DisTRaC: Distributed Transient Ram Ceph

Accelerating High-Performance Data Processing

Gabryel Mason-Williams

gabryel.mason-williams@rfi.ac.uk

Engineering and
Physical Sciences
Research Council

diamond

University of
BRISTOL

The Rosalind
Franklin Institute

About the Rosalind Franklin Institute and me

- The Rosalind Franklin Institute:
  - A United Kingdom Research Institute dedicated to developing new technologies to tackle important health research challenges. Based in Harwell Campus, Didcot and funded by the UKRI ESPRC.
  - **5 Themes**: Artificial Intelligence and Informatics, Biological Mass Spectrometry, Correlated Imaging, Next Generation Chemistry and Structural Biology
- Me:
  - Currently studying an MSc in Artificial Intelligence at Queen Mary University of London and working as a Research Software Associate at The Rosalind Franklin Institute
  - Looking for a PhD

UKRI Engineering and Physical Sciences Research Council

diamond

University of BRISTOL

The Rosalind Franklin Institute

Contents

- Background
- The Problem
- What is DisTRaC
- Solving the Problem
- Case Studies
- Ongoing Work
- Conclusion
- Acknowledgments
- Questions

Engineering and Physical Sciences Research Council

diamond

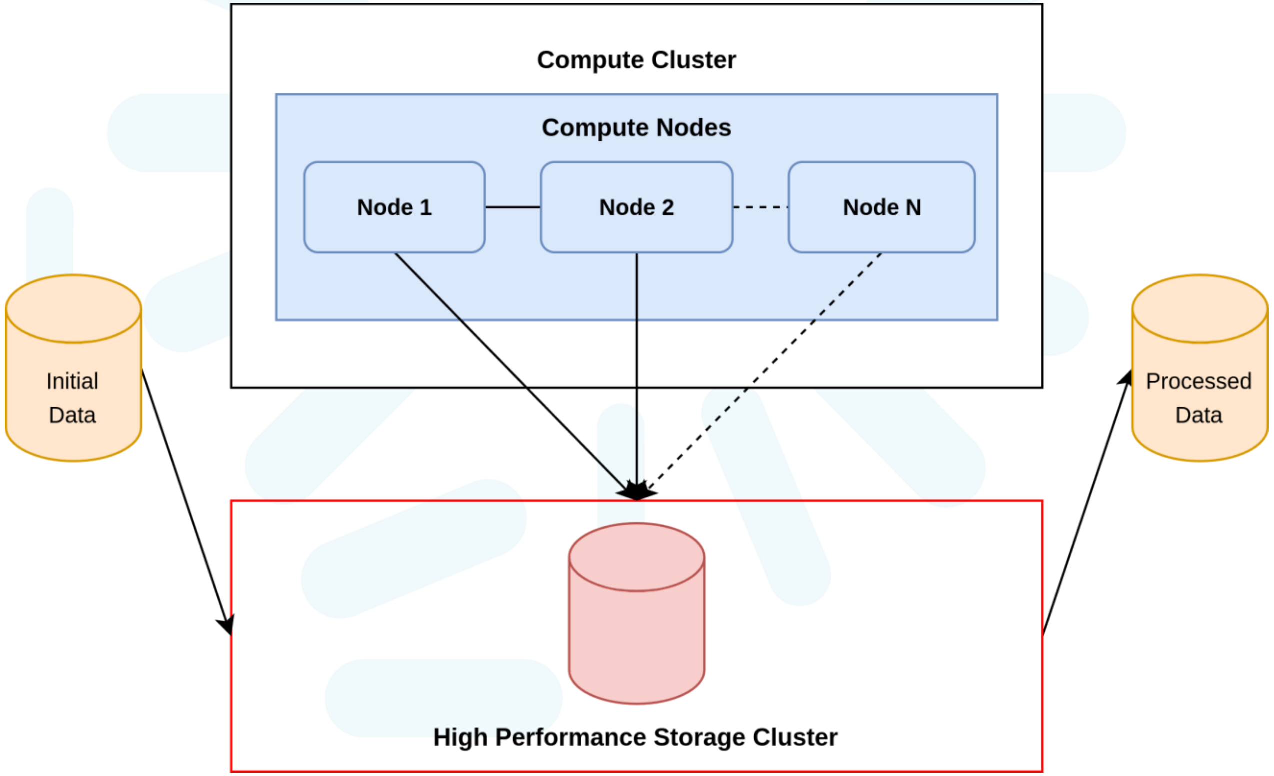
University of BRISTOL

The Rosalind Franklin Institute

Background

High Performance Compute Cluster

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

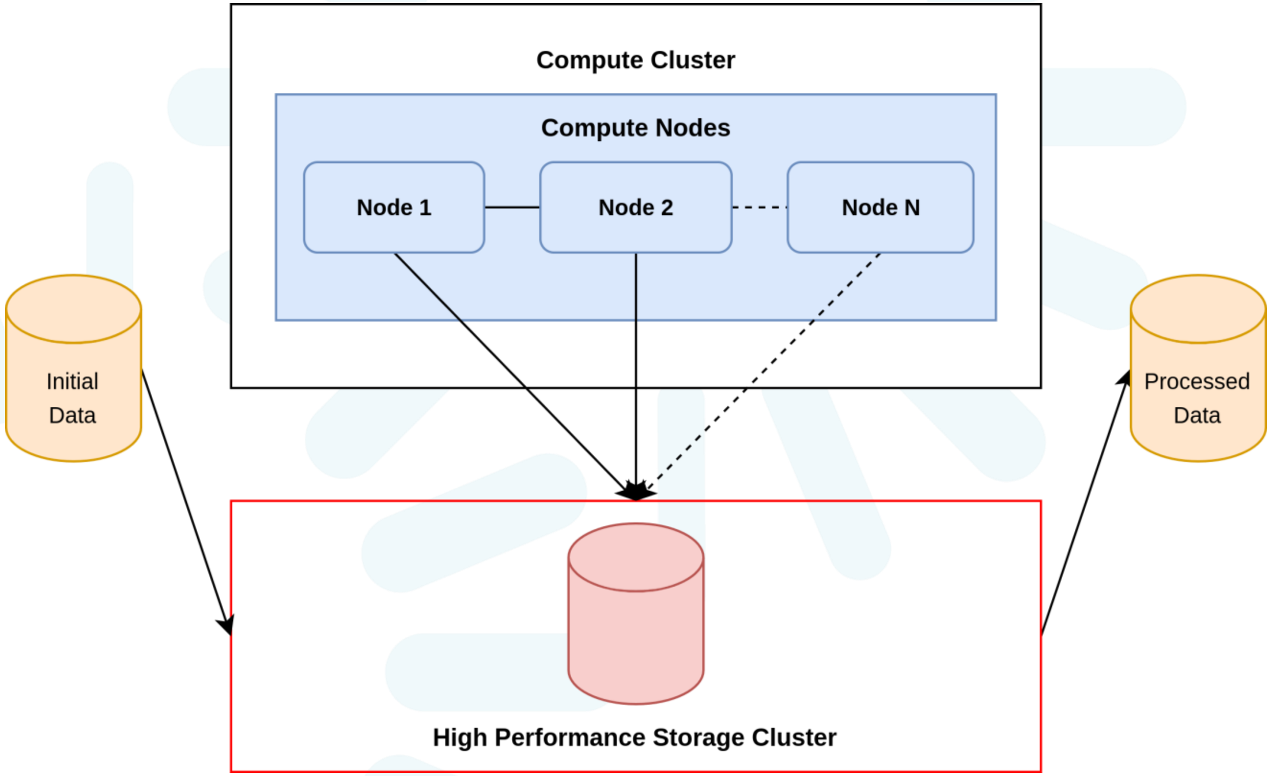


Traditional High Performance Compute Cluster Setup

What is the Problem

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

Engineering and Physical Sciences Research Council

diamond

University of BRISTOL

The Rosalind Franklin Institute

Solution?

DisTRaC

https://github.com/rosalindfranklininstitute/DisTRaC

Engineering and Physical Sciences Research Council

diamond

University of BRISTOL

The Rosalind Franklin Institute

What is DisTRaC?

- Distributed Transient Ram Ceph
- A program for deploying a transient Ceph [1] 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.

