KISTI Report on EOS operations for ALICE Experiment

EOS Workshop

14-15 March 2024

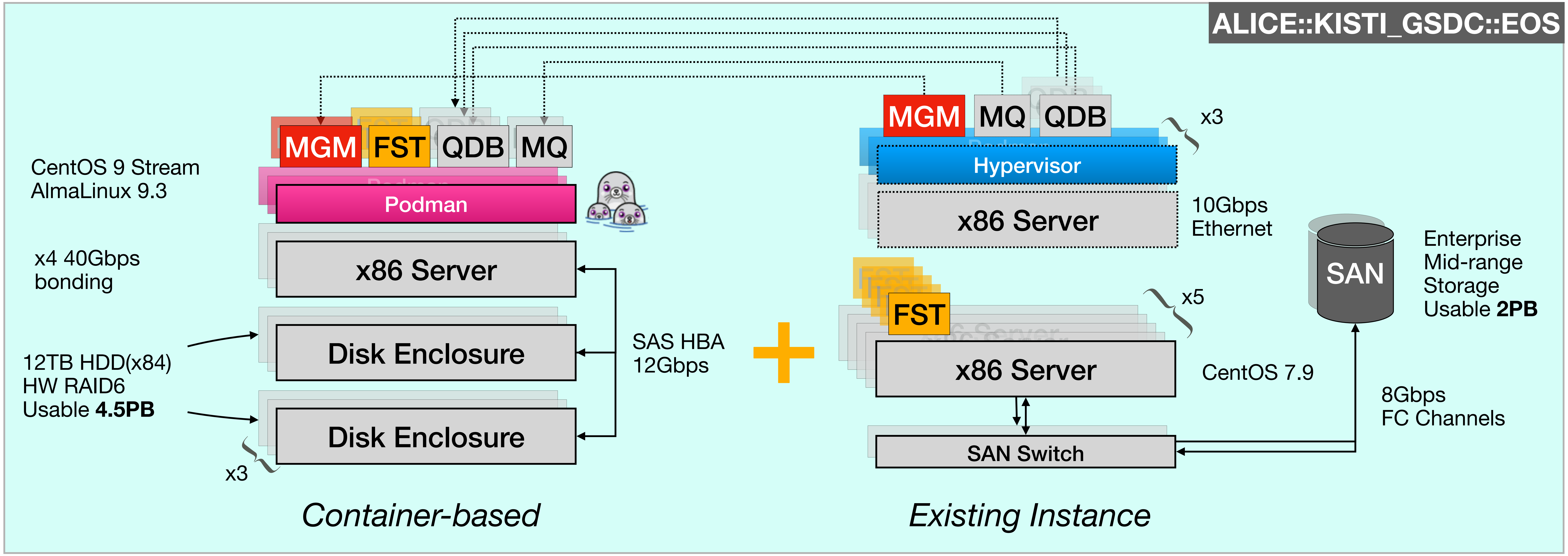
Jeongheon Kim, Sang-Un Ahn

Contents

- EOS Deployments Overview
- Operations for ALICE Experiment
- EOS v5 Deployment on EL9 using Podman
- Plan
- Summary

EOS Deployments (1/2)

