東京大学
素粒子物理国際研究センター
International Center for Elementary Particle Physics
The University of Tokyo

Experiences from ICEPP migration from DPM to dCache

GDB meeting

22nd March 2023

Masahiko Saito (UTokyo, ICEPP)

Tokyo regional center

- Support for ATLAS VO in WLCG (Tier2) and provide ATLAS-Japan dedicated resources (Tier3)
  - Only one WLCG site in Japan for ATLAS VO

- Hardware is leased and replaced every three years. The current (6th) system started in January 2022. (next: Jan. 2025)

- Tier2 (WLCG)
  - Worker nodes (ARC/HTCondor): ~11,000 cores
  - Storage (**DPM**): ~15 PB
- Tier3 (ATLAS-Japan)
  - Interactive nodes: ~ 200 cores
  - Worker nodes (HTCondor): ~ 1,800 cores
  - Storage (GPFS): 3 PB
  - GPU resources: V100, T4

History of the Tokyo site with DPM

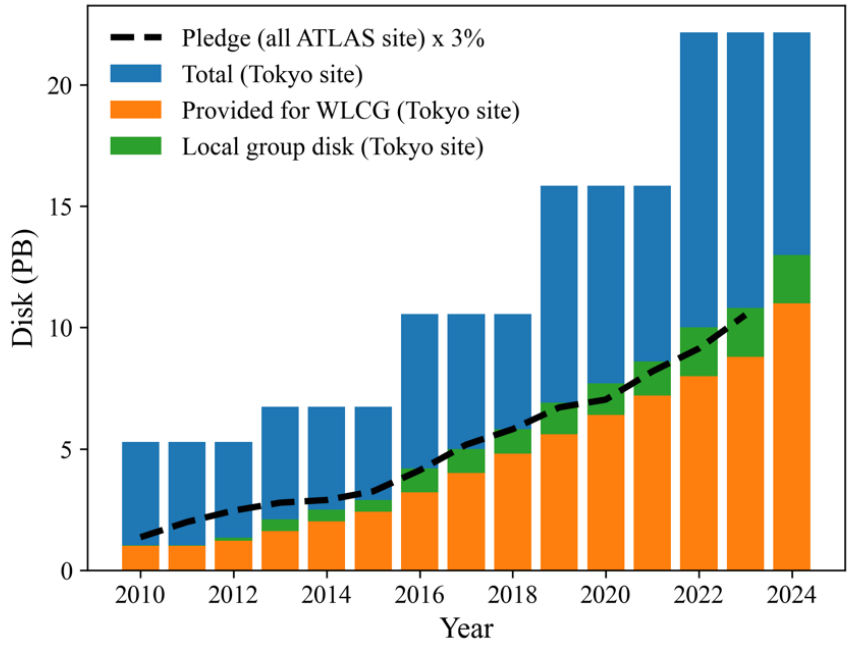
- We've been using DPM since the beginning (2006~)
  - Provided 40 TB in 2006, 200 TB in 2007
  - Storage volume has increased year on year
- In 2023 (Feb.), we provide
  - 8000 TB for pledge (~3% of all ATLAS resources)
  - 2000 TB for the Japanese users
  - Probably one of the biggest DPM user

From Tokyo tier2 talk at ISGC2008

Grid services

- gLite 3.0 (SLC3) in production
  - gLite 3.1 (SLC4, 32bit) WNs since last year
  - CE: Torque+Maui
    - LSF is used for non-Grid CPU servers
  - SE/SRM: DPM with MySQL
    - Disk servers: SLC4 x86_64, XFS (6Tbytes/filesystem)
    - Only one pool for ~190Tbytes
    - SRMv2.2: Space token deployed since February
      - For ATLAS FDR-1 and CCRC-1
  - BDII, MON, UI, RB, LFC, MyProxy