Engineering and Physical Sciences Research Council

diamond

University of BRISTOL

The Rosalind Franklin Institute

Why DisTRaC and not another deployment tool

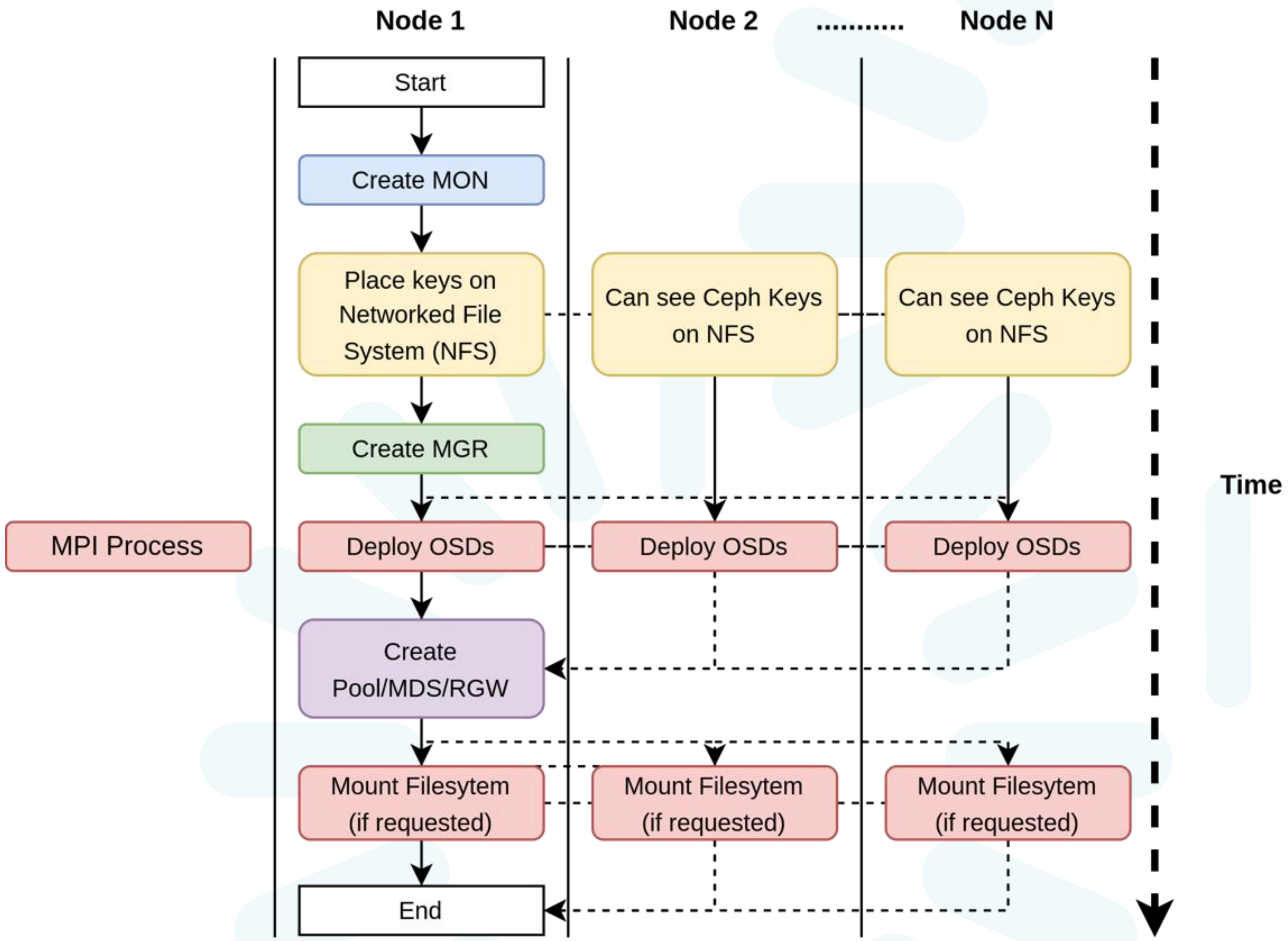
Current Deployment tools

- 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

DisTRaC Deployment



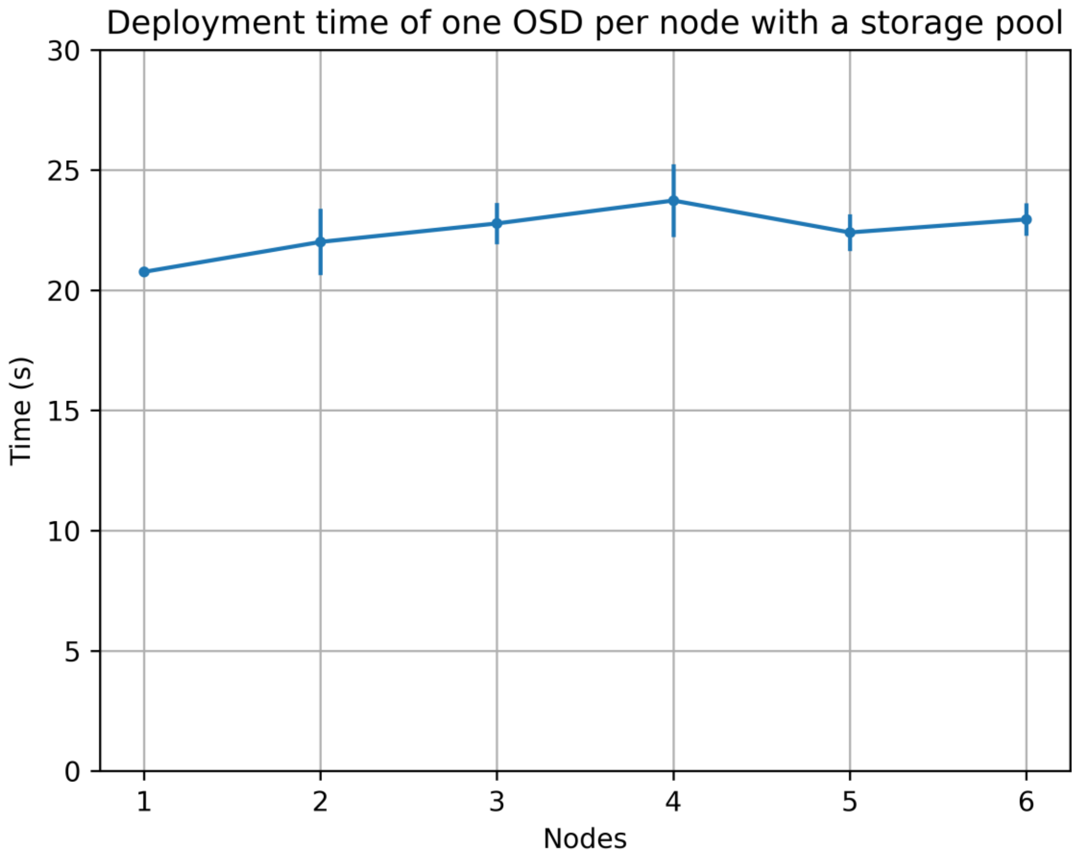
Engineering and Physical Sciences Research Council

