Science and Technology Facilities Council

Technical challenges of tape instance consolidation at RAL

Tom Byrne, Alison Packer, George Patargias, Mahalakshmi Agilandamurthy, Tim Folkes

Overview

- This is a story about one challenge we faced when moving to CTA from our previous tape system – CASTOR
  - I am not sure this will be directly useful for anybody, but I thought people might enjoy hearing about the journey

CASTOR @ RAL

- In 2022, there were two production CASTOR instances at RAL
  - For WLCG VOs – **“WLCG Tape”** CASTOR
  - For our local facilities users – **“Facilities”** CASTOR
- Each CASTOR instance had exclusive use of one of our two tape libraries
  - Each had ~30 drives
  - And ~100PB of stored data
- With CASTOR being no longer supported – our priority was getting the WLCG VOs off CASTOR comfortably before Run3

WLCG Tape migration to CTA @ RAL

- **The "WLCG Tape" CASTOR migration to our new CTA instance at RAL (Antares) was completed in mid 2022**
  - Talked about at the previous EOS workshop – was generally a smooth experience
- Our focus was now on how best to migrate the remaining CASTOR instance.
- Creating another EOS disk instance on Antares for the facilities namespace seemed like the best option
  - Closer to the model CERN use – Many EOS disk instances for one CTA instance gives opportunities to share 'Tape resources' between instances while maintaining separate namespaces, buffer capacities and authn/z methods

But one major hurdle – the possibility of namespace ID clashes between the two RAL namespaces

Tape file IDs in CTA/CASTOR

- A file on tape has a unique ID
  - This ID links the file in the namespace to a tape, file size and offset (i.e. info to retrieve the file)
  - In CASTOR this was the ns_file_id, in CTA it is the archive_file_id
- In both cases, this ID is stored on the tape for double checking/disaster recovery purposes
  - This is problematic when considering tape instance consolidation
  - Clashes need to be ‘physically’ resolved, not just metadata changes
- Since both RAL CASTOR instances started at ID 0, there is significant chance clashes exist.

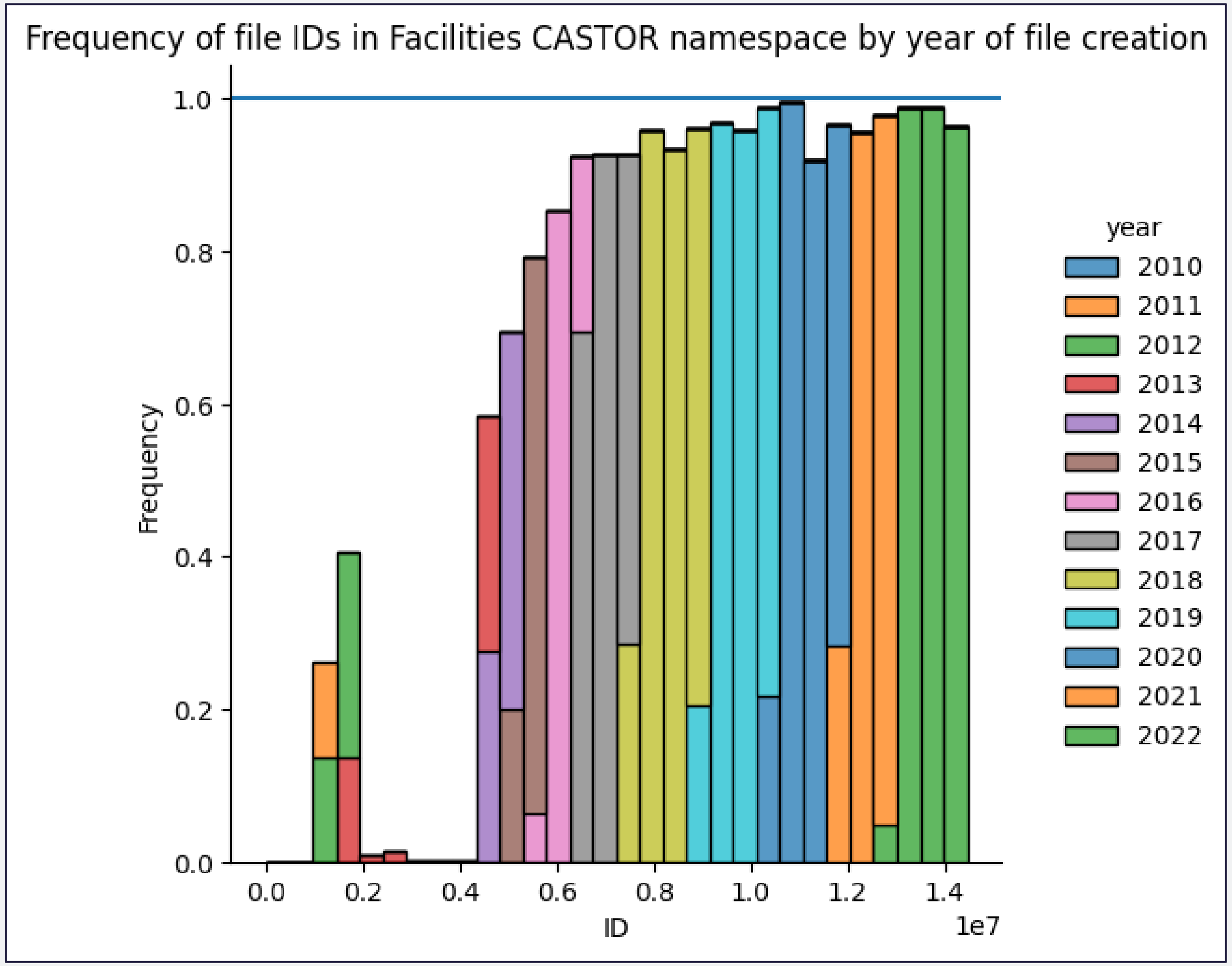
Clash resolution options

1. Copy all data from Facilities CASTOR to Antares
    - As this is a rewrite, files will get a new CTA tape file ID
    - This will be a long operation, and managing user access in the transition will be challenging
2. Resolve clashes on a case-by-case basis
    - Potentially much less work, but required understanding the scope of the clashing problem
    - Conceptually more complex than option one, multiple steps and more opportunity for mistakes

In late 2022 I carried out the analysis of the namespaces to determine the feasibility of option 2.

Facilities namespace overview

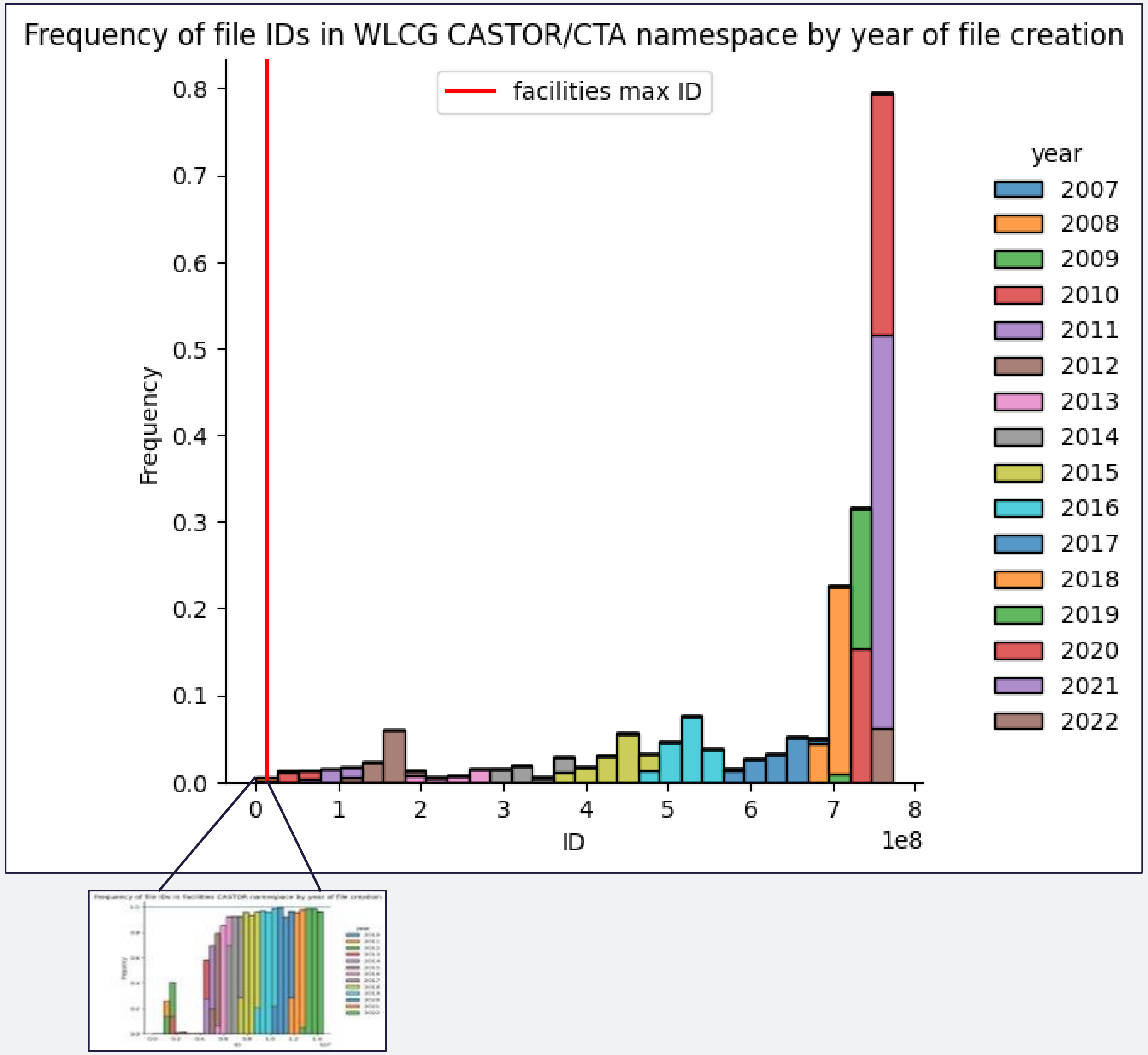
- For files created after 2017 in Facilities CASTOR, there has been essentially no churn
    - >90% of created files still exist
    - The namespace has been growing at around 1.5 million files/10PB a year since 2017
- This namespace density means that any WLCG file in the overlapping ID space is likely to clash



Science and Technology Facilities Council

WLCG CASTOR namespace overview

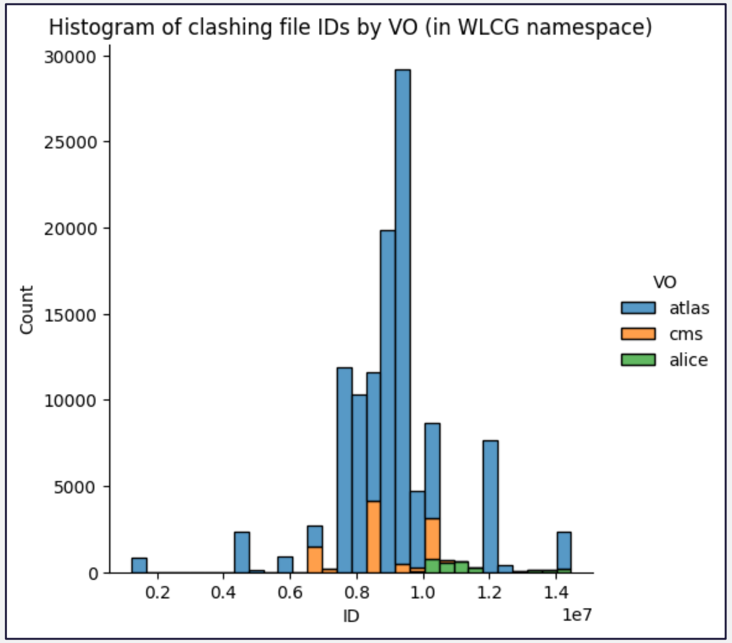
- The WLCG namespace is generally very sparse
  - This is due in part to CASTOR's dual purpose as disk storage for many years
  - Note the increase in density after 'disk only CASTOR' was removed
- Despite similar data volumes, the WLCG namespace covers a significantly larger ID space than the Facilities namespace
  - Files in the WLCG namespace in the overlapping area were created in 2007, 2008, and early 2009

