CTA at AARNET

Frozen Data Storage at AARNET
Denis Lujanski, David Jericho, Crystal Chua, Michael Usher,
Michael D'Silva, Suzana Zahri

Introduction

Who we are:

- AARNet - Australia's Academic and Research Network, EST 1989
- Australia's NREN

What we do:

- Internet
- Cyber Security
- Data Storage, Collaboration and Movement
- Mirror

A Special Thank You

We'd like to extend a sincere thank you to the CTA and EOS teams for their ongoing support with running EOS and CTA infrastructure. A few honourable mentions for the CTA project:

- Michael Davis
- Julien Leduc
- Steven Murray
- Georgios Kaklamanos
- Oliver Keeble
- Roberto Valverde
- Jakub Moscicki

Why CTA

- Open source with excellent track record of managing gigantic datasets
- Uses EOS, which we are familiar with
- Decoupled from choice of backup tool

Server Hardware

- 3 x Dell PowerEdge R7525 Servers
- Dual socket AMD 128 threads
- 512G RAM
- 2 x 500G SSD for OS
- 6 x 500G SSD for Ceph OSD
- 10 x 15T SSD for EOS FST (Archive staging space)
- 2 x 2T NVMe for app storage
- 50G Mellanox Cards
- 16G Fibre Channel Card (8G switch)

- 1 x Supermicro SSG-6048R-E1CR60L (old)
- Dual socket Intel 56 threads
- 256G RAM
- 2 x 800G SSD for OS
- 48 x 6T HDD for EOS FST (Retrieve staging space)
- 2 x 800G NVMe for app storage
- 40G Mellanox Cards
- 16G Fibre Channel Cards (8G switch)

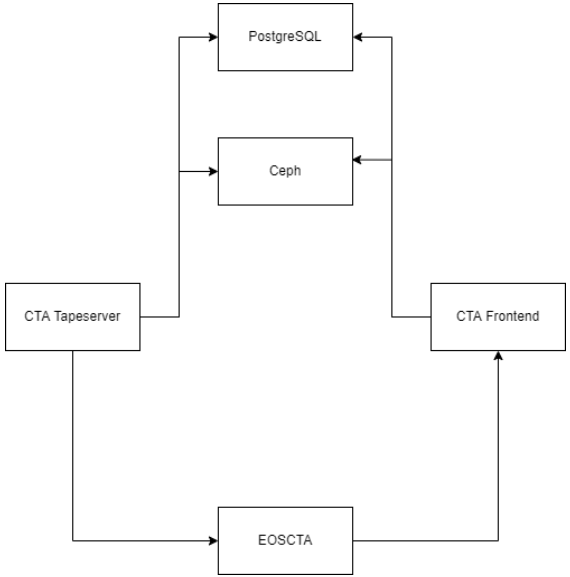
Tape Hardware

- IBM TS4500 L55 with S55 expansion frame
- 9 x LTO 7 Tape Drives
- 1283 x LTO7 Tapes (6T)

Deployment: Orchestration

- 3 node HA Kubernetes cluster, managed by Rancher v2
- Components deployed using Helm, or custom operators (eg: Rook ceph)

