

DisTRaX: Transient Ceph On RAM

Gabryel Mason-Williams

gabryel.mason-williams@rfi.ac.uk



Engineering and
Physical Sciences
Research Council

<https://github.com/rosalindfranklininstitute/DisTRaX>



The Rosalind
Franklin Institute

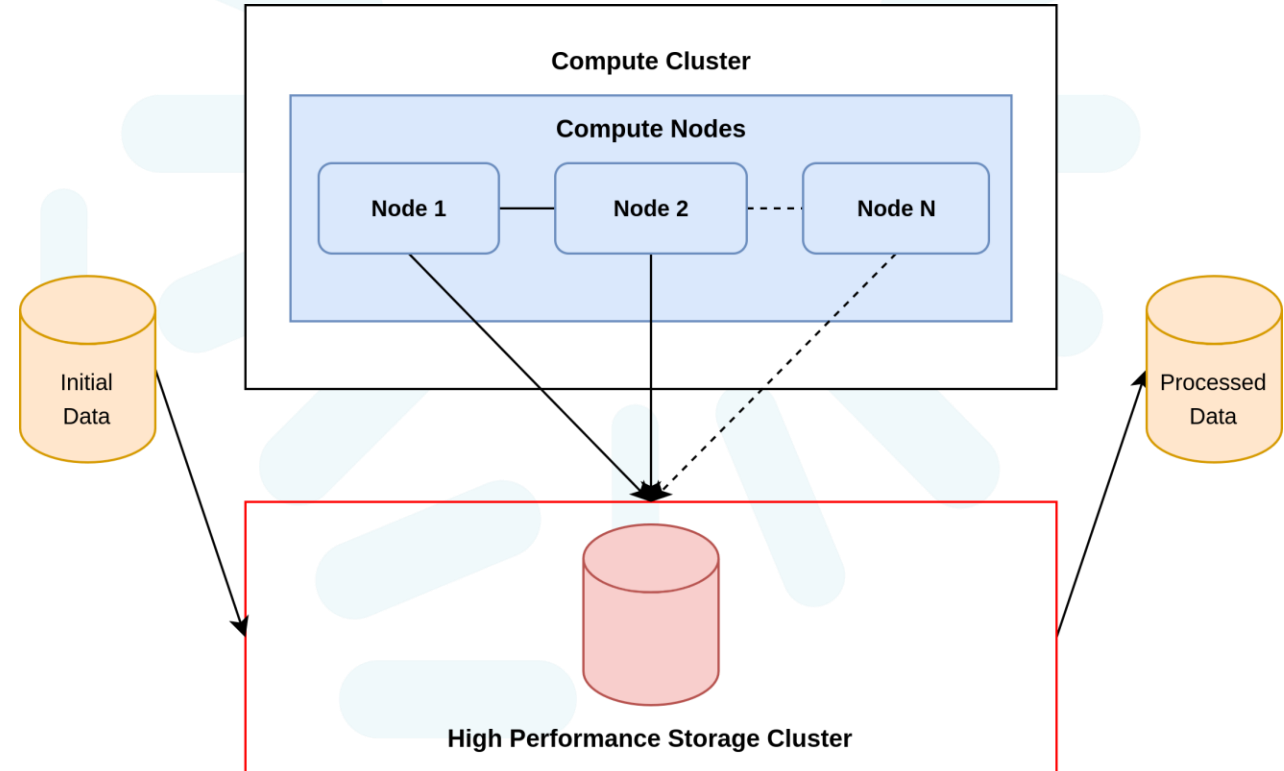
Contents

- Background
- The Problem
- What Is DisTRaX
- How DisTRaX Works
- How DisTRaX Solves the Problem
- Case Study
- Conclusion
- Questions

Background

High Performance Compute Cluster

- Job scheduler
- Compute - lots of RAM
- Storage
- Networking - Storage and Interconnect