diamond

University of BRISTOL

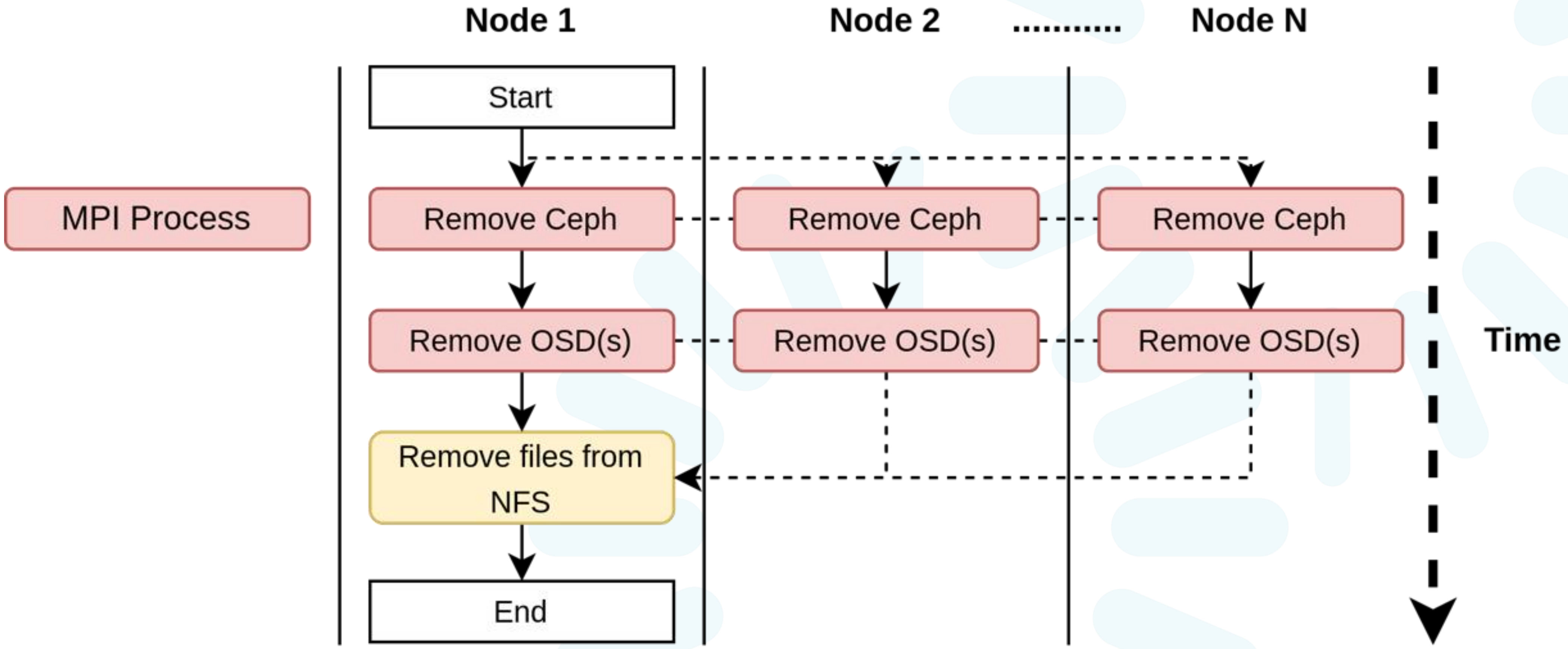
The Rosalind Franklin Institute

Deployment Time



Version 1 of DisTRaC, Ceph Luminous

DisTRaC Removal



Node 1

Node 2

....

Node N

MPI Process

Start

Remove Ceph

Remove OSD(s)