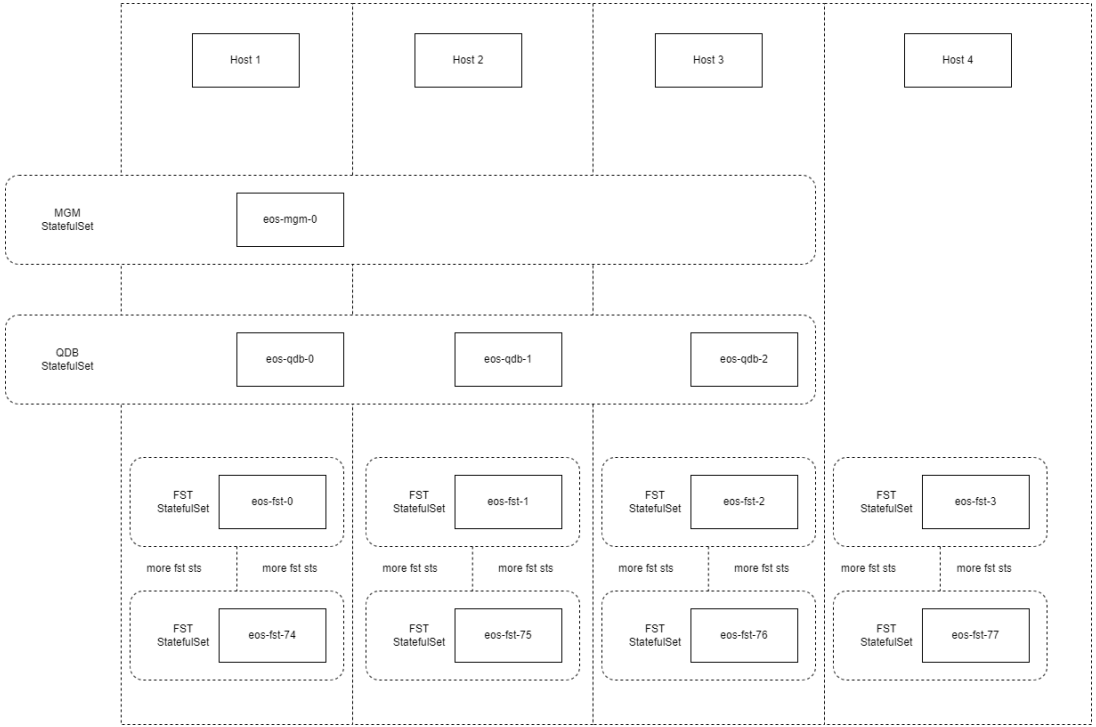
Deployment: Basic architecture



Deployment: EOSCTA

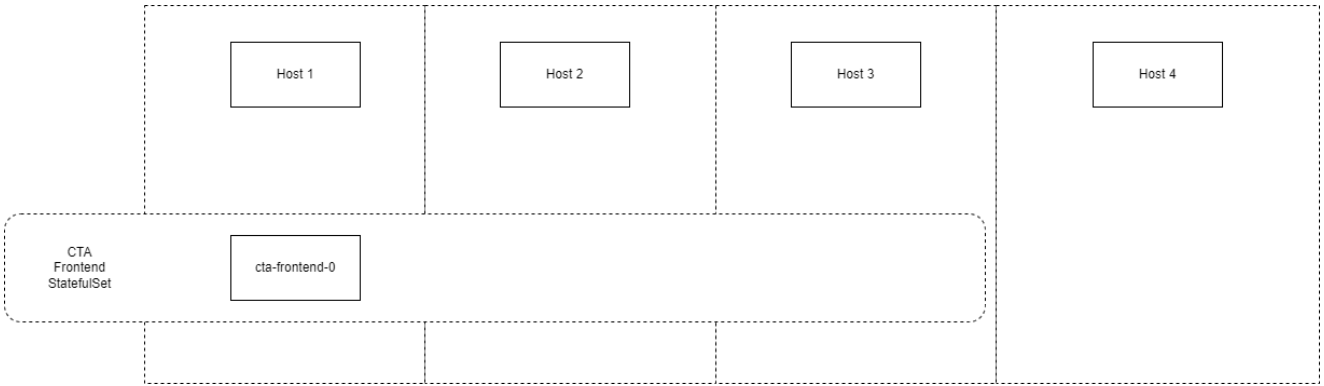
- Deployment tool: Helm
- Separate charts for:
  - MGM
  - FST
  - QDB
- MGM can be a Deployment

due to stateless nature



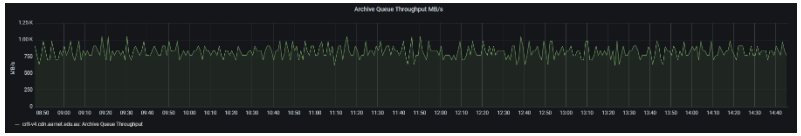
Deployment: CTA Frontend

- Deployment tool: Helm
- Frontend could potentially also be a Deployment (untested)



