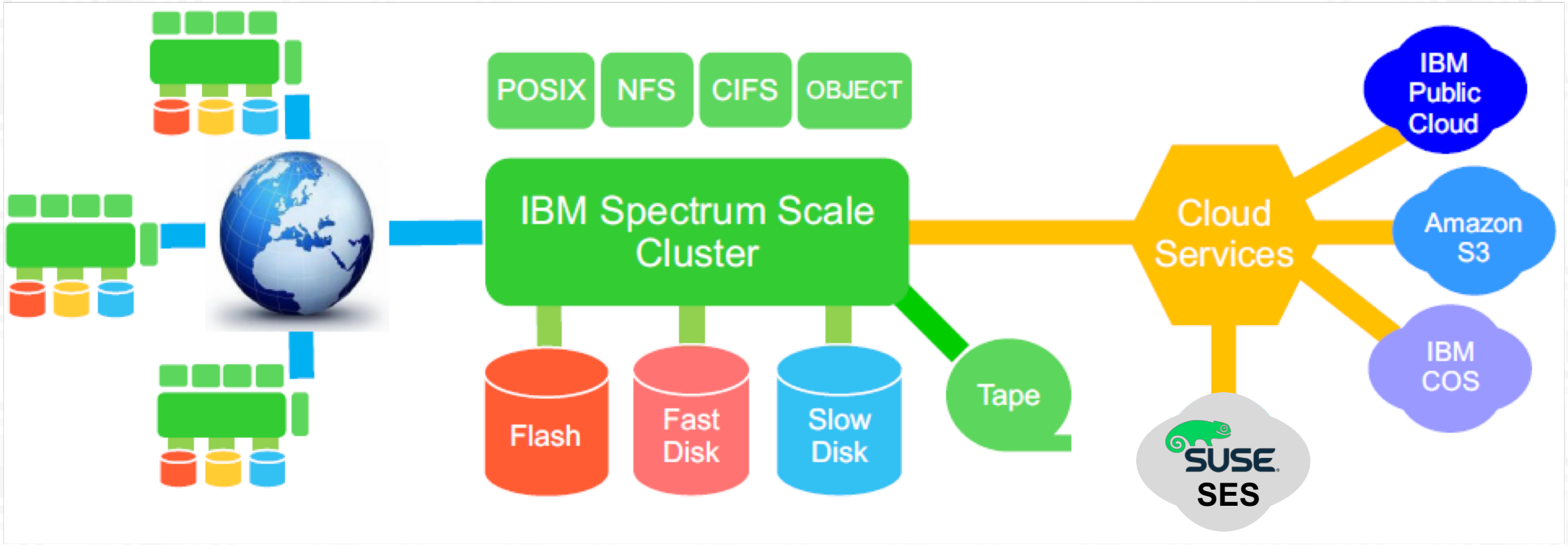
Ceph as a Spectrum Scale Storage Tier

SC18 Spectrum Scale User Group Meeting

Michael Hennecke | Sunday 11-Nov-2018

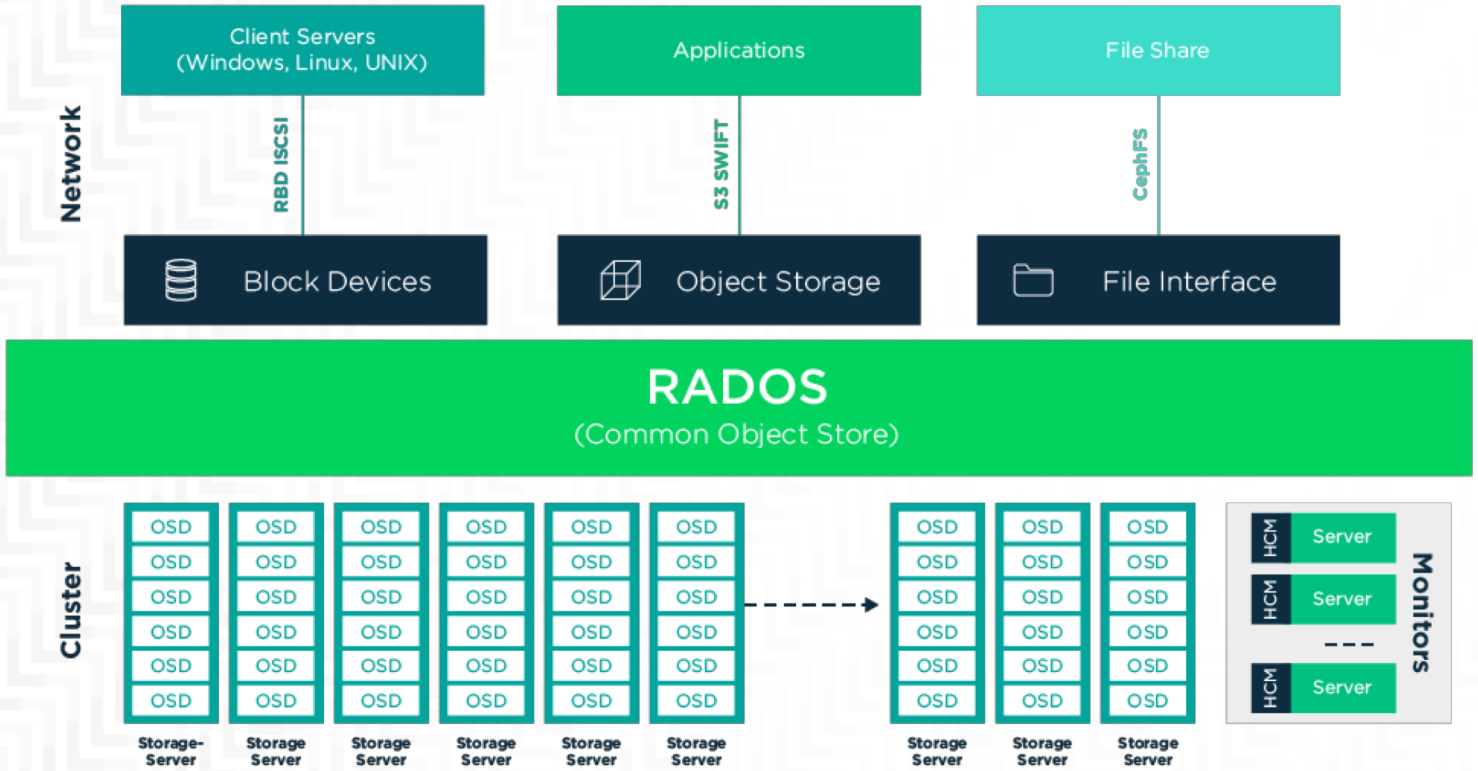
Spectrum Scale Data Management Infrastructure

Spectrum Scale‘s Information Lifecycle Management (ILM) rules can be used to control initial placement, migration, and deletion of data across storage pools.



SUSE Enterprise Storage (powered by Ceph)

- Minimum 4x Object Storage Nodes and 1x Management Node
- Ceph monitors (3x), gateway nodes, metadata servers can reside on OSD servers, but for demanding workloads they should run on separate nodes

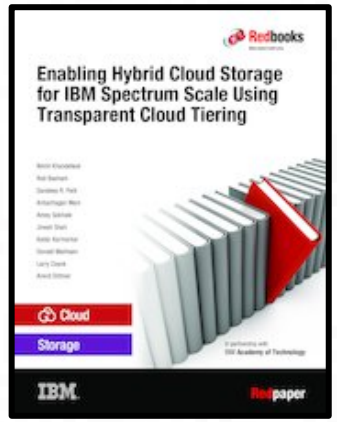


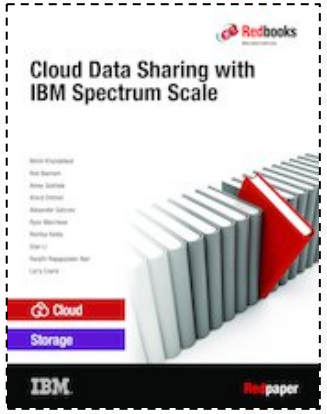
Hardware used for the POC:

