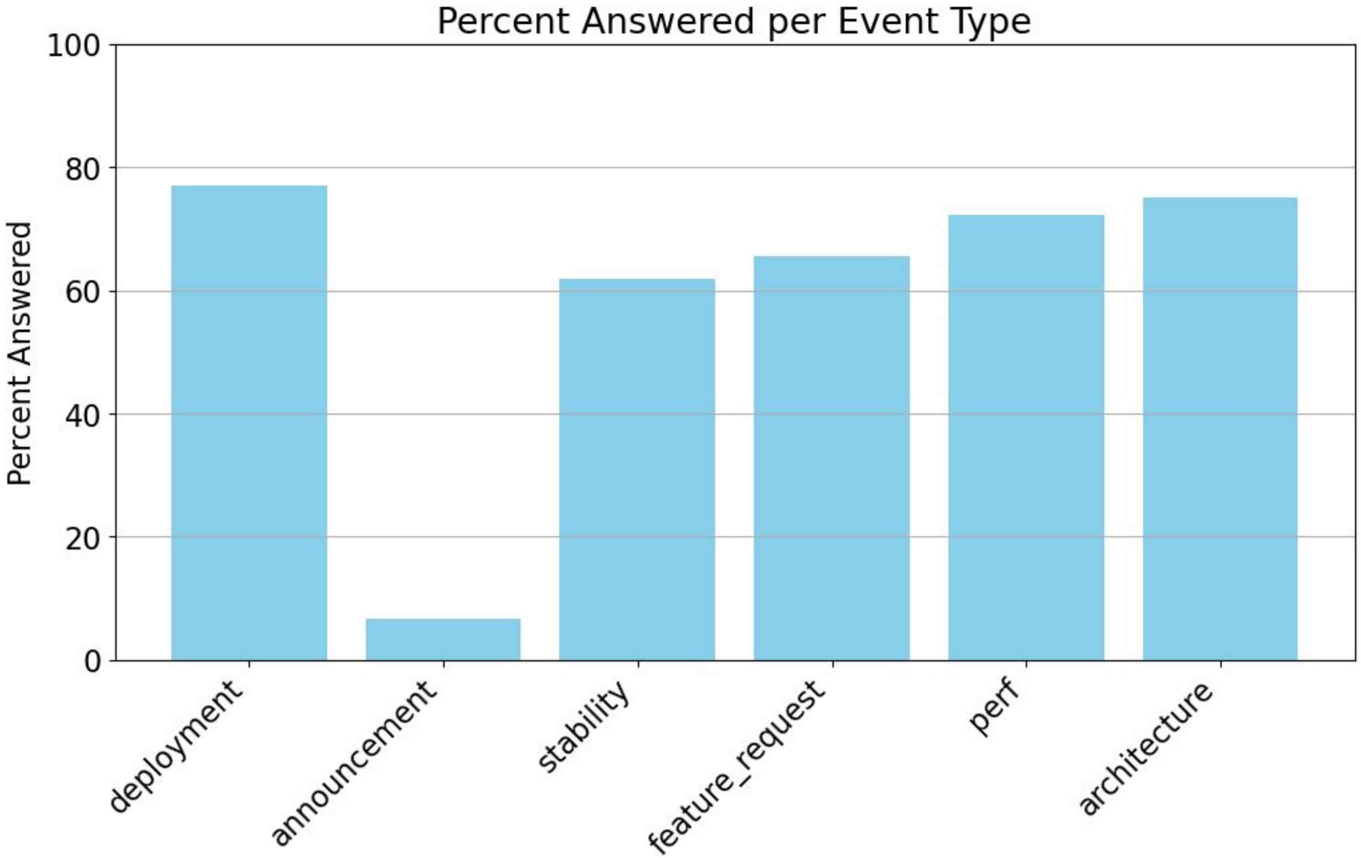
Around 1.8.X, Lustre became mature.