Remove files from NFS

End

Remove Ceph

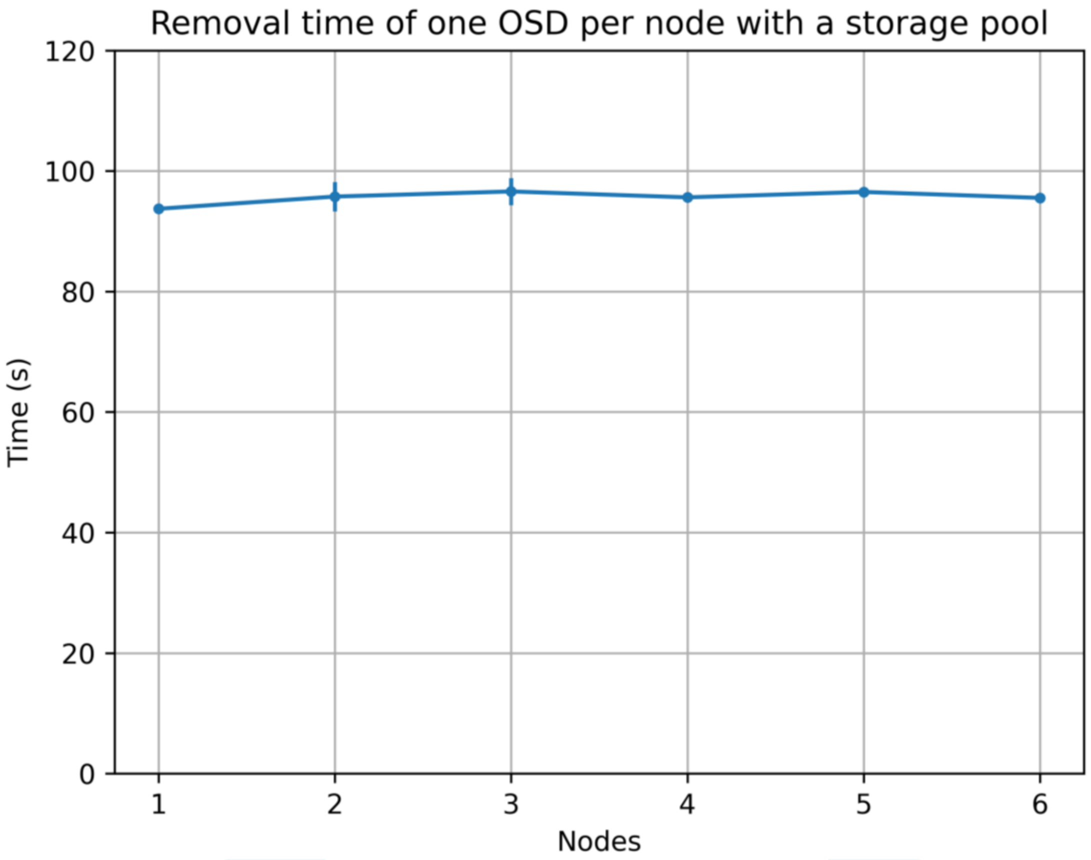
Remove OSD(s)

Remove Ceph

Remove OSD(s)

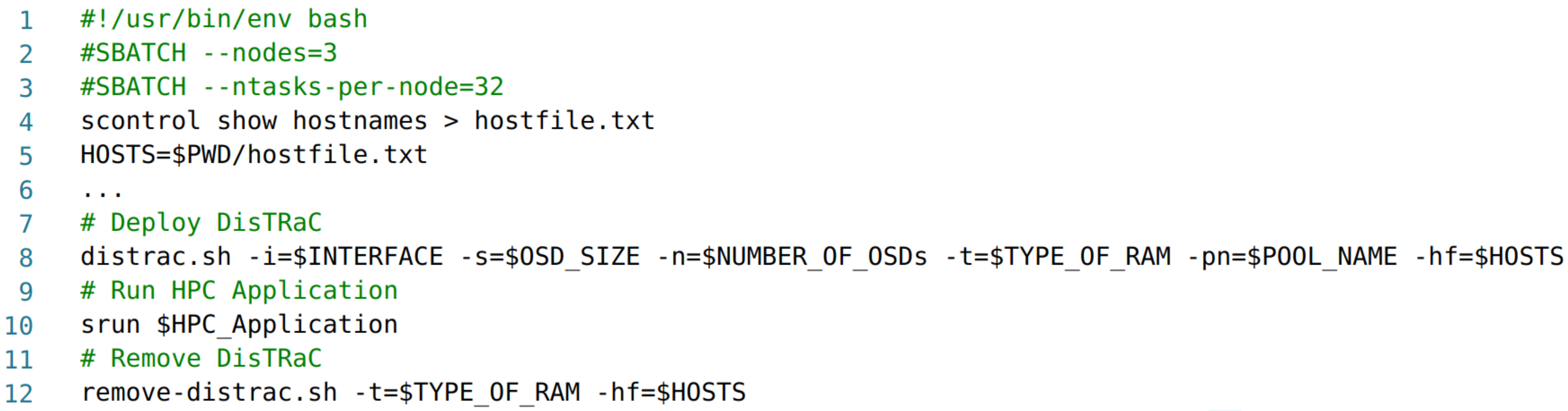
Time

Removal Time



Version 1 of DisTRaC, Ceph Luminous

Example Deployment Script



```
#!/usr/bin/env bash
#SBATCH --nodes=3
#SBATCH --ntasks-per-node=32
scontrol show hostnames > hostfile.txt
HOSTS=$PWD/hostfile.txt
...
# Deploy DisTRaC
distrac.sh -i=$INTERFACE -s=$OSD_SIZE -n=$NUMBER_OF_OSDs -t=$TYPE_OF_RAM -pn=$POOL_NAME -hf=$HOSTS
# Run HPC Application
srun $HPC_Application
# Remove DisTRaC
remove-distrac.sh -t=$TYPE_OF_RAM -hf=$HOSTS
```