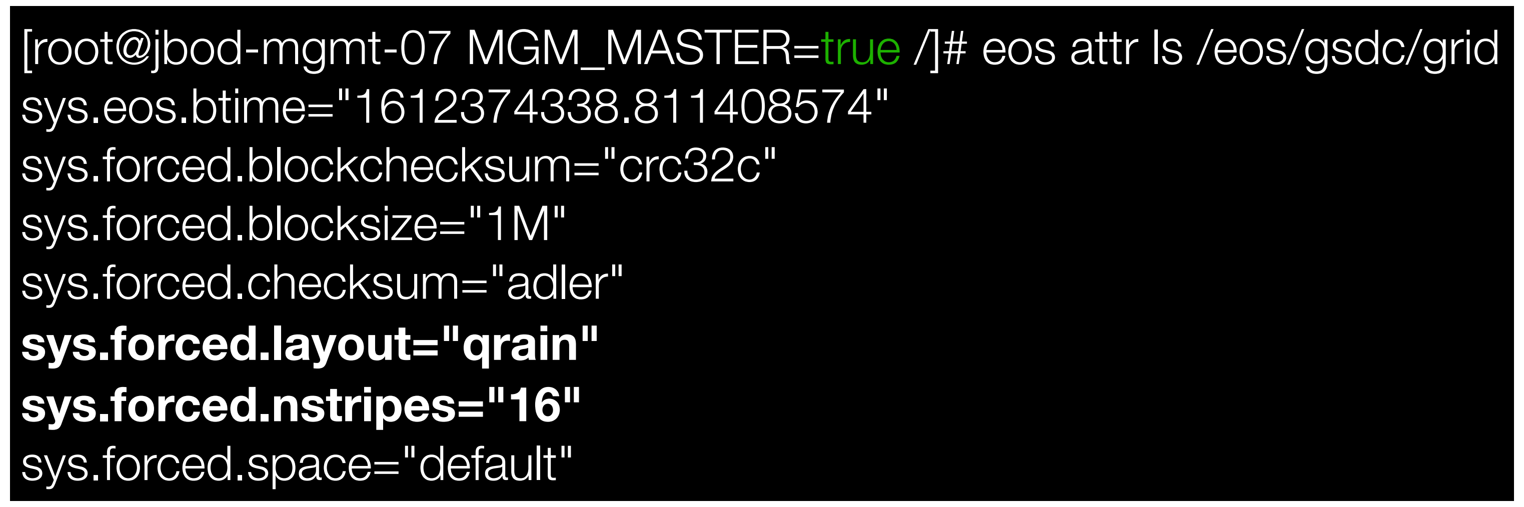
Disk Storage



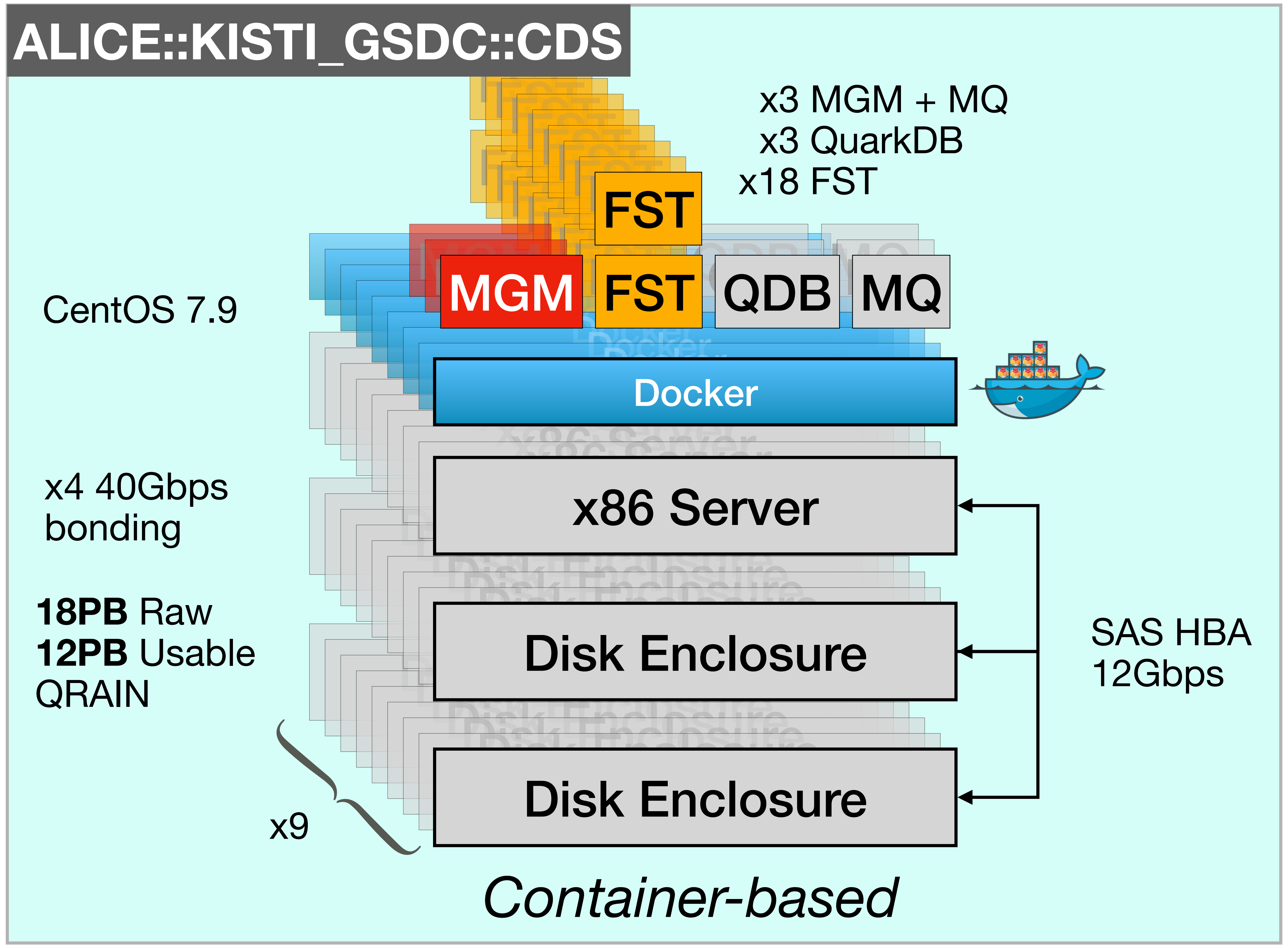
- Transparent transition of MGM and QuarkDB clusters from VMs to Containers
- EOS upgrade from 5.1.22 to 5.2.16 for existing setup, FMD migration from LevelDB completed beforehand
- Expanded to 6.5PB