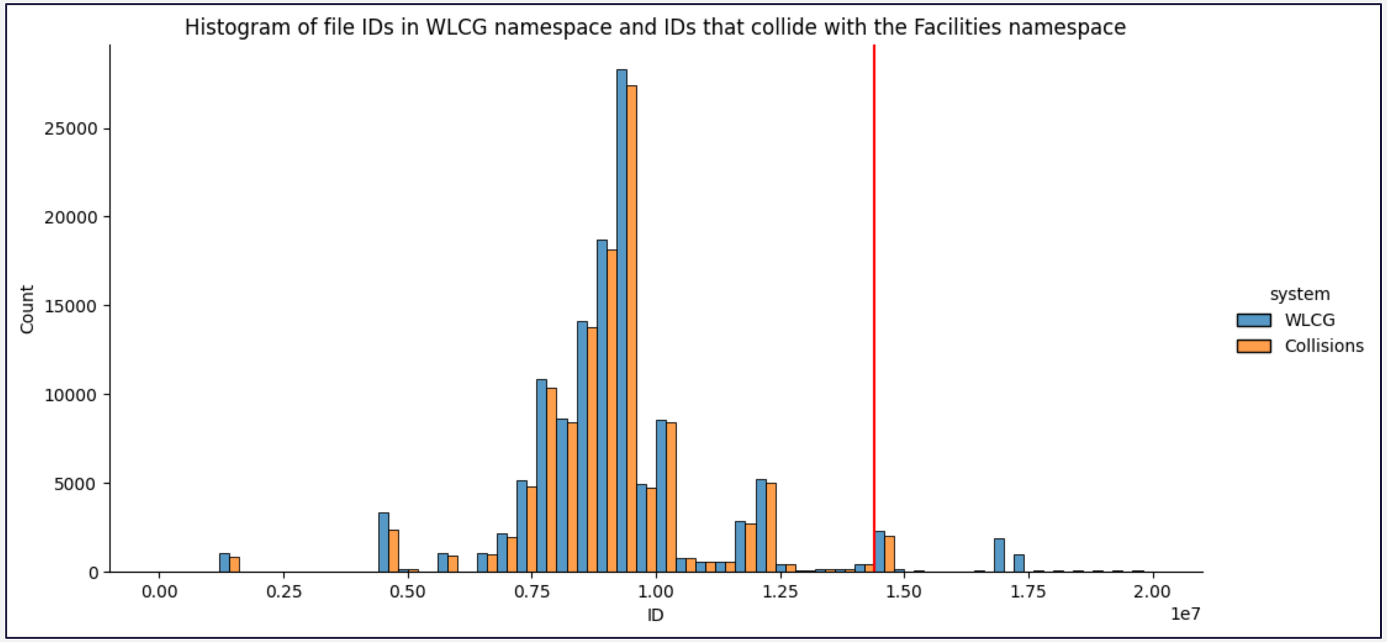


UKRI Science and Technology Facilities Council

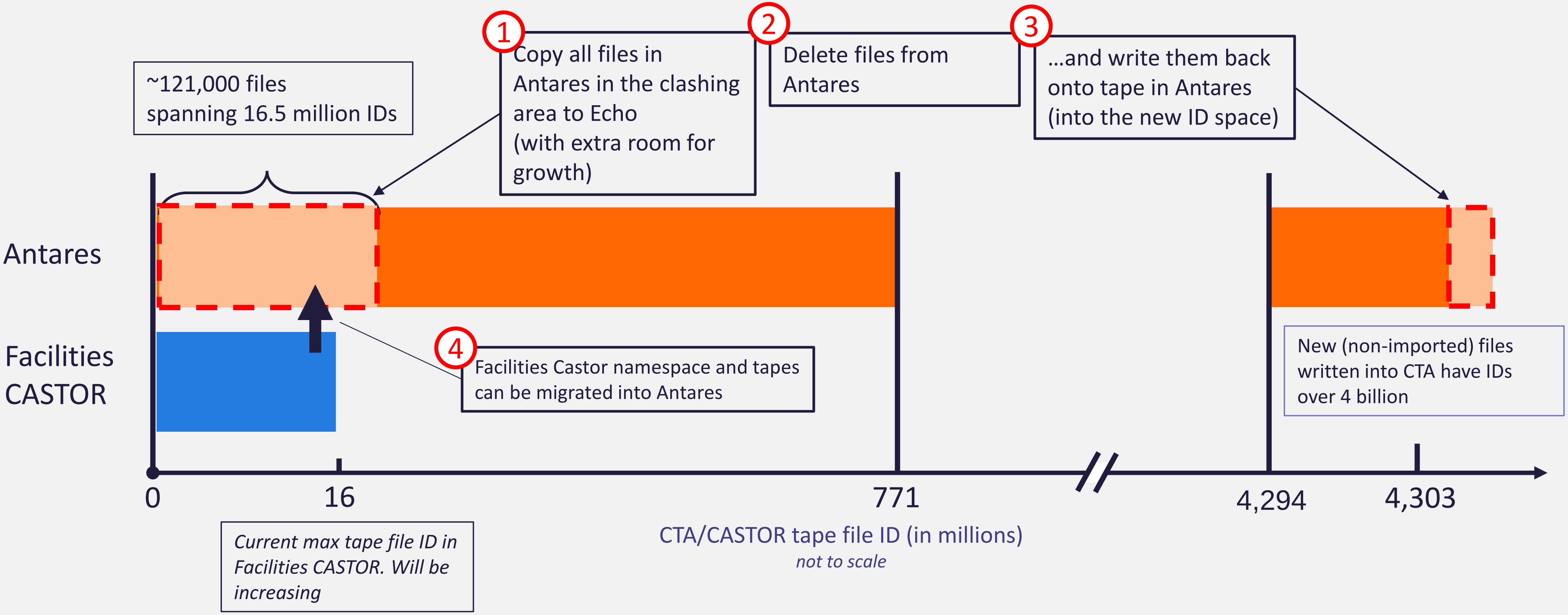
Clash analysis

- As suspected, most files in the WLCG namespace clash with the Facilities namespace
  - Dealing with all WLCG (Antares) files in the overlapping region is a negligible overhead (~3%) and logically simpler to verify there are no more collisions
- Files in the area of the WLCG namespace that clash are from ALICE, ATLAS and CMS tape pools
  - 121k files/200TB spread over 38 tapes
  - These numbers include leaving room for the facilities namespace to grow into before migration
- A very reasonable amount of clashes to deal with individually





The plan…



~121,000 files spanning 16.5 million IDs

1 Copy all files in Antares in the clashing area to Echo (with extra room for growth)

2 Delete files from Antares

3 …and write them back onto tape in Antares (into the new ID space)

4 Facilities Castor namespace and tapes can be migrated into Antares

New (non-imported) files written into CTA have IDs over 4 billion

Current max tape file ID in Facilities CASTOR. Will be increasing

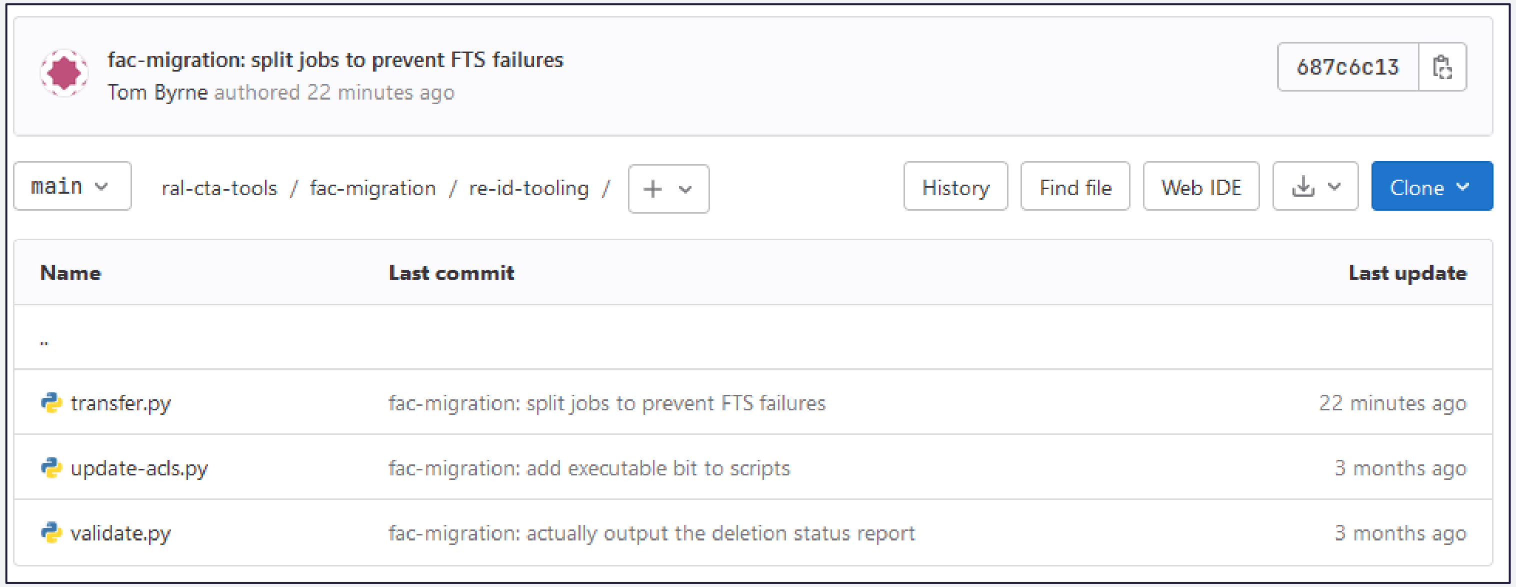
Re-ID tooling

Deleting user data from your archive is spooky business

- I wrote some tooling to try and make it as safe as possible
- All parts of the procedure operate on a tapes worth of files at a time to minimise chance of errors

1. Bulk ACL update to allow a migration user to stage, copy files to/from our disk storage and delete the files
2. FTS used to stage, transfer, and validate files moving between Echo (disk) and Antares
3. Removal step tied to validation of file in the other storage element prior to deletion

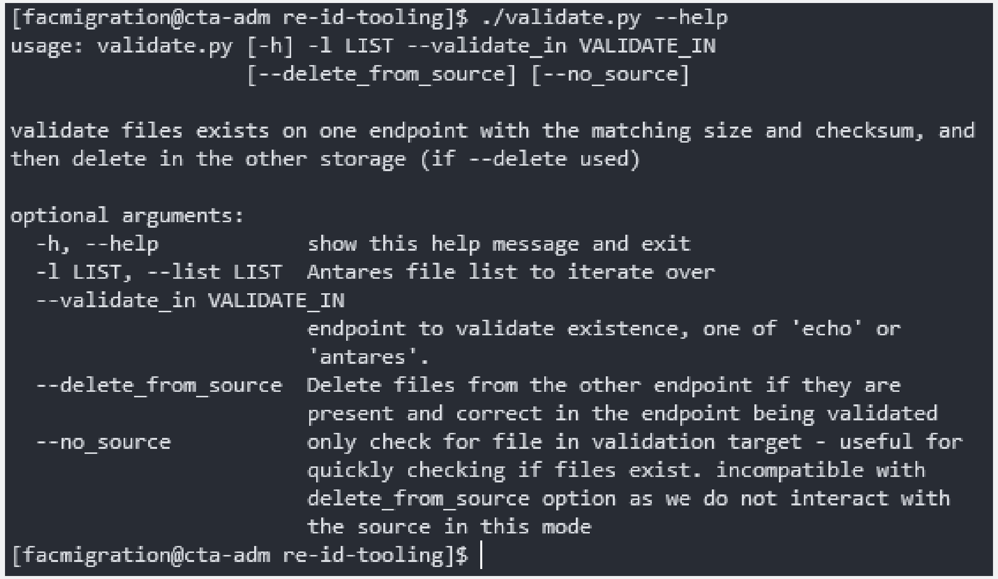
https://gitlab.stfc.ac.uk/tape/ral-cta-tools/-/tree/main/fac-migration/re-id-tooling



fac-migration: split jobs to prevent FTS failures
Tom Byrne authored 22 minutes ago

687c6c13

Name	Last commit	Last update
..		
transfer.py	fac-migration: split jobs to prevent FTS failures	22 minutes ago
update-acls.py	fac-migration: add executable bit to scripts	3 months ago
validate.py	fac-migration: actually output the deletion status report	3 months ago



```
[facmigration@cta-adm re-id-tooling]$ ./validate.py --help
usage: validate.py [-h] -l LIST --validate_in VALIDATE_IN
                   [--delete_from_source] [--no_source]

validate files exists on one endpoint with the matching size and checksum, and
then delete in the other storage (if --delete used)

optional arguments:
  -h, --help            show this help message and exit
  -l LIST, --list LIST  Antares file list to iterate over
  --validate_in VALIDATE_IN
                        endpoint to validate existence, one of 'echo' or
                        'antares'.
  --delete_from_source  Delete files from the other endpoint if they are
                        present and correct in the endpoint being validated
  --no_source           only check for file in validation target - useful for
                        quickly checking if files exist. incompatible with
                        delete_from_source option as we do not interact with
                        the source in this mode
[facmigration@cta-adm re-id-tooling]$
```

The re-ID’ing process

- Went relatively smoothly – took a few weeks in total as a background process
  - Going tape by tape made it simple to manage contention with production work
  - Keeping track of where we were in the process was interesting, having robust error checking removed a lot of the stress
- Building on the shoulders of giants is nice
  - Interacting with grid storage as a ‘user’ was surprisingly painless, FTS and XRootD python clients are well documented and very quick to get started on
  - FTS made it incredibly easy to manage transfers between our storage endpoints, dealing with staging, parallel transfers and retries
  - Successful dog food eating exercise ✓

CASTOR:

```
MAX(FILEID)
-----------
   15742892
```

CTA:

```
MIN(ARCHIVE_FILE_ID)
--------------------
            16513467
```

Science and
Technology
Facilities Council

Closing thoughts

- Tape archives can live a long time
  - 15 years in the case of CASTOR at RAL
  - Plenty of time to make decisions you (or your successors) will regret ☺
  - Given that legacy decisions are likely to cause unforeseen problems, the important thing is dealing with them with enthusiasm!
- We’re looking forward to a smooth end to the era of CASTOR at RAL
  - Thanks to all the support from everyone at CERN over the years

“Castor Canadensis ready for retirement”

Photo by Steve from washington, dc, usa - American Beaver, CC BY-SA 2.0, https://commons.wikimedia.org/w/index.php?curid=3963858

Science and Technology Facilities Council

Questions?