Traditional High Performance Compute Cluster Setup

What Is the Problem



Engineering and
Physical Sciences
Research Council

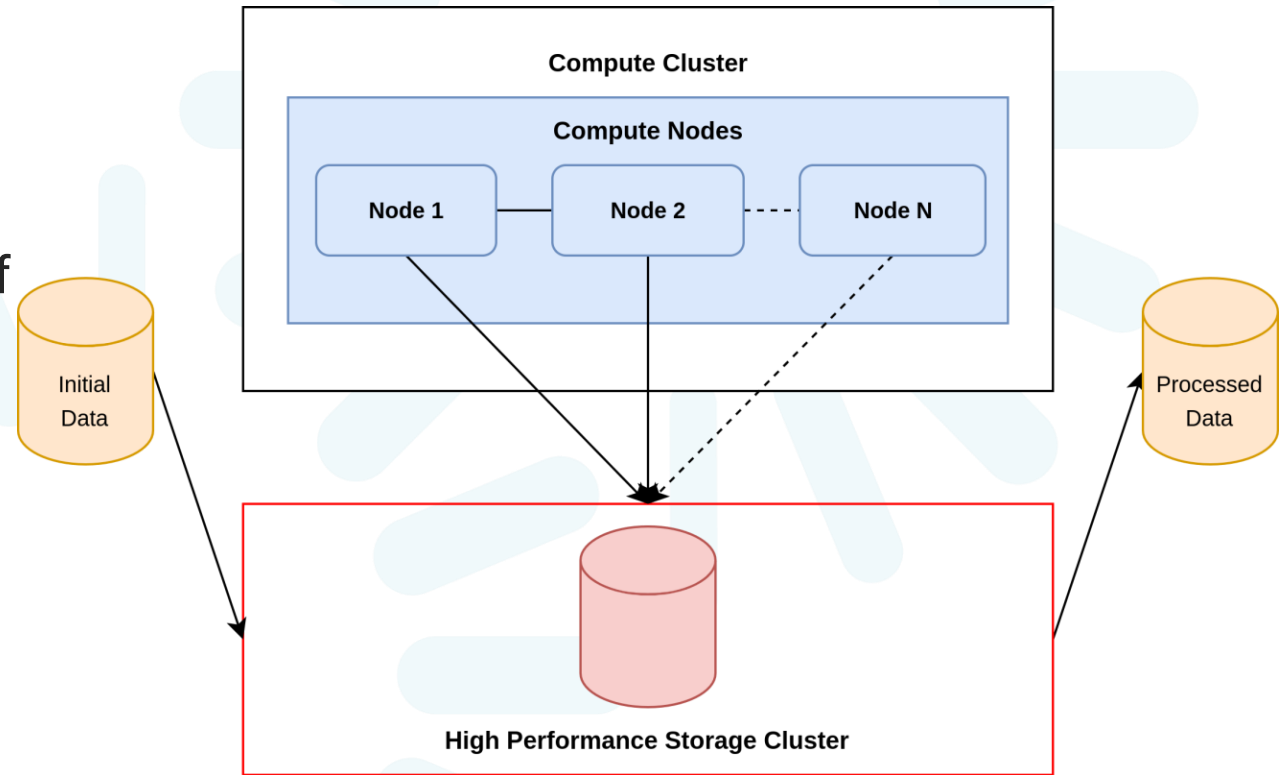
<https://github.com/rosalindfranklininstitute/DisTRaX>



The Rosalind
Franklin Institute

The Problem

- Network connection limits IO Bound Applications
- Shared Storage Resourced
 - Users' can affect the performance of others
- Storing of Intermediate Data
- Storage clusters are expensive, hard to maintain and set up, especially in cloud
- Inefficient use of resources



Traditional High Performance Compute Cluster Setup

Solution?



Engineering and
Physical Sciences
Research Council

<https://github.com/rosalindfranklininstitute/DisTRaX>



The Rosalind
Franklin Institute

DisTRaX

<https://github.com/rosalindfranklininstitute/DisTRaX>



Engineering and
Physical Sciences
Research Council

<https://github.com/rosalindfranklininstitute/DisTRaX>



The Rosalind
Franklin Institute

What Is DisTRaX?

