Containerized dCache on Ceph

Mathias Lindberg & Hugo U.R. Strand

Chalmers Centre for
Computational Science
and Engineering

Oct 21, 2020
NeIC/NordForsk
NDGF all hands

Ceph

C3SE Cephyr system

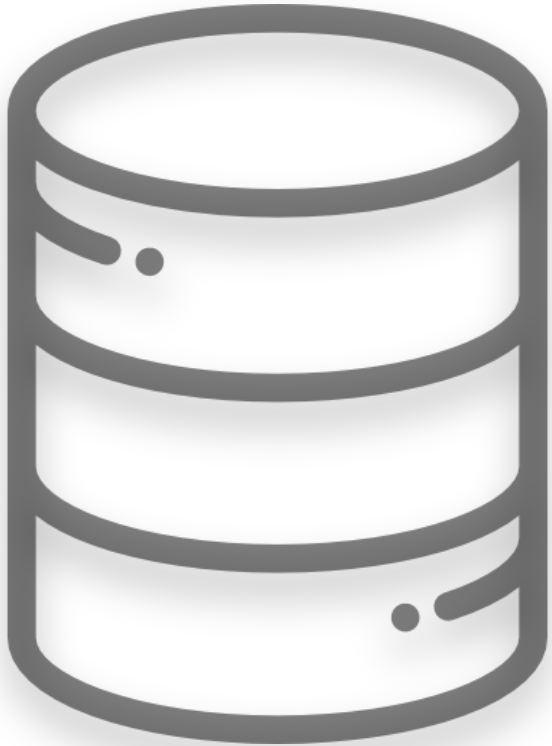
SOUTH POLE_

- Storage Hardware
  - Initial: 7x 2u 12x10TB spinning disk + 2 x 1.6TB NvME 4x25GbE Mellanox Cx4-lx
  - Expansion: 6x 4u 24x10TB spinning disk + 2 x 1.6TB NvME 4x25GbE Mellanox Cx4-lx
- Network
  - 2x Huawei 6860 series 48x25GbE 6x100GbE
  - 2x 100Gb interconnect
- NVMe partitioning
  - Ceph-volume lvm 60GB for WAL+DB per OSD
  - RocksDB size increments 3GB, 30GB, 300GB
  - OSD's WAL+DB split evenly between NVMe's for redundancy

ceph

Ceph Block Device Storage

- Rados Block Devices (RBDs)
  - *reliable autonomic distributed object store (RADOS)*
  - Ceph data pool k=6, m=2
  - 6x 60TB dCache pools
    on separate RBD-devices
- Features
  - striping, data-pool, exclusive-lock
  - Object size 16M stripe unit 16M
  - Stripe-count 6
- Performance (R/W SOB)
  - 1.4 GB/s per device
  - 5.3 GB/s (42.4 Gbit/s) per host



dCache

dCache Hardware

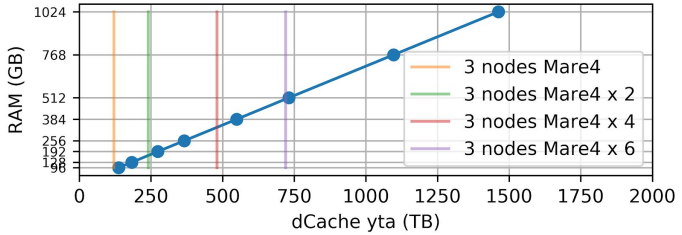
Extrapolation using NEIC recommendations

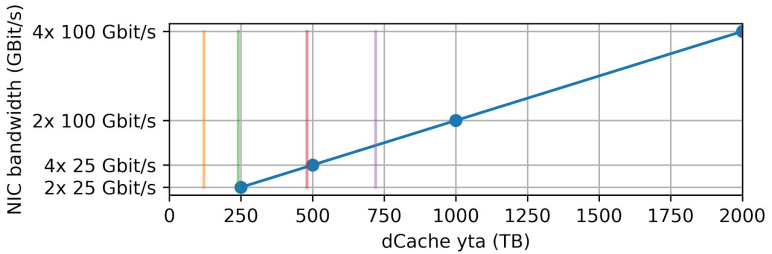
https://wiki.neic.no/wiki/DCache_Pool_Hardware