Science and Technology Facilities Council

Thank you

Science and Technology Facilities Council @STFC_Matters Science and Technology Facilities Council

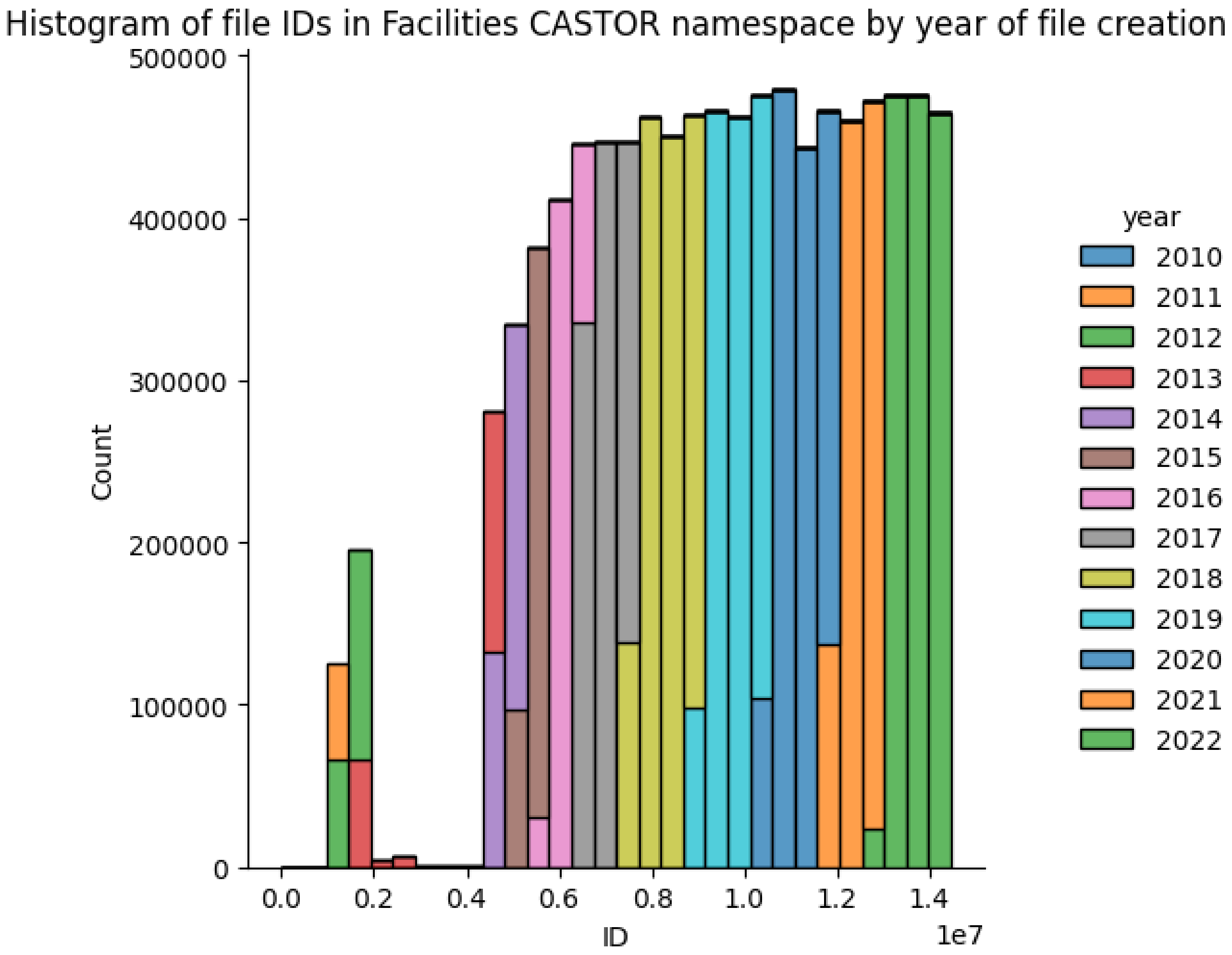
Backups

Science and
Technology
Facilities Council

Background info and caveats

- CTA and CASTOR both have 'tape file IDs'
  - CASTOR generally calls these 'file IDs', CTA uses the term 'archive file IDs'
  - These are recorded on tape in both systems
  - Any clashes will prevent the merging of namespaces (i.e. the facilities migration)
- New files will use the lowest unused ID available
  - IDs are not reused, so deleted files will leave gaps in the ID space
- The current plan is to read out and rewrite WLCG files in CTA to clear the ID space for the facilities migration
  - The tape a file was on will have to be repacked subsequently
- All analysis was done on namespace dumps taken on the 29th of September 2022
  - Things may change with analysis of future data
- I am not so familiar with the data I am analysing, or the analysis tooling I used
  - I have done my best to check my workings, but there may be errors. Please say if you see something that looks off.

Facilities namespace



Histogram of file IDs in Facilities CASTOR namespace by year of file creation

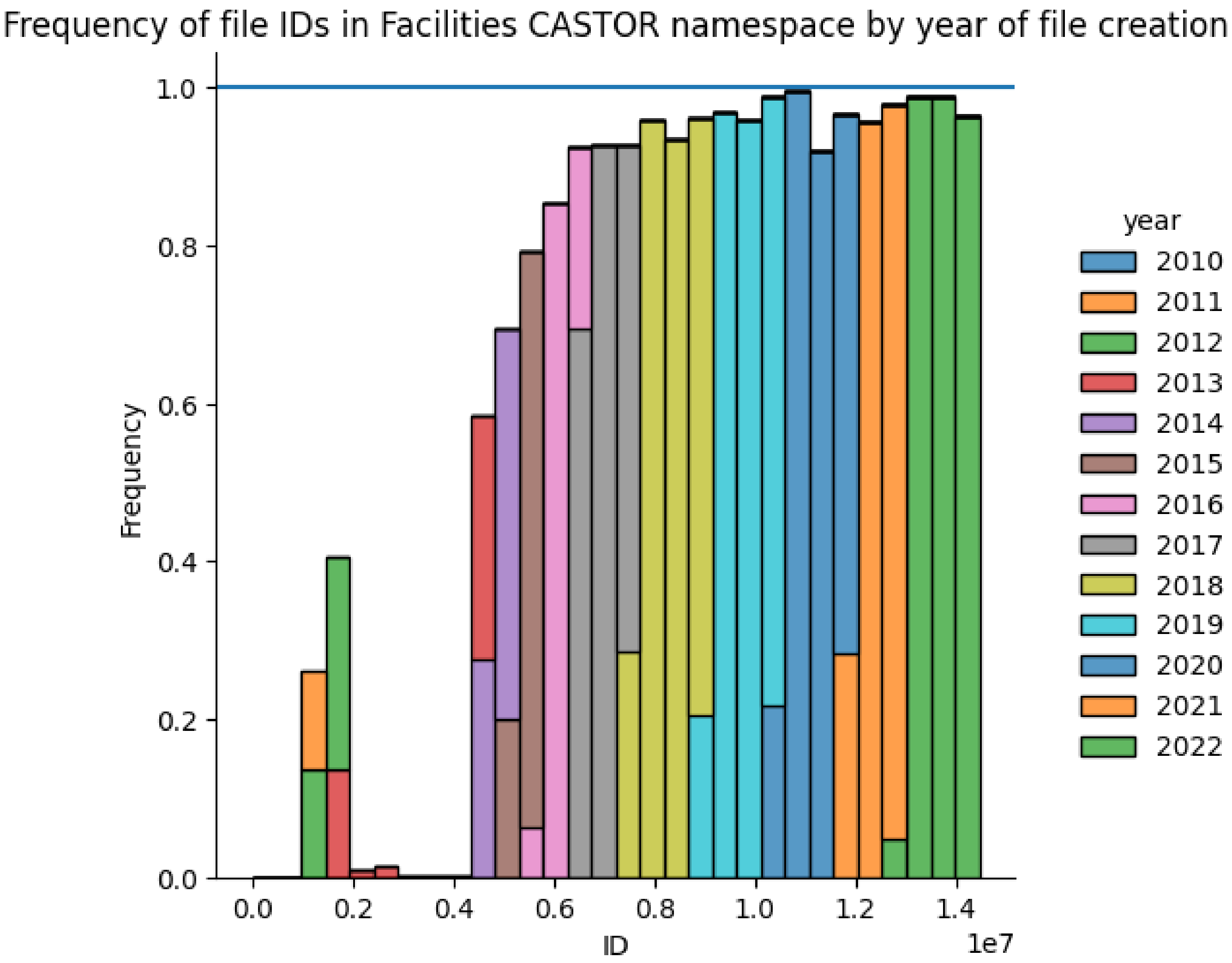
Year	Num. Files	ID range	Occupancy	Avg. File size	Total data
2010	314	1.1 million	0.03%	505.7 MB	158.8 GB
2011	59.9 thousand	197.2 thousand	30.40%	2.8 GB	165.7 TB
2012	195.8 thousand	285.4 thousand	68.62%	5.6 GB	1.1 PB
2013	227.5 thousand	3.1 million	7.46%	7.0 GB	1.6 PB
2014	370.5 thousand	3.6 million	10.25%	5.2 GB	1.9 PB
2015	448.5 thousand	560.6 thousand	80.00%	5.9 GB	2.7 PB
2016	551.9 thousand	637.5 thousand	86.57%	7.4 GB	4.1 PB
2017	1.1 million	1.2 million	92.69%	7.0 GB	7.6 PB
2018	1.4 million	1.5 million	94.68%	7.7 GB	10.9 PB
2019	1.4 million	1.4 million	97.00%	8.1 GB	11.3 PB
2020	1.4 million	1.4 million	96.27%	9.1 GB	12.3 PB
2021	1.0 million	1.1 million	96.59%	9.5 GB	9.9 PB
2022*	1.4 million	1.5 million	97.98%	7.2 GB	10.3 PB

* Incomplete year

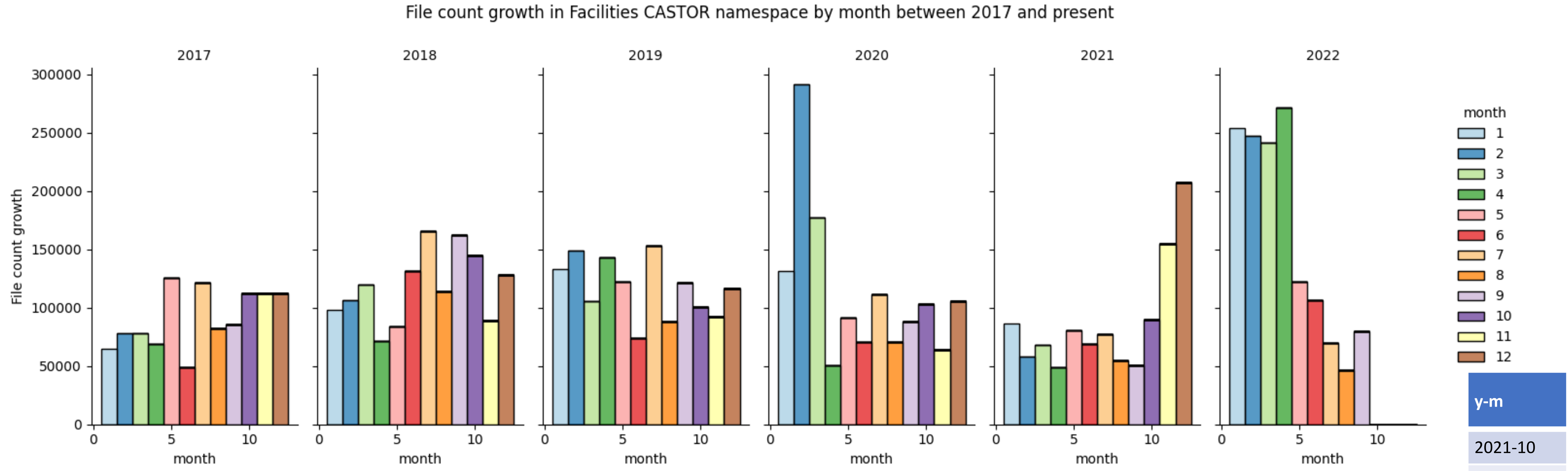
Current max ID: 14,456,235
(14.4 million)

Facilities namespace density

- Looking at the frequency gives a better idea of the namespace density
  - Frequency is defined as the number of observations in the bin divided by the bin width
- A frequency of 1 indicates that every ID in the bin range is present (i.e. the namespace is fully occupied)
- For files created after 2017 in Facilities CASTOR, there has been essentially no churn
  - >90% of created files still exist
- This namespace density means that any WLCG file in the overlapping ID space is likely to clash