Los Alamos NATIONAL LABORATORY

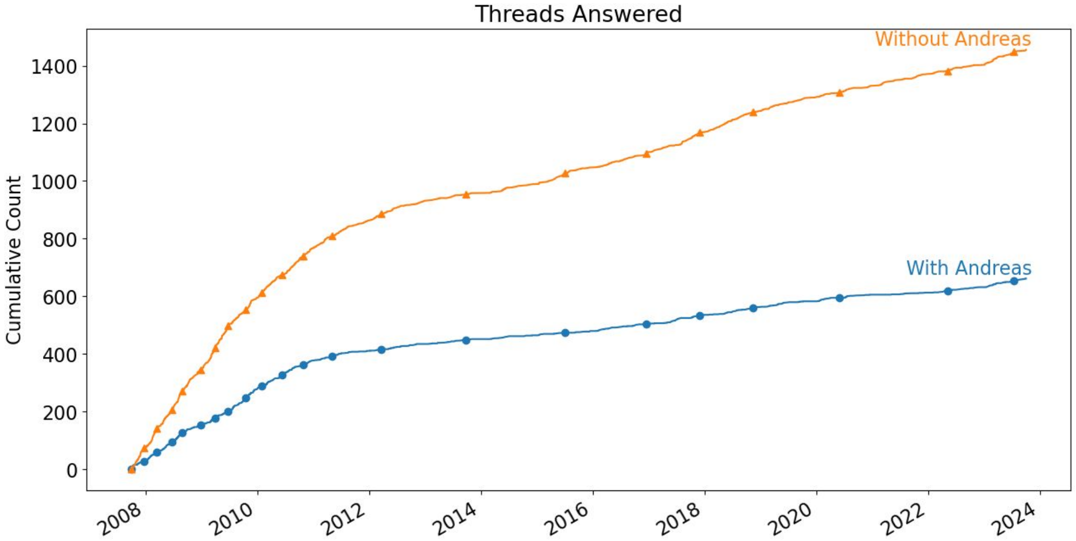
For Those of Us Who Like Pie

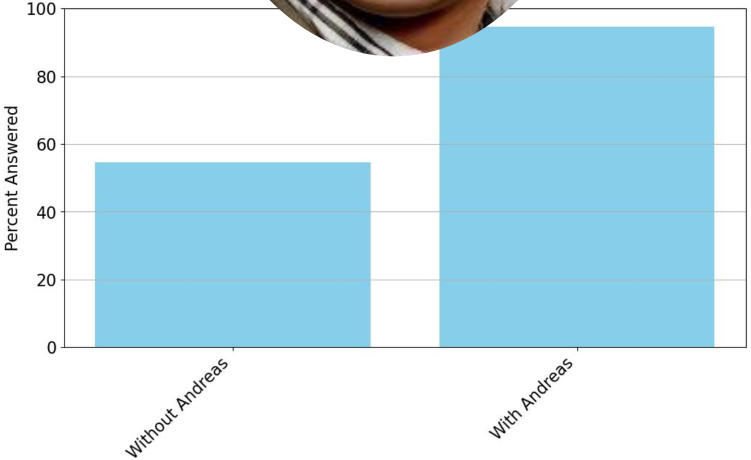


For Those of Us Who Like Bars



For Those of Us Who Like Andreas





Summary of Mailing List Analysis

This is not to bash Lustre!

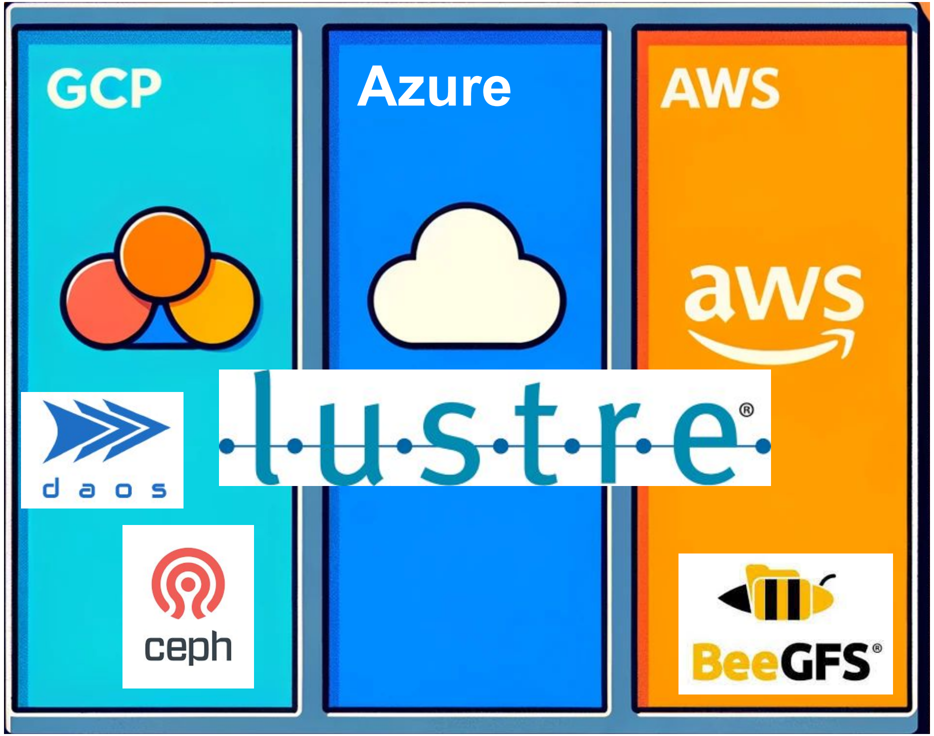
Lustre is great!

I want to expand the analysis for more systems.

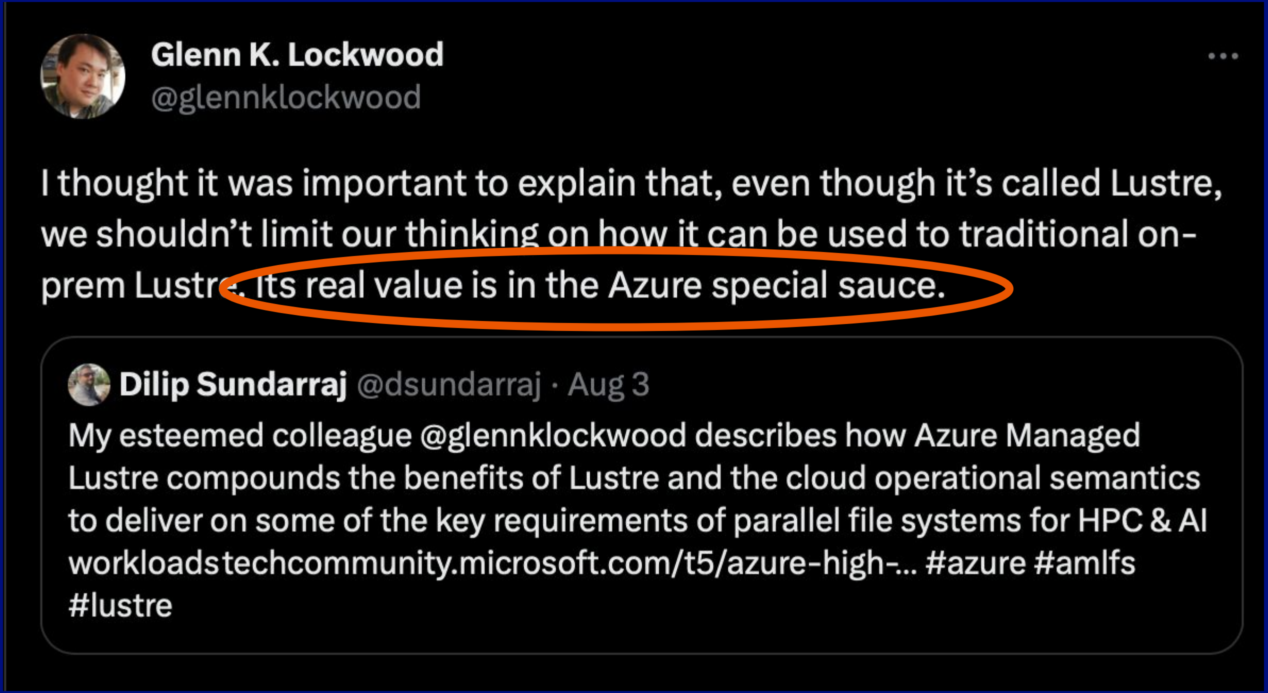
Configuration and deployment are challenging.