- Distributed Transient Ram Storage (Ceph)
- A tool for deploying a transient storage (Ceph) cluster onto HPC infrastructure utilising RAM in a scalable and efficient manner.
- Creating a job persistent and isolated in-memory file/object store for HPC applications.

Why Another Deployment Tool

Current Deployment tools

- Designed to build long-lasting maintainable clusters
 - Lots of safety checks
 - Slow to deploy and remove clusters
- Sequential
- Require passwordless SSH

We need something quick and efficient

- Compute should be used for compute not setting up storage

How DisTRaX Works

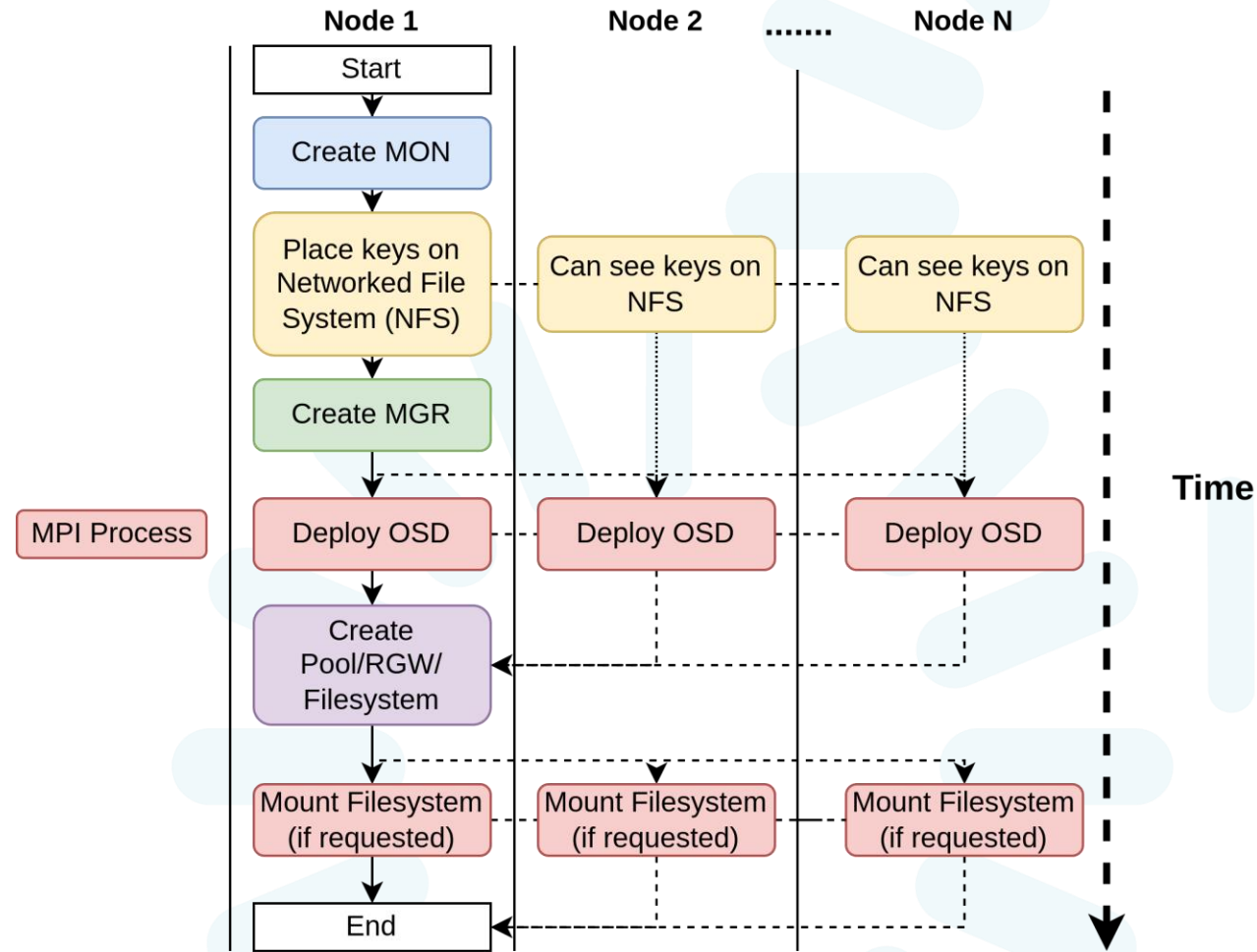


Engineering and
Physical Sciences
Research Council

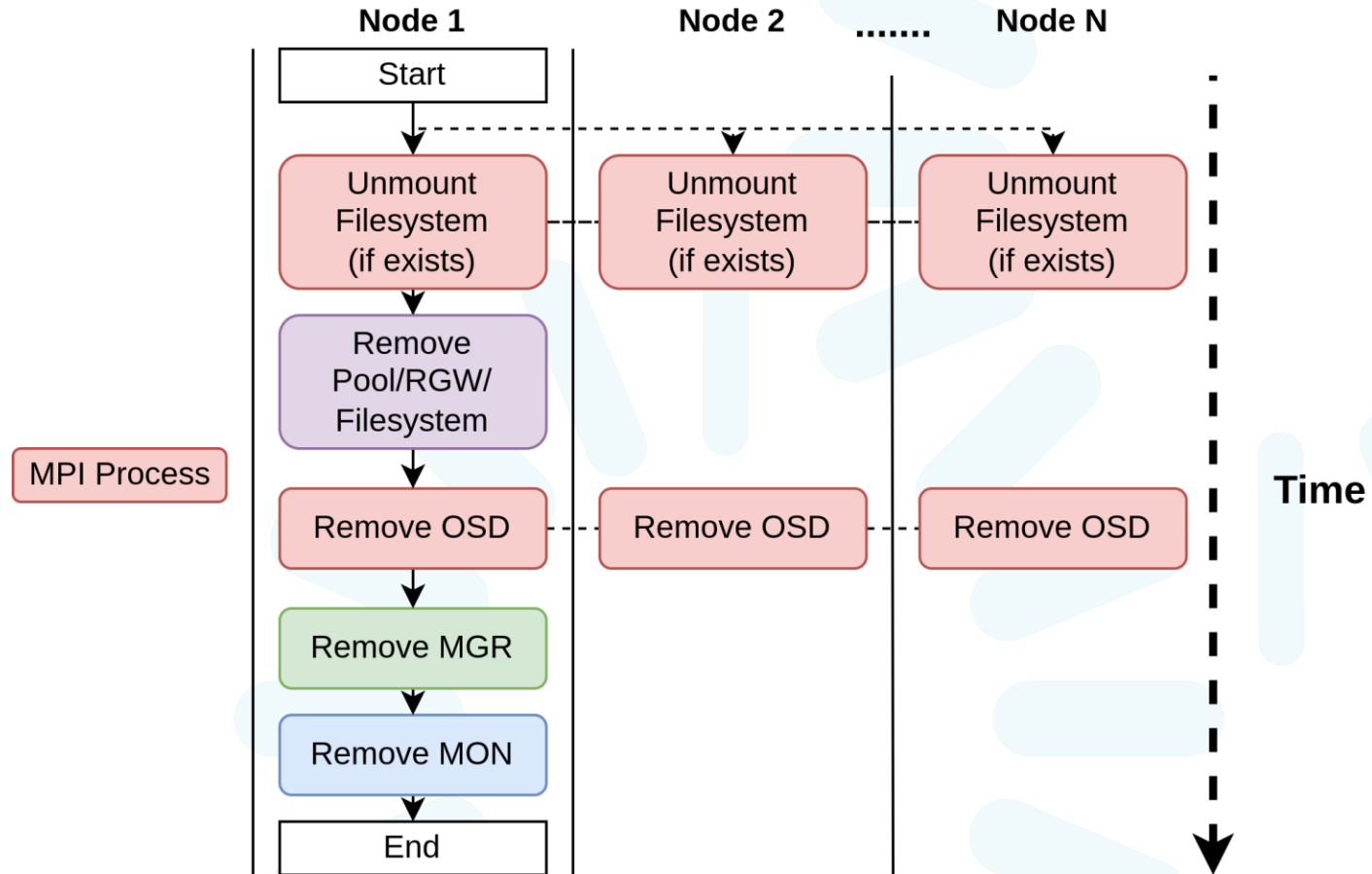
<https://github.com/rosalindfranklininstitute/DisTRaX>