Facilities namespace growth since 2017

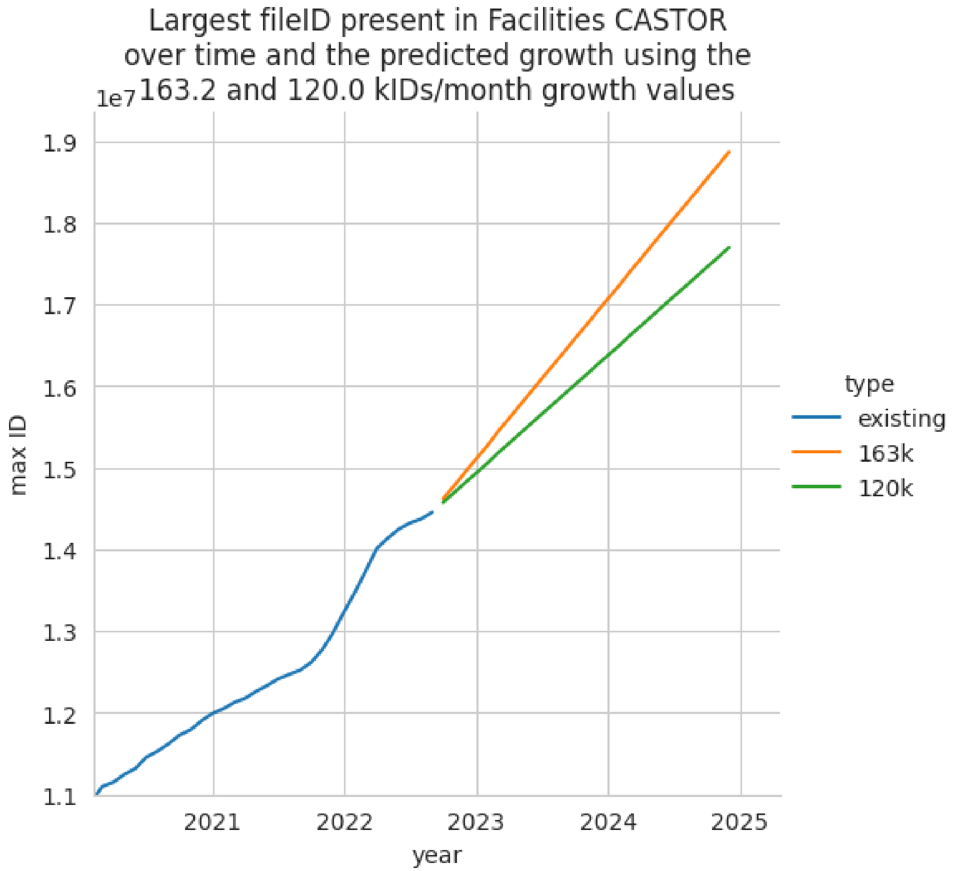

File count growth in Facilities CASTOR namespace by month between 2017 and present

y-m	File ID range (thousands)
2021-10	93.5
2021-11	158.3
2021-12	210.0
2022-01	257.5
2022-02	250.0
2022-03	244.6
2022-04	283.4
2022-05	146.3
2022-06	109.7
2022-07	72.8
2022-08	49.7
2022-09	83.5
mean	**163.2**

- Monthly growth has significant variance
  - There appears to be a visible COVID dip in 2020/2021, however...
  - The winter 2021 - spring 2022 period was extremely active.
    - Is this a post COVID experimental boom?
  - The latter part of 2022 seems to be closer to the expected rate based on previous years
  - The average growth over the entire 2017-present period is 120kfiles/month
- The average growth of the ID space over of the last 12 months is **163.2 kIDs/month**
  - This is a higher, but more probably more representative (and therefore safer) figure to use as predicted monthly growth of the Facilities namespace going forward in this analysis

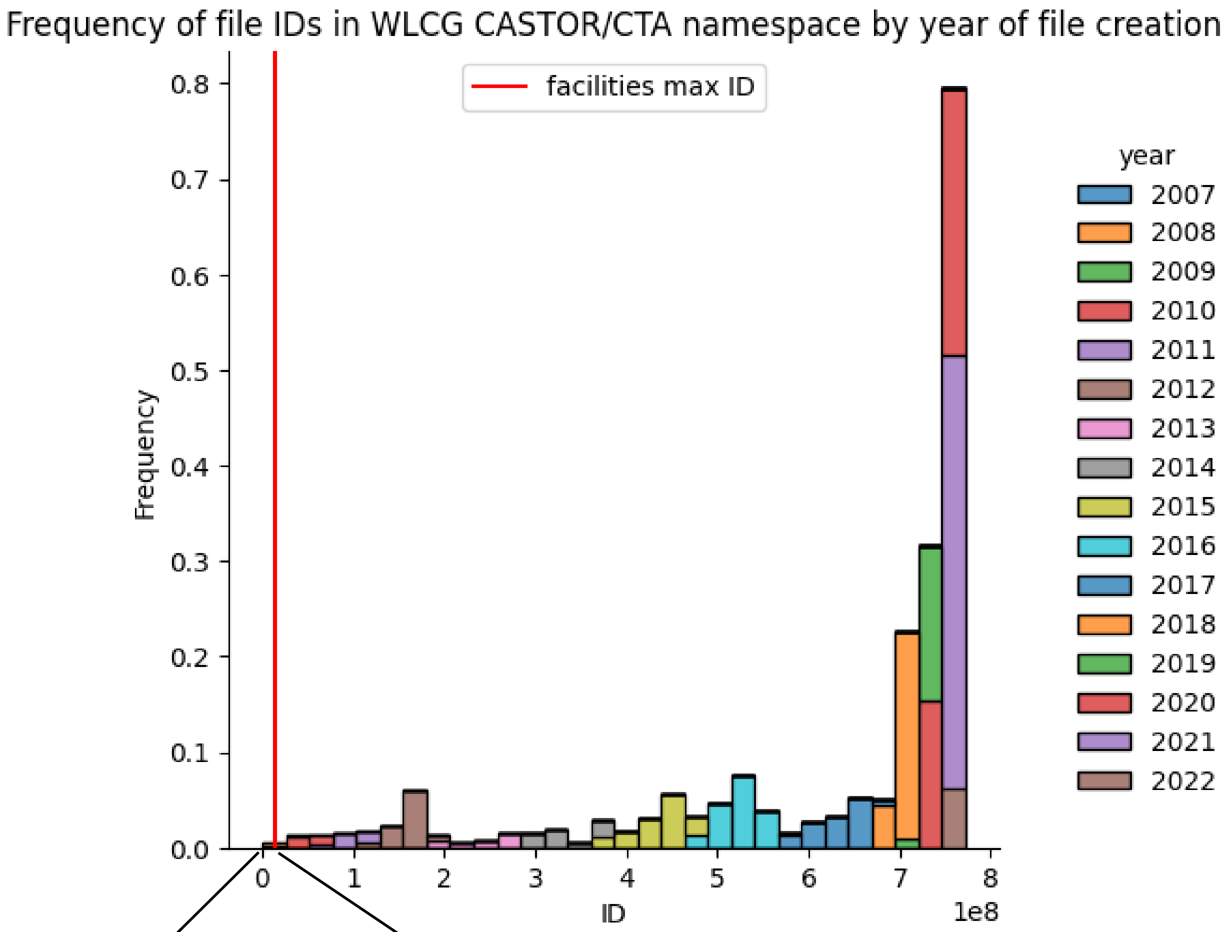
Future growth of the facilities namespace

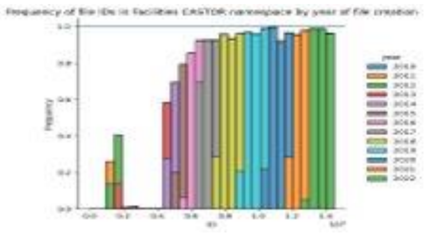
- Depending on the predicted growth value used, the facilities namespace will reach...
  - the ID 16 million between June and October 2023
  - the ID 17 million between December 2023 and May 2024
- We want to be migrated long before any of these dates, but we should make sure our planning has given us sufficient contingency
  - Delays out of our control, more facilities files than expected, etc.
  - We really don't want to be cutting it close!
- I propose that we should be prepared for the Facilities namespace to reach 16 million before migration
  - If we can prepare for it to reach higher IDs (16.5mil, 17mil) without too much effort, we should.



WLCG namespace density (or, how many clashes are we talking?)

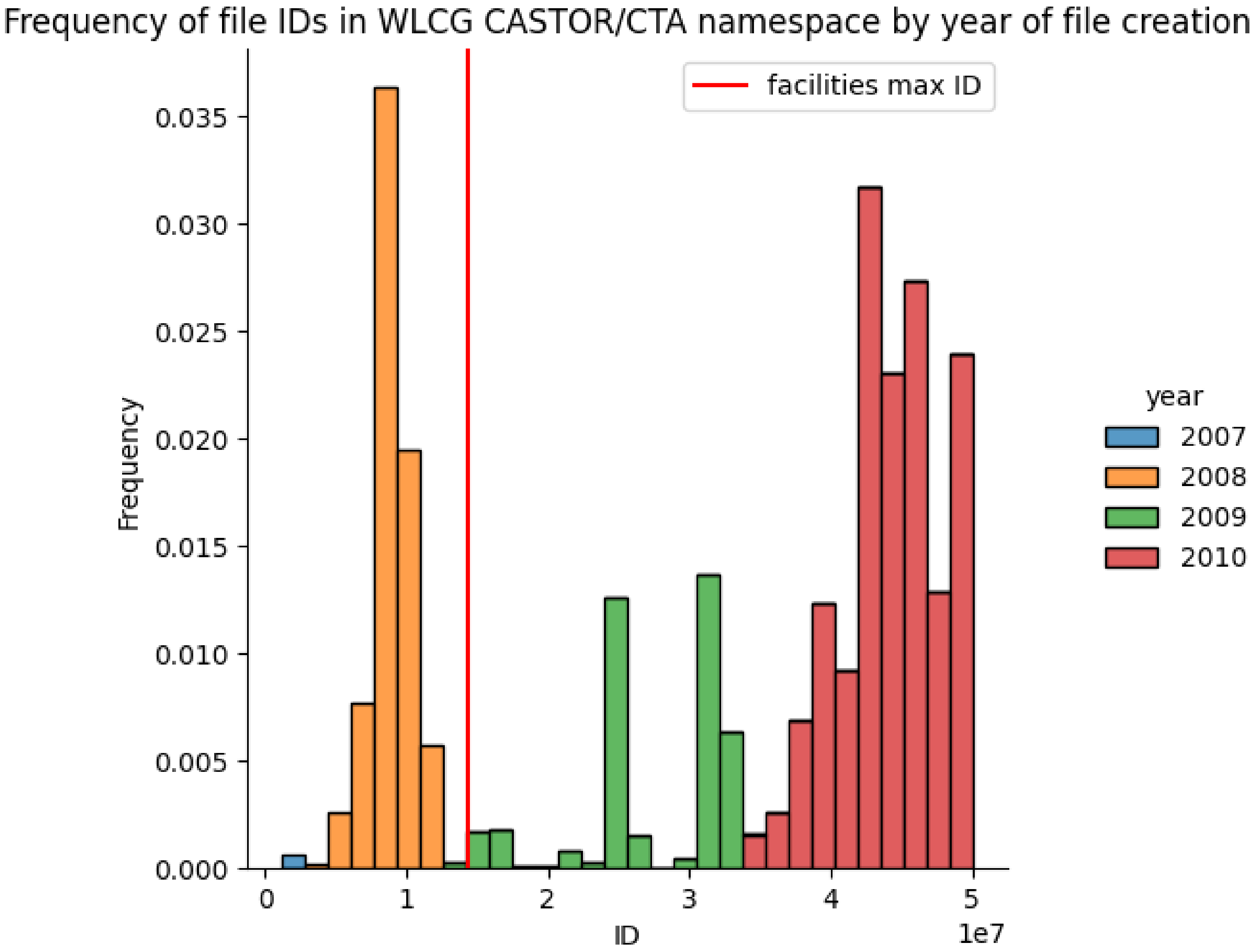
- The WLCG namespace is generally very sparse
- This may be due in part to CASTOR's dual purpose as disk storage for many years
    - The increase in density after 'disk only CASTOR' was removed adds weight to this claim
- The WLCG namespace covers a significantly larger ID space than the Facilities namespace