Why Isn't This a Solved Problem?

- Armies of CSP developers
- Armies of Enterprise developers
  - DDN, HPE, IBM, MinIO, Panasas, Pure, VAST, Weka



The Armies Seem to be Producing Closed Solutions



Glenn K. Lockwood
@glennklockwood

I thought it was important to explain that, even though it's called Lustre, we shouldn't limit our thinking on how it can be used to traditional on-prem Lustre. Its real value is in the Azure special sauce.

Dilip Sundarraj @dsundarraj · Aug 3

My esteemed colleague @glennklockwood describes how Azure Managed Lustre compounds the benefits of Lustre and the cloud operational semantics to deliver on some of the key requirements of parallel file systems for HPC & AI workloadstechcommunity.microsoft.com/t5/azure-high-... #azure #amlfs #lustre

The full article is a quick informative read: LINK

How Can We Improve the Lives of Current Admins?

TASSI: Tool for Agile Scalable Storage Infrastructure

Automation for configuration, deployment, testing of distributed storage

“I find a bug in version V. I have a reproducer R. Someone submits patch P.”

"How can I simply spin up a small system
running version V, apply patch P, and test with R?”

– Thomas Bertschinger, LANL

TASSI: Basic Workflow

TASSI

- Control agent

Admin

- Human in control

Git Repo

- Store configs, tests, outputs

Lustre.org and ZFS

- Software repos

Virtual Cluster

- Connected set of VMs

libvirt

- VM management software

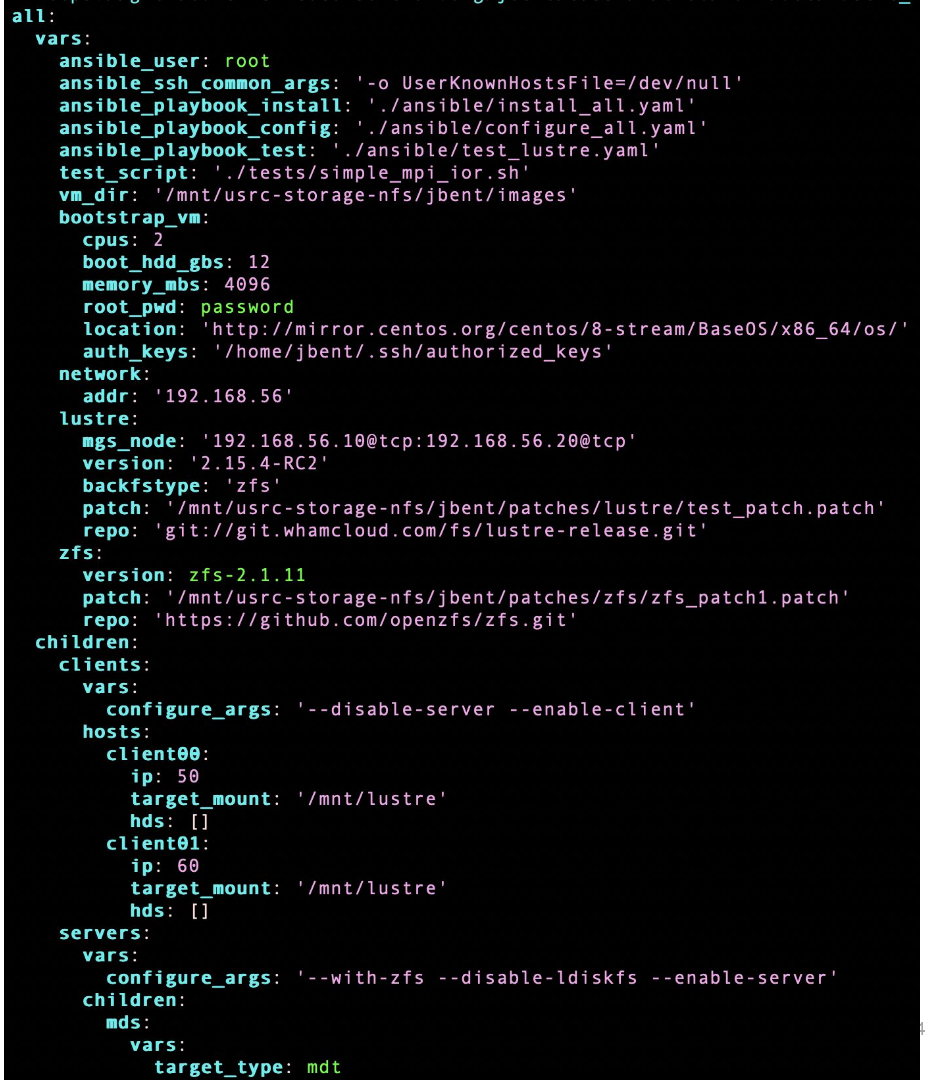
NFS

- Stash precreated images

TASSI
Admin
Git repo
Lustre.org
ZFS
Virtual Cluster
libvirt
NFS

TASSI Current Actual Config File

Currently a bit obfuscated



```yaml
all:
  vars:
    ansible_user: root
    ansible_ssh_common_args: '-o UserKnownHostsFile=/dev/null'
    ansible_playbook_install: './ansible/install_all.yaml'
    ansible_playbook_config: './ansible/configure_all.yaml'
    ansible_playbook_test: './ansible/test_lustre.yaml'
    test_script: './tests/simple_mpi_ior.sh'