- **6x Ceph OSD nodes – "Capacity"**
  - SR650, single-socket (4110)
  - **12x HDD** 3.5" SATA 4TB
  - 4-port 10GbE
  - 128GB memory
  - ➔ Ceph storage pool is 3WayReplicated
- **3x Ceph MON nodes**
  - SR530, single-socket (3104)
  - 2-port 10GbE
  - 16GB memory
- **2x `radosgw`** roles, colocated on 2x OSDs
  - Dedicated 1x 10GbE links to „WAN"
- Maximum **„WAN" bandwidth** between SES cluster and Spectrum Scale TCT nodes:
  - **2x 10GbE Ethernet**

Transparent Cloud Tiering (TCT) – Documentation

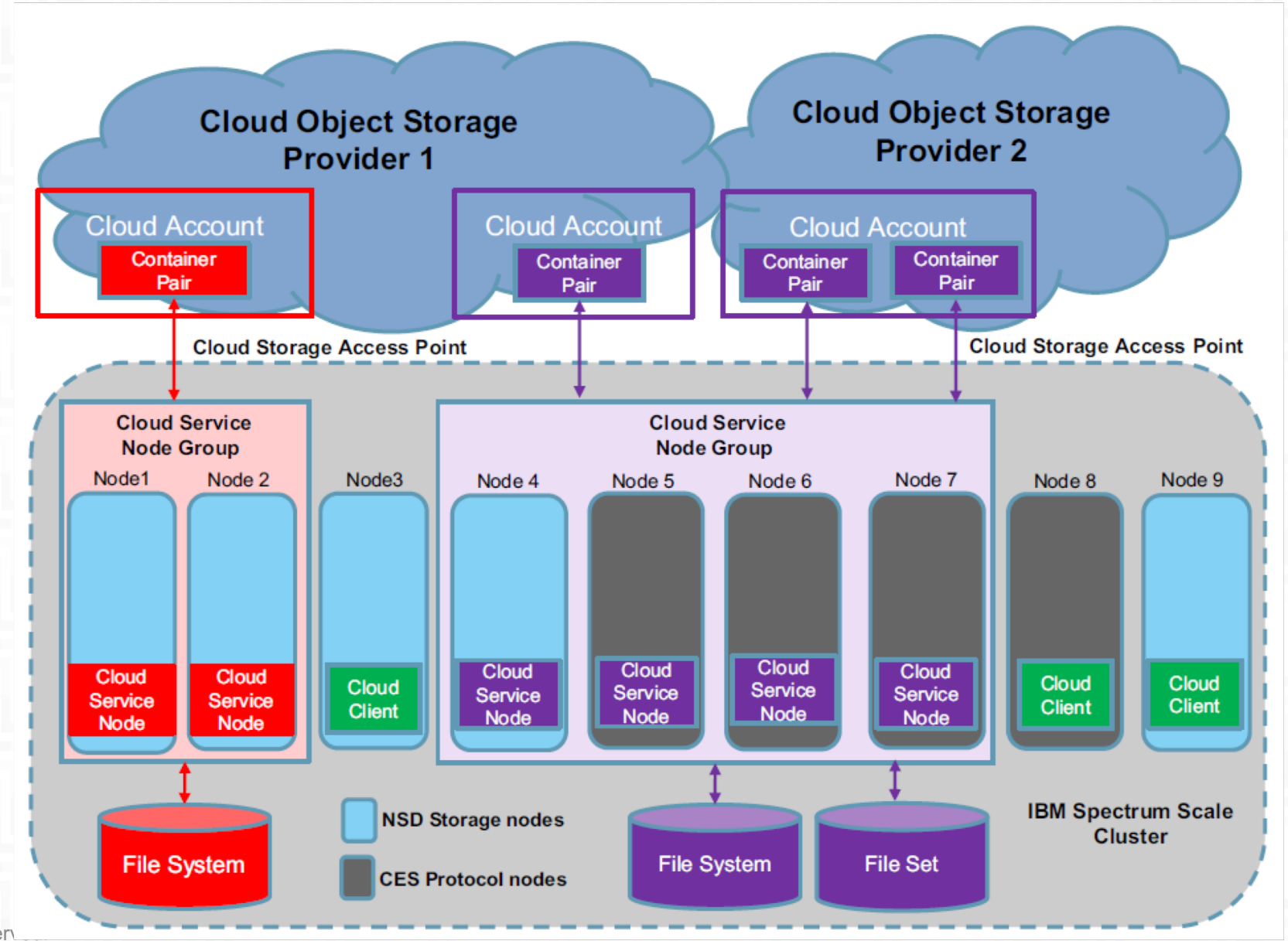
- Spectrum Scale 5.0.2 Product Documentation:
  - Installation Guide:
    Chapter 7. Installing Cloud services on IBM Spectrum Scale nodes. pp387-392
  - Administration Guide:
    Chapter 6. Configuring and tuning your system for Cloud services. pp91-122
  - Command and Programming Reference: mmcloudgateway command. pp239-266
- Spectrum Scale Knowledge Center: „Planning for Cloud services”
  - https://www.ibm.com/support/knowledgecenter/STXKQY_5.0.2/com.ibm.spectrum.scale.v5r02.doc/bl1ins_planning_MCS.htm
- IBM Redbooks (one for **tiering**, one for data sharing):
  - **Enabling Hybrid Cloud Storage for IBM Spectrum Scale Using Transparent Cloud Tiering**
    http://www.redbooks.ibm.com/abstracts/redp5411.html?Open
  - Cloud Data Sharing with IBM Spectrum Scale
    http://www.redbooks.ibm.com/abstracts/redp5419.html?Open