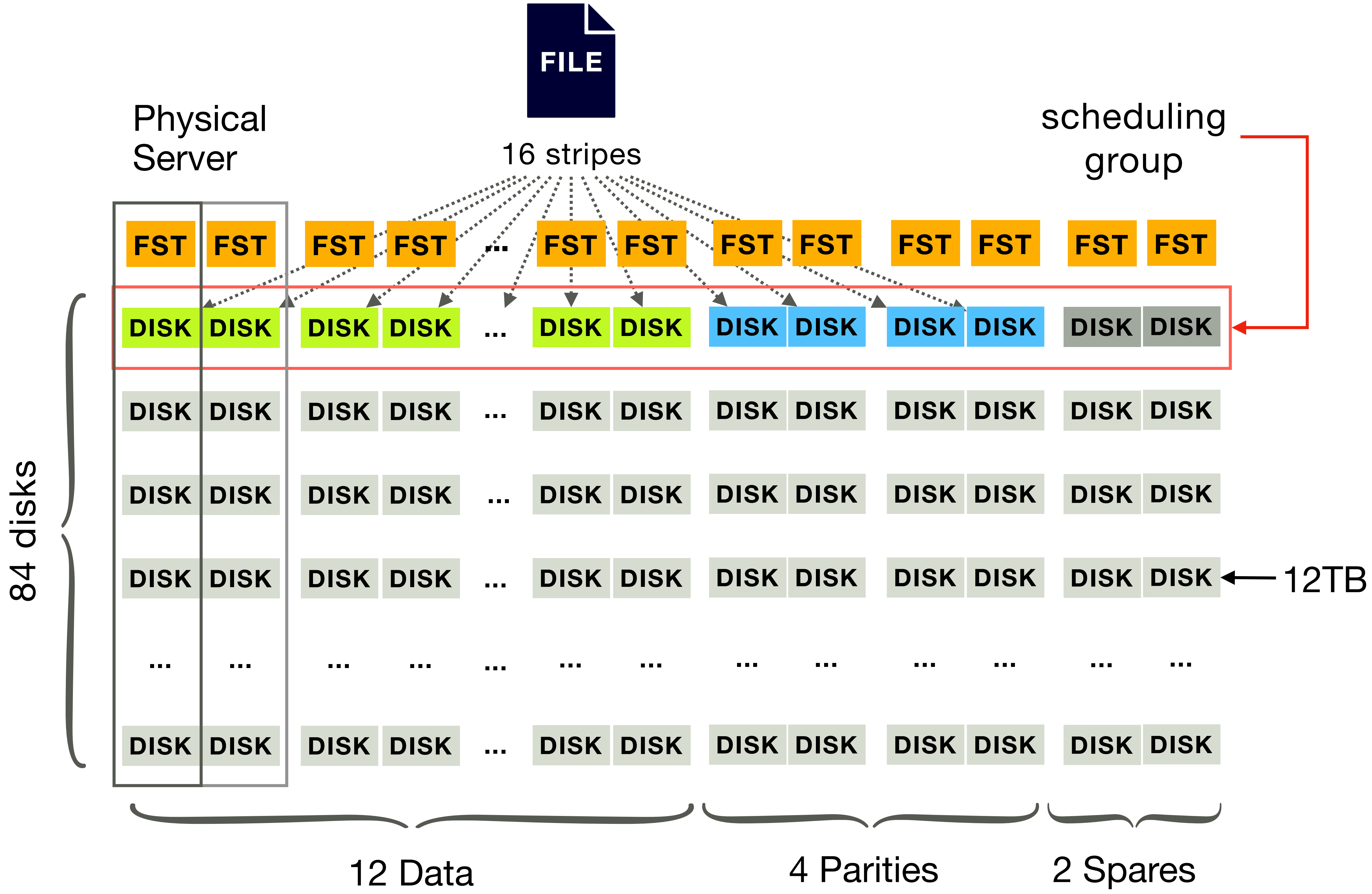
EOS Deployments (2/2)

Custodial Storage



```
[root@jbod-mgmt-07 MGM_MASTER=true /]# eos attr ls /eos/gsdc/grid
sys.eos.btime="1612374338.811408574"
sys.forced.blockchecksum="crc32c"
sys.forced.blocksize="1M"
sys.forced.checksum="adler"
sys.forced.layout="qrain"
sys.forced.nstripes="16"
sys.forced.space="default"
```

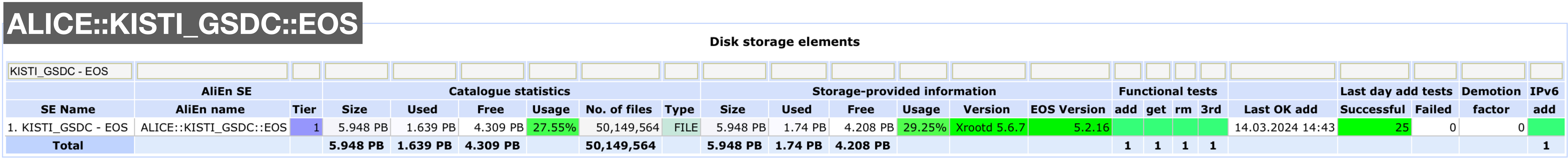




- Disk-based Raw Archive storage for ALICE in production since 2021 deployed using Docker Container
- Comparable level of data protection provided by QRAIN Layout (12 stripes + 4 parities + 2 spares)
- Successful upgrade to v5.1.22 from v4.8.82 (May 2023)