    vm_dir: '/mnt/usrc-storage-nfs/jbent/images'
    bootstrap_vm:
      cpus: 2
      boot_hdd_gbs: 12
      memory_mbs: 4096
      root_pwd: password
      location: 'http://mirror.centos.org/centos/8-stream/BaseOS/x86_64/os/'
      auth_keys: '/home/jbent/.ssh/authorized_keys'
    network:
      addr: '192.168.56'
    lustre:
      mgs_node: '192.168.56.10@tcp:192.168.56.20@tcp'
      version: '2.15.4-RC2'
      backfstype: 'zfs'
      patch: '/mnt/usrc-storage-nfs/jbent/patches/lustre/test_patch.patch'
      repo: 'git://git.whamcloud.com/fs/lustre-release.git'
    zfs:
      version: zfs-2.1.11
      patch: '/mnt/usrc-storage-nfs/jbent/patches/zfs/zfs_patch1.patch'
      repo: 'https://github.com/openzfs/zfs.git'
  children:
    clients:
      vars:
        configure_args: '--disable-server --enable-client'
      hosts:
        client00:
          ip: 50
          target_mount: '/mnt/lustre'
          hds: []
        client01:
          ip: 60
          target_mount: '/mnt/lustre'
          hds: []
    servers:
      vars:
        configure_args: '--with-zfs --disable-ldiskfs --enable-server'
      children:
        mds:
          vars:
            target_type: mdt
```

TASSI Config File Essential Elements

Host OS

- Base image (e.g. CentOS 8)
- RAM

Lustre Software

- Repo location (e.g. Whamcloud)
- Version / tag
- Patches
- Backend (e.g. ZFS)

ZFS Software

- Repo location
- Version / tag
- Patches

Cluster

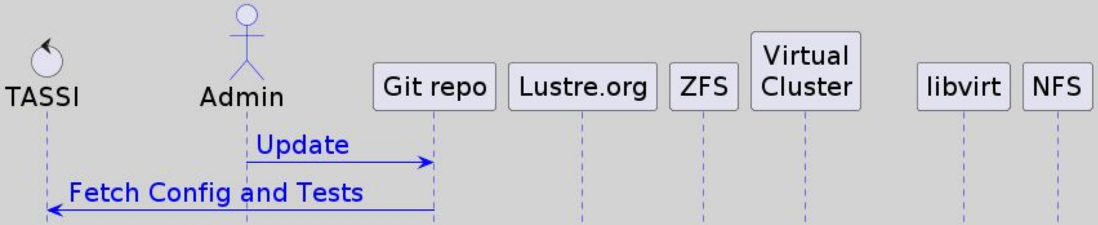
- Num clients, MDS, OSS, MDT, OST
- HDD sizes for each target

Test Script

TASSI: Basic Workflow

Step One

- Admin updates a config



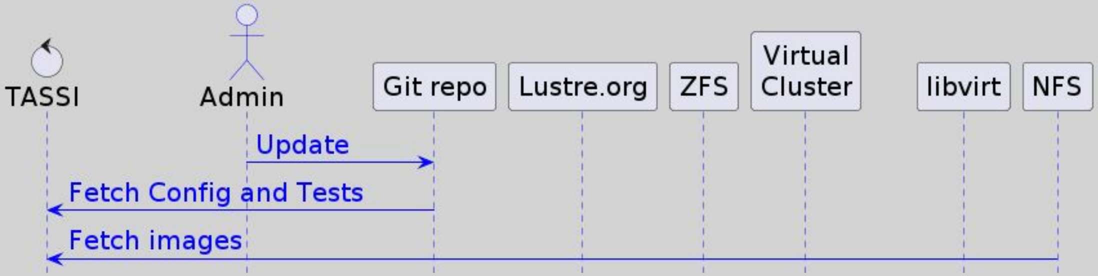
TASSI: Basic Workflow

Step One

- Admin updates a config

Step Two

- Fetch precreated images



TASSI: Basic Workflow

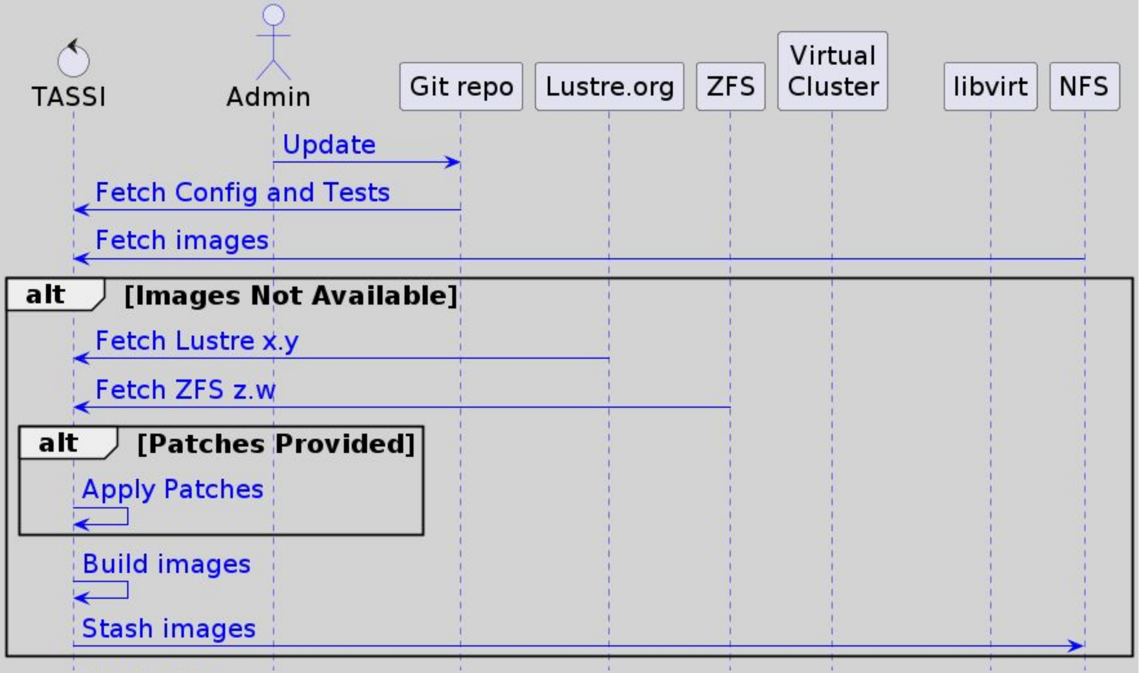
Step One

- Admin updates a config

Step Two

- Fetch precreated images
- Build if not precreated
  - Apply any specified patches

