

Recent Developments in the CernVM-FS Server Backend

René Meusel

Jakob Blomer, Gerardo Ganis,
Predrag Buncic, Seppo Heikkila

ACAT 2014 - Prague, 4th of September



1

Usage Statistics and Adoption

2

New Challenges and Features

3

File System History

4

Garbage Collection

5

Smart Stratum1 Servers



What is CernVM-FS?

- **Scalable software distribution system**

- Infrequent atomic updates in a central location
- Read-only access on the clients

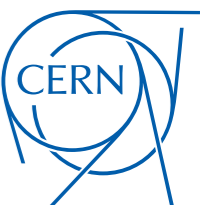
- **HTTP based global data transfer**

- Minimal protocol requirements
- Aggressive hierarchical cache strategy

- Assumption: Coherent working set on physically close nodes (cf. software vs. data distribution)

- **Accessible through a mounted file system (POSIX)**

- FUSE module, NFS exported FUSE volume or Parrot



Who Uses CernVM-FS?

- **All LHC experiments**



- **CernVM 3**



- Operating system in CernVM-FS



- **Others beyond the HEP community**

- Human Brain Project, BioMed, VLEMED, ...



- Stratum0s at CERN, RAL, NIKHEF, Fermilab, DESY, ...

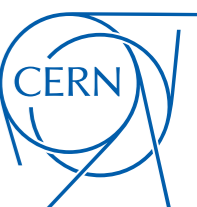


Repository Statistics

Repository	Files	Refer. Objects	Volume	Ø File Size	
atlas.cern.ch	34'500'000	3'700'000	2.1 TiB	66.2 kiB	Mainly Software
cms.cern.ch	30'600'000	4'800'000	0.9 TiB	33.1 kiB	
lhcb.cern.ch	13'600'000	4'600'000	0.5 TiB	41.9 kiB	
alice.cern.ch	5'900'000	240'000	0.5 TiB	90.7 kiB	
ams.cern.ch	2'900'000	1'900'000	1.9 TiB	0.7 MiB	Software + Conditions Data
alice-ocdb.cern.ch	700'000	700'000	0.1 TiB	0.2 MiB	Conditions Data
atlas-condb.cern.ch	8'000	7'800	0.5 TiB	60.8 MiB	

- *Files* and *Volume* as saved in the CernVM-FS catalogs
- Actual number of *Referenced Objects* is compressed and de-duplicated
- Based on latest revision - no history involved

(Effective: August 2014)



New Challenges



New Challenges



Large Files (> 200 MiB)



Long term data preservation



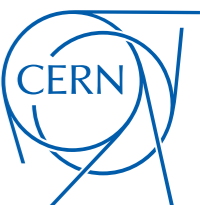
Rapidly changing repository content



Increasing configuration distribution effort



Instant repository update propagation

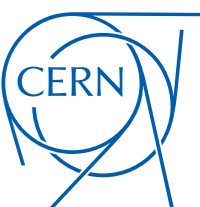


CernVM-FS Repository

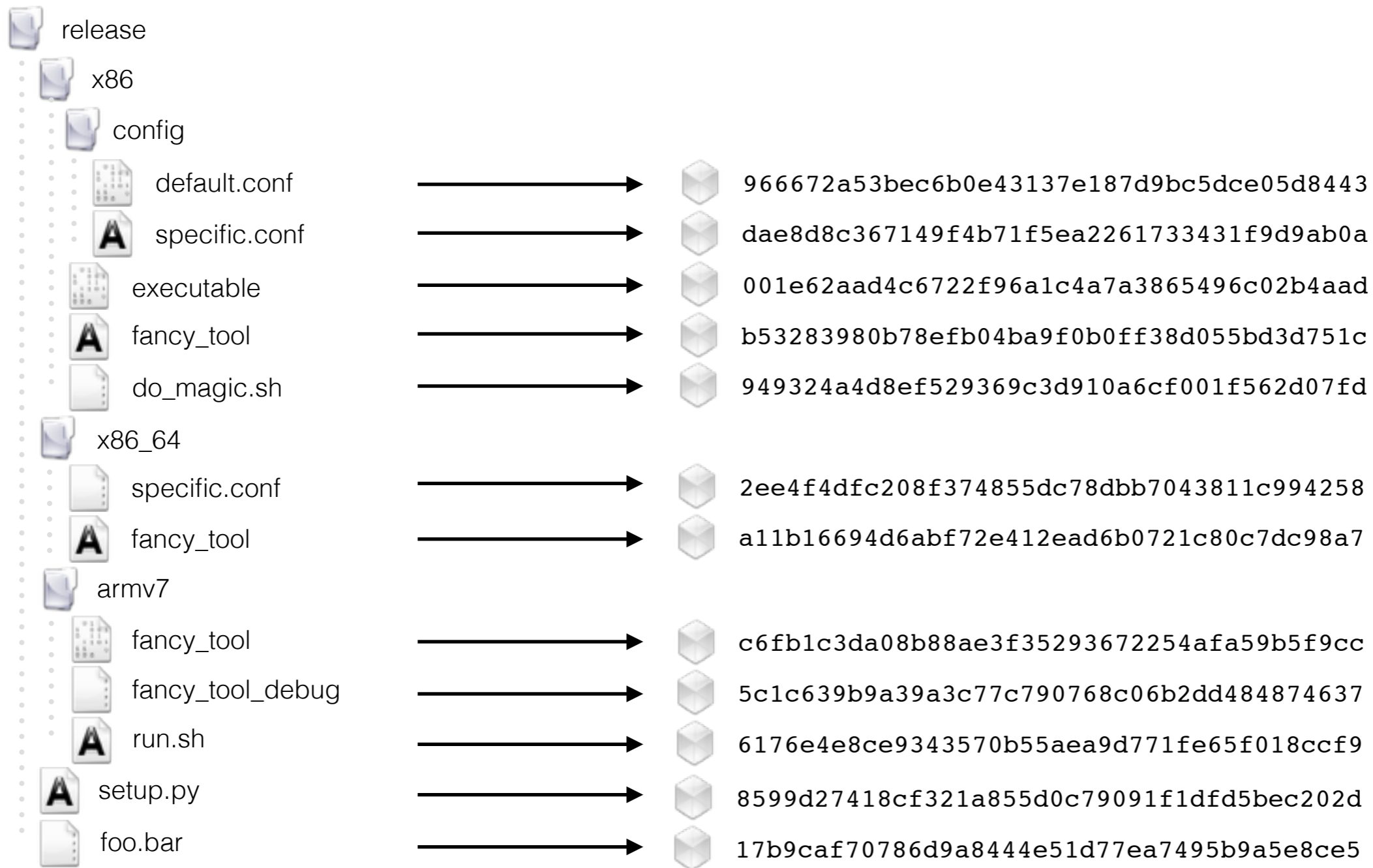
From POSIX File Filesystem to
Content-Addressable Objects