DisTRaC deploy and remove

Recap

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

UKRI Engineering and Physical Sciences Research Council

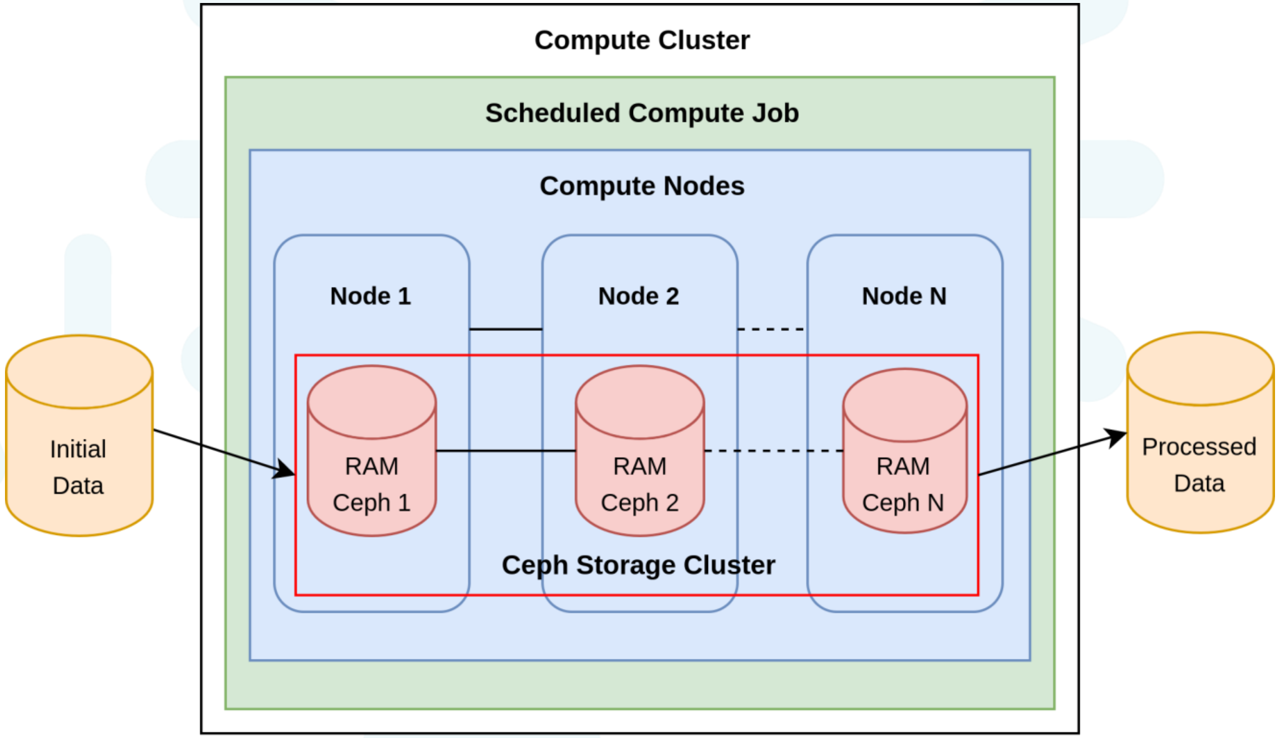
diamond

University of BRISTOL

The Rosalind Franklin Institute

How DisTRaC solves the problem

- The IO bottleneck is now the node interconnect
- Isolated resources
- Takes pressure off HP storage
- Can remove the need for HP storage
- Reduces HPC cluster costs, especially in the cloud.
- Helps HPC facilities move towards Net-Zero



DisTRaC Deployed High Performance Compute Cluster Setup

Engineering and Physical Sciences Research Council
diamond
University of BRISTOL
The Rosalind Franklin Institute

Case Studies

- RELION [2]
- SAVU [3]

Case Study: RELION

- RELION: A cryo-microscopy structure determination program used at The Rosalind Franklin Insitute
- 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
- DisTRaC: 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 HP file system in the cloud
- More efficient usage of existing hardware
- Helping towards Net-Zero Goals