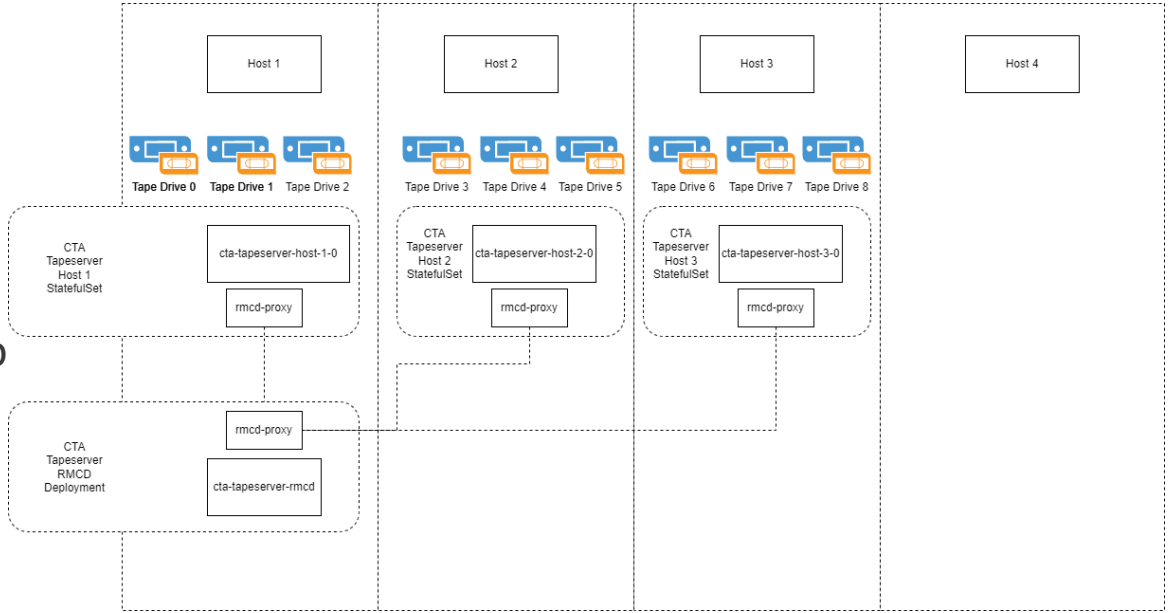
Deployment: CTA Tapeserver (Original)

- Deployment tool: Helm
- Problem: Each host can only do 8Gpbs to tape over FC. We have 9 tape drives, each capable of 2.4Gbps).



Deployment: CTA Tapeserver (New)

- Deployment tool: Helm
- Had to split between servers to avoid fibre channel bottleneck
- Rmcd listens on 127.0.0.1 only (for security reasons), so splitting tapeserver between hosts introduced complexity of using of double proxy



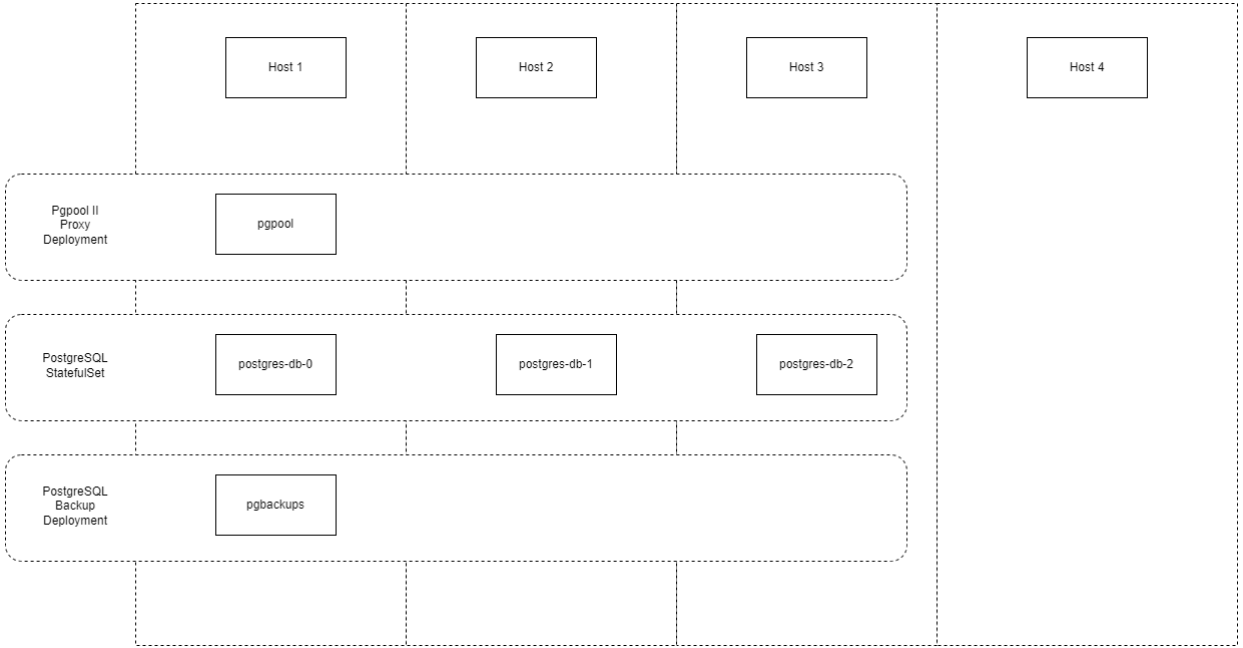
Deployment: Ceph

- Deployment tool: Rook (rook.io)
- 6 x OSD's per host
- 500G per OSD
- 9T of raw ceph storage
- 3 replica pool for CTA use
- Out of the box, except:
  - Runs in cta k8s namespace instead of default rook-ceph (sed)
  - rook-ceph-mgr k8s service exposed via NodePort instead of default ClusterIP (so that it can be monitored externally)