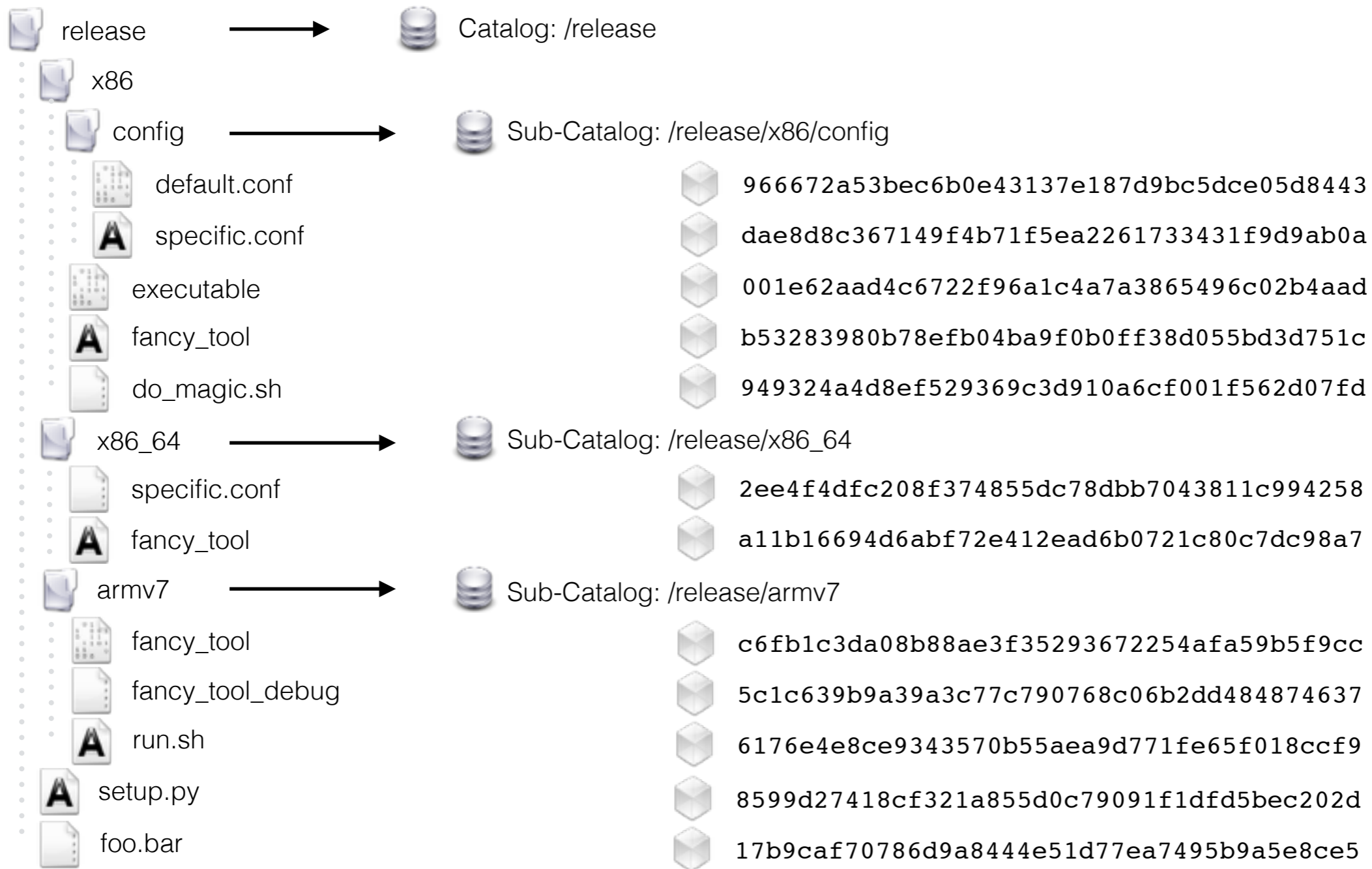
From POSIX to Blob-Objects



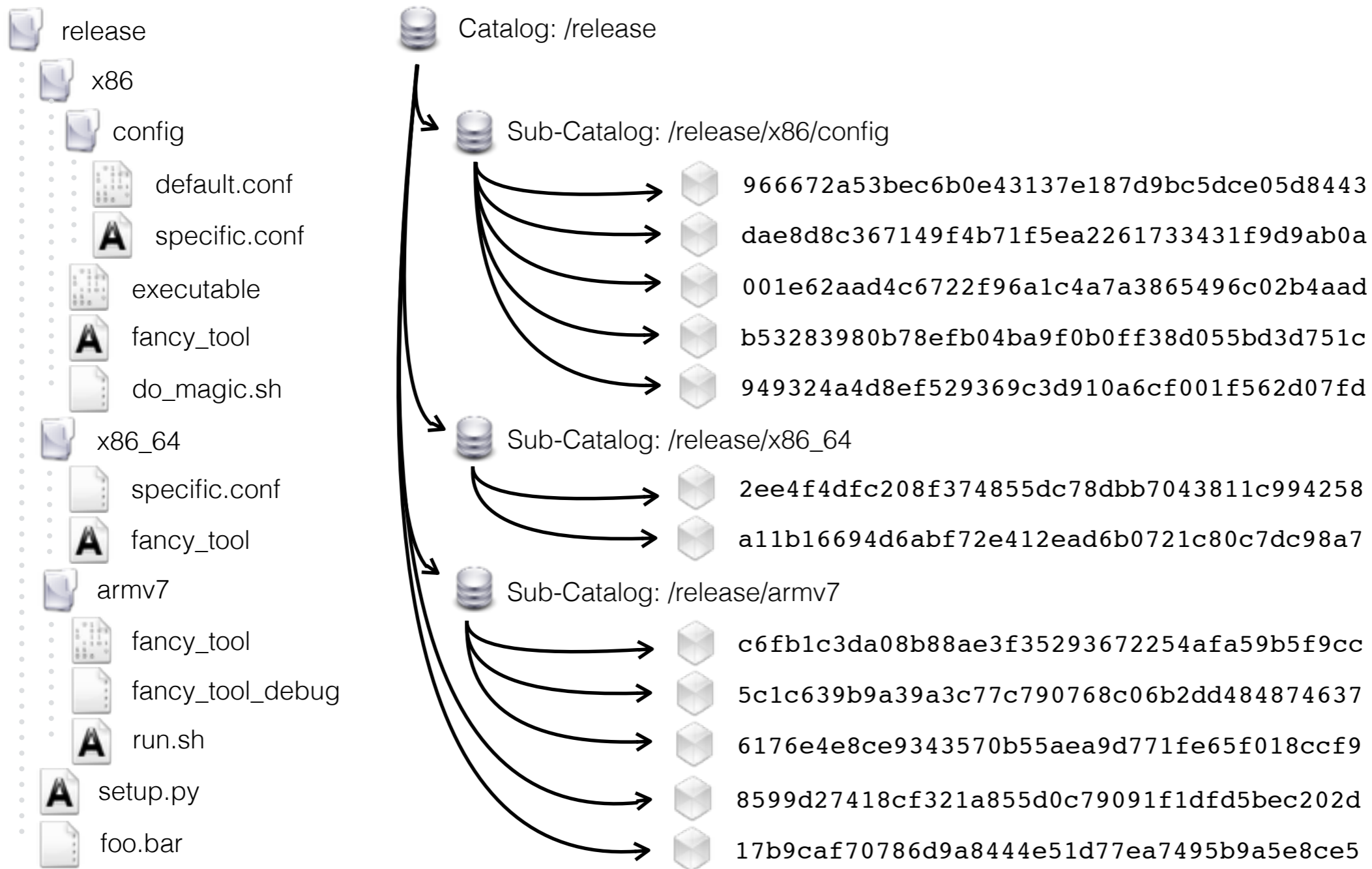
From POSIX to Blob-Objects



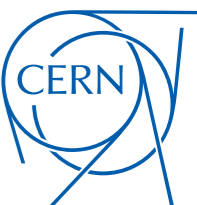
From POSIX to Blob-Objects



















From POSIX to Blob-Objects





From POSIX to Blob-Objects




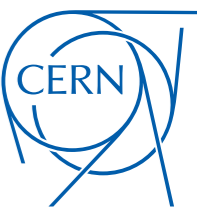
From POSIX to Blob-Objects

-  038f625d0790e06b0848a04bef90a51bd7b3ebecC
-  59bb67e545ac1951ac0f274ff63e8d2cc78ef420C
-  966672a53bec6b0e43137e187d9bc5dce05d8443
-  dae8d8c367149f4b71f5ea2261733431f9d9ab0a
-  001e62aad4c6722f96a1c4a7a3865496c02b4aad
-  b53283980b78efb04ba9f0b0ff38d055bd3d751c
-  949324a4d8ef529369c3d910a6cf001f562d07fd
-  11f58905f87d7ad5513aede20e21722b890eb9d6C
-  2ee4f4dfc208f374855dc78dbb7043811c994258
-  a11b16694d6abf72e412ead6b0721c80c7dc98a7
-  949324a4d8ef529369c3d910a6cf001f562d07fdC
-  c6fb1c3da08b88ae3f35293672254afa59b5f9cc
-  5c1c639b9a39a3c77c790768c06b2dd484874637
-  6176e4e8ce9343570b55aea9d771fe65f018ccf9
-  8599d27418cf321a855d0c79091f1dfd5bec202d
-  17b9caf70786d9a8444e51d77ea7495b9a5e8ce5

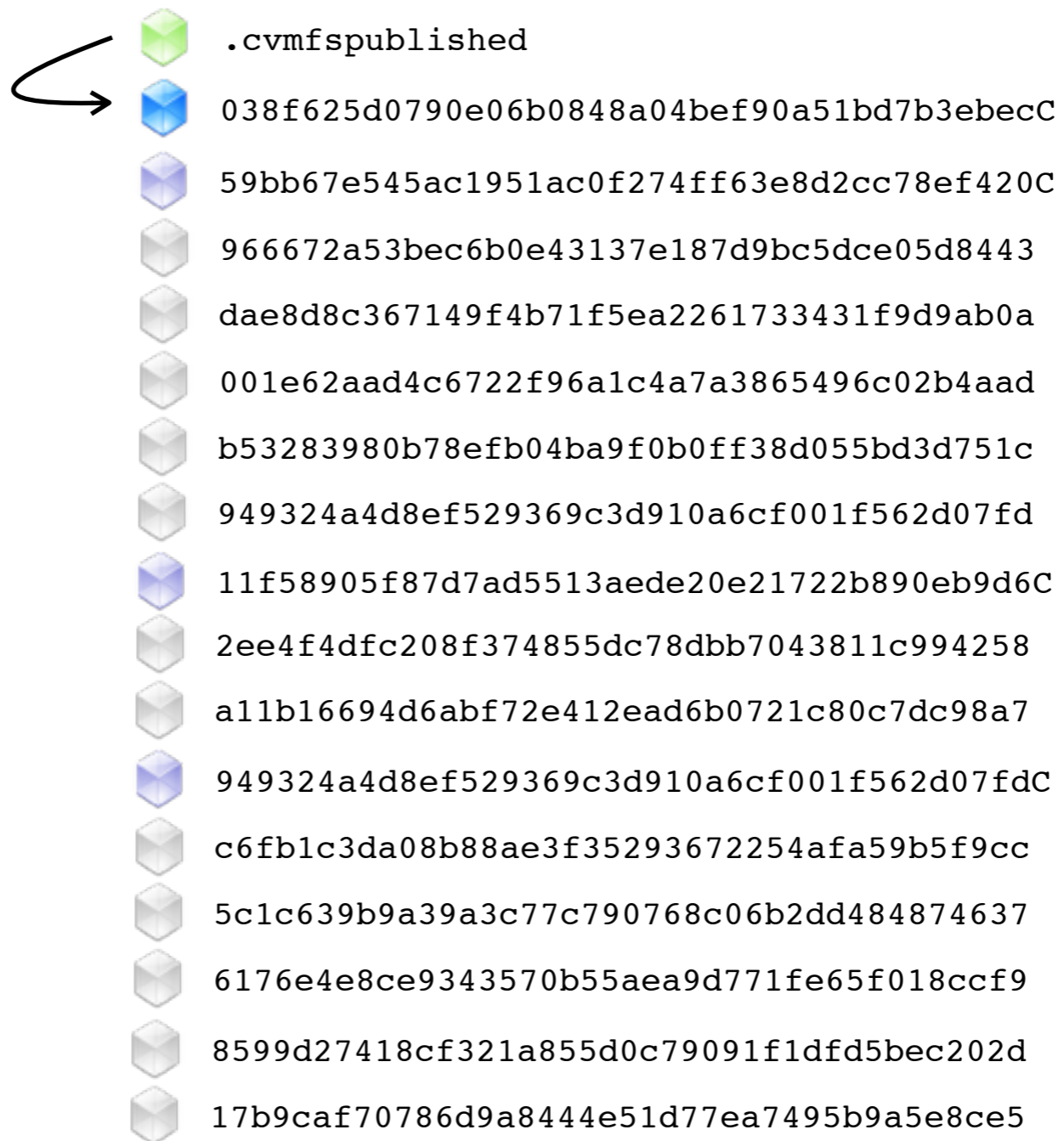
 Data Object





 Root Catalog

 Catalog



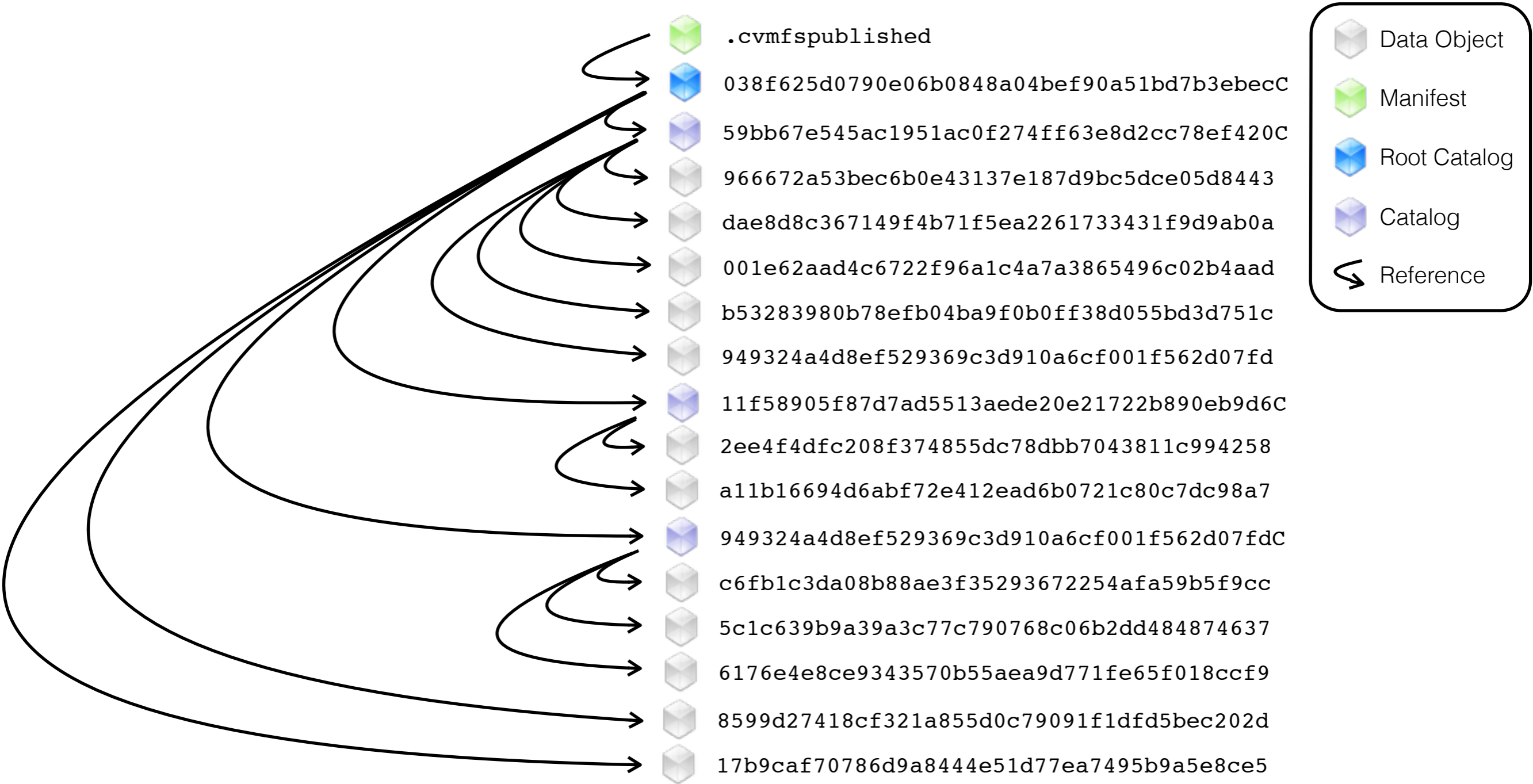
From POSIX to Blob-Objects



-  Data Object
-  Manifest
-  Root Catalog
-  Catalog



From POSIX to Blob-Objects



From POSIX to Blob-Objects

- **Merkle tree**


- only .cvmfspublished needs to be signed

- **Content-Addressable Storage**

- File de-duplication
- Trivial file integrity checks

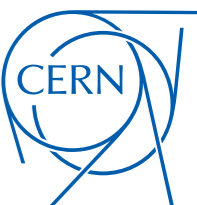
- **Flat Namespace**

- Perfect for HTTP caching
- Minimal storage API requirements (PUT, GET, [DELETE])

	.cvmfspublished
	038f625d0790...
	59bb67e545ac...
	966672a53bec...
	dae8d8c36714...
	001e62aad4c6...
	b53283980b78...
	949324a4d8ef...
	11f58905f87d...
	2ee4f4dfc208...
	a11b16694d6a...
	949324a4d8ef...
	c6fb1c3da08b...
	5c1c639b9a39...
	6176e4e8ce93...
	8599d27418cf...
	17b9caf70786...



Stratum0
(backend storage)



From POSIX to Blob-Objects

- **Merkle tree**

- only .cvmfspublished needs to be signed

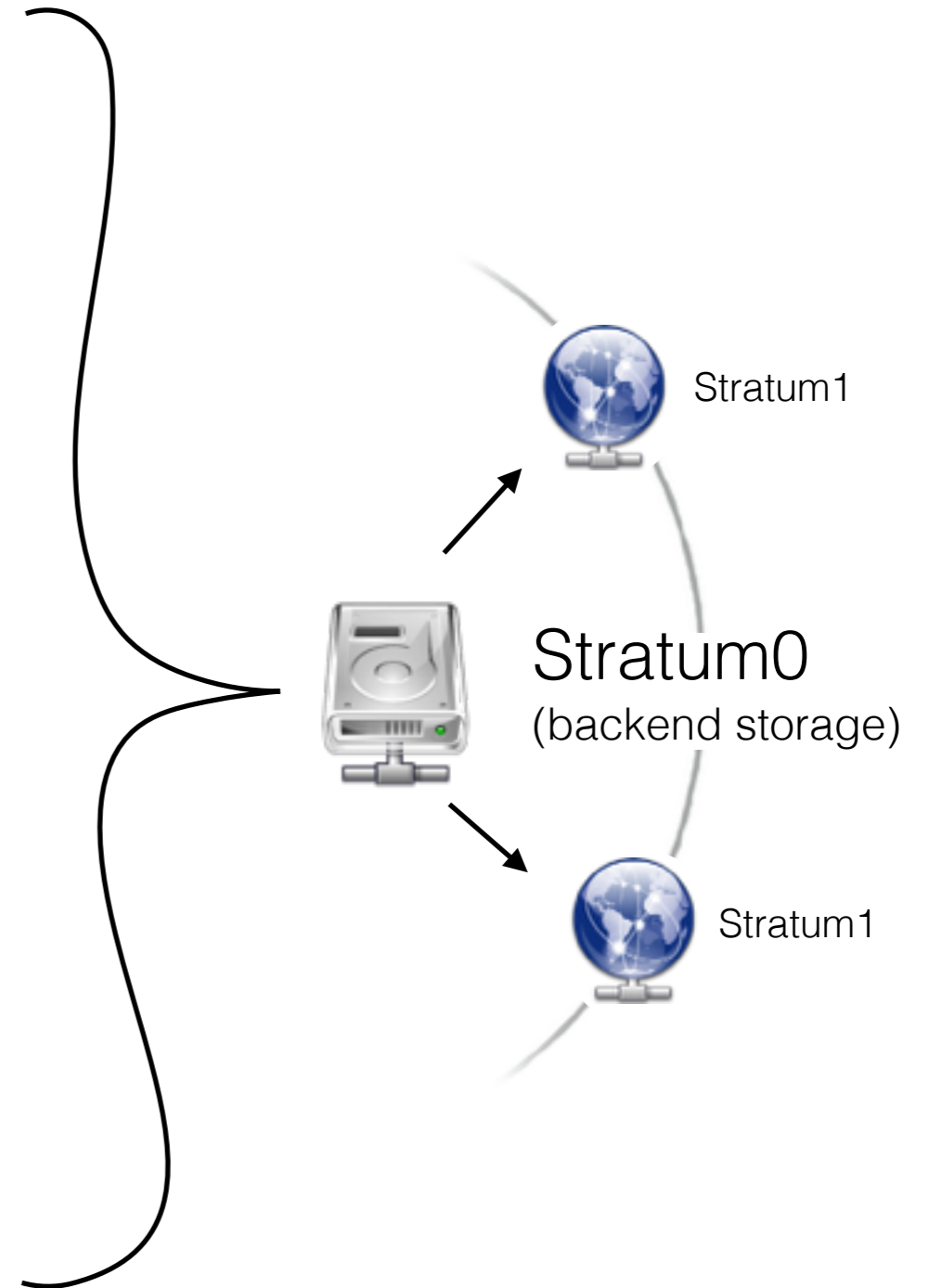
- **Content-Addressable Storage**

- File de-duplication
- Trivial file integrity checks

- **Flat Namespace**

- Perfect for HTTP caching
- Minimal storage API requirements (PUT, GET, [DELETE])

	.cvmfspublished
	038f625d0790...
	59bb67e545ac...
	966672a53bec...
	dae8d8c36714...
	001e62aad4c6...
	b53283980b78...
	949324a4d8ef...
	11f58905f87d...
	2ee4f4dfc208...
	a11b16694d6a...
	949324a4d8ef...
	c6fb1c3da08b...
	5c1c639b9a39...
	6176e4e8ce93...
	8599d27418cf...
	17b9caf70786...