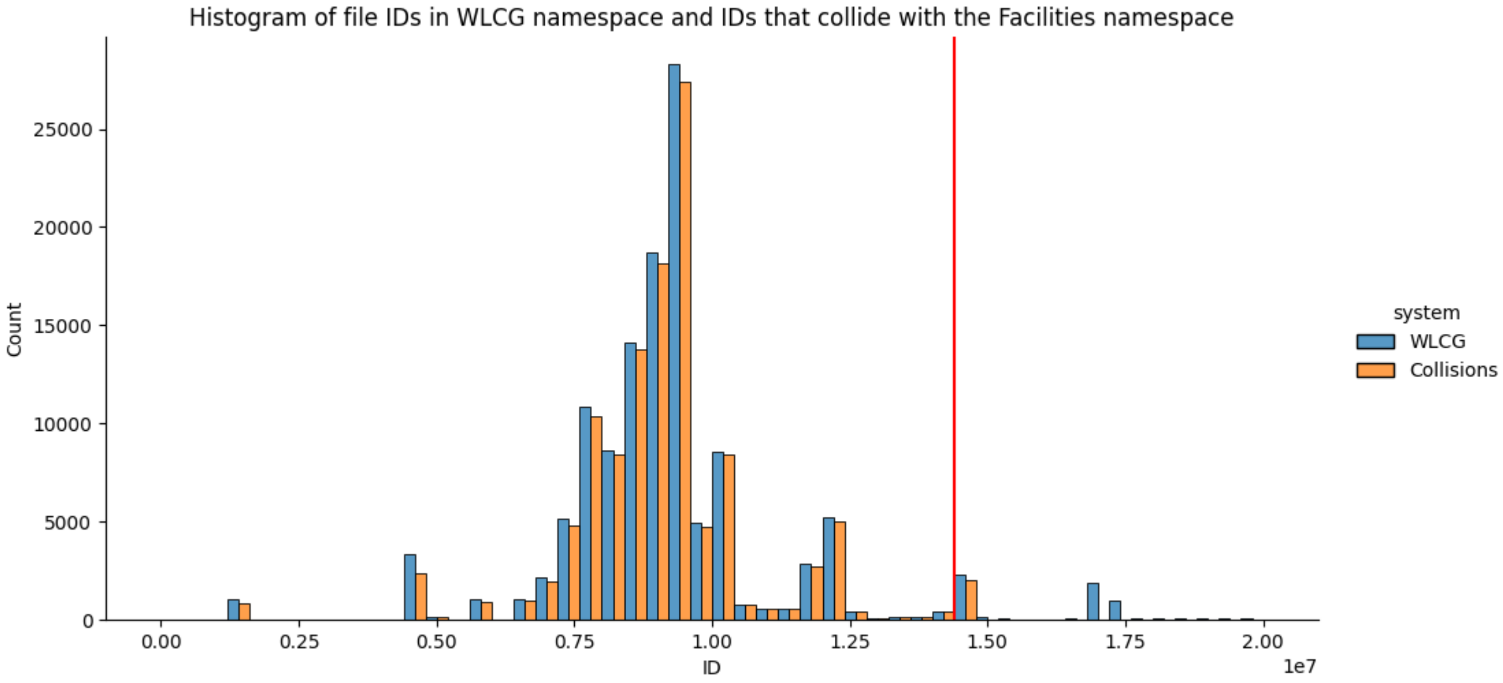
WLCG namespace density

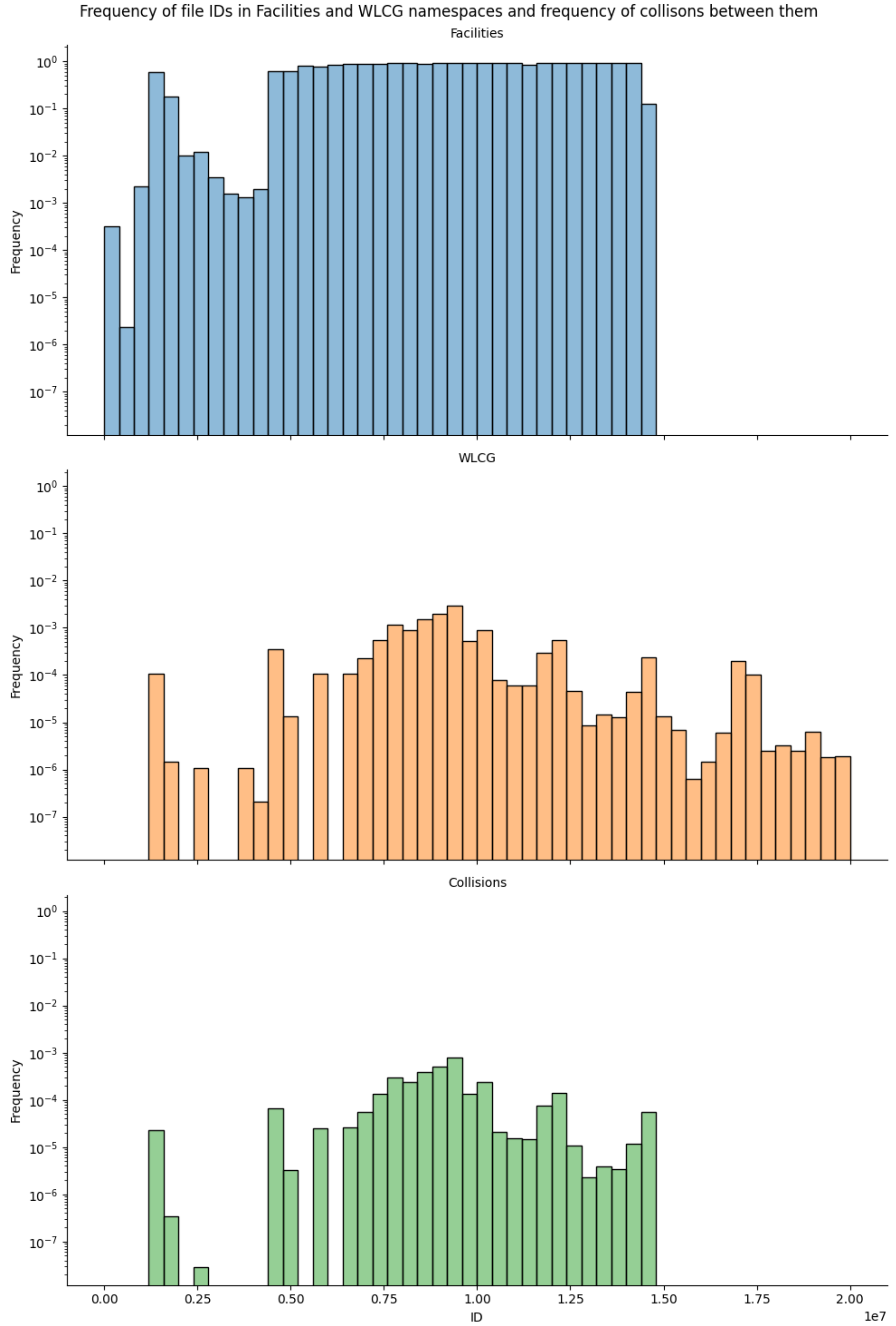
- The Facilities namespace is currently the size of the WLCGs namespace in early 2009
- Existing clashes are almost entirely with 2007 and 2008 WLCG files, and all new clashes will be with 2009 files
- Facilities CASTOR would have to double in number of files created before clashes with WLCG 2010 files are seen


Frequency of file IDs in WLCG CASTOR/CTA namespace by year of file creation

Clashes

- The dense nature of the Facilities namespace results in most WLCG files in the overlapping space to clash
- **It's worth considering re-ID'ing *all* WLCG files in the overlapping namespace area**





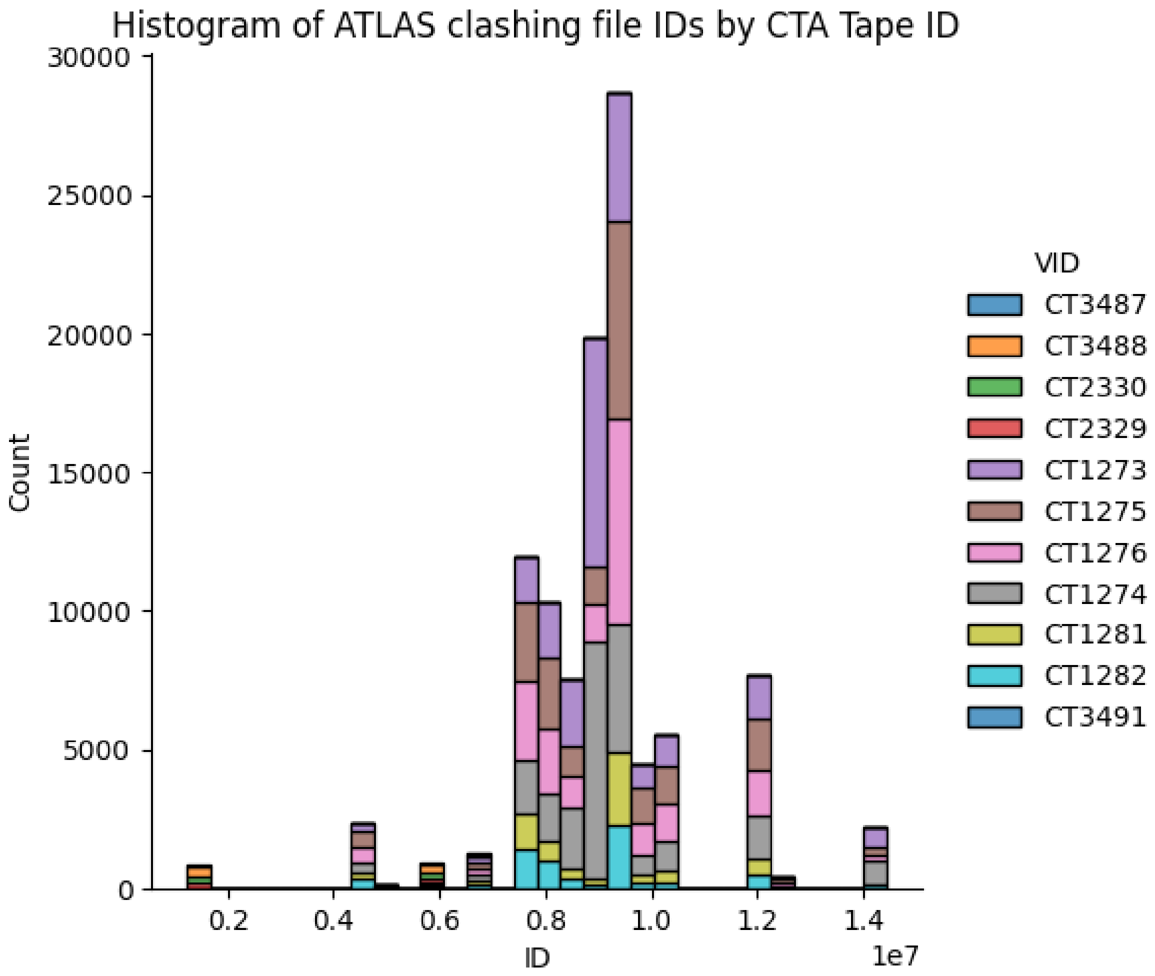
Comparison of clashing files vs all files in the overlapping part of the WLCG namespace

	Clashing		WLCG namespace overlap area		Ratios	
VO	Files	Tapes	Files	Tapes	Files	Tapes
alice	2831	17	2901	17	97.59%	100.00%
atlas	103925	11	106652	11	97.44%	100.00%
cms	8924	10	9194	10	97.06%	100.00%
other	0	0	0	0	N/A	N/A
Totals	**115680**	**38**	**118747**	**38**	**97.42%**	**100.00%**

- Re-ID'ing the entire overlapping area instead of just the clashes will require re-ID'ing 3% more files, and involve no extra tapes.
  - This removes the overhead of identifying clashes before dealing with them
- It will be easier to verify we have completed the re-ID'ing operation and have no clashes
  - i.e. **is the largest facilities ID smaller than the smallest WLCG ID?**

Tape distribution in clashes

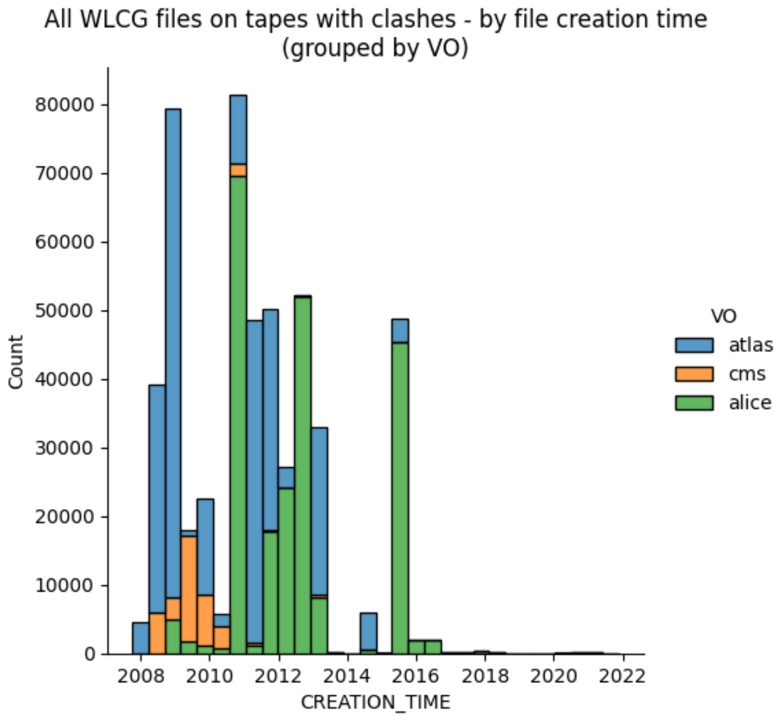
- Individual tapes have a wide range of IDs on them
  - This seems at odds with how tapes are written
- Is this due to repack packing files from across the ID space onto a tape?

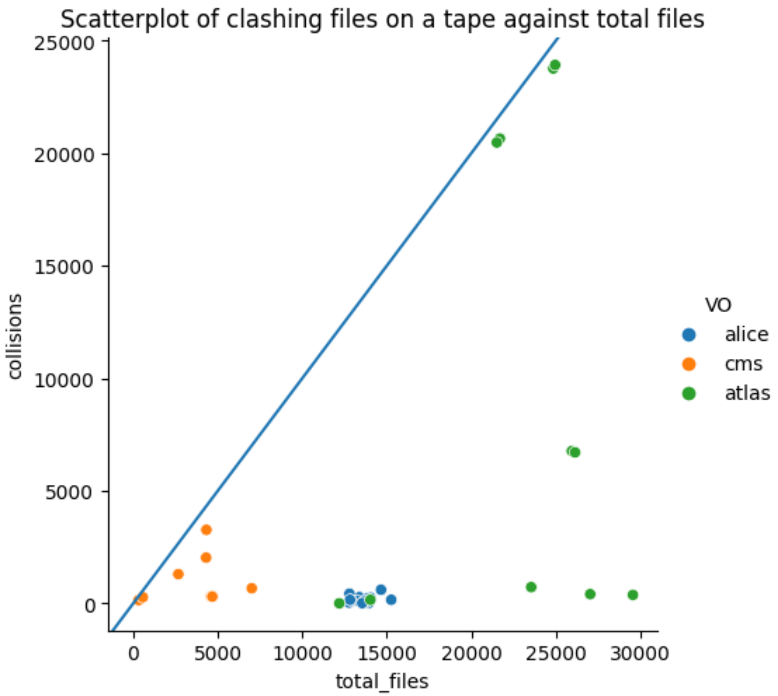


Characteristics of files on tapes with clashes

VO	clashing files	size of clash	tapes with clashes	total files on tapes	total data on tapes	percentage of files clashing	pools with clashes
alice	2.83k	6.1GB	17	231k	345.0TB	1.22%	alice
atlas	103k	151.1TB	11	251k	310.9TB	41.34%	atlasraw
cms	8.92k	39.4TB	10	37.8k	147.5TB	23.59%	cms2008-2009all, cms2010all
Totals	**115k**	**190.5TB**	**38**	**521k**	**803.5TB**	**22.05%**	