EOS @ KISTI for ALICE

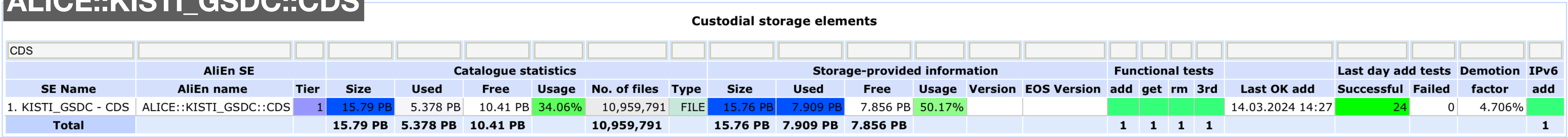
ALICE::KISTI_GSDC::EOS



Disk storage elements

	AliEn SE		Catalogue statistics						Storage-provided information						Functional tests					Last day add tests		Demotion	IPv6
KISTI_GSDC - EOS																							
SE Name	AliEn name	Tier	Size	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage	Version	EOS Version	add	get	rm	3rd	Last OK add	Successful	Failed	factor	add
1. KISTI_GSDC - EOS	ALICE::KISTI_GSDC::EOS	1	5.948 PB	1.639 PB	4.309 PB	27.55%	50,149,564	FILE	5.948 PB	1.74 PB	4.208 PB	29.25%	Xrootd 5.6.7	5.2.16	🟩	🟩	🟩	🟩	14.03.2024 14:43	25	0	0	🟩
Total			**5.948 PB**	**1.639 PB**	**4.309 PB**		**50,149,564**		**5.948 PB**	**1.74 PB**	**4.208 PB**				**1**	**1**	**1**	**1**					**1**

ALICE::KISTI_GSDC::CDS



Custodial storage elements

	AliEn SE		Catalogue statistics						Storage-provided information						Functional tests					Last day add tests		Demotion	IPv6
CDS																							
SE Name	AliEn name	Tier	Size	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage	Version	EOS Version	add	get	rm	3rd	Last OK add	Successful	Failed	factor	add
1. KISTI_GSDC - CDS	ALICE::KISTI_GSDC::CDS	1	15.79 PB	5.378 PB	10.41 PB	34.06%	10,959,791	FILE	15.76 PB	7.909 PB	7.856 PB	50.17%			🟩	🟩	🟩	🟩	14.03.2024 14:27	24	0	4.706%	🟩
Total			**15.79 PB**	**5.378 PB**	**10.41 PB**		**10,959,791**		**15.76 PB**	**7.909 PB**	**7.856 PB**				**1**	**1**	**1**	**1**					**1**

- IPv4/IPv6 Dual Stack
- ALICE-Specific Token Authentication/Authorization, HTTP(S), Third-Party Copy enabled
- MLSensor (successor of EOS Apmon) deployed for monitoring (not yet for CDS)