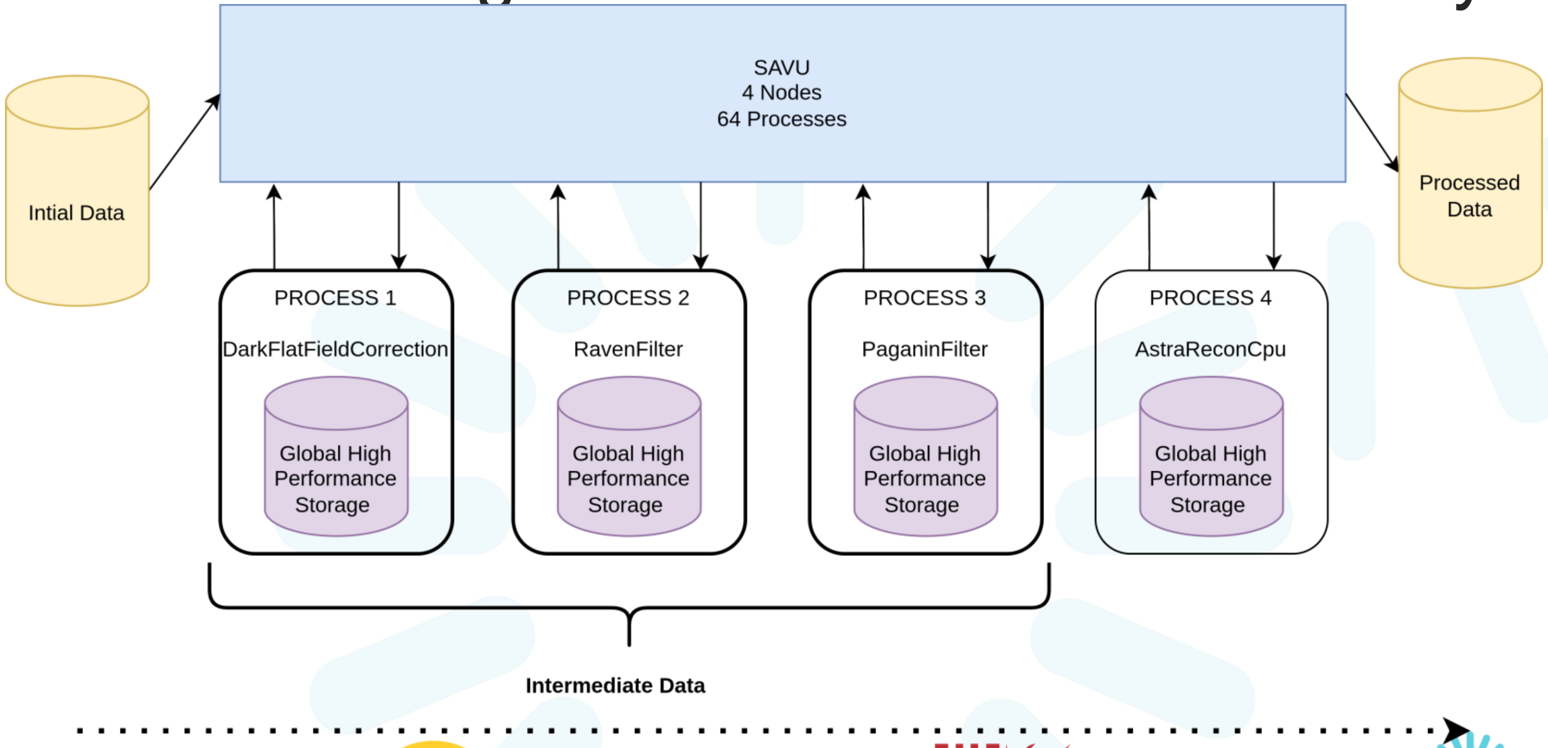
UKRI Engineering and Physical Sciences Research Council diamond University of BRISTOL The Rosalind Franklin Institute

Case Study: SAVU

- SAVU: Tomography Reconstruction and Processing Pipeline used at Diamond Light Source and The Rosalind Franklin Insitute.
- Runs using whole node cluster allocation
- Can run at network rate
- Produces intermediate files.
- Capable of saturating access to a parallel file system
- Can negatively impact other users' jobs

SAVU: Setup At Diamond Light Source

- Dataset: Diamond Light Source Visit NT23252 Dataset [5]
- Baseline: 4 nodes utilizing the GPFS Central HP File system