From POSIX to Blob-Objects

- **Merkle tree**





- only .cvmfspublished needs to be signed

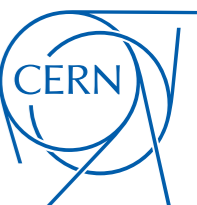
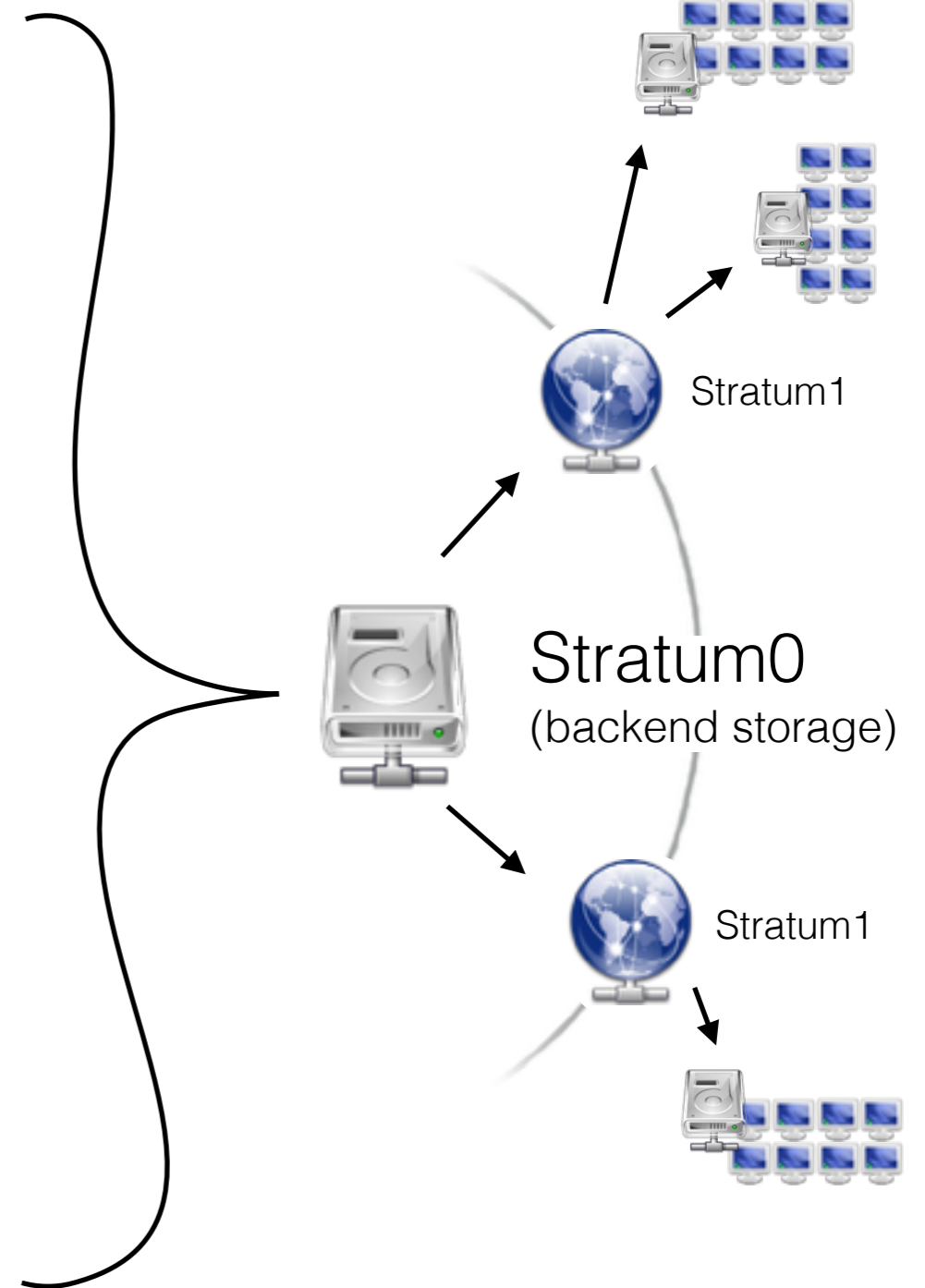
- **Content-Addressable Storage**

- File de-duplication
- Trivial file integrity checks

- **Flat Namespace**

- Perfect for HTTP caching
- Minimal storage API requirements (PUT, GET, [DELETE])

	.cvmfspublished
	038f625d0790...
	59bb67e545ac...
	966672a53bec...
	dae8d8c36714...
	001e62aad4c6...
	b53283980b78...
	949324a4d8ef...
	11f58905f87d...
	2ee4f4dfc208...
	a11b16694d6a...
	949324a4d8ef...
	c6fb1c3da08b...
	5c1c639b9a39...
	6176e4e8ce93...
	8599d27418cf...
	17b9caf70786...



Repository Storage

- **RESTful backend storage** - PUT, GET, (DELETE)

- POSIX compliant file systems

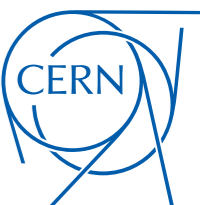


Stratum0
(backend storage)

- Key-Value stores (using S3 API - Seppo Heikkila)

- **Pluggable storage connector implementation**

- Facilitate implementation of native Key-Value store APIs
- Like: Basho Riak, Ceph, Amazon Dynamo, ...



File System History

Named Repository Snapshots

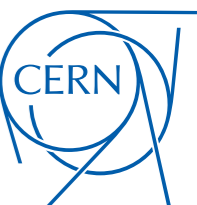
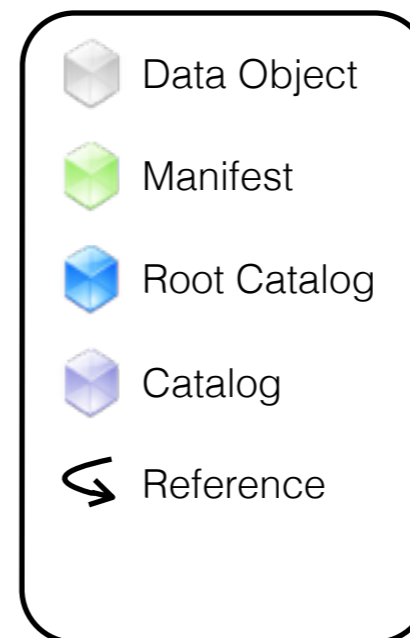


File System History

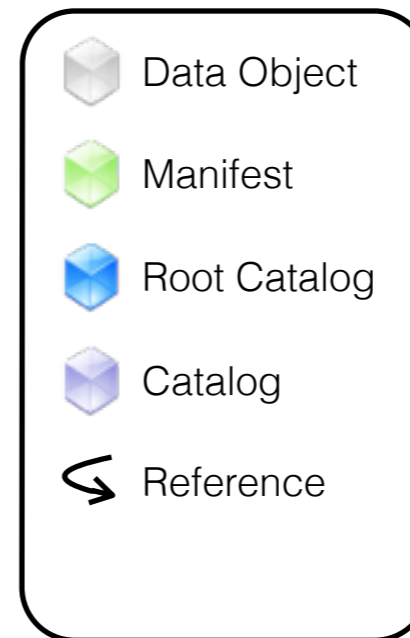
- **Client:** Mount historic revisions of a repository
 - Use legacy software in its contemporary environment
 - Long term data preservation
- **Stratum0:** Rollback to previous revisions
 - Undo - Overwrite broken revisions with previous version