DisTRaX Deployment



DisTRaX Removal



Example Deployment Script

```
[setup]
backend = ceph
folder = my_ceph
interface = eth0
number_of_hosts = 2
service = filesystem
```

```
[ram]
type = brd
number = 1
size_in_gb = 120
```

distrax.cfg

```
1  #!/bin/env bash
2  #SBATCH --job-name=distrax_example
3  #SBATCH --nodes=2
4  #SBATCH --ntasks-per-node=2
5  #SBATCH --cpus-per-task=1
6  #SBATCH --time=10:00
7  # Get hostnames
8  hostnames=`scontrol show hostnames | sed -n -e 'H;${x;s/\n/,/g;s/^,/,;p;}'`
9  # Get number of hosts
10 number_of_hosts=`scontrol show hostnames | wc -l`
11 # Create Storage System
12 mpirun -np $number_of_hosts --host $hostnames distrax_mpi -c distrax.cfg -a create
13 # Move data to storage cluster
14 rsync input_data_location /mnt/distrax/
15 # Run Application
16 ./Heavy-IO-application --input_dir /mnt/distrax --output_dir /mnt/distrax
17 # Move data off storage cluster
18 rsync /mnt/distrax permanent_output_data_location
19 # Remove Storage System
20 mpirun -np $number_of_hosts --host $hostnames distrax_mpi -c distrax.cfg -a remove
```