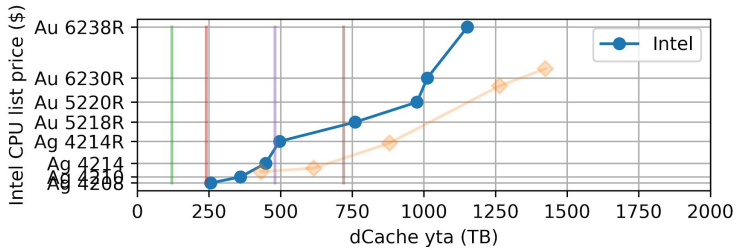
Projected dCache size 320TB

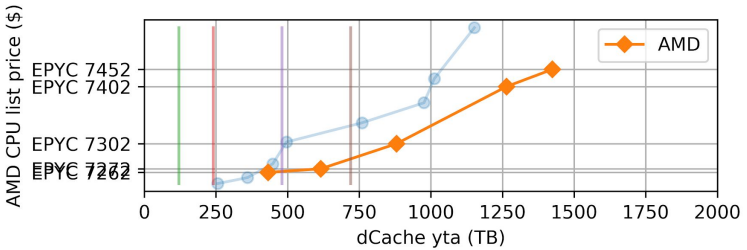
- 120TB on 3x hosts
- Host requirements
  - Memory (RAM)
  - Network bandwidth (NIC)
  - CPU Intel/AMD

Settle for 4x larger









dCache Hosts

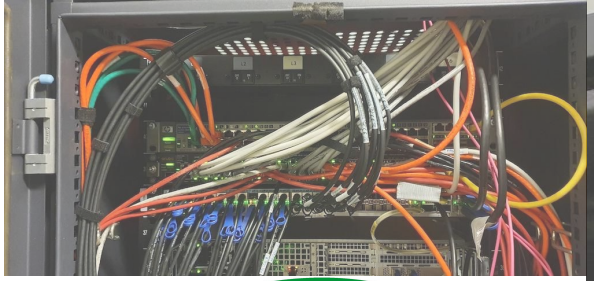
3x Super Micro Servers *navis[1-3]*

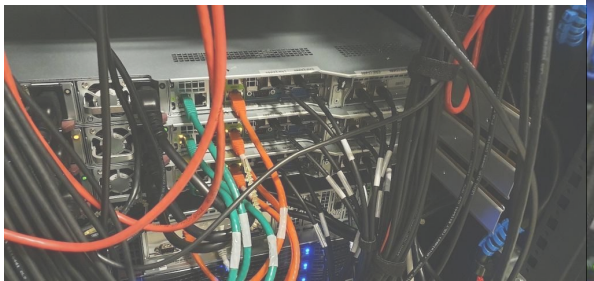
Hardware

- 2x AMD Epyc 7302 (16 cores)
- 512 GB RAM
- 2x Mellanox NICs 2x 25Gbit/s in one 100Gbit/s bond/team

Software

- CentOS 8
- RADOS kernel module
- Cobbler + Ansible (config)





Containers

dCache Containers

- GitLab CI
  - Build containers
  - Host container repository
- Software configuration
  - CentOS 8
  - iptables, ganglia-gmond, cronie, logrotate, openssh
  - dcache 5.2.x (and java 1.8.0)
- Runtime, network, storage
  - Podman
  - External IP using CNI-ipvlan
  - Bind-mount RBD devices
- Management
  - inhouse cli: mare4

GitLab CI

podman

CNI

Cli Management

Poor-man's kubernetes

- Simple python wrapper
- Map & mount rbd, start container
- Stop container, umount & unmap
- Aggregated status
  - ps, stats, mounts, maps

```
root@mgmt:~# mare4 mounts
Host navis1
------------------------------------------------------------------
Filesystem      Size  Used Avail Use% Mounted on
/dev/rbd0        60T   54T  6.6T  90% /mnt/mare4/mare040
/dev/rbd1        10G 1014M  9.1G  10% /mnt/mare4/mare040_meta
/dev/rbd2        60T   53T  7.1T  89% /mnt/mare4/mare041
/dev/rbd3        10G  993M  9.1G  10% /mnt/mare4/mare041_meta

Host navis2
------------------------------------------------------------------
Filesystem      Size  Used Avail Use% Mounted on
/dev/rbd0        60T   53T  7.2T  89% /mnt/mare4/mare042
/dev/rbd1        10G  1.0G  9.0G  11% /mnt/mare4/mare042_meta
/dev/rbd2        60T   53T  7.1T  89% /mnt/mare4/mare043
/dev/rbd3        10G  988M  9.1G  10% /mnt/mare4/mare043_meta

Host navis3
------------------------------------------------------------------
Filesystem      Size  Used Avail Use% Mounted on
/dev/rbd0        60T   54T  7.0T  89% /mnt/mare4/mare044
/dev/rbd1        10G  935M  9.1G  10% /mnt/mare4/mare044_meta
                 60T   52T  8.4T  87% /mnt/mare4/mare045
                 10G  923M  9.1G  10% /mnt/mare4/mare045_meta
```

```
root@mgmt:~# mare4 stats
host    pool     cpu_percent    mem_usage          netio              blocki
------  -------  -------------  -----------------  -----------------  -----------------
navis1  mare040  --             4.809GB / 53.69GB  262.9MB / 193.5MB  1.422GB / 609.9MB
navis1  mare041  --             5.07GB / 53.69GB   209.8MB / 180.4MB  1.4GB / 480.2MB
navis2  mare042  --             4.594GB / 53.69GB  217MB / 178.5MB    1.447GB / 688.7MB
navis2  mare043  --             4.96GB / 53.69GB   202.1MB / 178MB    1.393GB / 502.6MB
navis3  mare044  --             4.448GB / 53.69GB  217.2MB / 178.3MB  1.3GB / 553.1MB
navis3  mare045  --             4.389GB / 53.69GB  216.9MB / 178.8MB  1.282GB / 679MB
```