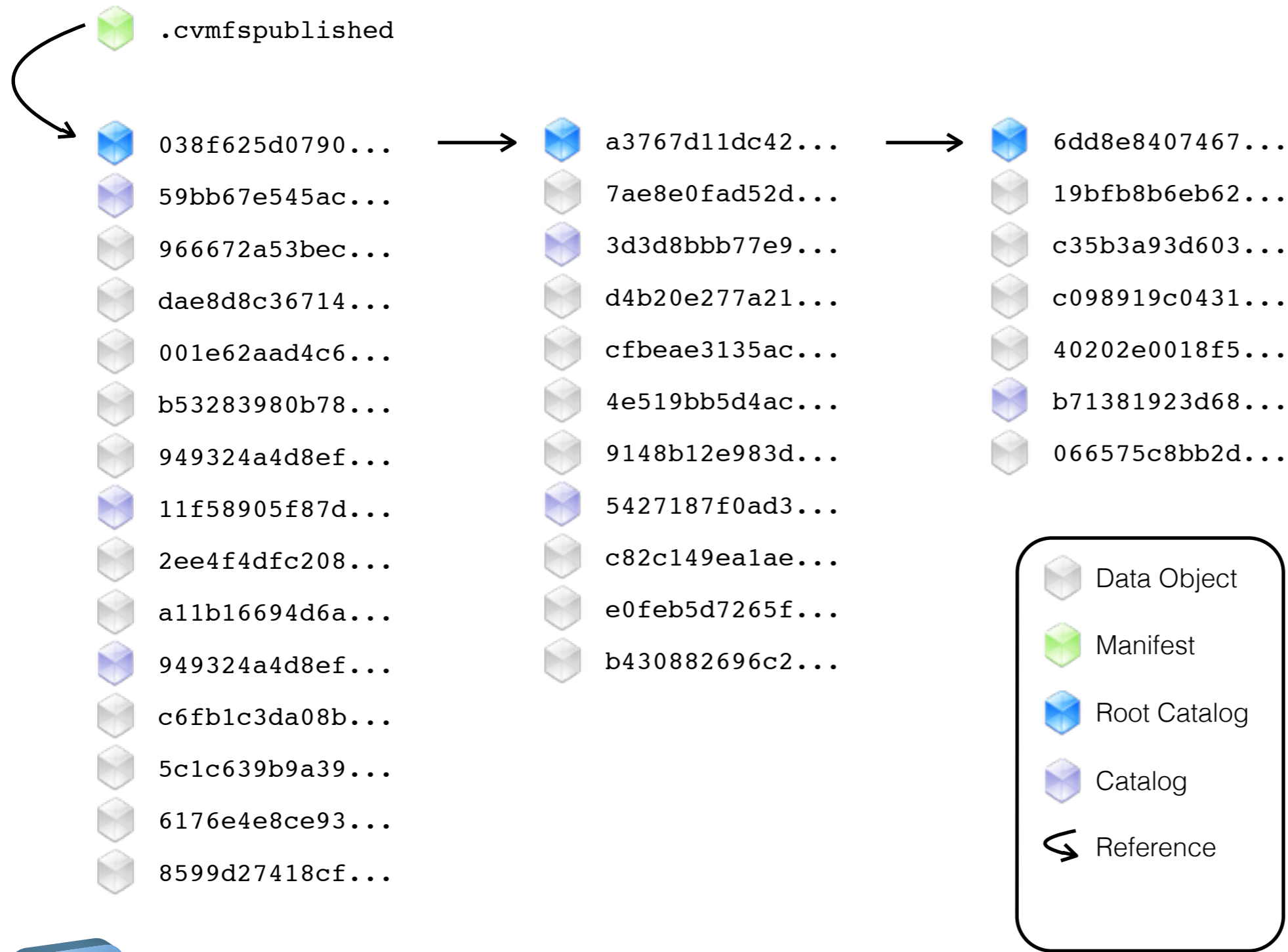
File System History



File System History



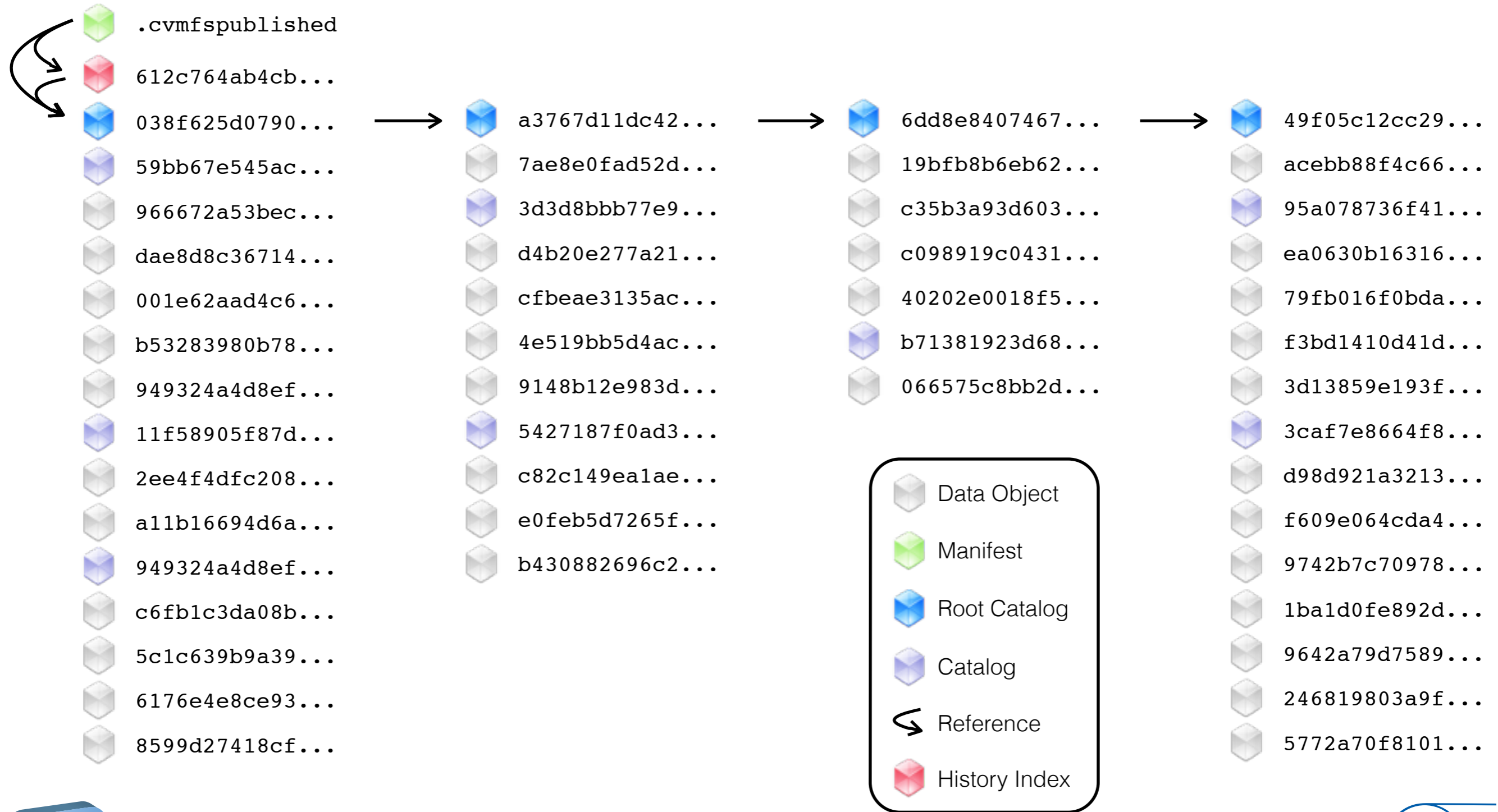
File System History



File System History



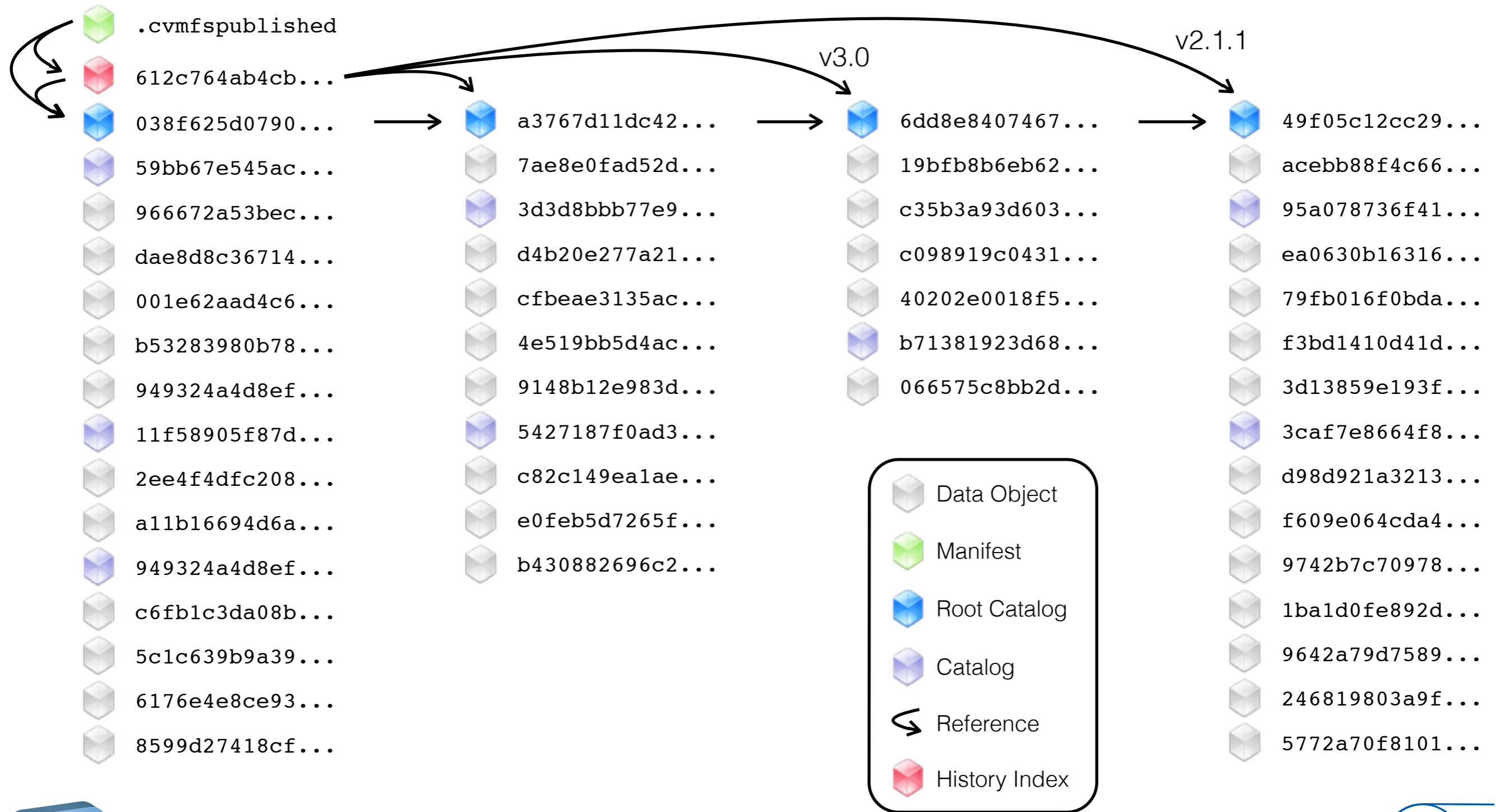
File System History



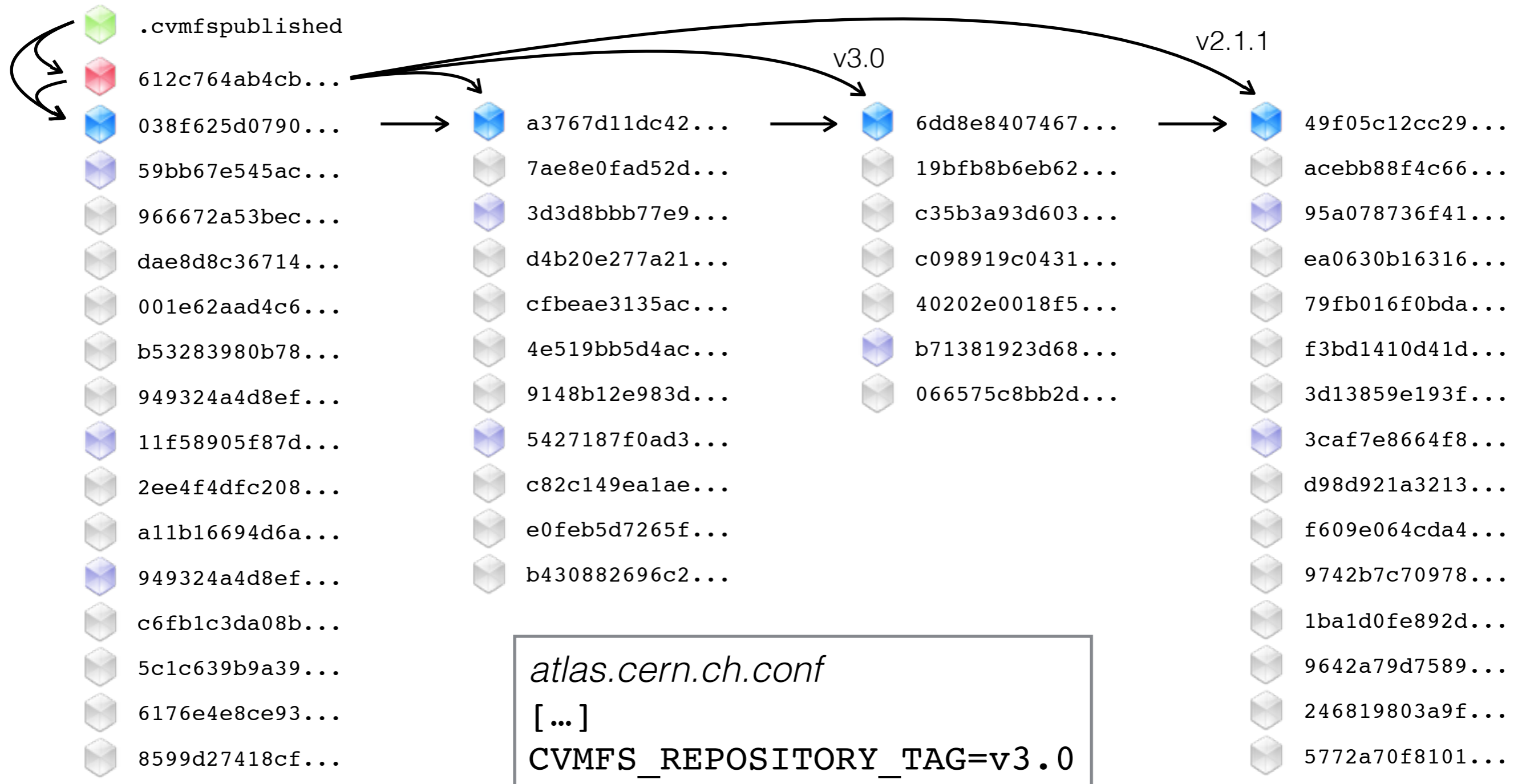
File System History



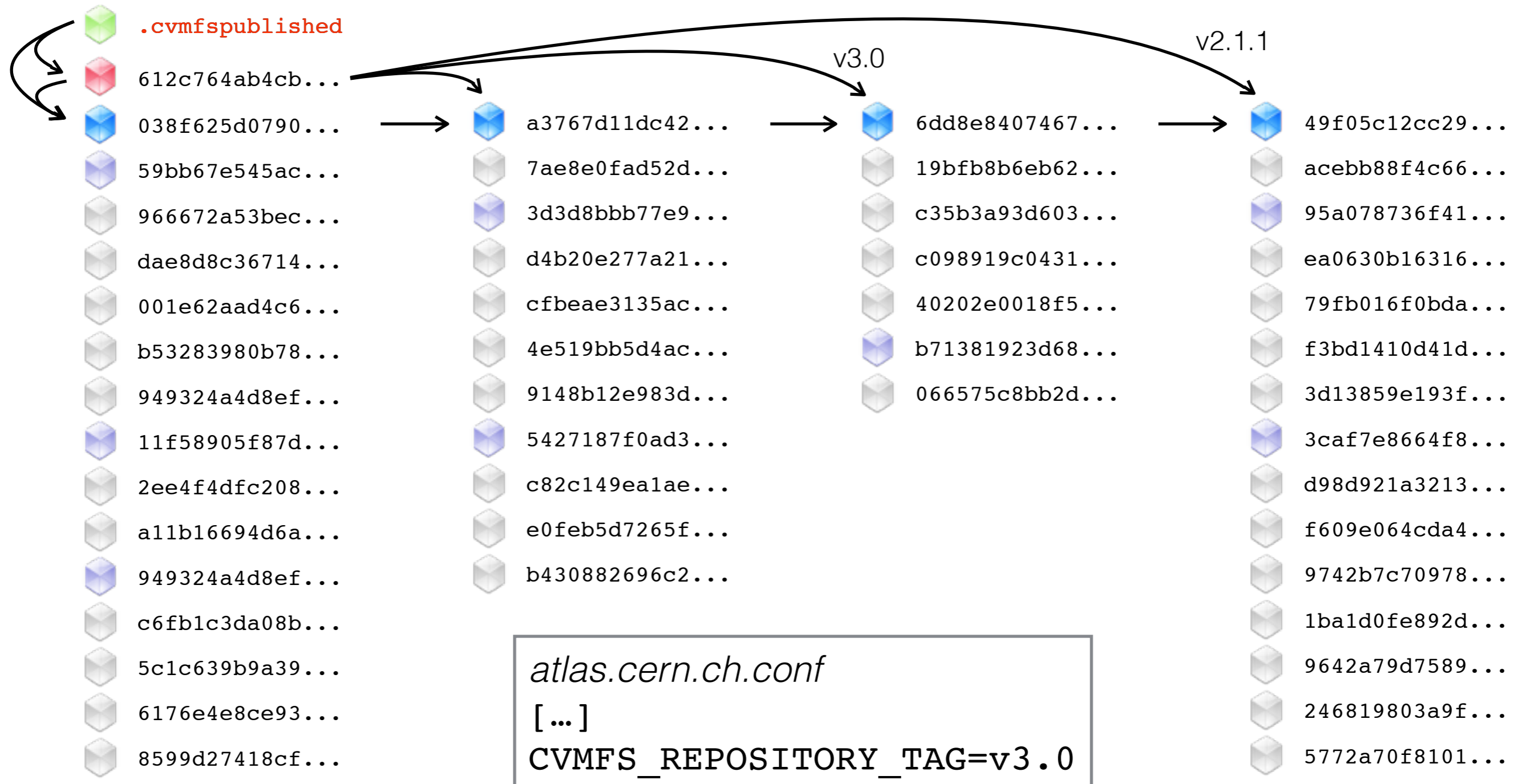
File System History



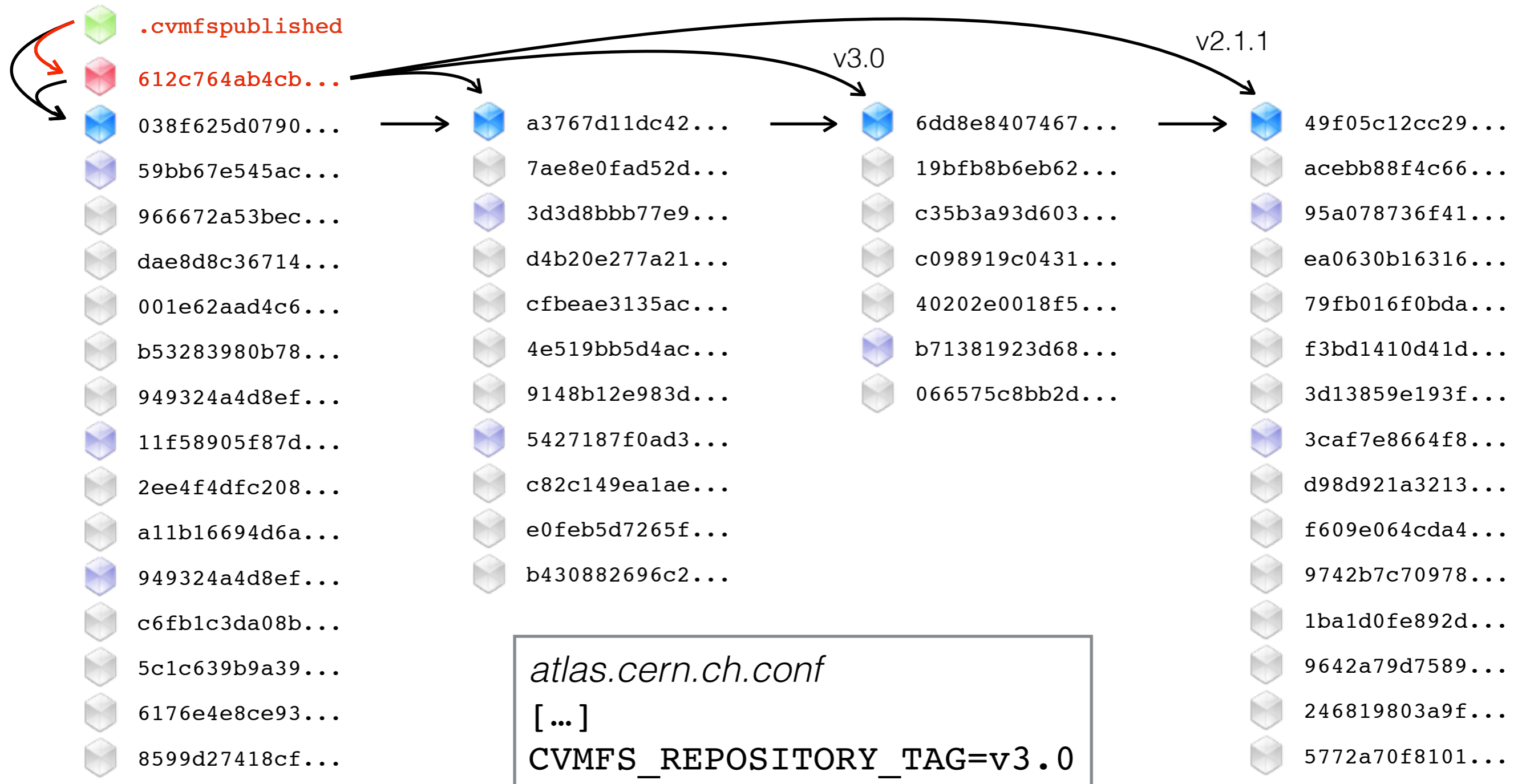