Storage resources



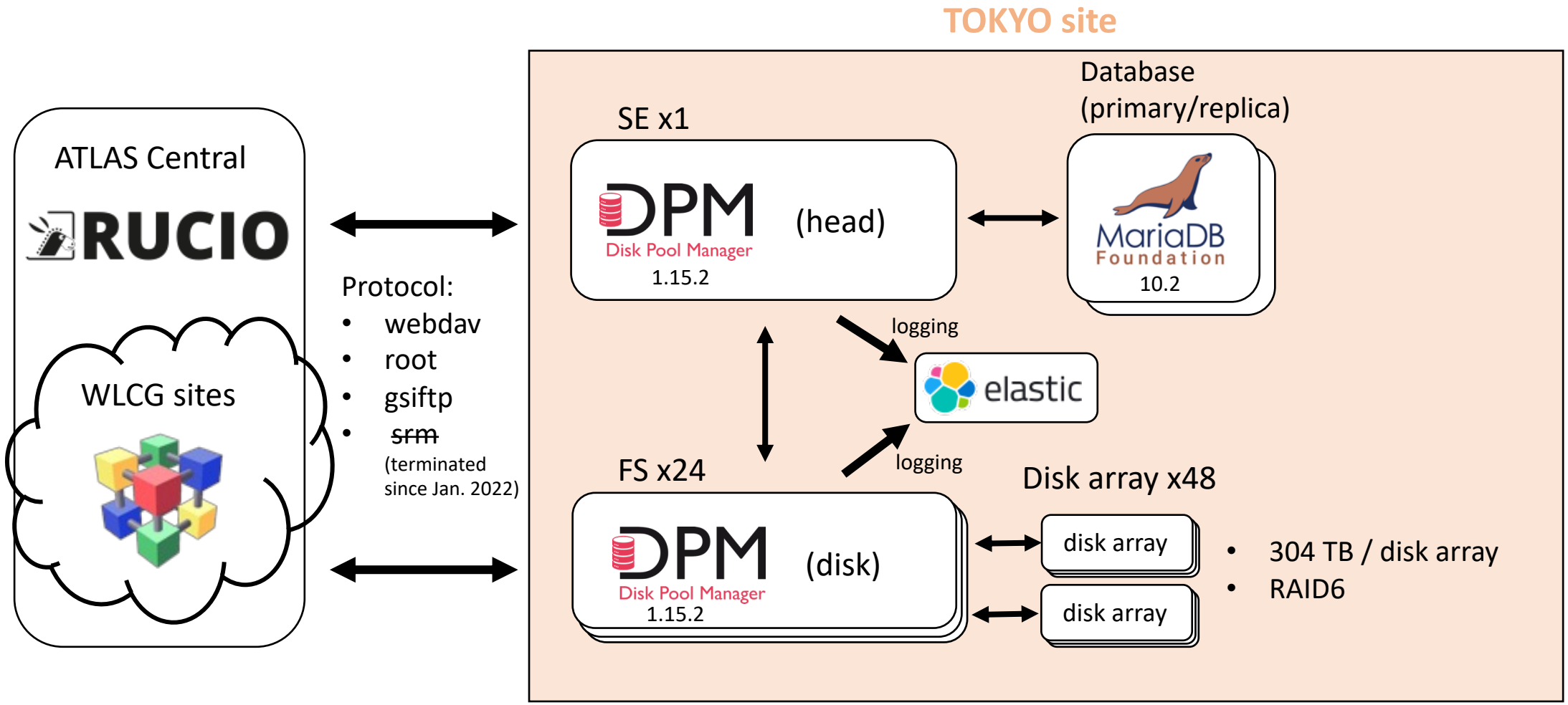
Where to migrate after DPM retirement

- DPM EOL is summer 2024. We need to migrate to other storage middleware.
- Requirements:
  - **In-place** migration: Our next system replacement is Jan 2025. No additional servers can be prepared.
  - **Easy** migration: We don't have enough person-power to get involved in the dedicated migration.