(not shown, spin up initial bootstrap VM)



TASSI: Basic Workflow

Step One

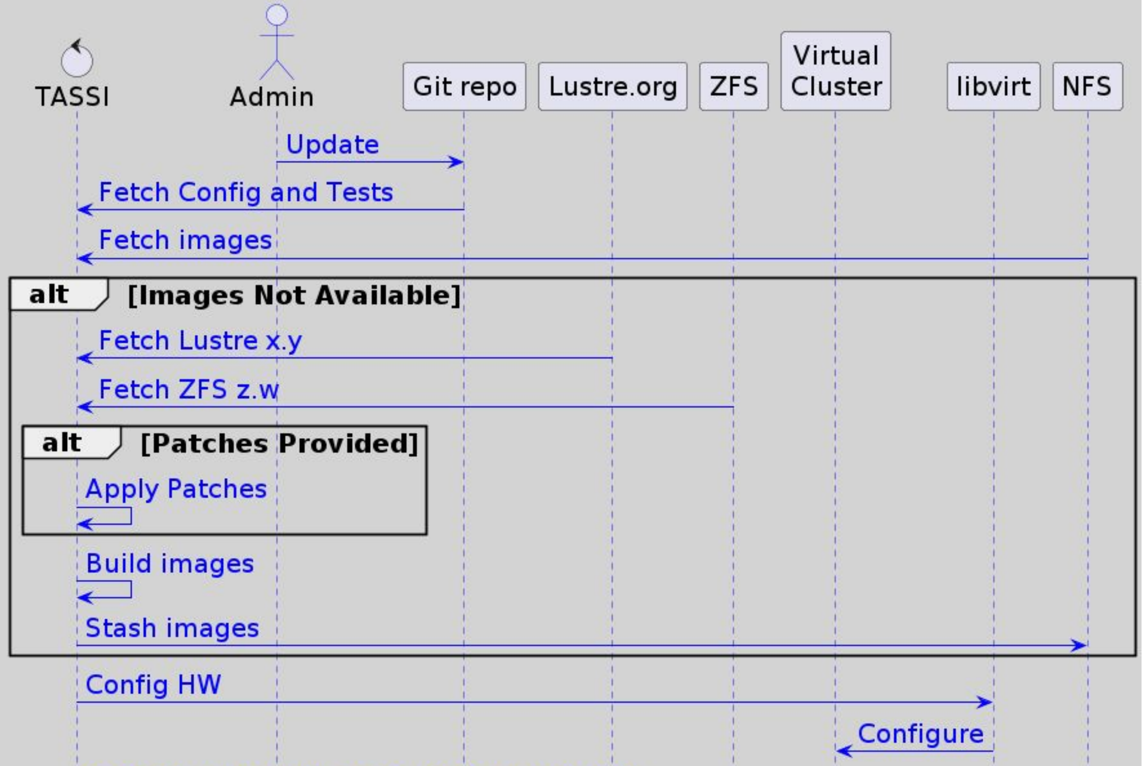
- Admin updates a config

Step Two

- Fetch precreated images
- Build if not precreated
  - Apply any specified patches

Step Three

- Setup the “physical” cluster



TASSI: Basic Workflow

Step One

- Admin updates a config

Step Two

- Fetch precreated images
- Build if not precreated
  - Apply any specified patches