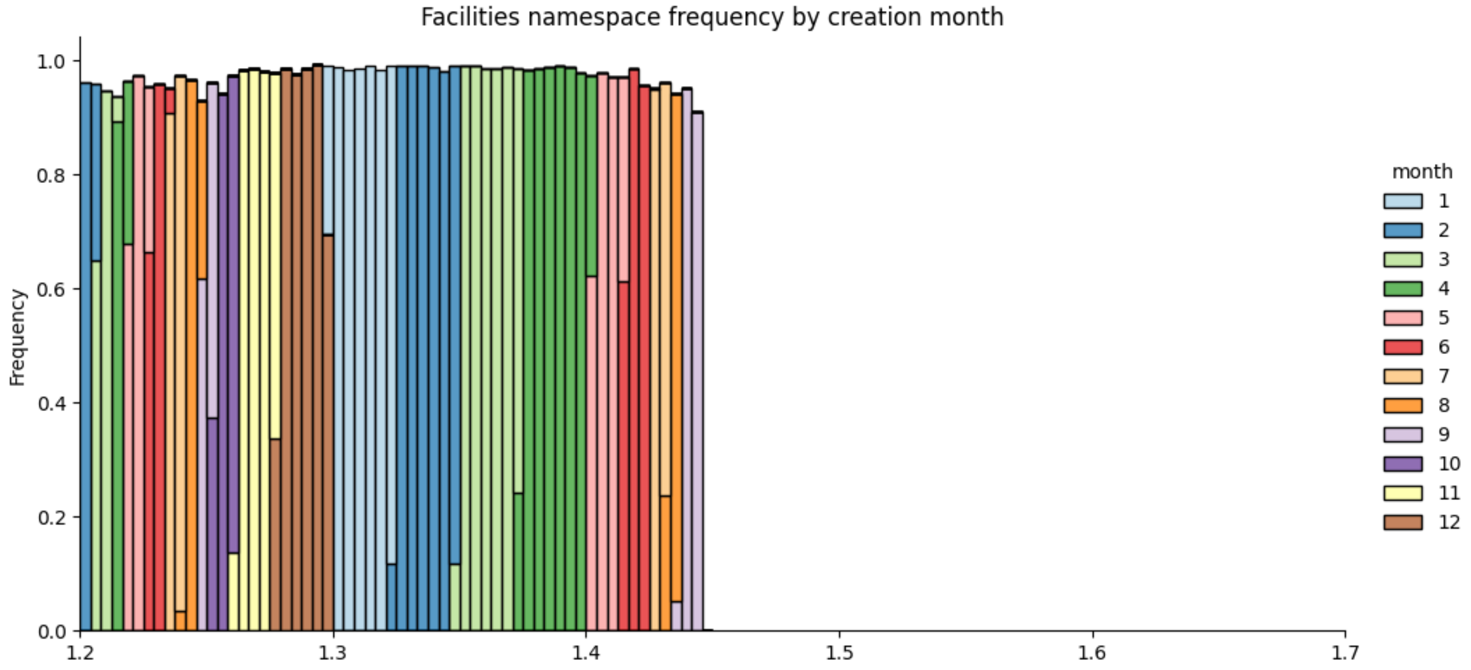
- These 38 clashing tapes have significant numbers of files outside of the overlap
    - i.e. files that were created after 2009
    - there are files on these tapes that were created in 2022
- These tapes will probably require repacking some
    - We would have to re-ID 4x the number of files if we dealt with all the files on clashing tapes

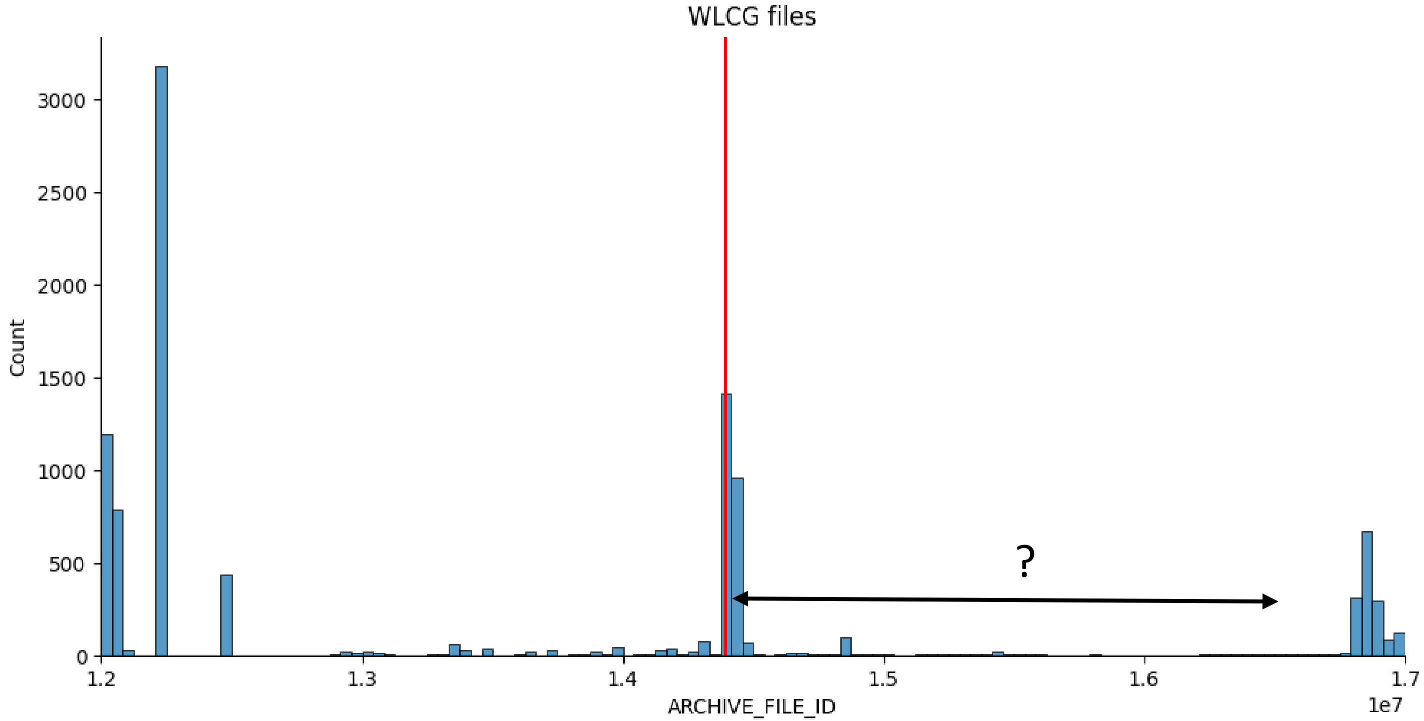




Growth of facilities namespace and implications on further clashes

- Given our understanding of the facilities CASTOR growth rates, the facilities namespace growth in the next year will be into a relatively sparse area in the WLCG namespace
- It looks like the area up to 16.5 million is very sparse
  - What happens in the WLCG namespace in this area?

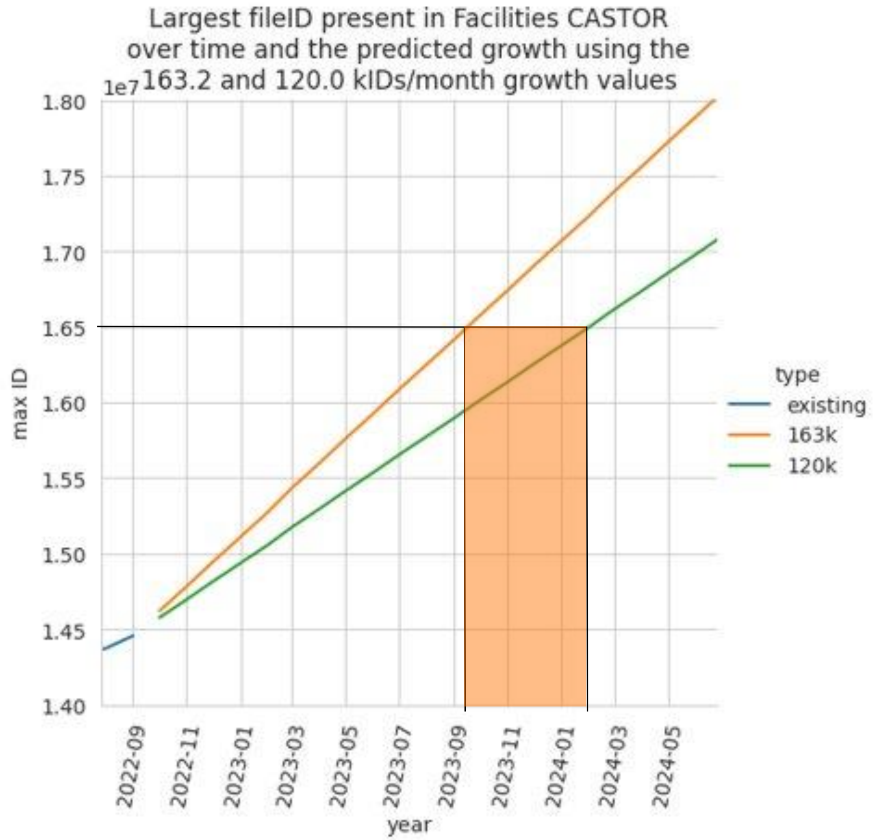




Area between current max facilities ID and 16.5 million

- There are under 2.5k files in that area
  - On 19 tapes
- This seems like a reasonable set of files to propose to add for re-ID'ing
- This should prepare us for all future clashes until late 2023
  - Which gives us plenty of contingency!

VO	files	total size	tapes	pools
alice	279	1.4GB	13	alice
atlas	2212	4.6TB	6	atlasraw
Total	**2491**	**4.6TB**	**19**	



Reanalysing the WLCG namespace between ID 0 and ID 16.5 million

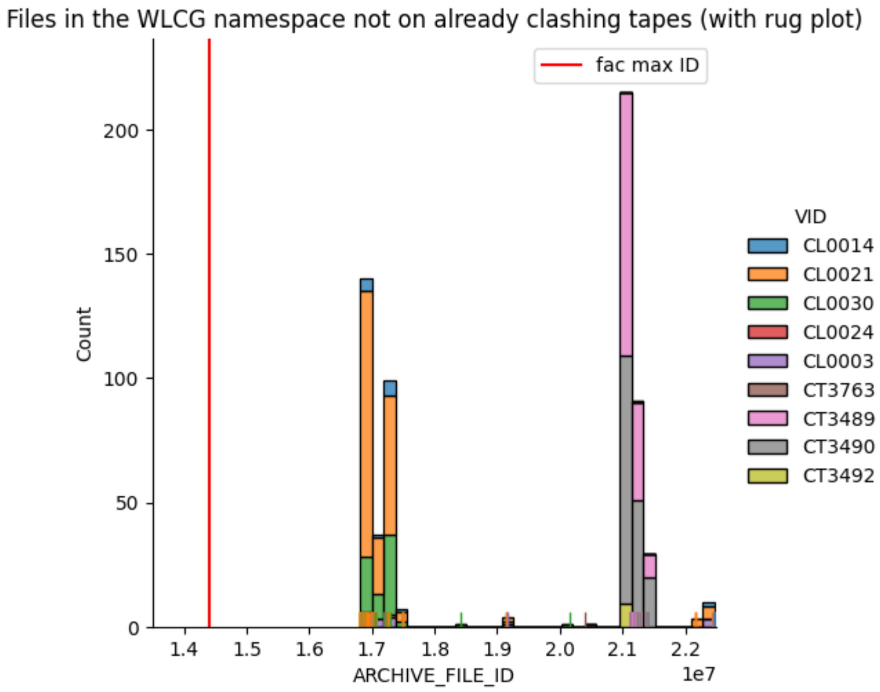
VO	files	average size	total size	tapes	pools
alice	3.18k	2.4MB	7.6GB	17	alice
atlas	108k	1.4GB	157.6TB	11	atlasraw
cms	9.19k	4.4GB	40.4TB	10	cms2008-2009all, cms2010all
Total	**121k**	**1.6GB**	**198.1TB**	**38**	

- This seems like a good set of files to target for re-ID'ing
- No extra tapes are added with the addition of the 'max fac ID to 16.5m' area
  - Which does raise the question...