File System History



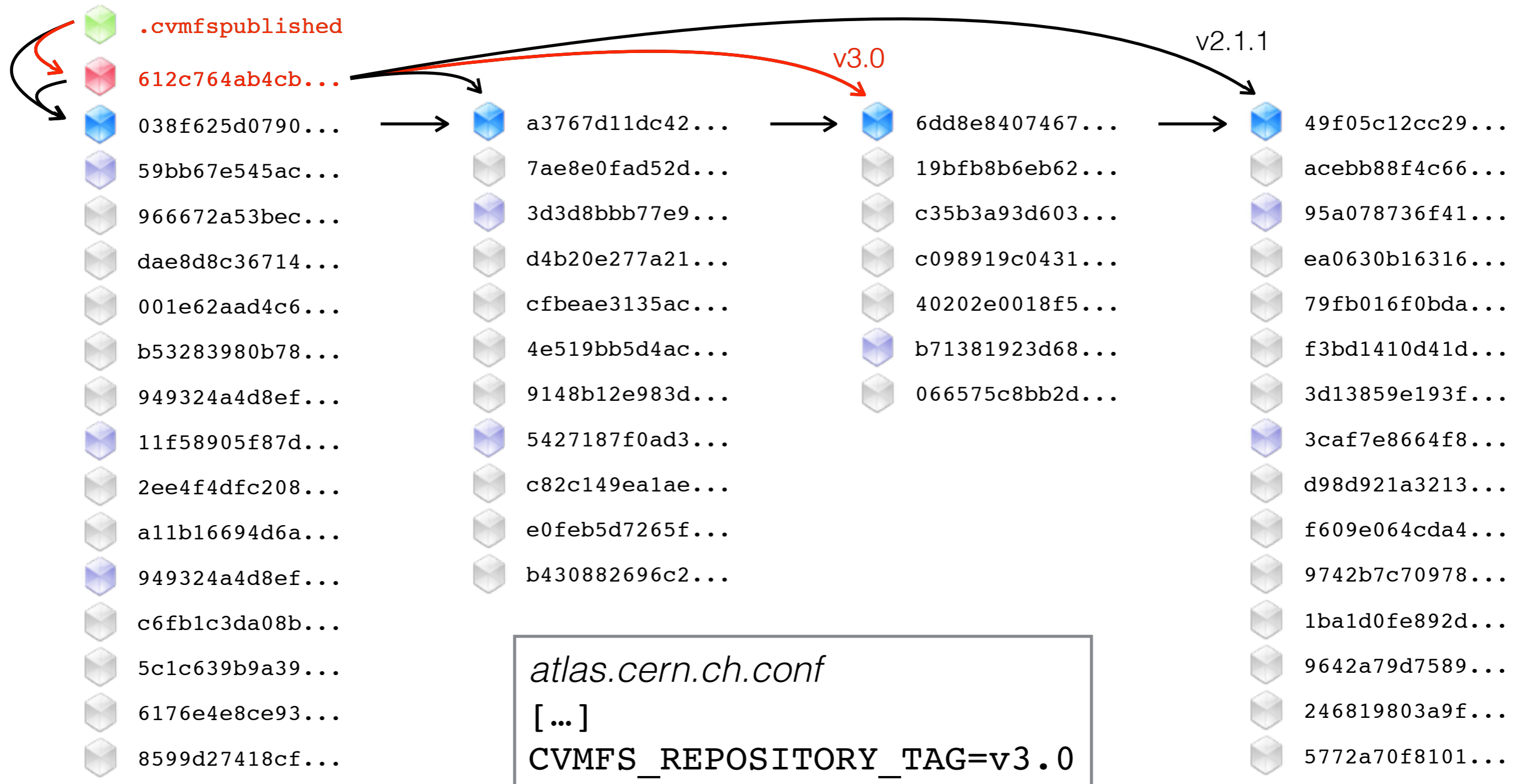
File System History



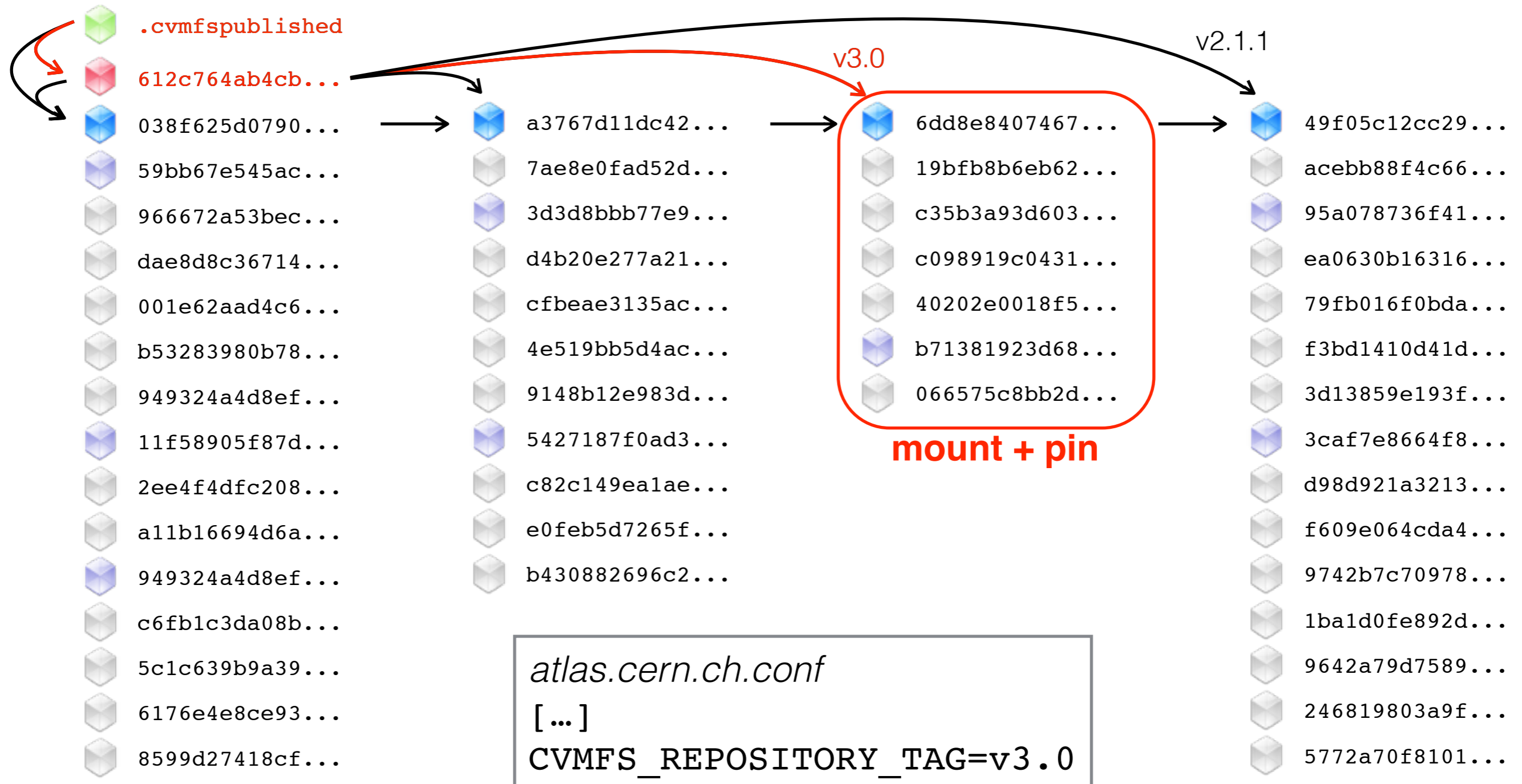
File System History



File System History



File System History



Garbage Collection

Permanently Removing
Overwritten or Deleted Files
in Volatile Repositories

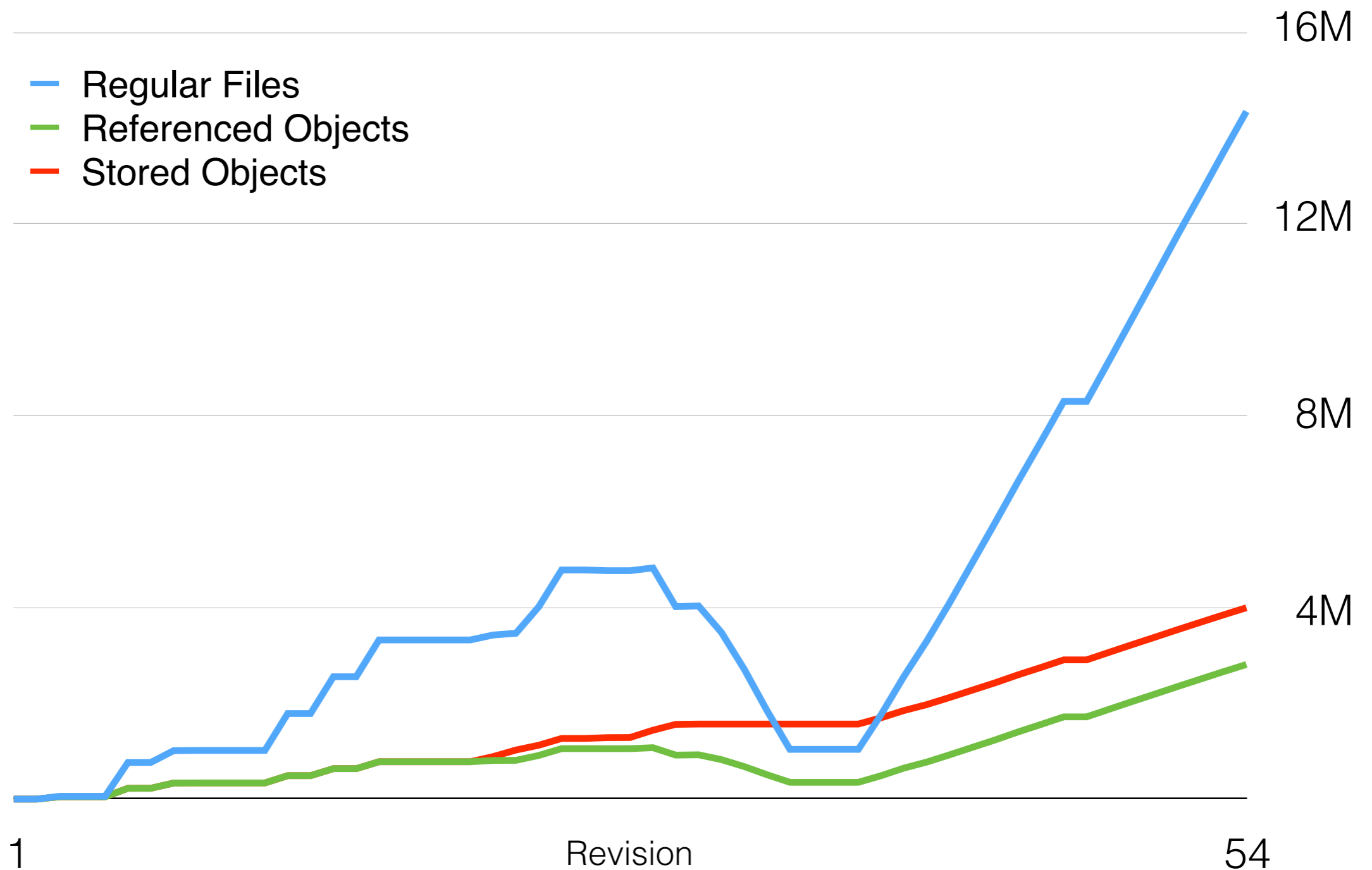


Garbage Collection

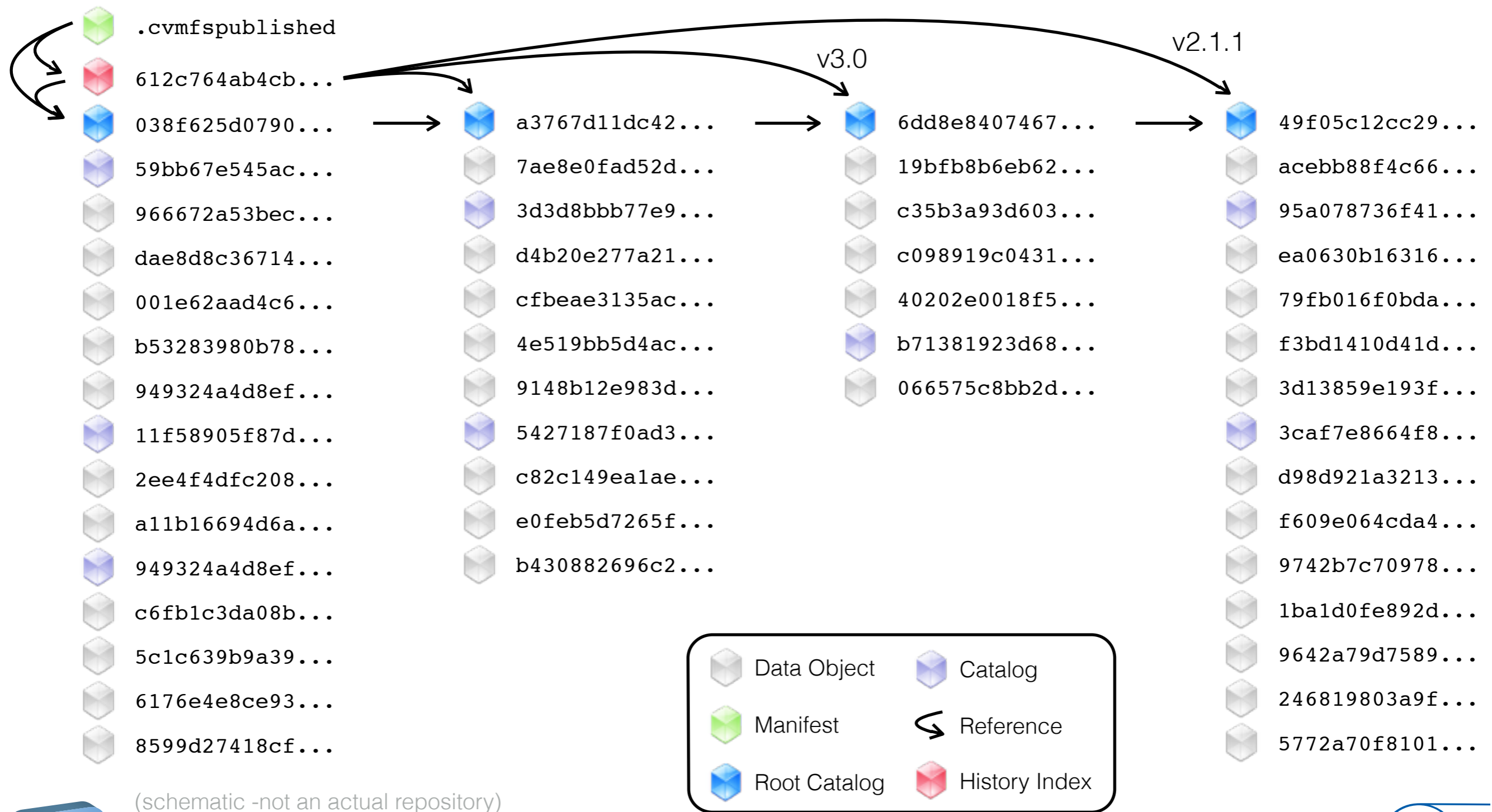