Job submission



Engineering and
Physical Sciences
Research Council

<https://github.com/rosalindfranklininstitute/DisTRaX>

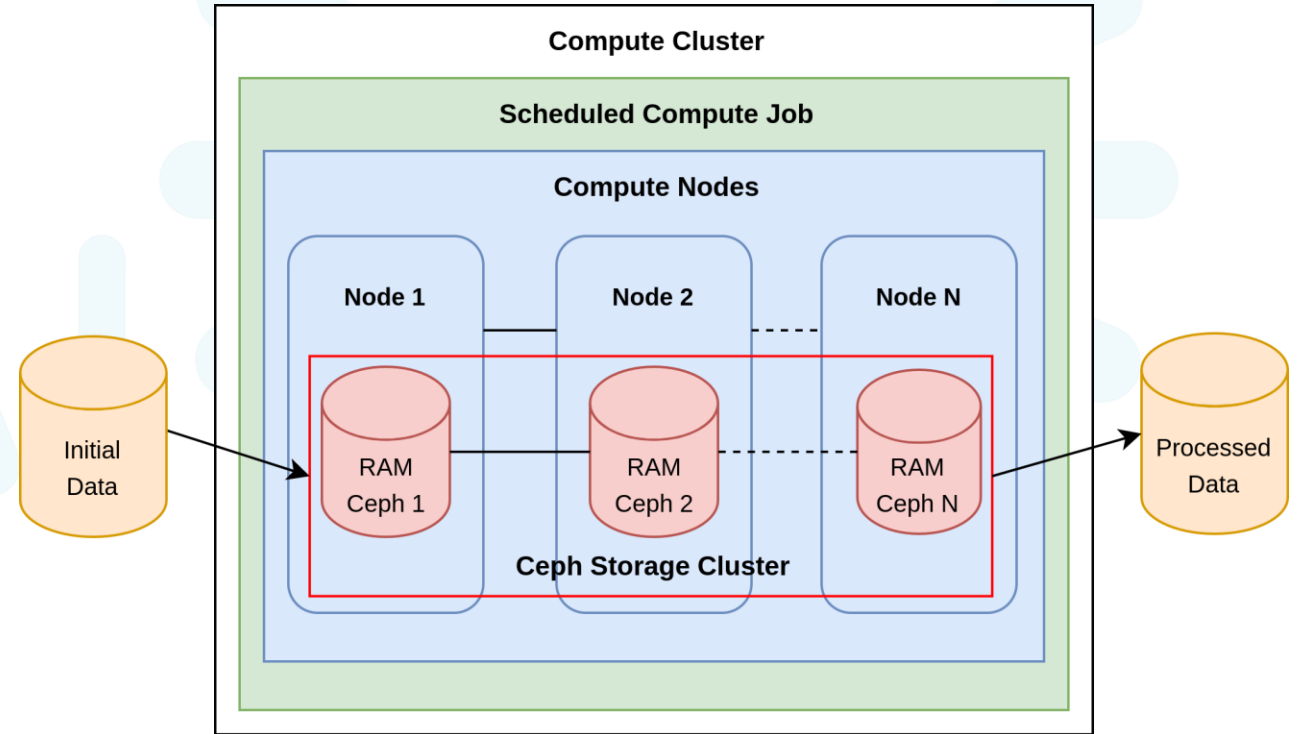


Recap

- We can create a Ceph cluster in fast and scalable way
- We can use DisTRaX within a job submission script
- But how does this solve the problem?

How DisTRaX Solves the Problem

- The IO bottleneck is now the node interconnect
- Isolated resources
- Takes pressure off high performance storage
- Can remove the need for high performance storage
- Reduces HPC cluster costs, especially in the cloud.



DisTRaX Deployed High Performance Compute Cluster Setup

Case Study: RELION

- RELION: A cryo-microscopy structure determination program used at The Rosalind Franklin Institute
- Compute Bound Application
- Runs using whole node cluster allocation
- Produces small intermediate files.
- Can negatively impact other users' jobs

RELION: Benchmark Setup

- Dataset provided by Cambridge[4]
- Baseline: 2xg4dn8xlarge nodes utilizing the EBS file system provided by AWS
- DisTRaX: 2xg4dn8xlarge nodes utilizing 96 Gib of RAM split into 6-16 Gib OSDS 3 on each host

RELION: Results

- Reduction of processing time: **5.51%**
- Total time reduction: **4.37%**
- Reduction in IO overhead: **100%**
- Removed the cost and need for running of running high performance file system in the cloud
- More efficient usage of existing hardware

Ongoing Work

- Adding support for Heterogenous Clusters
- Adding support for on-device drives.

Conclusion

- DisTRaX is a deployment tool that creates a hyper-converged HPC cluster for the duration of the job by utilising the RAM of the Compute Nodes.
- DisTRaX reduces the I/O overhead of the networked filesystem and offers a potential data processing performance increase.
- Helps better utilise existing hardware to improve the performance of HPC applications

Thank you

This project has spanned many years and has had many people involved:

Diamond Light Source - Scientific Computing:

- Dave Bond
- Mark Basham

STFC Ceph User Group:

- Tom Byrne

Rosalind Franklin Institute Artificial Intelligence theme:

- Mark Basham
- Laura Shemilt
- Joss Whittle

University of Bristol:

- Matthew Williams
- Christopher Woods

References

- [1] Weil, S.A., Brandt, S.A., Miller, E.L., Long, D.D. and Maltzahn, C., 2006, November. Ceph: A scalable, high-performance distributed file system. In *Proceedings of the 7th symposium on Operating systems design and implementation* (pp. 307-320).
- [2] Scheres, S.H., 2012. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *Journal of structural biology*, 180(3), pp.519-530.
- [3] Wong, W., Bai, X.C., Brown, A., Fernandez, I.S., Hanssen, E., Condrón, M., Tan, Y.H., Baum, J. and Scheres, S.H., 2014. Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *Elife*, 3, p.e0308

Thank You for Listening

Questions?

gabryel.mason-williams@rfi.ac.uk



Engineering and
Physical Sciences
Research Council

<https://github.com/rosalindfranklininstitute/DisTRaX>



The Rosalind
Franklin Institute