Step Three

- Setup the “physical” cluster

Step Four

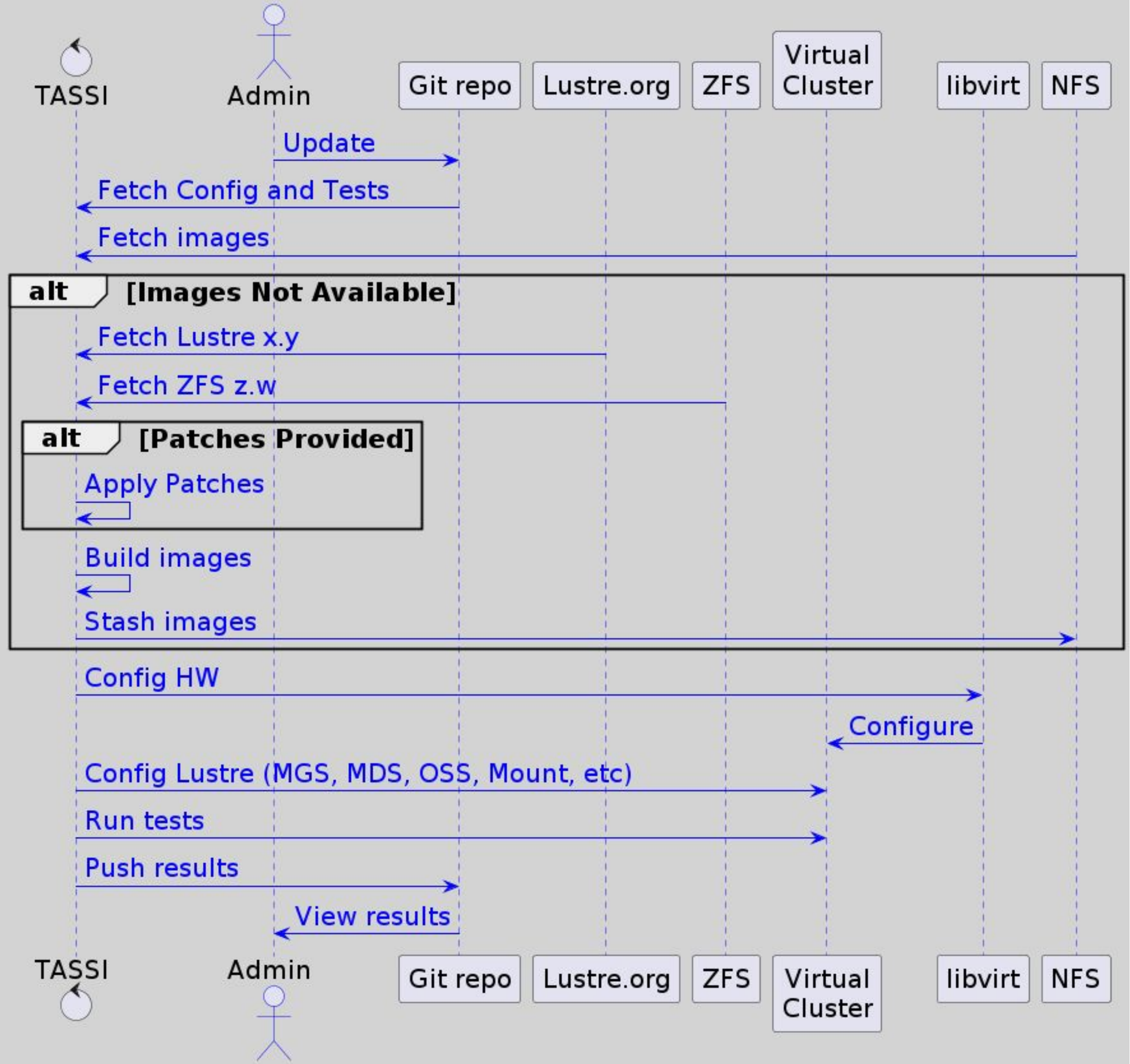
- Setup Lustre

Step Five

- Run the specified tests

Step Six

- Commit the outputs



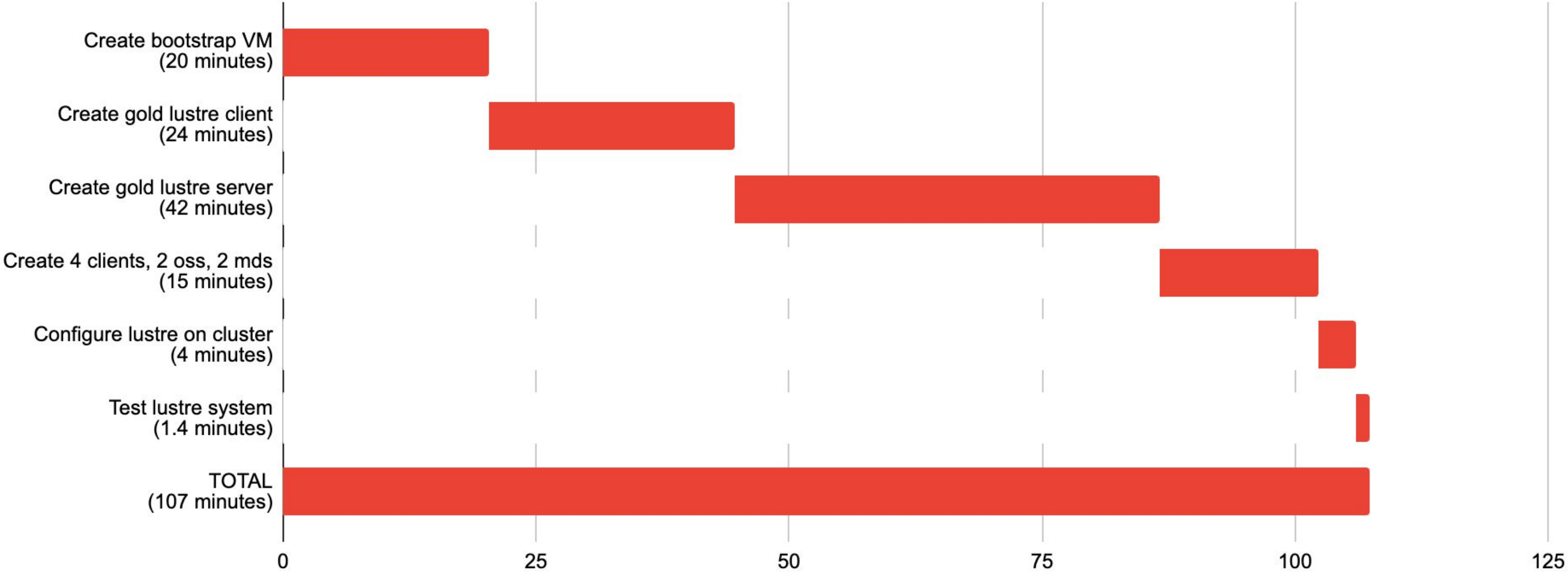
Los Alamos NATIONAL LABORATORY

TASSI Tools Used

- Control agent
  - Python
- Virtual Machine Management
  - libvirt/KVM/Qemu
- Lustre and ZFS installation
  - Ansible
- Config repository
  - Git
- Image stashing
  - NFS