- CernVM-FS backend initially designed as **insert-only**
- **New use-case:** LHC experiment's nightly build repositories
 - high update rate (up to twice a day)
 - large volume of newly staged files (10-100GiB)
 - short lived revisions (stay online max. two weeks)
- Insert-only quickly fills up backend storage!



LHCb Nightly Builds



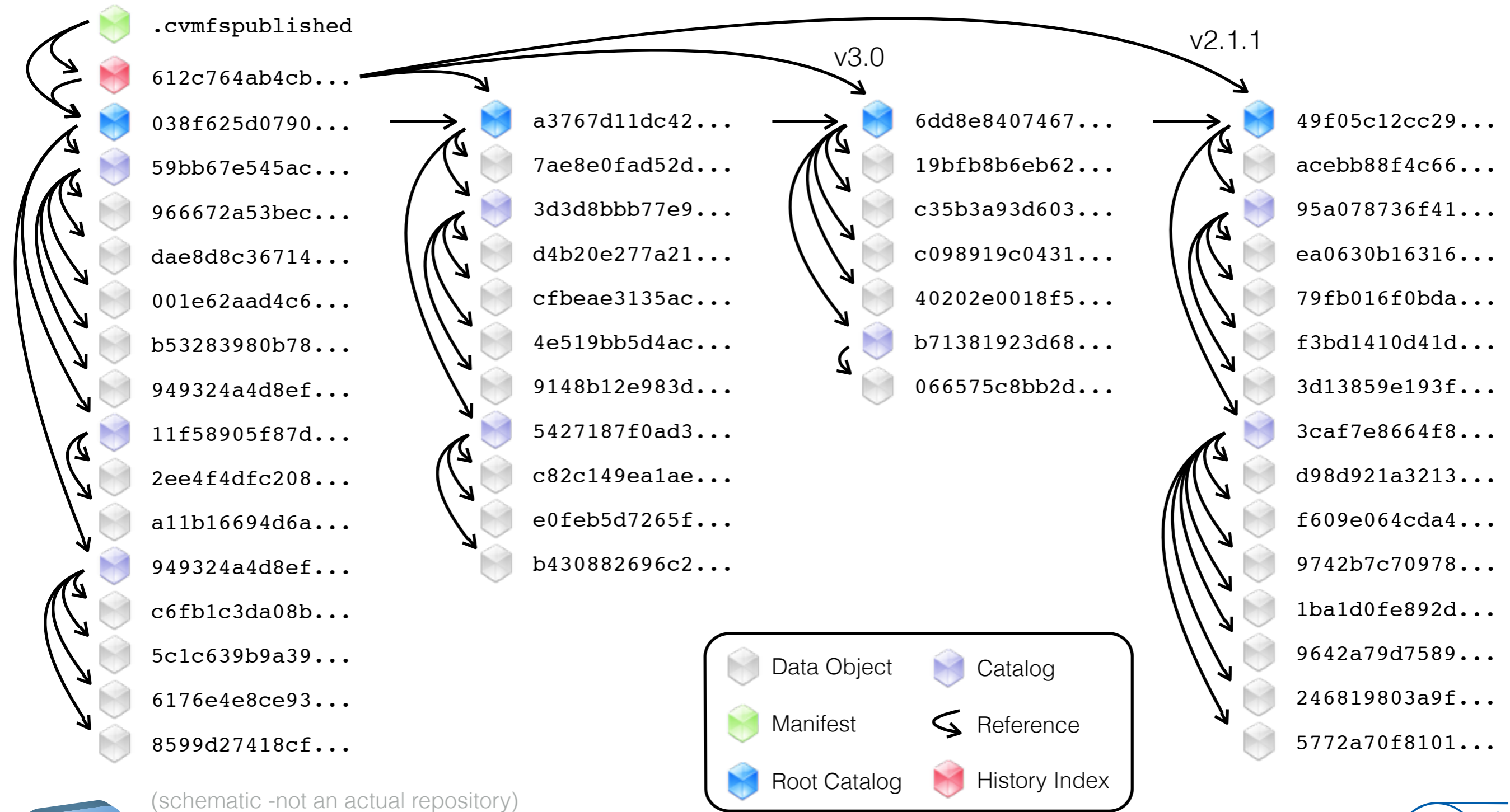
Garbage Collection



(schematic -not an actual repository)