→ Decided to move into dCache

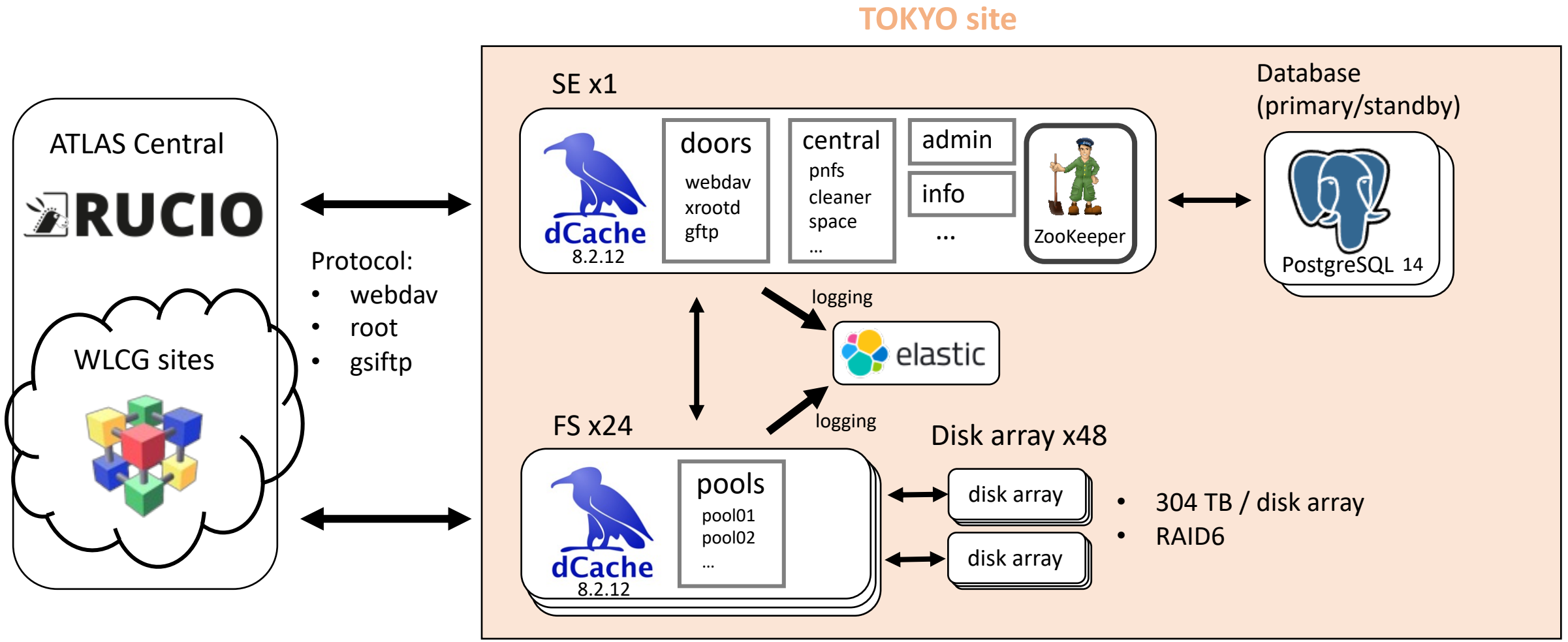
  - We don't need to copy files and to prepare additional (many) servers.
  - Good guidance document (wiki) and the mature migration script ("migrate.py")

Overview of Tokyo Tier2 storage element (SE)



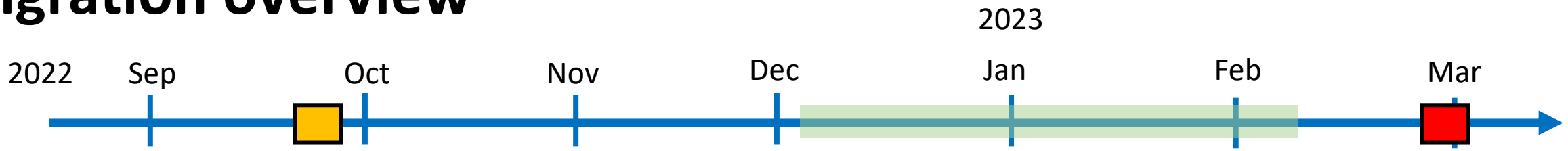
- Storage volume: 14.6 PB (provided 8 + 2 PB), 70M objects are stored
- Database size: 22 GB (mysqldump output size)

Overview of Tokyo Tier2 storage element (SE)



- Use the same servers (head/disk/db) with DPM
- Transparent to end users except for SRR URL

Migration overview



24 Sep – 26 Sep

3 days downtime
(for annual power equipment maintenance)
- Updated DPM to v1.15.2,
  which includes a migration script (migrate.py)
- Applied DB inconsistency check (see next page)

27 Feb – 1 Mar

3 days downtime for the migration (actual ~55h)
- Most of the work was completed
  in the first 2 days, but it took another day
  to understand/fix miscellaneous small issues
  in our cron scripts for monitoring, ACL etc.