TASSI Timings - 4 Clients - No Gold: 107 minutes

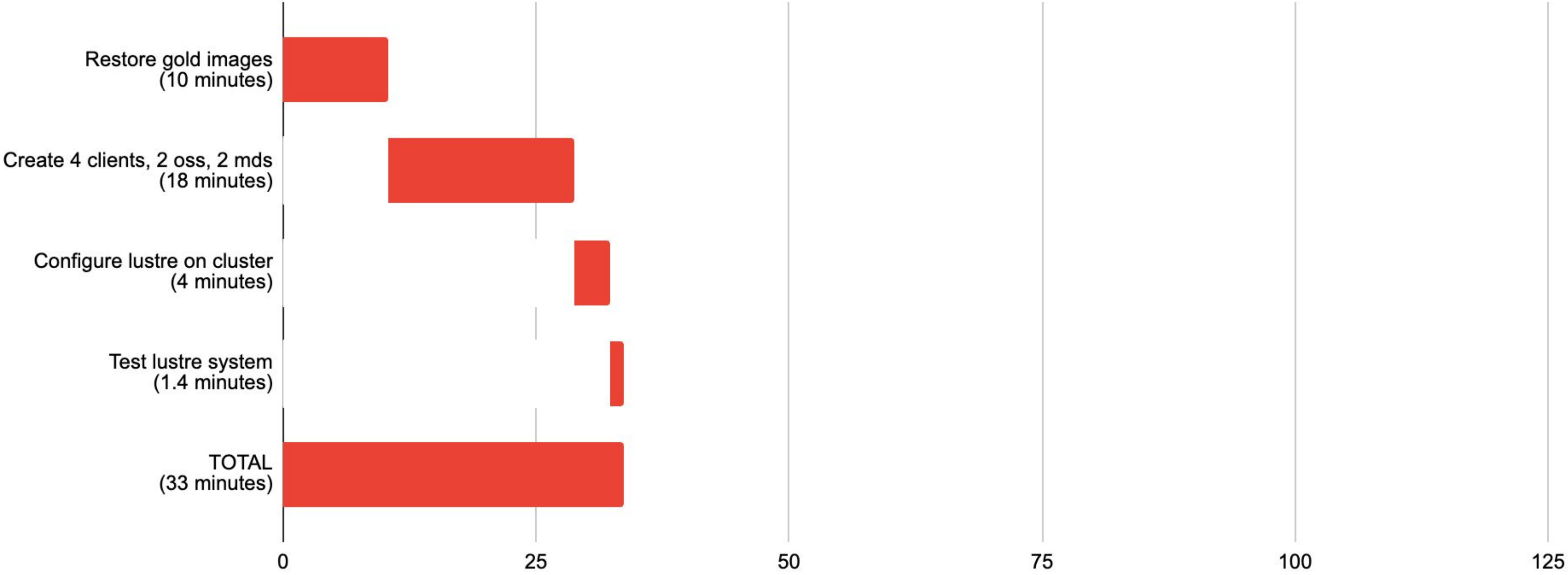
2 MDS, 2 OSS



Los Alamos NATIONAL LABORATORY

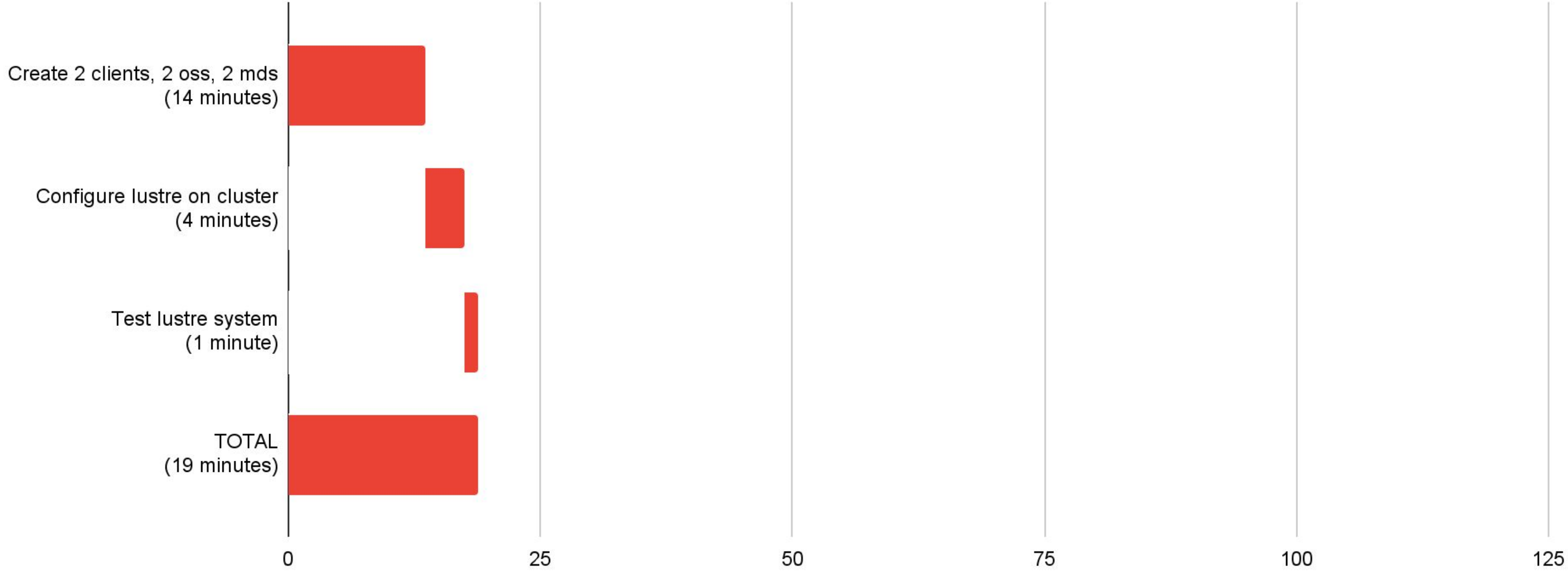
TASSI Timings - 4 Clients - Cold Gold: 33 minutes