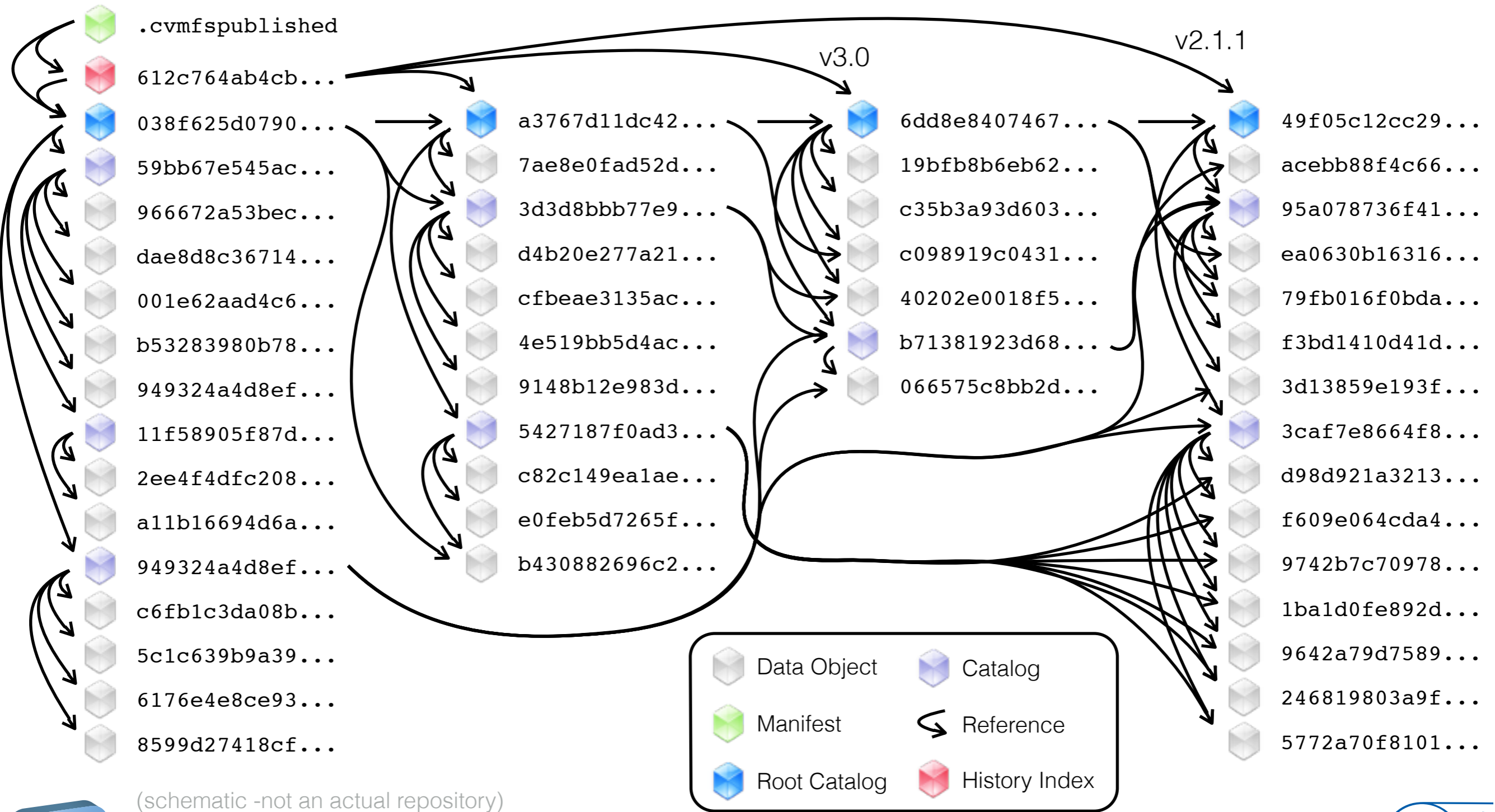
Garbage Collection



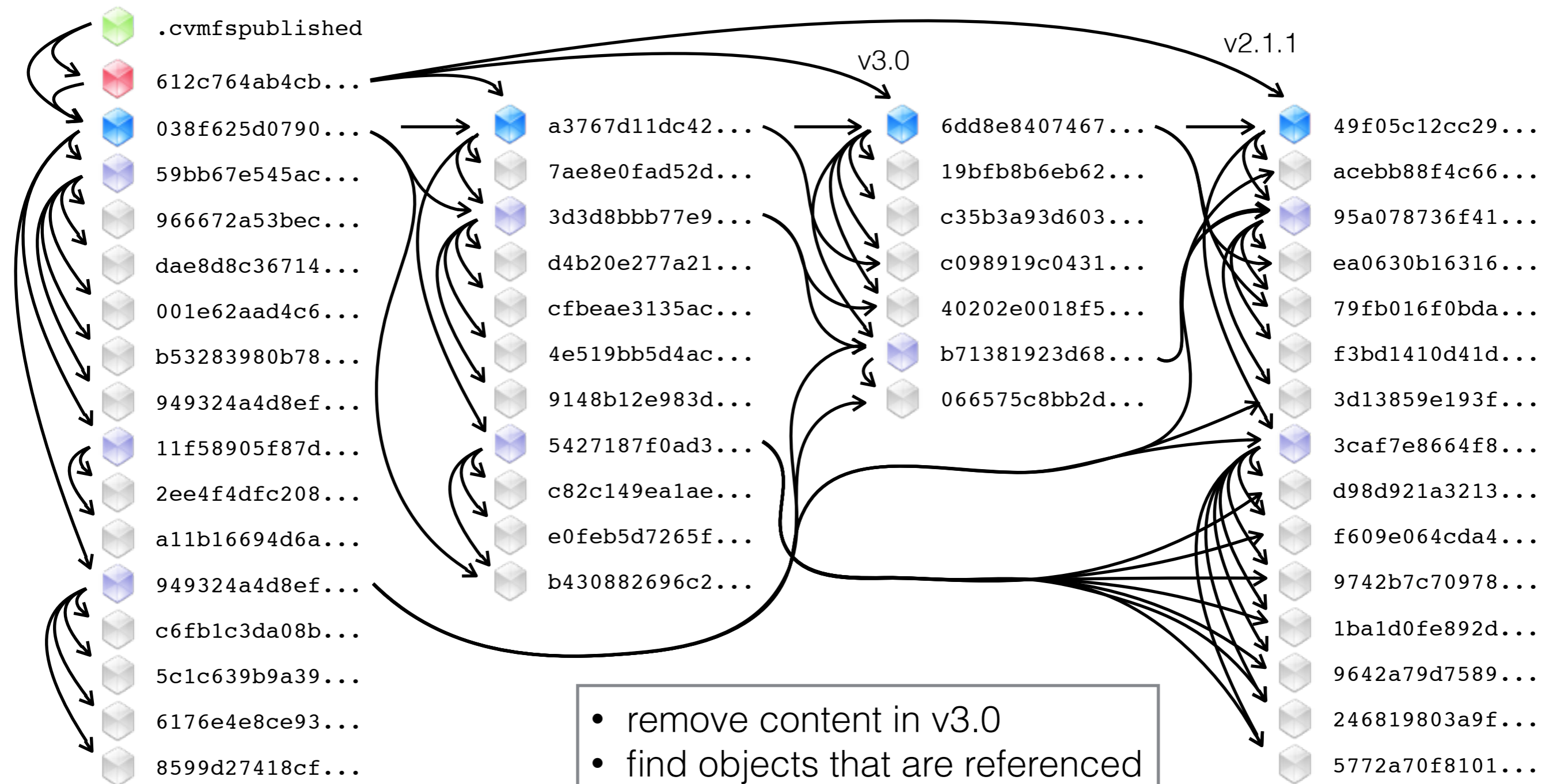
(schematic -not an actual repository)



Garbage Collection



Garbage Collection

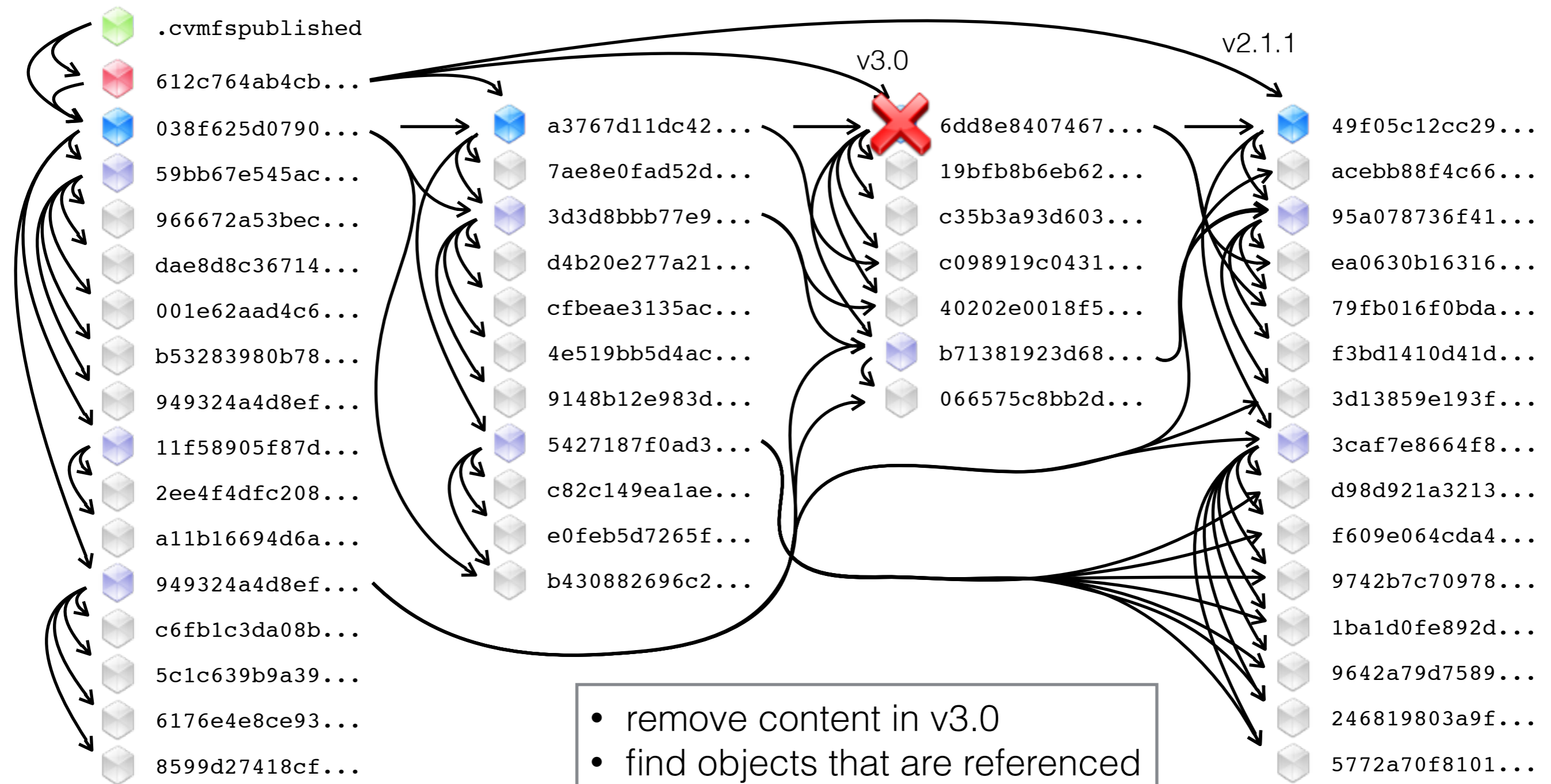


(schematic -not an actual repository)

- remove content in v3.0
- find objects that are referenced nowhere else



Garbage Collection

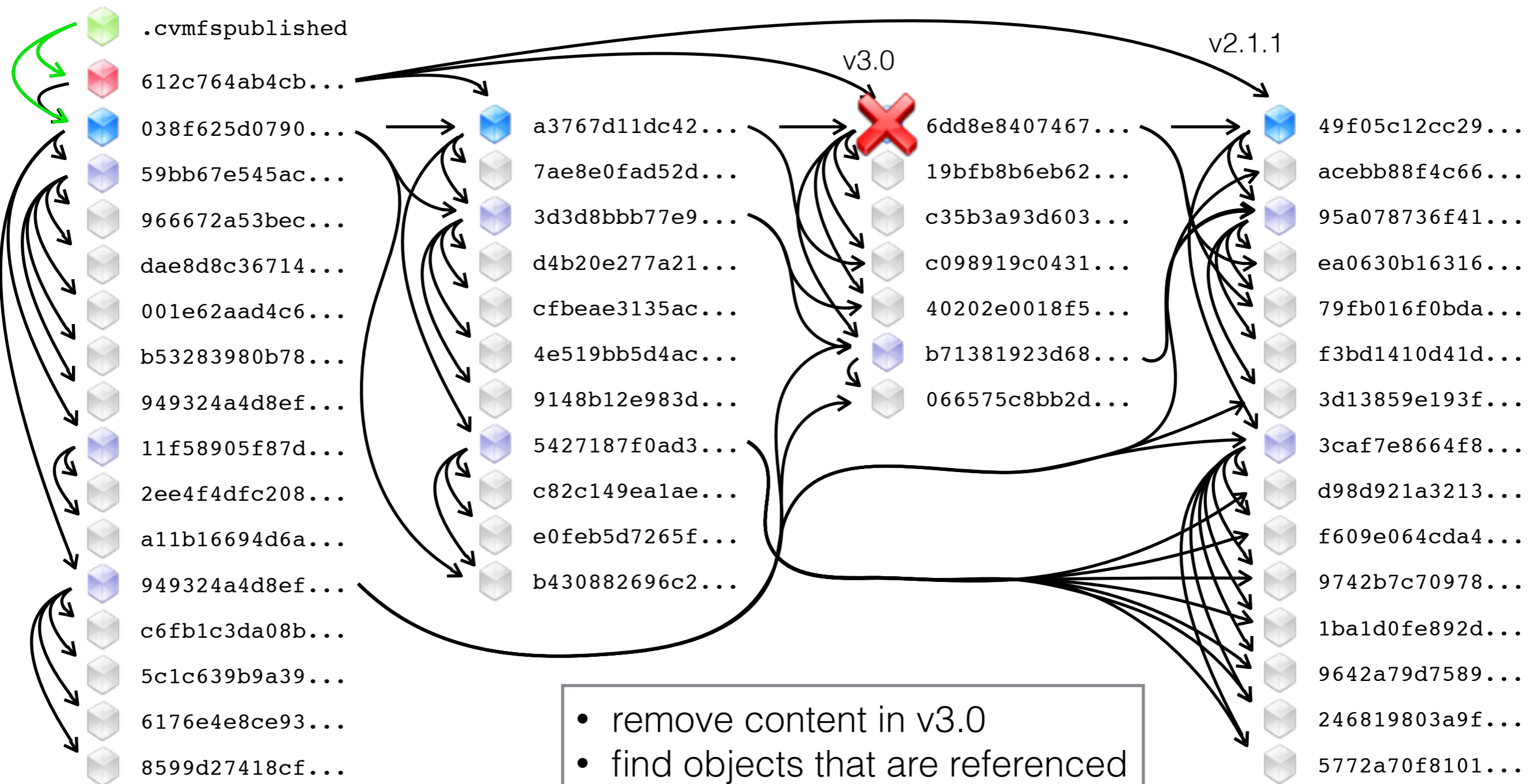


(schematic -not an actual repository)

- remove content in v3.0
- find objects that are referenced nowhere else



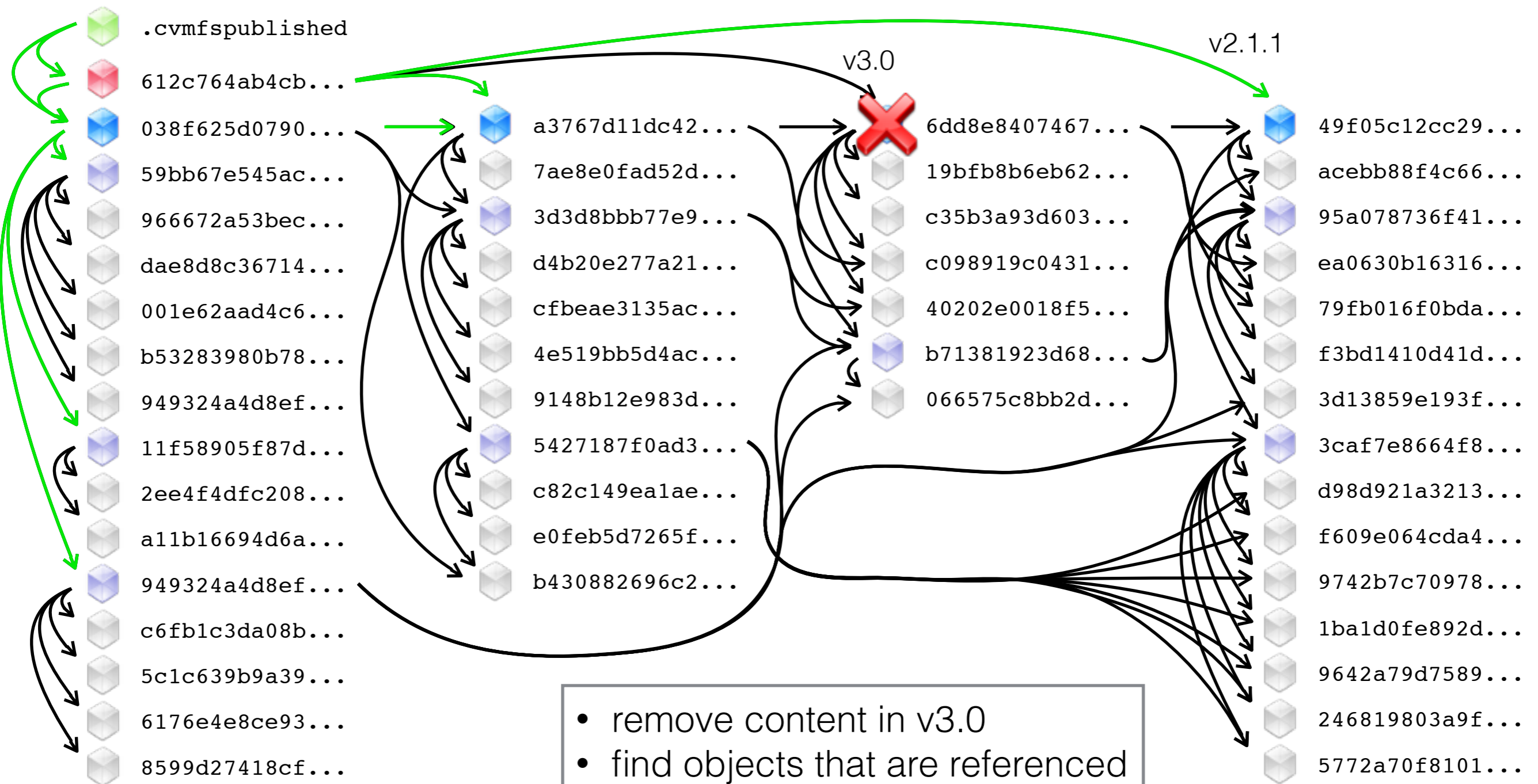
Garbage Collection



(schematic -not an actual repository)