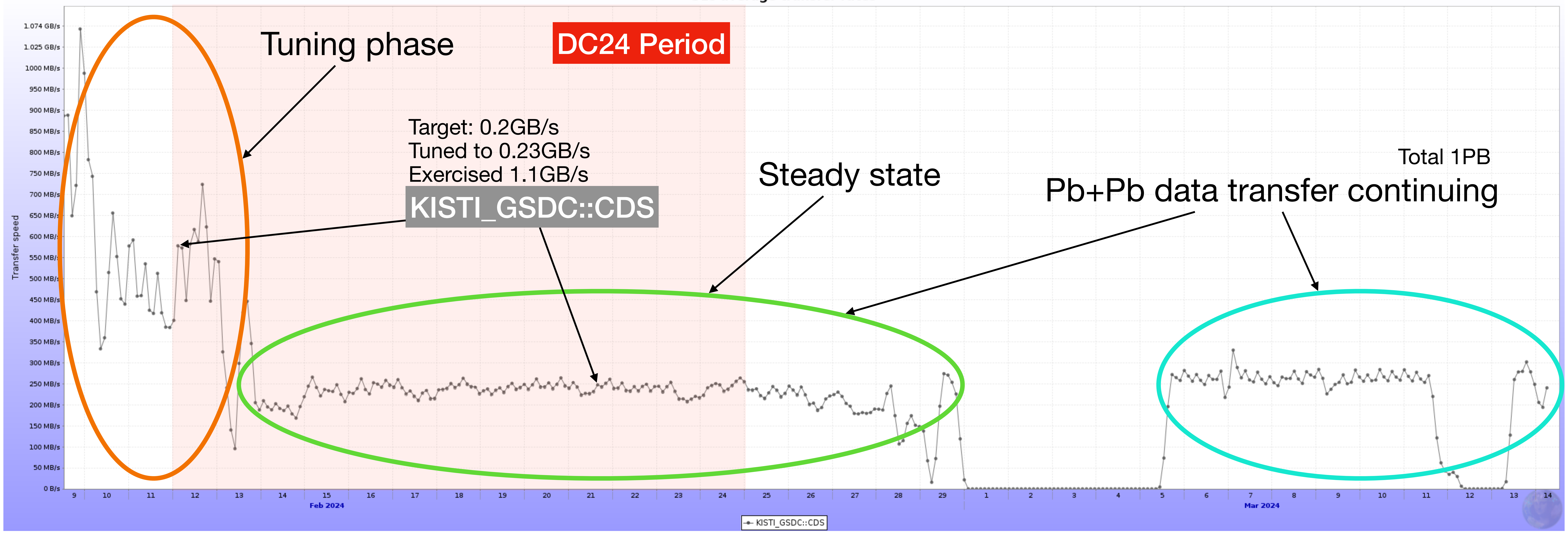
WLCG Data Challenge 24

CDS Participation as a Tape

Centre	Target rate GB/s	Average achieved GB/s
CNAF	0.8	0.98 (+20%)
IN2P3	0.4	0.6 (+40%)
KISTI	0.2	0.25 (+22%)
GridKA	0.6	1.12 (+90%)
NDGF	0.3	0.35 (+15%)
NL-T1	0.1	0.25 (+150%)
RAL	0.1	0.58 (+500%)
CERN	*10*	*14.2 (+40%)*

- Transfer of real Pb+Pb data collected in 2023, 34PB in total
- 1PB of data being transferred after the challenge, ETA end of March