2 MDS, 2 OSS



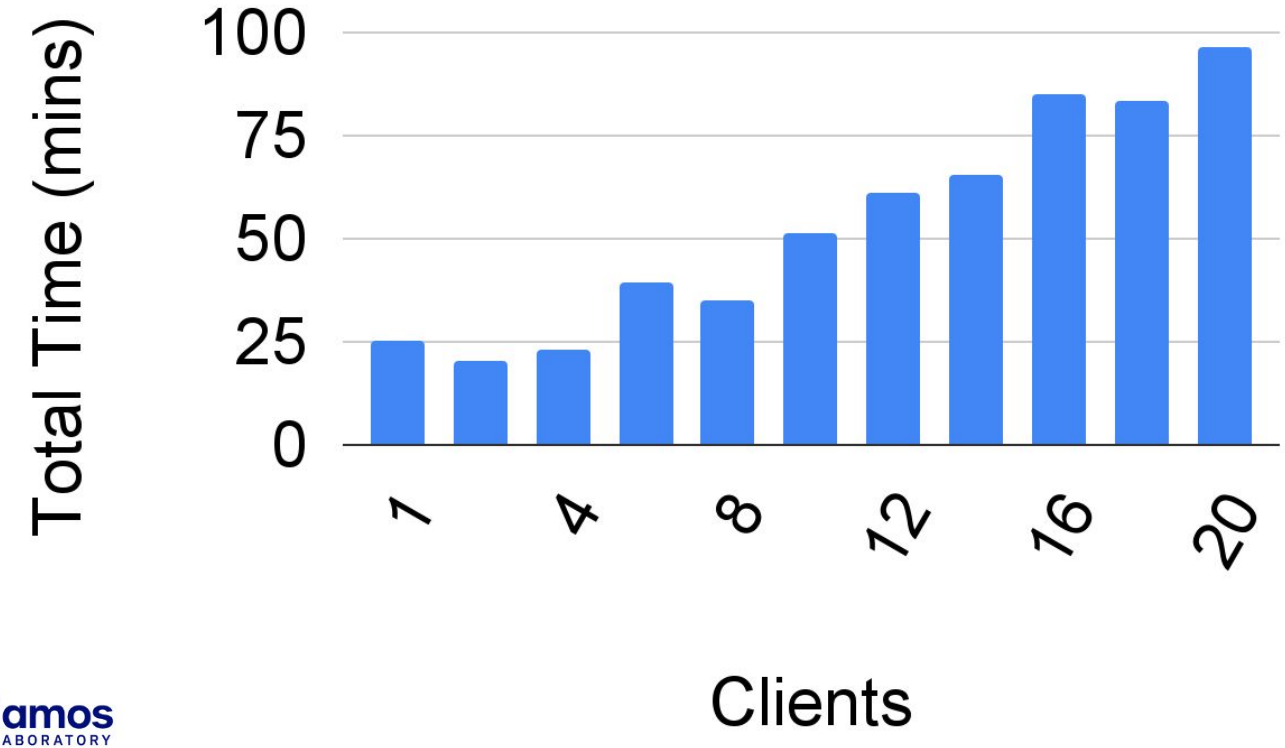
TASSI Timings - 4 Clients - Warm Gold: 19 minutes

2 MDS, 2 OSS



TASSI Timings: Cold Gold Client Scaling

2 MDS, 2 OSS



Los Alamos NATIONAL LABORATORY

TASSI Future Work

- Lustre specification
  - E.g. test with ldiskfs backend also
- Performance
  - E.g. parallelize node creation, use multiple physical nodes
- CI/CD
  - E.g. convert current git manual trigger to actual CI/CD mechanism
- Cluster deployment / configuration
  - Use CSPs, LANL OCHAMI, etc to dynamically provision/configure cluster
- File system installation, configuration
  - E.g. hardware accelerators, specify disk locations
- File system support
  - Add more (HammerSpace? BeeGFS? DAOS?)
- Storage system support
  - Add KV as opposed to just file system (KV-CSD? RocksDB?)
- Decomposability and disaggregation
  - Via NVMeoF for example

TASSI Conclusions

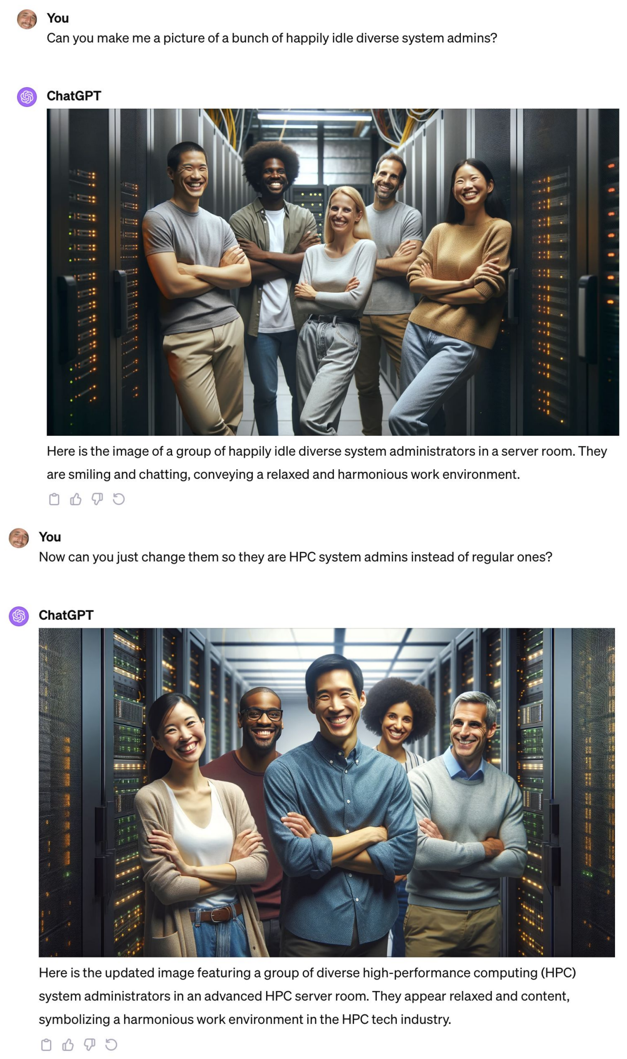
- Open source tools
- Flexible agile dynamic storage
- Enabling confident happy sys admins

Creating high performance, reliable storage systems for end-users

Email jbent@newmexicoconsortium.org and dmanno@lanl.gov for more info, feedback.

Or even to get involved!

Los Alamos
NATIONAL LABORATORY



You

Can you make me a picture of a bunch of happily idle diverse system admins?

ChatGPT

Here is the image of a group of happily idle diverse system administrators in a server room. They are smiling and chatting, conveying a relaxed and harmonious work environment.

You

Now can you just change them so they are HPC system admins instead of regular ones?

ChatGPT

Here is the updated image featuring a group of diverse high-performance computing (HPC) system administrators in an advanced HPC server room. They appear relaxed and content, symbolizing a harmonious work environment in the HPC tech industry.