Engineering and Physical Sciences Research Council

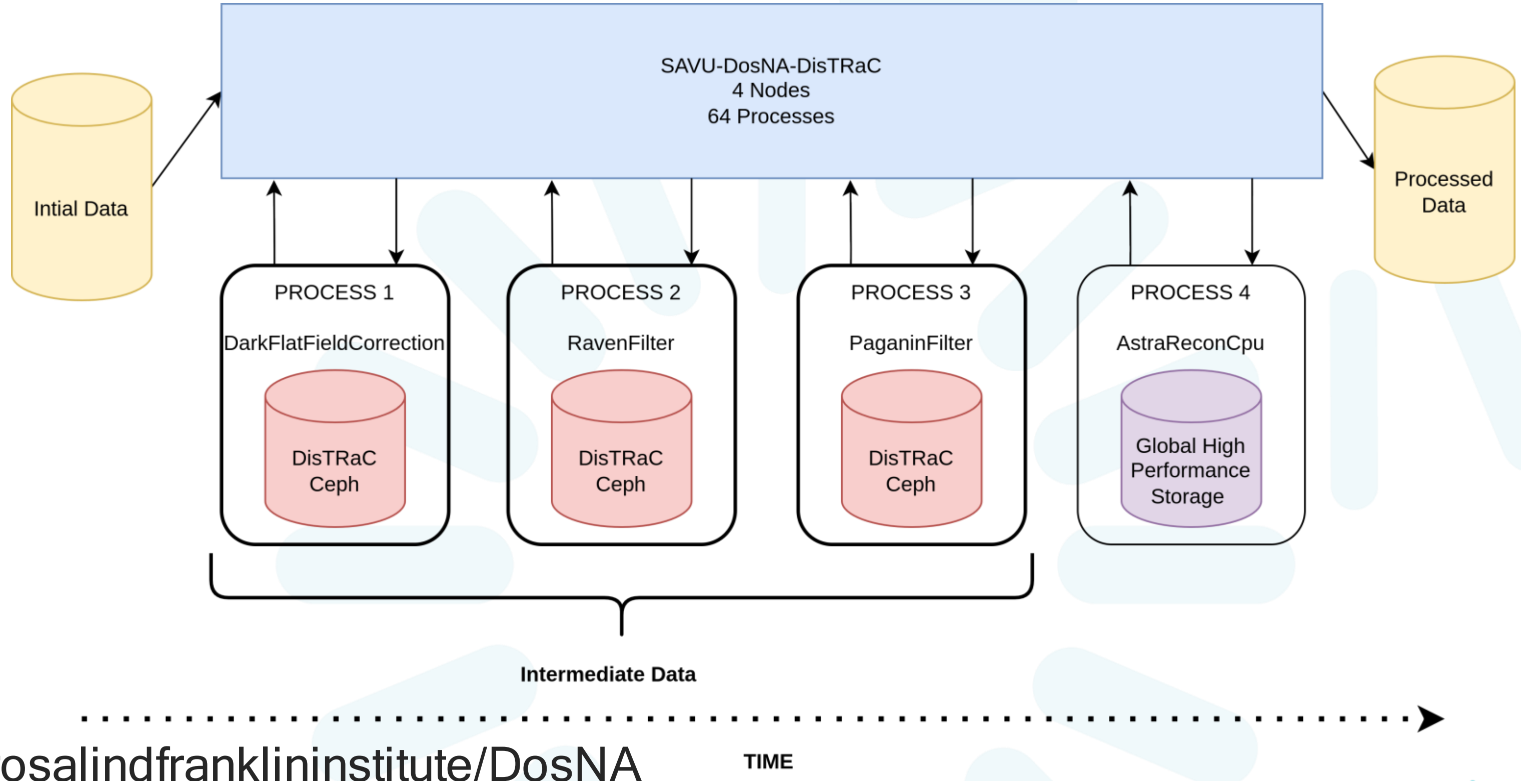
diamond

University of BRISTOL

The Rosalind Franklin Institute

SAVU: Setup At Diamond Light Source

- DisTRaC: 4 nodes utilizing Ceph via DosNA[1]



[1] https://github.com/rosalindfranklininstitute/DosNA

SAVU: Results

- Total time reduction: **8.32%**
- Reduction in IO overhead: **81.04%**
- Reduce impact of SAVU on other users
- Prevented storing of intermediate data.
- More efficient usage of existing hardware
- Helping towards Net-Zero Goals

SAVU: Benchmark Setup At AWS

- Dataset: Diamond Light Source Visit NT23252 Dataset
- Baseline: 4 nodes utilizing the EBS AWS File system
- DisTRaC: 4 nodes utilizing Ceph Via DosNA

SAVU: Results

- Total time reduction: **67.53%**
- Reduction in IO overhead: **81.04%**
- Reduce costs of AWS
- Makes the cloud more viable for HPC
- Helping towards Net-Zero Goals

UKRI Engineering and Physical Sciences Research Council

diamond

University of BRISTOL

The Rosalind Franklin Institute

Ongoing work

- DisTRaC – Intergration into Cluster-In-The-Cloud[1]
- Adding support for Heterogenous Clusters
- Adding support for NVME deployment
- DisTRaX- removing the Ceph requirement making it extensible to other storage mechanisms.

[1] https://cluster-in-the-cloud.readthedocs.io/en/latest/

Conclusion

- DisTRaC is a Ceph deployment tool that creates a hyper-converged HPC cluster for the duration of the job by utilising the RAM of the Compute Nodes.
- DisTRaC 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
- Moves us closer to sustainable and net-zero HPC

UKRI Engineering and Physical Sciences Research Council

diamond

University of BRISTOL

The Rosalind Franklin Institute

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

UKRI Engineering and Physical Sciences Research Council

diamond

University of BRISTOL

The Rosalind Franklin Institute

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] Wadeson, N. and Basham, M., 2016. Savu: a Python-based, MPI framework for simultaneous processing of multiple, N-dimensional, large tomography datasets. *arXiv preprint arXiv:1610.08015*.
- [4] Wong, W., Bai, X.C., Brown, A., Fernandez, I.S., Hanssen, E., Condron, 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.e03080.
- [5] Mark Basham, Nghia Vo, Avery Pennington, Win Tun, Olly King, & Gabryel Mason-Williams. (2020). Diamond Light Source Visit NT23252 Dataset [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4030687

UKRI Engineering and Physical Sciences Research Council

diamond

University of BRISTOL

The Rosalind Franklin Institute

Thank you for listening

Questions?

UKRI Engineering and Physical Sciences Research Council

diamond

University of BRISTOL

The Rosalind Franklin Institute