SEs average transfer rates



Tuning phase

DC24 Period

Target: 0.2GB/s
Tuned to 0.23GB/s
Exercised 1.1GB/s

KISTI_GSDC::CDS

Steady state

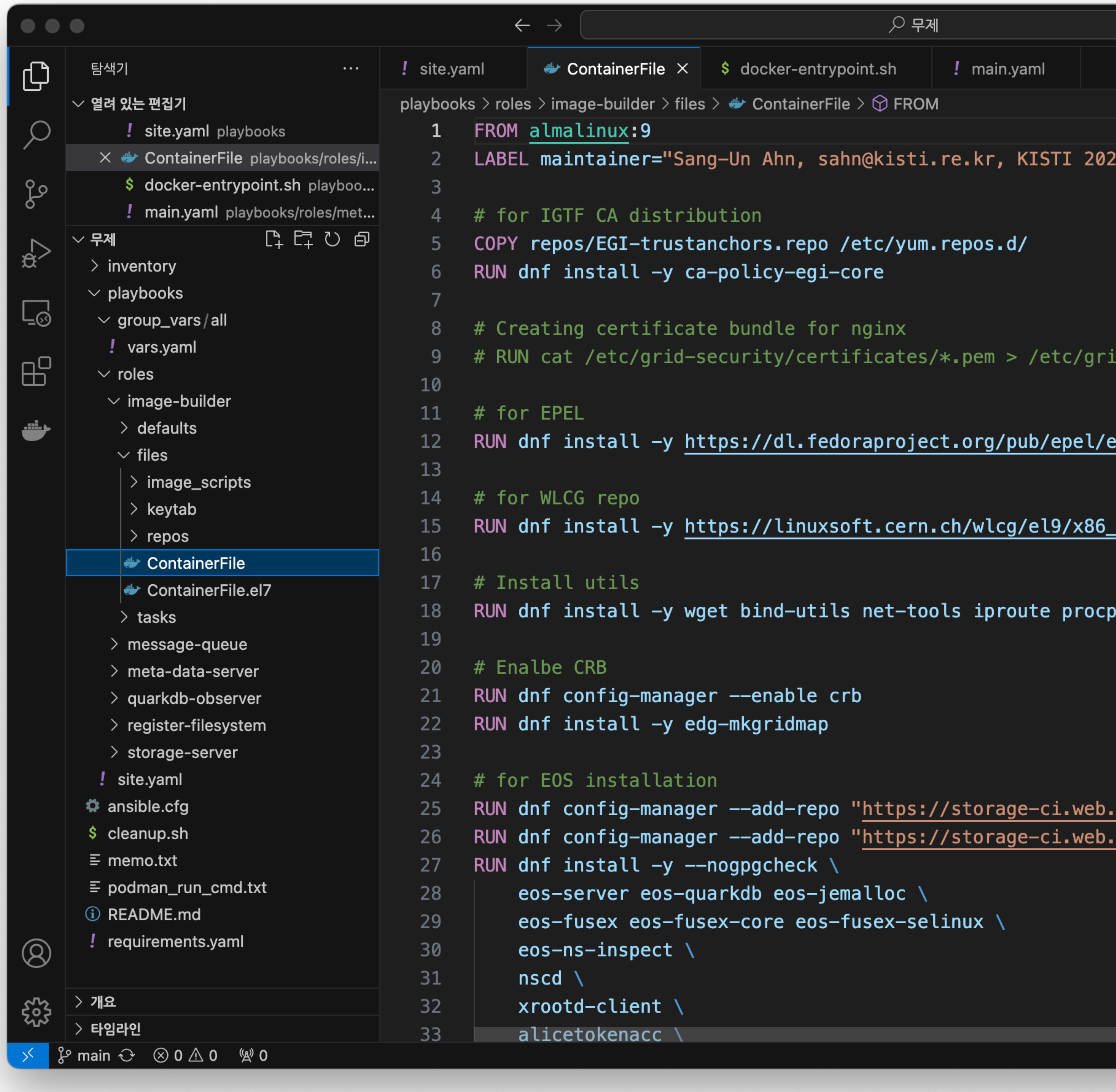
Total 1PB

Pb+Pb data transfer continuing

EOS v5 Container on EL9: Practices (1/3)

Podman container runtime

- Exploiting native support of EL9 for Podman (daemonless)
- ContainerFile (cf. *Dockerfile*)
    - EL9 base images - { CentOS 9 Stream | AlmaLinux 9.3 }
    - EOS v5 EL9 release installation
    - Reused Container entry script for CDS deployment (Docker based)
        - A few modification made to accommodate different monitoring scheme of ALICE: *eosapmond →mlsensor*

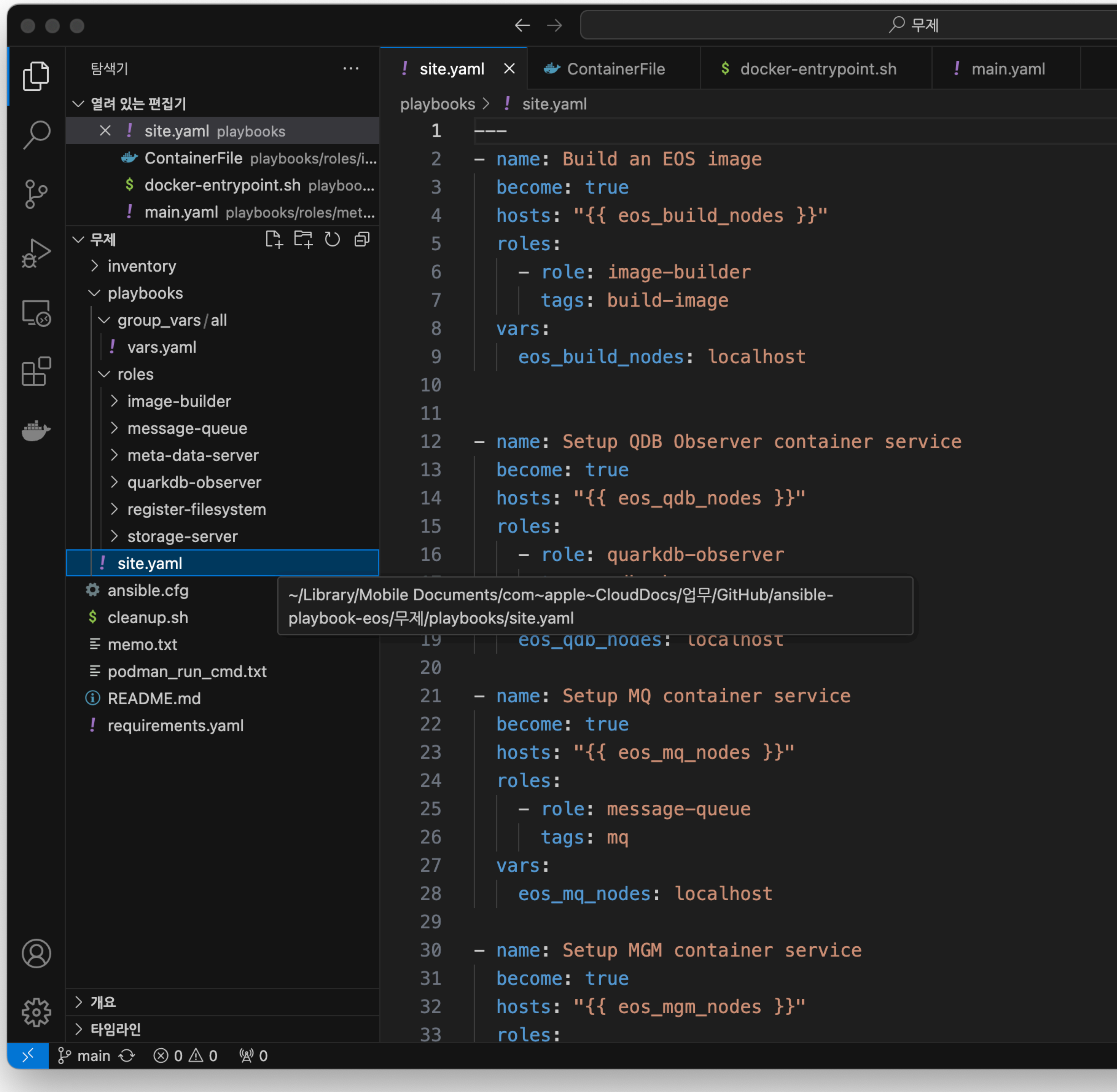


```
FROM almalinux:9
LABEL maintainer="Sang-Un Ahn, sahn@kisti.re.kr, KISTI 202

# for IGTF CA distribution
COPY repos/EGI-trustanchors.repo /etc/yum.repos.d/
RUN dnf install -y ca-policy-egi-core

# Creating certificate bundle for nginx
# RUN cat /etc/grid-security/certificates/*.pem > /etc/gri

# for EPEL
RUN dnf install -y https://dl.fedoraproject.org/pub/epel/e

# for WLCG repo
RUN dnf install -y https://linuxsoft.cern.ch/wlcg/el9/x86_

# Install utils
RUN dnf install -y wget bind-utils net-tools iproute procp

# Enalbe CRB
RUN dnf config-manager --enable crb
RUN dnf install -y edg-mkgridmap

# for EOS installation
RUN dnf config-manager --add-repo "https://storage-ci.web.
RUN dnf config-manager --add-repo "https://storage-ci.web.
RUN dnf install -y --nogpgcheck \
    eos-server eos-quarkdb eos-jemalloc \
    eos-fusex eos-fusex-core eos-fusex-selinux \
    eos-ns-inspect \
    nscd \
    xrootd-client \
    alicetokenacc \
```