What are the first files in the WLCG namespace not on the 38 clashing tapes

- **Some files from the alice pool with IDs ~16.8 million**
- Very limited files and tapes in the ID space above 16.5 million
- No danger of large amounts of extra repack needed, even if we go over 16.5 million

ARCHIVE_FILE_ID	SIZE_IN_BYTES	CREATION_TIME	VID	Pool	VO	year
16813424	7392	2009-02-22	CL0014	alice	alice	2009
16815517	7392	2009-02-22	CL0021	alice	alice	2009
16815527	7392	2009-02-22	CL0021	alice	alice	2009
16815581	7392	2009-02-22	CL0021	alice	alice	2009
…	…	…	…	…	…	…



Namespace analysis conclusions

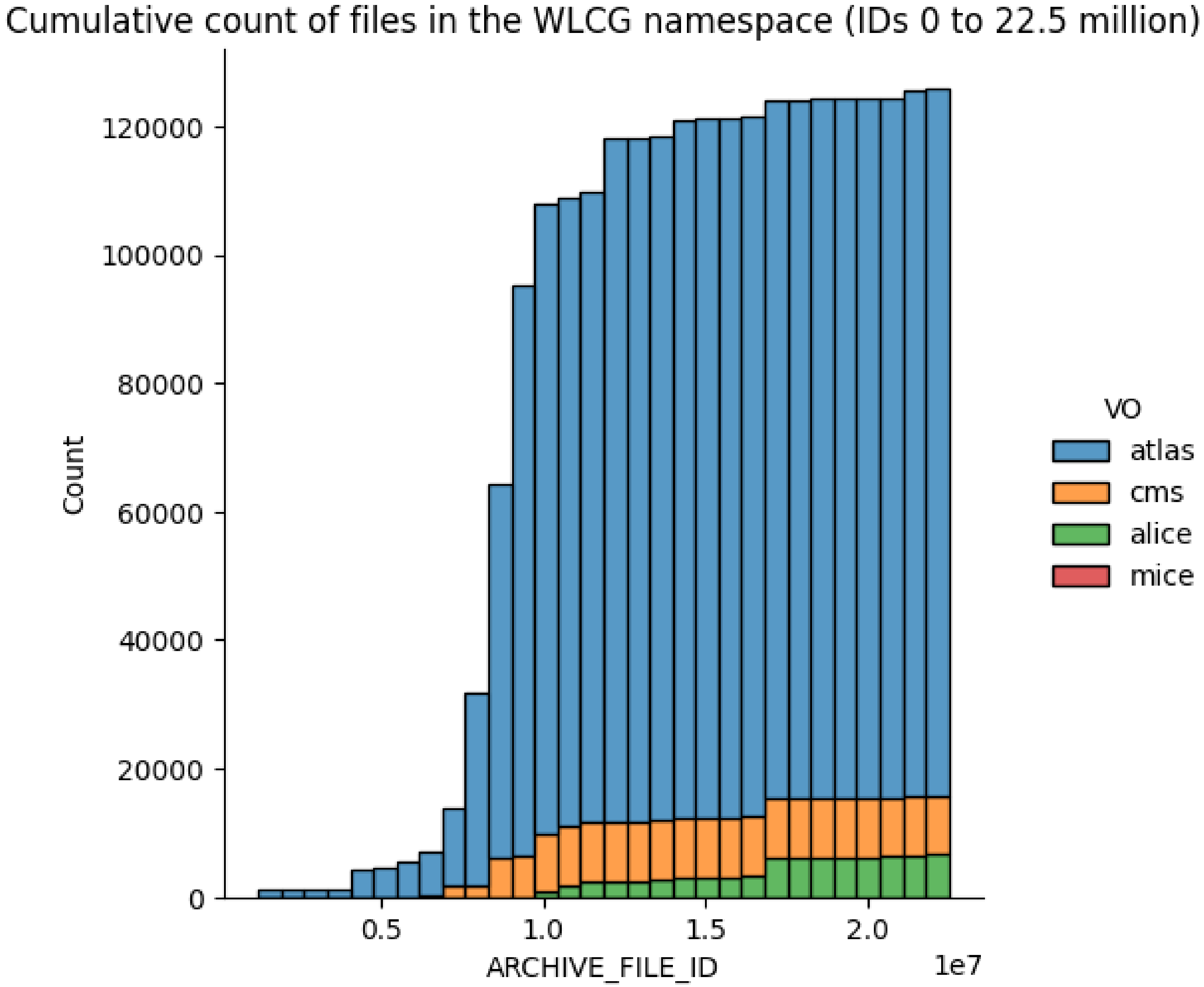
- The facilities namespace is generally very dense, while the WLCG namespace in the overlapping areas is sparse
  - the density of the facilities namespace means approximately every file in the overlapping part of the WLCG namespace will need to be re-ID'd
  - it is not necessarily worth searching for collisions, looking for WLCG files with IDs lower than the largest facilities file ID has a minimal overhead (~3% more files)
    - This will make confirming all clashes are dealt with *much* easier
- There were 115680 clashing files found in this current analysis
  - split between ALICE, ATLAS and CMS and across 38 tape
  - the 38 tapes also contain significantly more recent WLCG files, so some amount of repack will needed
    - 121k files to re-ID (198TB)
    - 400k files to repack (803TB)
- Clearing the WLCG namespace up to 16.5 million gives us room for facilities growth that will happen before migration
  - based on the Facilities growth seen in previous years we will hit 16.5m in late 2023 to early 2024
  - this will only require re-ID'ing another ~2500 WLCG files due to the sparsity of the WLCG namespace in this area

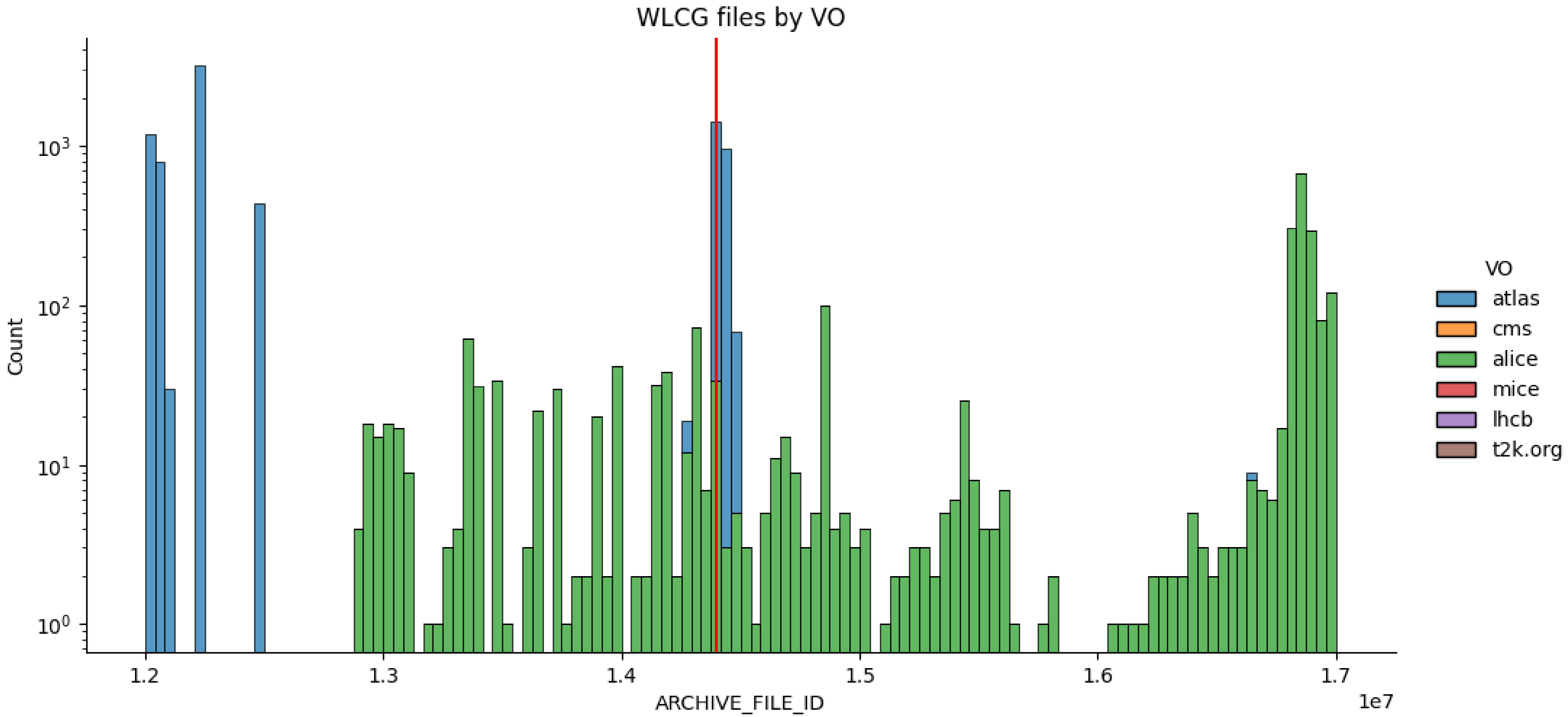
Re-ID'ing and repack high level strategy

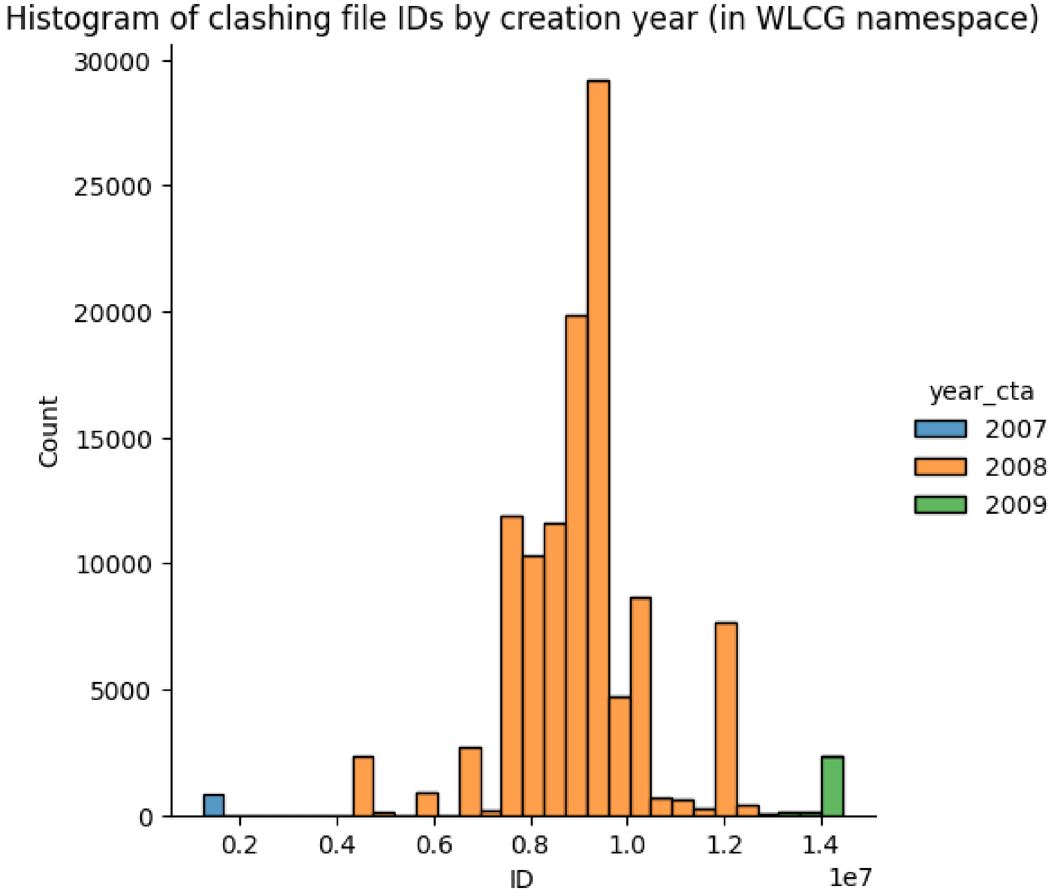
1. Pick one of the 38 tapes with clashes
2. Generate a list of all files on that tape with ID < 16.5 million
3. Read all files on the list out of Antares and onto Echo
4. Delete files in Antares
5. Write files back into Antares from Echo
6. Repack the original tape
7. Repeat process on next tape

Re-ID'ing thoughts and considerations

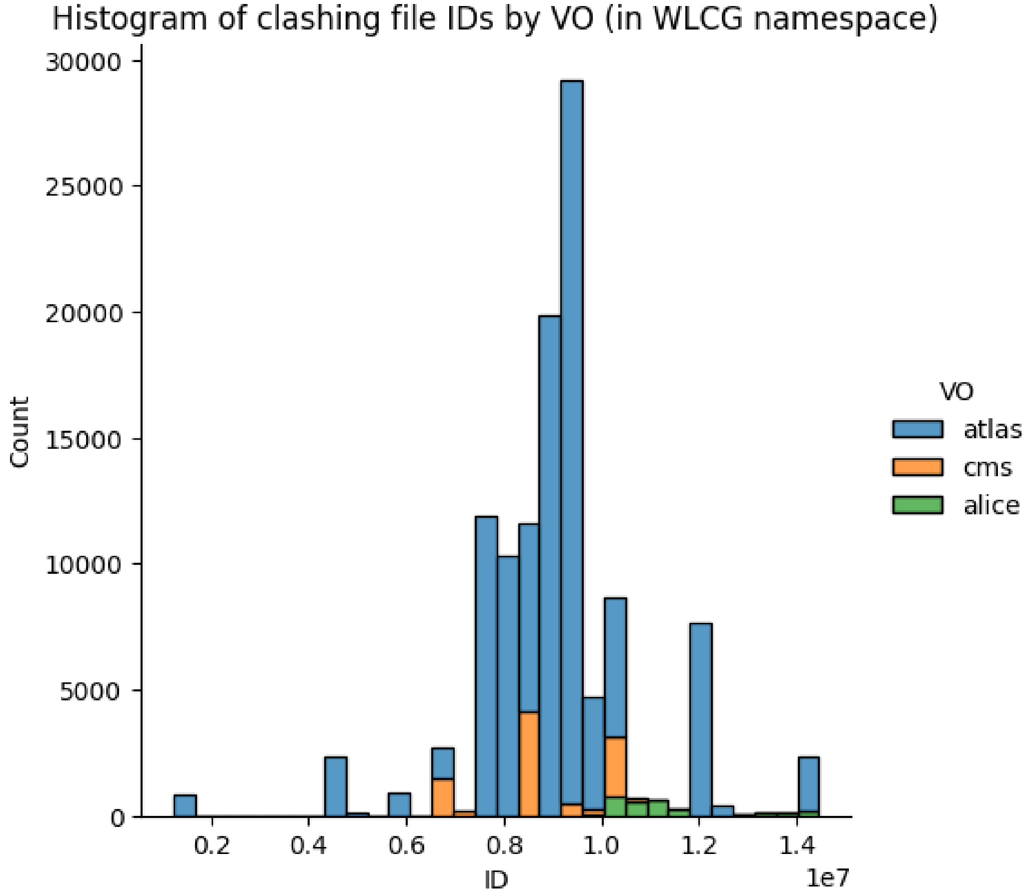
- We must make sure we have re-ID'd all files that we want to before repacking a tape
    - If we decide we need to move any more files we will have to repack the entire tape again
    - This means we should make sure we don't re-ID a 'too small' range
        - **Is 0 - 16.5 million the right ID range to target?**
- The movement between Echo and Antares should probably be handled by FTS
    - This is what it is designed to do, and it should make our lives easier
- Some questions about how we will retrieve and reinject files into Antares
    - We can easily map to Atlas/CMS, but spoofing ourselves as Alice without appropriate credentials will be a little harder
        - Not impossible, just need to make sure we have thought it through
        - May require some ACL changes to allow a specific user for the operation to recall and write to their areas