Deployment: PostgreSQL Overview

- Deployment tool: Helm (fork of https://github.com/bitnami/charts/tree/master/bitnami/postgresql-ha)
- Moving to Crunchy Data PGO (https://access.crunchydata.com/documentation/postgres-operator/4.7.4/)
- Lots of changes to make production ready
- PgPool II bug related to multiple retrieve requests
- Automatic failover buggy

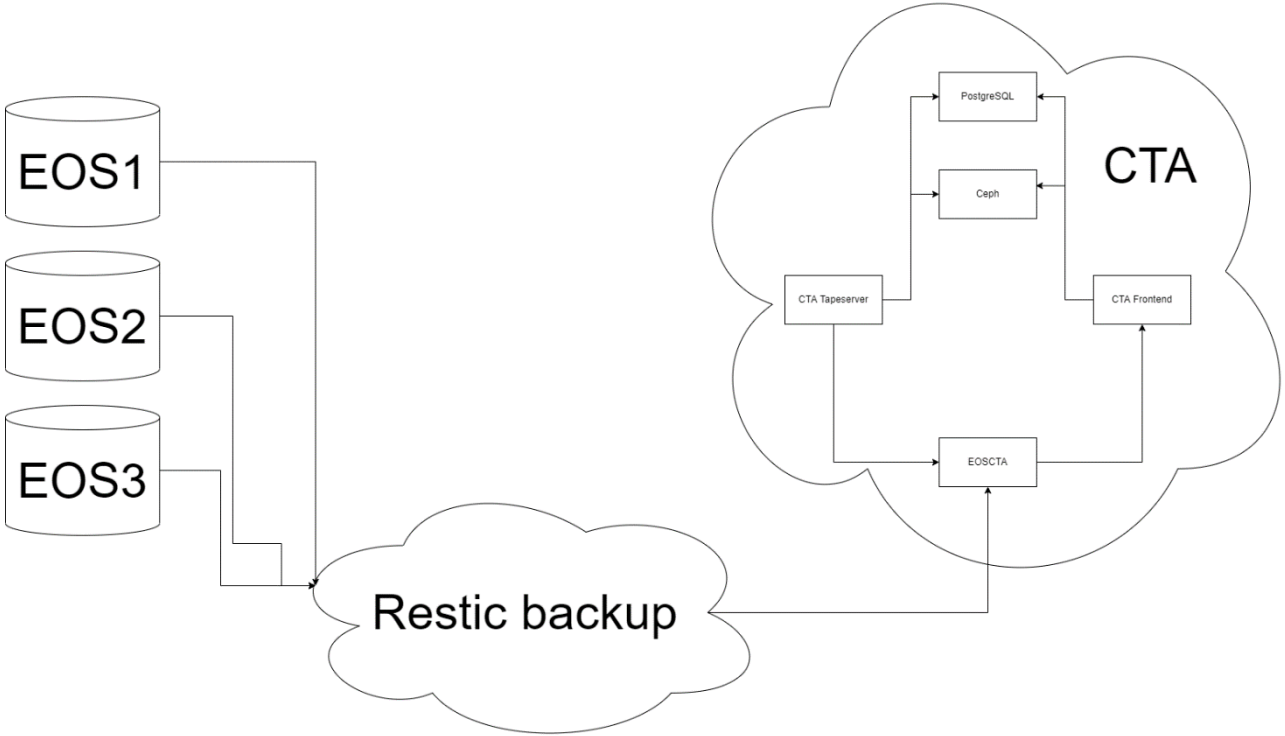
Deployment: PostgreSQL



Stats

- 747T EOSCTA space (460T SSD for Archive staging (single replica), 287T HDD for Retrieve (single replica))
- 4.3P Data archived to tape
- 4.6M Files archived to tape
- 21.6Gbps aggregate throughput to tape

Bigger Picture



What’s next

- Move PostgreSQL to Crunchy Data PGO
- Double retrieve staging space to EOSCTA, more tape drives
- Review cta package build process / CI
- More tape libraries in other DC’s, multiple tape copies
- Make restic tape-aware and forgo having to script restores and make snapshot deletion viable
- Implement s3 access for CTA

Questions?

Denis.Lujanski@aarnet.edu.au