EOS v5 Container on EL9: Practices (2/3)

Automation via Ansible Playbook

- Playbook structure:
    - *site.yaml* - tags defined to perform specific operation (role) in an automated way
    - group_vars
        - *vars.yaml* - key-value variables for group parameters such as *eos_instance_name*, *eos_geotag*, ports, master/slave MGM FQDNs, QDB cluster and FST data directories
    - roles
        - *image-builder* | *message-queue* (MQ) | *meta-data-server* (MGM) | *quarkdb-observer* (QDB) | *register-filesystem* | *storage-server* (FST)
        - *handlers* defined to invoke *firewalld* policy implementation and *systemd* integration
        - Creating essential configuration files by templating *xrd.cf.{qdb|mq|mgm|fst}*, *eos_env*, *scitokens*, ALICE-specific (*TkAuthz.Authorization & mlsensor)*, etc.



```yaml
---
- name: Build an EOS image
  become: true
  hosts: "{{ eos_build_nodes }}"
  roles:
    - role: image-builder
      tags: build-image
  vars:
    eos_build_nodes: localhost


- name: Setup QDB Observer container service
  become: true
  hosts: "{{ eos_qdb_nodes }}"
  roles:
    - role: quarkdb-observer
    eos_qdb_nodes: localhost

- name: Setup MQ container service
  become: true
  hosts: "{{ eos_mq_nodes }}"
  roles:
    - role: message-queue
      tags: mq
  vars:
    eos_mq_nodes: localhost

- name: Setup MGM container service
  become: true
  hosts: "{{ eos_mgm_nodes }}"
  roles:
```

~/Library/Mobile Documents/com~apple~CloudDocs/업무/GitHub/ansible-playbook-eos/무제/playbooks/site.yaml