Cumulative count of files in the WLCG namespace (IDs 0 to 22.5 million)
Count
0
20000
40000
60000
80000
100000
120000
0.5
1.0
1.5
2.0
1e7
ARCHIVE_FILE_ID
VO
atlas
cms
alice
mice

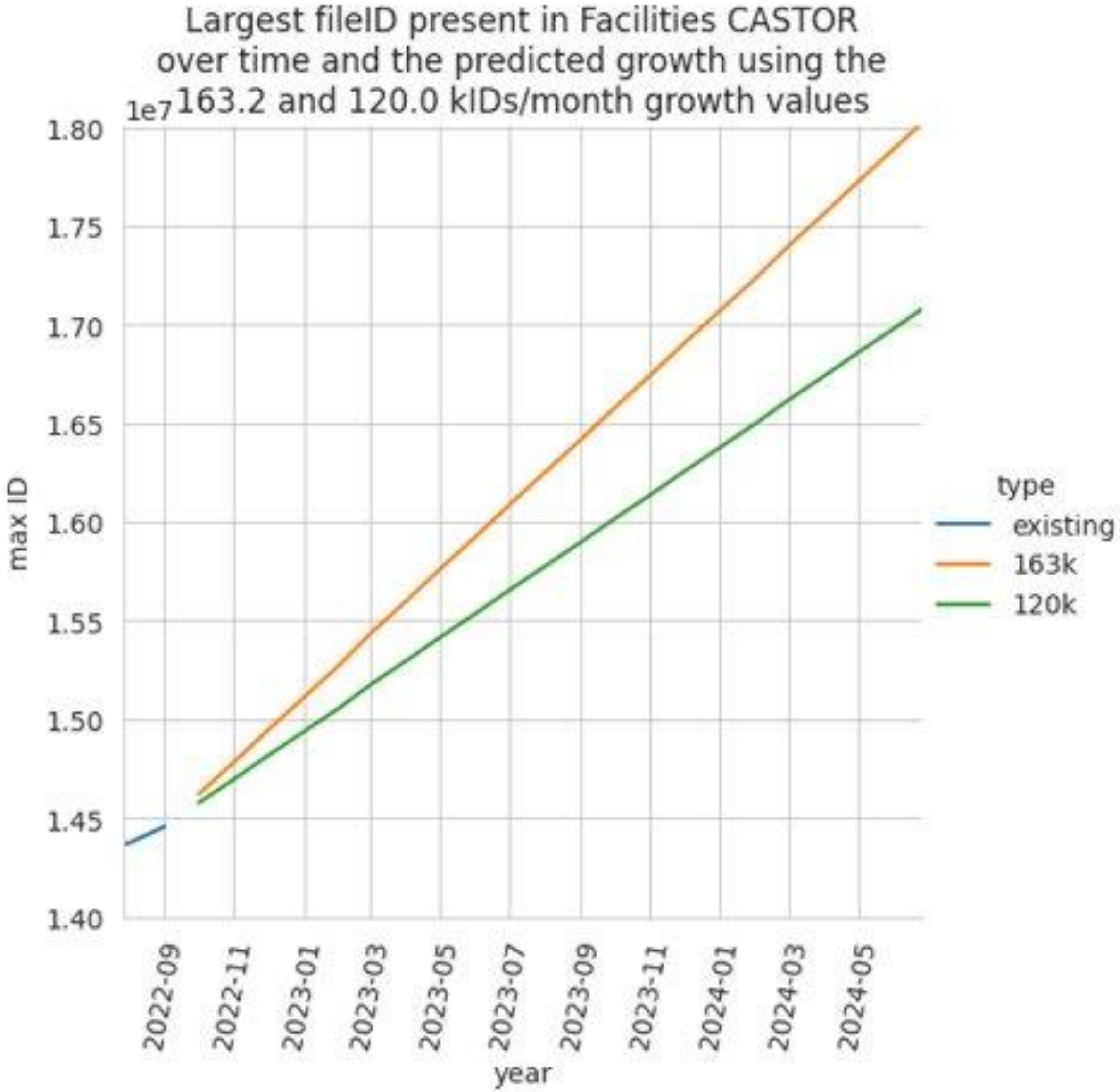

WLCG files by VO
Count
10^3
10^2
10^1
10^0
1.2
1.3
1.4
1.5
1.6
1.7
1e7
ARCHIVE_FILE_ID
VO
atlas
cms
alice
mice
lhcb
t2k.org



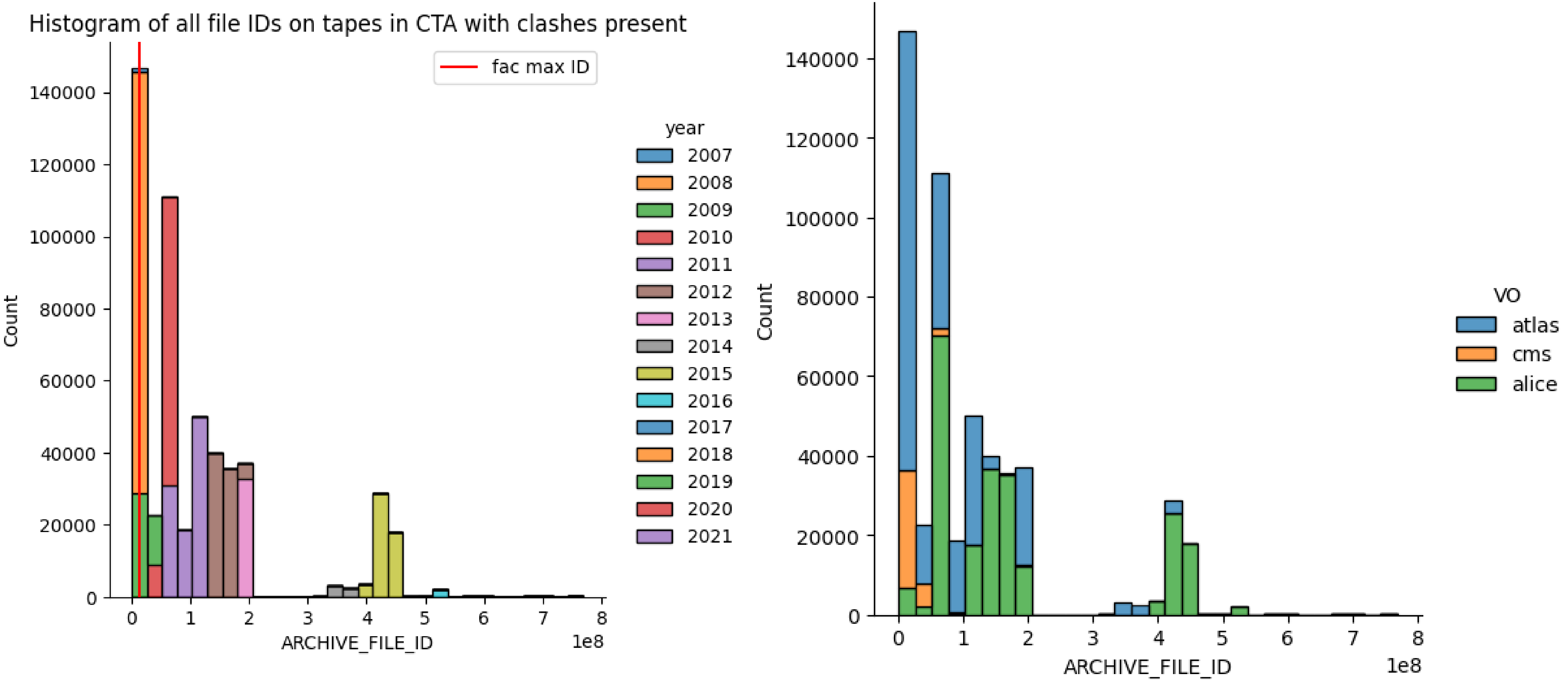
Histogram of clashing file IDs by creation year (in WLCG namespace)



Histogram of clashing file IDs by VO (in WLCG namespace)



Largest fileID present in Facilities CASTOR over time and the predicted growth using the 163.2 and 120.0 kIDs/month growth values



Histogram of all file IDs on tapes in CTA with clashes present

Facilities CASTOR namespace stats

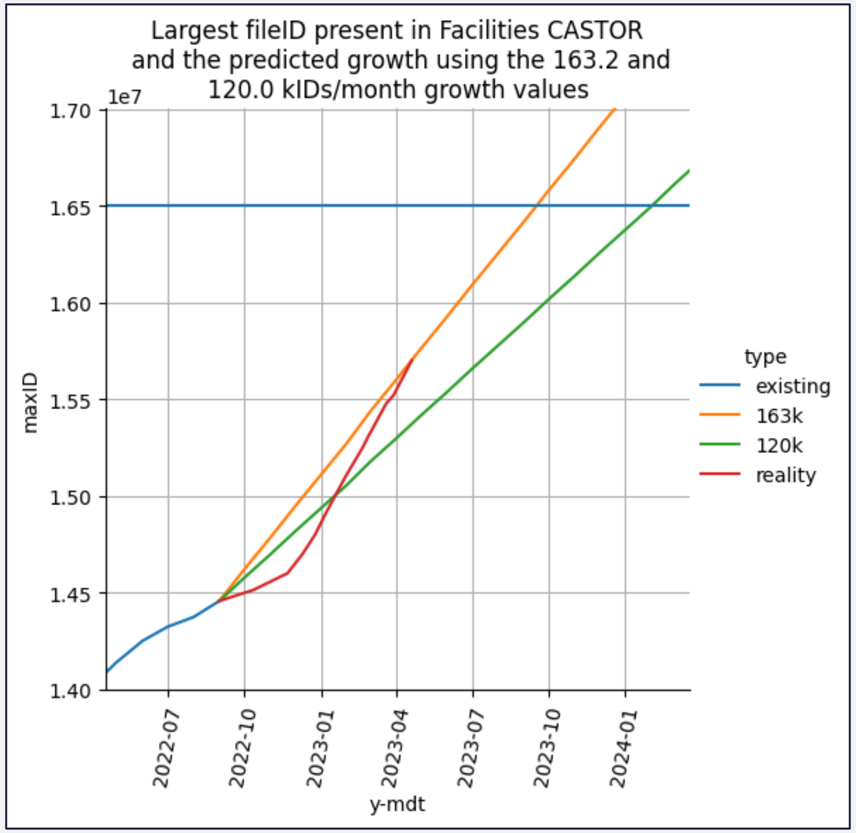
Year	Num. Files	ID range	Occupancy	Avg. File size	Total data
2010	314	1.1 million	0.03%	505.7 MB	158.8 GB
2011	59.9 thousand	197.2 thousand	30.40%	2.8 GB	165.7 TB
2012	195.8 thousand	285.4 thousand	68.62%	5.6 GB	1.1 PB
2013	227.5 thousand	3.1 million	7.46%	7.0 GB	1.6 PB
2014	370.5 thousand	3.6 million	10.25%	5.2 GB	1.9 PB
2015	448.5 thousand	560.6 thousand	80.00%	5.9 GB	2.7 PB
2016	551.9 thousand	637.5 thousand	86.57%	7.4 GB	4.1 PB
2017	1.1 million	1.2 million	92.69%	7.0 GB	7.6 PB
2018	1.4 million	1.5 million	94.68%	7.7 GB	10.9 PB
2019	1.4 million	1.4 million	97.00%	8.1 GB	11.3 PB
2020	1.4 million	1.4 million	96.27%	9.1 GB	12.3 PB
2021	1.0 million	1.1 million	96.59%	9.5 GB	9.9 PB
2022*	1.4 million	1.5 million	97.98%	7.2 GB	10.3 PB

* Incomplete year

Files to re-ID

- With room for growth – the number of WLCG files that need to be re-ID'd are
  - ~121k files
  - ~200TB
  - 38 tapes

VO	files	total size	tapes	pools
alice	3.18k	7.6GB	17	alice
atlas	108k	157.6TB	11	atlasraw
cms	9.19k	40.4TB	10	cms2008-2009all, cms2010all
Total	**121k**	**198.1TB**	**38**	


Largest fileID present in Facilities CASTOR
and the predicted growth using the 163.2 and
120.0 kIDs/month growth values
1e7
1.70
1.65
1.60
1.55
1.50
1.45
1.40
maxID
2022-07
2022-10
2023-01
2023-04
2023-07
2023-10
2024-01
y-mdt
type
existing
163k
120k
reality