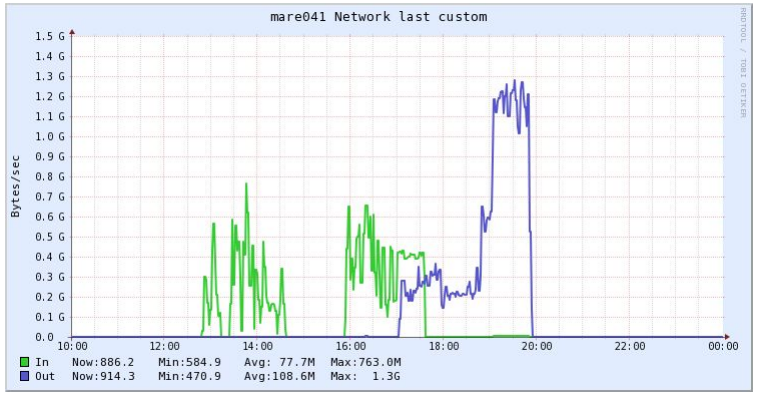
dCache Performance

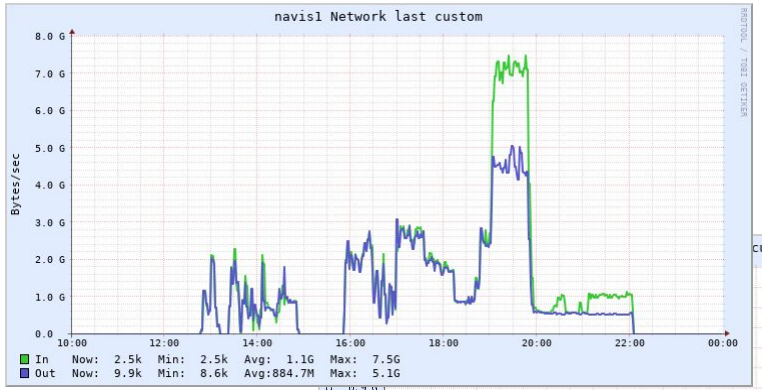
IO Benchmarks

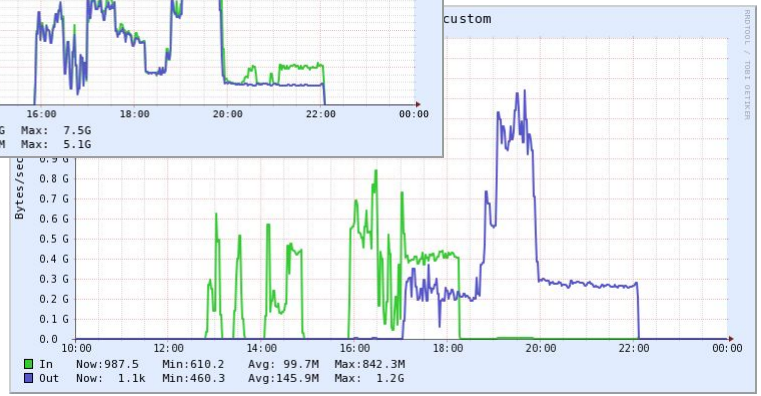
- Pools: Read 1.2 GB/s (9.6 Gbit/s)
- Hosts (navis[1-3])
  - In > 7 GB/s (56 Gbit/s)
  - Out > 5 GB/s (40 Gbit/s)

File removal (slow)

- Berkley DB in /pool/meta
- Mitigation (5x improvement)
  - /pool/meta to separate RBD
  - erasure code ➜ replicated_hdd
  - object size 4K







Q&A