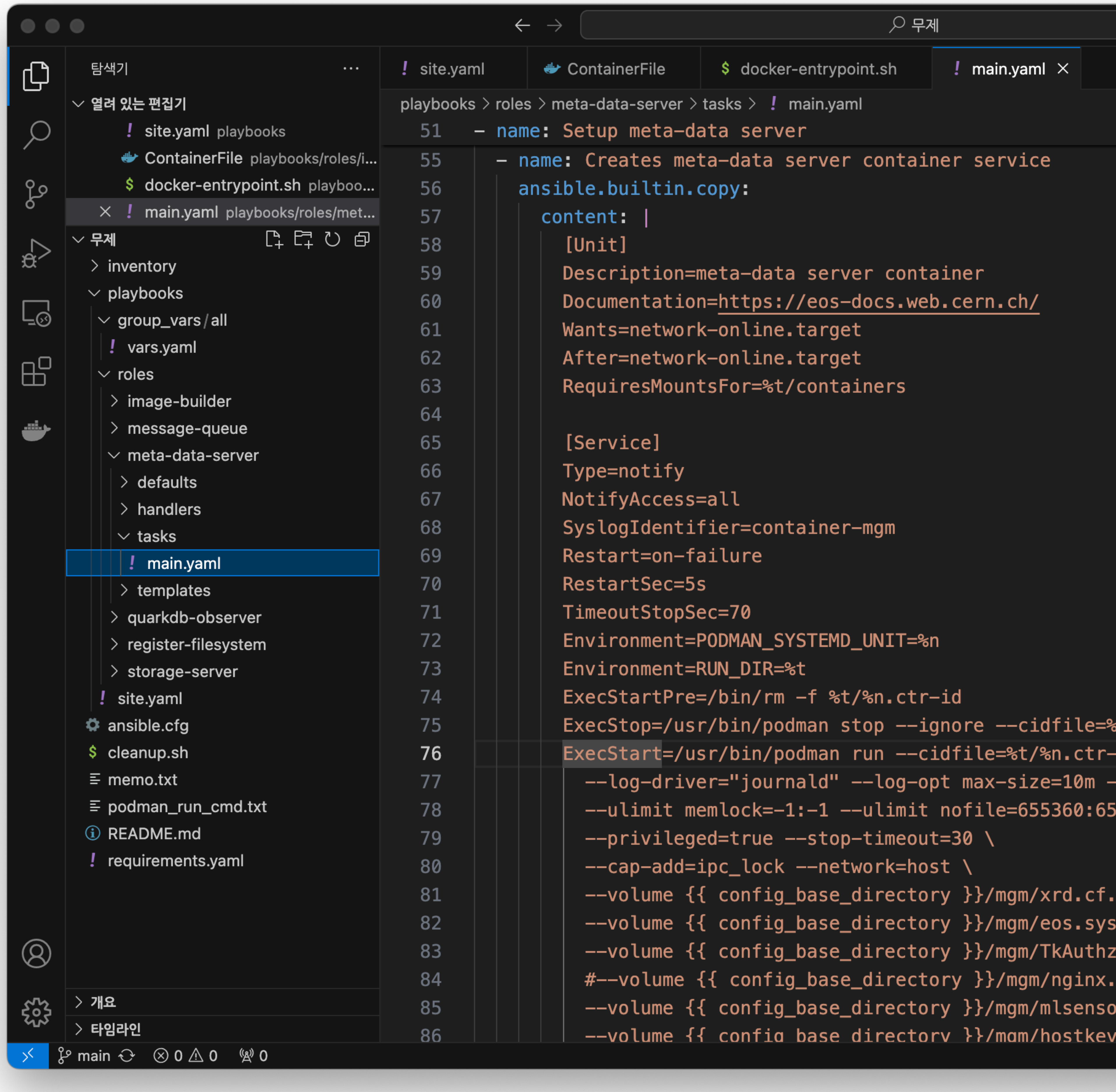
EOS v5 Container on EL9: Practices (3/3)

Systemd Integration

- Systemd service file for each of EOS components manipulating podman commands in such a way that it invokes *podman {run|rm|stop|...}*
    - E.g. */etc/systemd/system/{qdb|mq|fst|mgm}-container.service*

      <...>
      ExecStart=/usr/bin/podman run < parameters >
      ExecStop=/usr/bin/podman stop
      ExecStopPost=/usr/bin/podman rm
      <...>

    - *systemdctl {start|stop|restart} {qdb|mq|fst|mgm}-container.service*
        - *syslog* (*journalctl*) traces container logs (= podman logs)
    - Service update as well as roll-back can be quick and easy
        - Update images (pulling from registries or uploading from local one)
        - *systemctl restart *-container.service*



```
- name: Setup meta-data server
  - name: Creates meta-data server container service
    ansible.builtin.copy:
      content: |
        [Unit]
        Description=meta-data server container
        Documentation=https://eos-docs.web.cern.ch/
        Wants=network-online.target
        After=network-online.target
        RequiresMountsFor=%t/containers

        [Service]
        Type=notify
        NotifyAccess=all
        SyslogIdentifier=container-mgm
        Restart=on-failure
        RestartSec=5s
        TimeoutStopSec=70
        Environment=PODMAN_SYSTEMD_UNIT=%n
        Environment=RUN_DIR=%t
        ExecStartPre=/bin/rm -f %t/%n.ctr-id
        ExecStop=/usr/bin/podman stop --ignore --cidfile=%
        ExecStart=/usr/bin/podman run --cidfile=%t/%n.ctr-
          --log-driver="journald" --log-opt max-size=10m -
          --ulimit memlock=-1:-1 --ulimit nofile=655360:65
          --privileged=true --stop-timeout=30 \
          --cap-add=ipc_lock --network=host \
          --volume {{ config_base_directory }}/mgm/xrd.cf.
          --volume {{ config_base_directory }}/mgm/eos.sys
          --volume {{ config_base_directory }}/mgm/TkAuthz
          #--volume {{ config_base_directory }}/mgm/nginx.
          --volume {{ config_base_directory }}/mgm/mlsenso
          --volume {{ config_base_directory }}/mgm/hostkev
```

Plan

- Further work on EOS deployment playbook to run on AWX system
- Expanding EOS Disk for ALICE further up to 7PB or more to meet pledges
    - FST nodes running on bare metals (2PB) to be decommissioned
        - Group draining could help to vacate there FSTs
- Updating EOS CDS to v5 as well as upgrading to EL9
    - Heavy revisions required on CDS Docker deployment

Summary

- Two EOS instances are operated for different purposes
- EOS components are deployed upon containers using Ansible playbook

Thank you