Q1 2023

Test of the migration process and the dCache behaviour on test servers
- Prepared Ansible playbooks to deploy dCache, Zookeepers, PostgreSQL
  - dCache config files were prepared with reference to those generated by migrate.py
- Tuning of the dCache namespace import
  - It took 24h 20m on the first try, because of 70 M objects.
  - Improved by ~10% (21h 20m) with PyPy and “commit size=100” (for PostgreSQL query)

DB inconsistency check

Sep 2022: 1st iteration (during downtime)

(Our DPM has 70M objs (43M dirs + 27M files))

1. Fixing lost and dark data — ~ 1h
   Found 20k dark objects & 750 lost objects.
   Most of lost files were old files (pre-2013) or test files.

2. Fixing internal database inconsistencies — ~ 1h x 2 (dry run → apply)
   e.g. 23k no-replica,
   900k incorrect-spacetoken,
   3k replica-type, etc.

3. Correct pool for spacetoken
   Found 38 files (created before 2008).
   Since replica move was failed, manually deleted them.

4. Calculate missing checksums — ~ 6h
   275k files, 16.8 TB

Many inconsistencies accumulated over 15 years were found and fixed.

DB inconsistency check

(Our DPM has 70M objs (43M dirs + 27M files))

Sep 2022: 1st iteration (during downtime)
Feb 2023: 2nd iteration (during production)

1. Fixing lost and dark data
   - ~ 1h — Found 20k dark objects & 750 lost objects. Most of lost files were old files (pre-2013) or test files.
   - ~ 2h — Found only 40 dark files
2. Fixing internal database inconsistencies
   - ~ 1h x 2 (dry run → apply) — e.g. 23k no-replica, 900k incorrect-spacetoken, 3k replica-type, etc.
   - ~ 1h x 2 — No inconsistencies (except for nlink, dir-size and spacetoken-size)
3. Correct pool for spacetoken

   Found 38 files (created before 2008).
   Since replica move was failed, manually deleted them.
4. Calculate missing checksums
   - ~ 6h — 275k files, 16.8 TB
   - ~ 20m — 15k files, 56 GB

Inconsistencies didn't increase so much in the 5 months of production

3rd iteration: Re-do just before the migration procedure (found 374 files, 56 GB)

Timeline in downtime

Steps to get working dCache

0) stop DPM
1) execute DPM export script
2) distribute dCache config files
3) run commands in admin shell
4) execute dCache import script
5) link existing files
6) start dCache
   (create metadata cache)

~ 3h 20m

namespace.csv: 35GB, 69M lines

distribute by ansible

~ 21h

import of 69430406 records
completed in 75633.7s (cpu 25876.0s, eff 34.2%):
dirs 42668585, files 26761821 (replicas 26761821),
links 0, unknown 0

~ 20m

~1M files per File server

~ 1h 30m

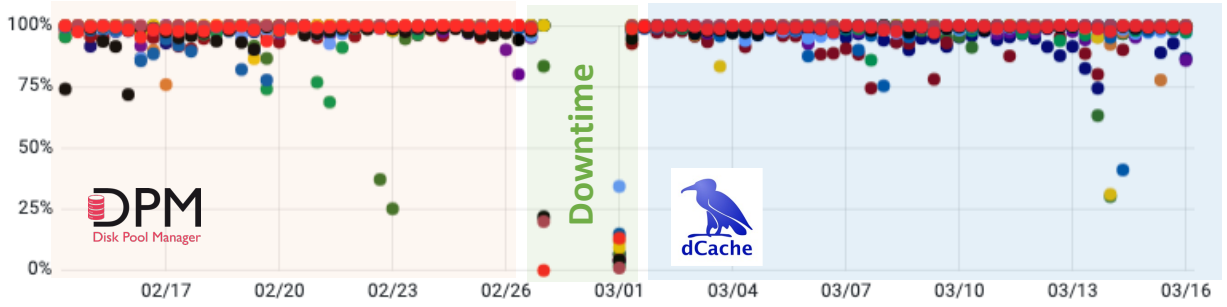
14 broken files are detected. Most of them were created around just before the downtime.

No critical issues happened. Very smooth migration!