- Spectrum Scale User Group Talks → www.spectrumscale.org

Transparent Cloud Tiering: Components and Terminology



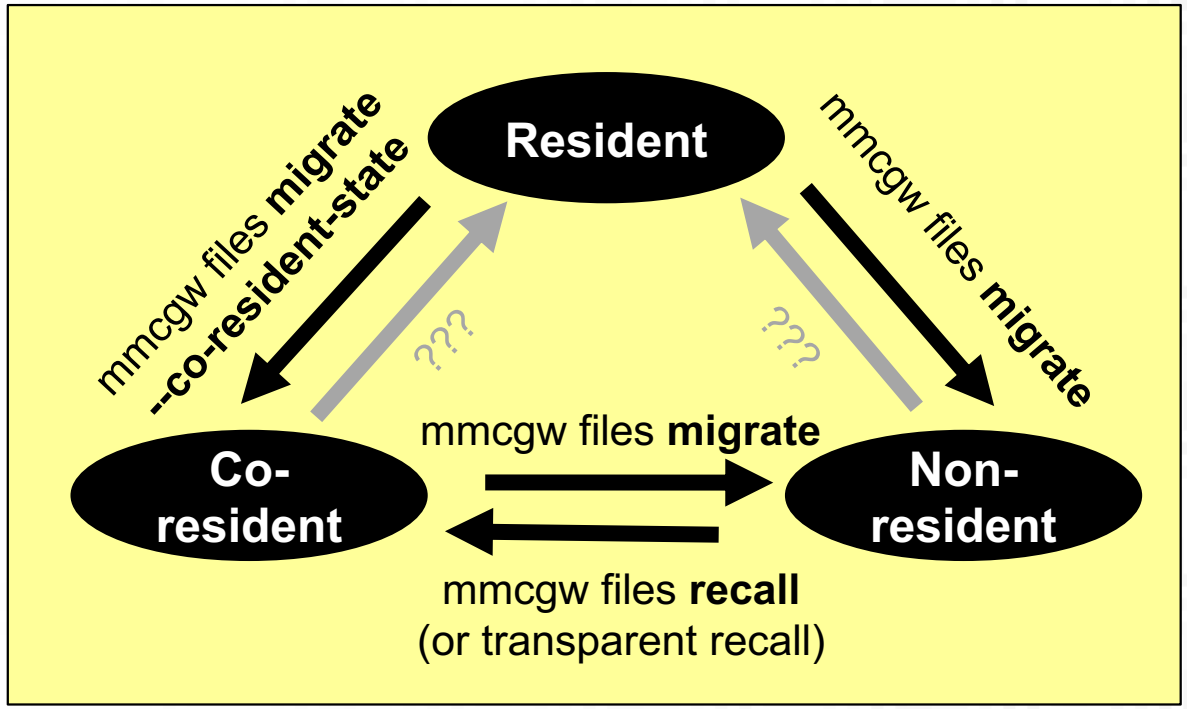
TCT Setup for Ceph, 1 of 2

Ask an IBM (or Lenovo) technical specialist for instructions how to enable TCT. It is not enabled by default, to ensure a review of the intended use case is done before using TCT.

```
mmcrnodeclass tct_group1 -N tct2201
mmchnode --cloud-gateway-enable -N tct_group1
mmcloudgateway service start -N tct_group1

echo "odgmAbQLZTwxS1h2EKkjaooO0KCFXhOZjYYkY4w1" > ~/ceph_user1_password.txt
mmcloudgateway account create
  --cloud-nodeclass tct_group1 \
  --account-name ceph_user1 \
  --cloud-type CLEVERSAFE-NEW \
  --username 91ESX93OCJ8D1SL5S4DG --pwd-file ~/ceph_user1_password.txt

mmcloudgateway cloudStorageAccessPoint create \
  --cloud-nodeclass tct_group1 \
  --account-name ceph_user1 \
  --cloud-storage-access-point-name csap1-ceph_user1 \
  --url "http://is-sr650a-tct.ses.eu.lenovo.com:7480"
```

TCT Setup for Ceph, 2 of 2

```
mmcloudgateway cloudService create \
  --cloud-nodeclass tct_group1 \
  --cloud-service-name tct_tiering1-ceph_user1 \
  --cloud-service-type Tiering \
  --account-name ceph_user1

mmcloudgateway containerPairSet create \
  --cloud-nodeclass tct_group1 \
  --container-pair-set-name tct_fs1_container1 \
  --cloud-service-name tct_tiering1-ceph_user1 \
  --scope-to-filesystem --path /gpfs/tct_fs1 \