Transfer efficiency/volume

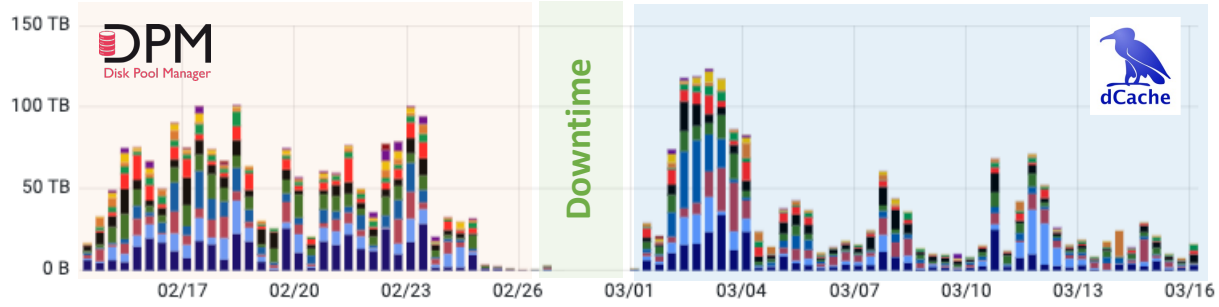
Transfers: Others → Tokyo

Transfer Efficiency



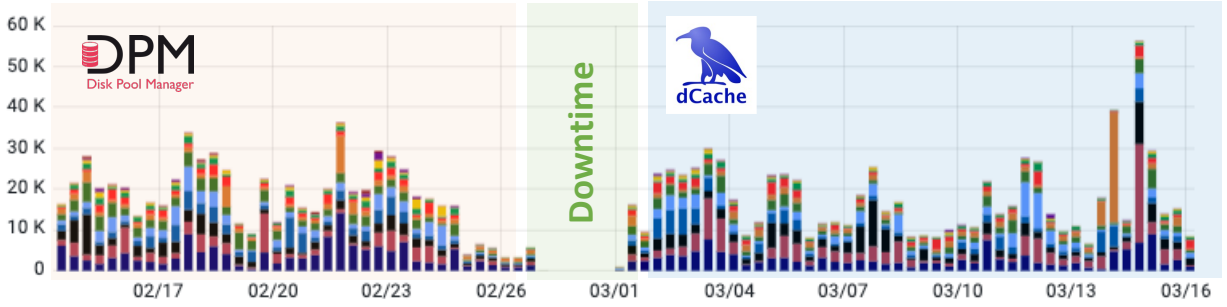
	min	max	avg	current
UK	6.71%	100%	94.4%	99.8%
TW	4.41%	100%	94.9%	100%
DE	4.16%	99.9%	95.4%	99.1%

Transfer Volume



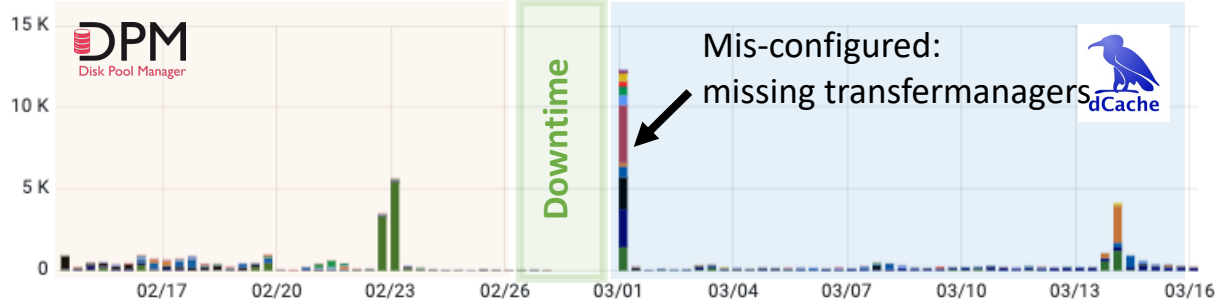
	avg	total
US	7.68 TB	699 TB
ND	4.90 TB	446 TB
CERN	4.77 TB	434 TB

Transfer Successes



	avg	total
US	3.04 K	277 K
CERN	2.36 K	214 K
DE	2.15 K	196 K

Transfer Failures



	avg	total
UK	163	14.8 K
US	97.5	8.87 K
DE	74.0	6.74 K

No issues for transfer

Summary

- Tokyo tier2 site migrated from DPM to dCache.
  - Storage volume: 8600 TB, 70 M objects
- Based on the guide, the migration was finished without critical issues.
  - DB inconsistency check was applied at the last Sep
    - It took < 12h. Many inconsistencies were found and fixed.
  - Took 55 hours of downtime to migrate 27th Feb - 1st Mar.
    - Database export/import took ~24 hours.
    - No showstoppers during migration
- Our dCache cluster is running smoothly, as is the DPM.