  --enc DISABLE \
  --transparent-recalls ENABLE

mmcloudgateway maintenance list
```




```
mmcloudgateway files \
  { migrate | recall | restore |
    delete | destroy | reconcile |
    list | cloudList | ...}
```

Performance Observations ... Work in Progress ;-)

- Always test the networking layer first ... using **`iperf3`** or similar tools
- Basic testing of the the Ceph **`radosgw`** performance can be done with **`s3cmd`**
  - Beware the default **`--multipart-chunk-size-mb=15`**, increase it to up to **`5120`** (5 GiB)
  - **`s3cmd put`** performance is unreliable: delay in beginning, stalls ... watch the network (**`dstat`**)
  - **`s3cmd get`** is more reliable; single **`get`** achieves ~300MiB/s, 4x in parallel saturate 10GbE link
- Explicit **`migrate`** and **`recall`** tests with **`mmcloudgateway files`**
  - **`recall`** runs at 10GbE wirespeed, even if only a single file is recalled (threaded multipart get?)
  - **`migrate`** is slower: ~100MiB/s for single file, scales up to ~300MiB/s for parallel migrates
    - Beware the CSAP's **`--mpu-parts-size`** max=default of 128MiB ... this needs to increase!
- Usage of a second **`cloudStorageAccessPoint`** is erratic
  - Single file always seems to go through single CSAP
  - For multiple files (e.g. parallel **`recall`** calls), 2nd CSAP is sometimes used, but unpredictable

thanks.

mhennecke @ lenovo.com