Garbage Collection

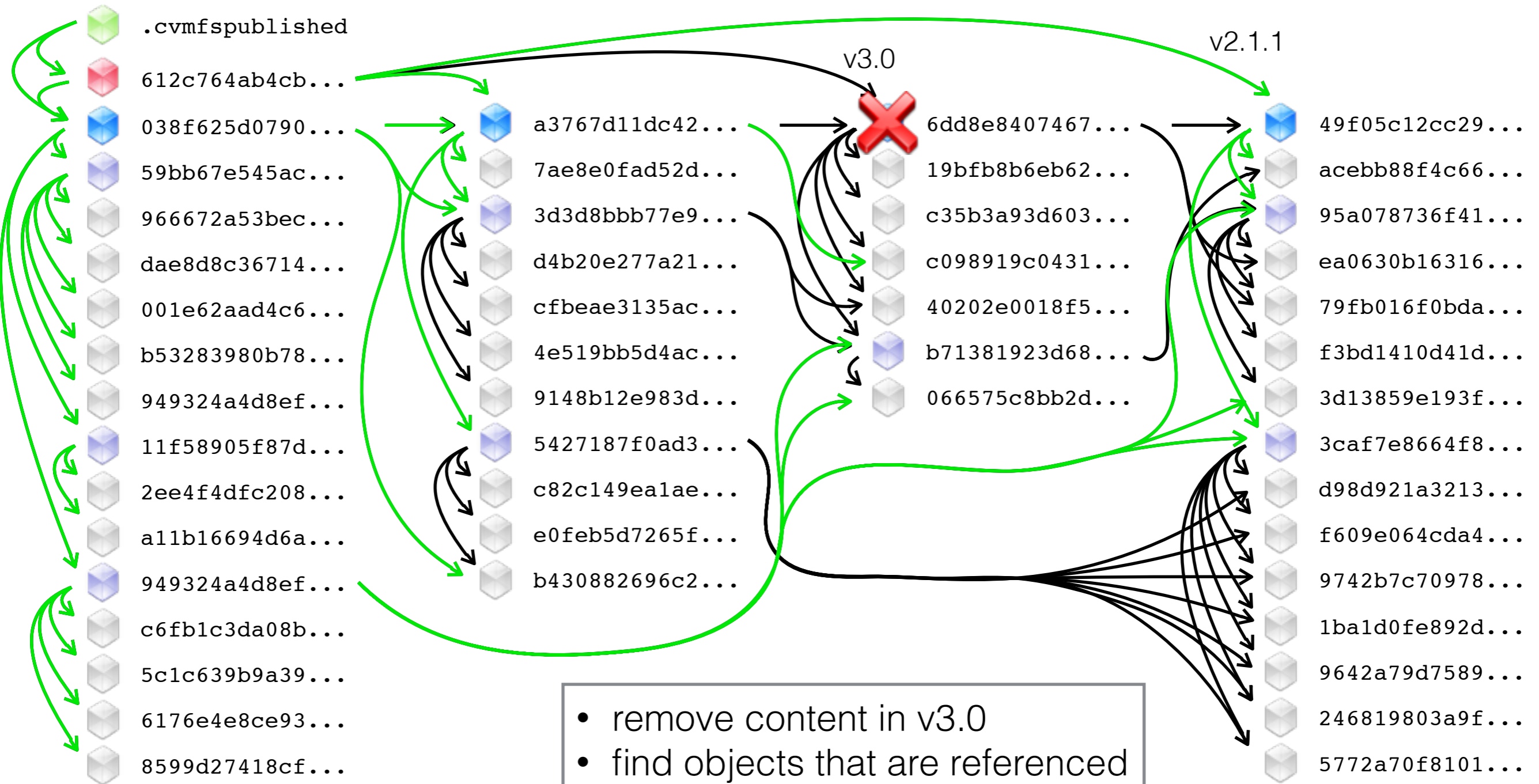


(schematic -not an actual repository)

- remove content in v3.0
- find objects that are referenced nowhere else



Garbage Collection



(schematic -not an actual repository)

- remove content in v3.0
- find objects that are referenced nowhere else



Garbage Collection



(schematic -not an actual repository)

- remove content in v3.0
- find objects that are referenced nowhere else



Garbage Collection



(schematic -not an actual repository)

- remove content in v3.0
- find objects that are referenced nowhere else



Garbage Collection



- remove content in v3.0
- find objects that are referenced nowhere else

(schematic -not an actual repository)



Garbage Collection

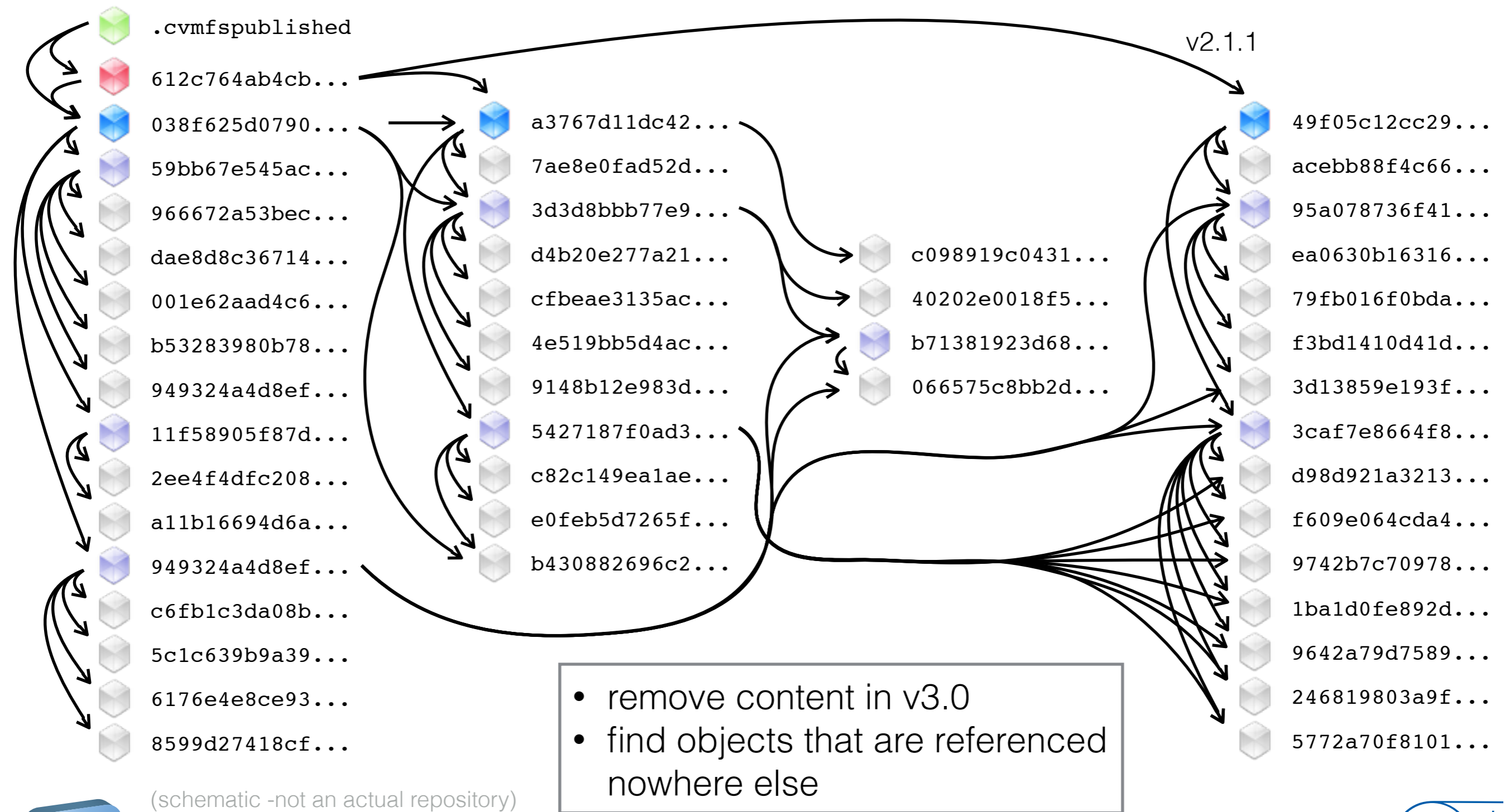


(schematic -not an actual repository)

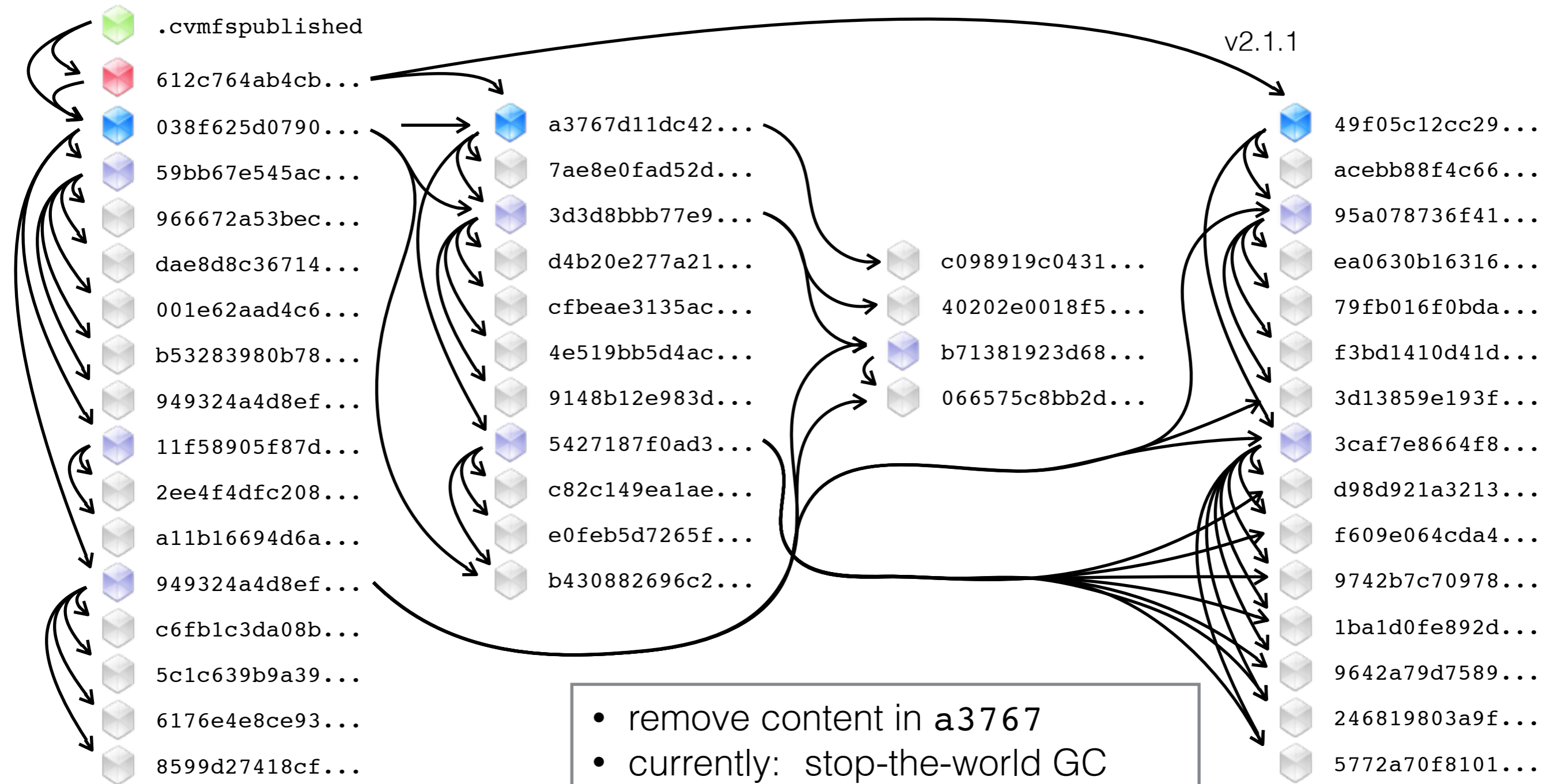
- remove content in v3.0
- find objects that are referenced nowhere else



Garbage Collection



Garbage Collection

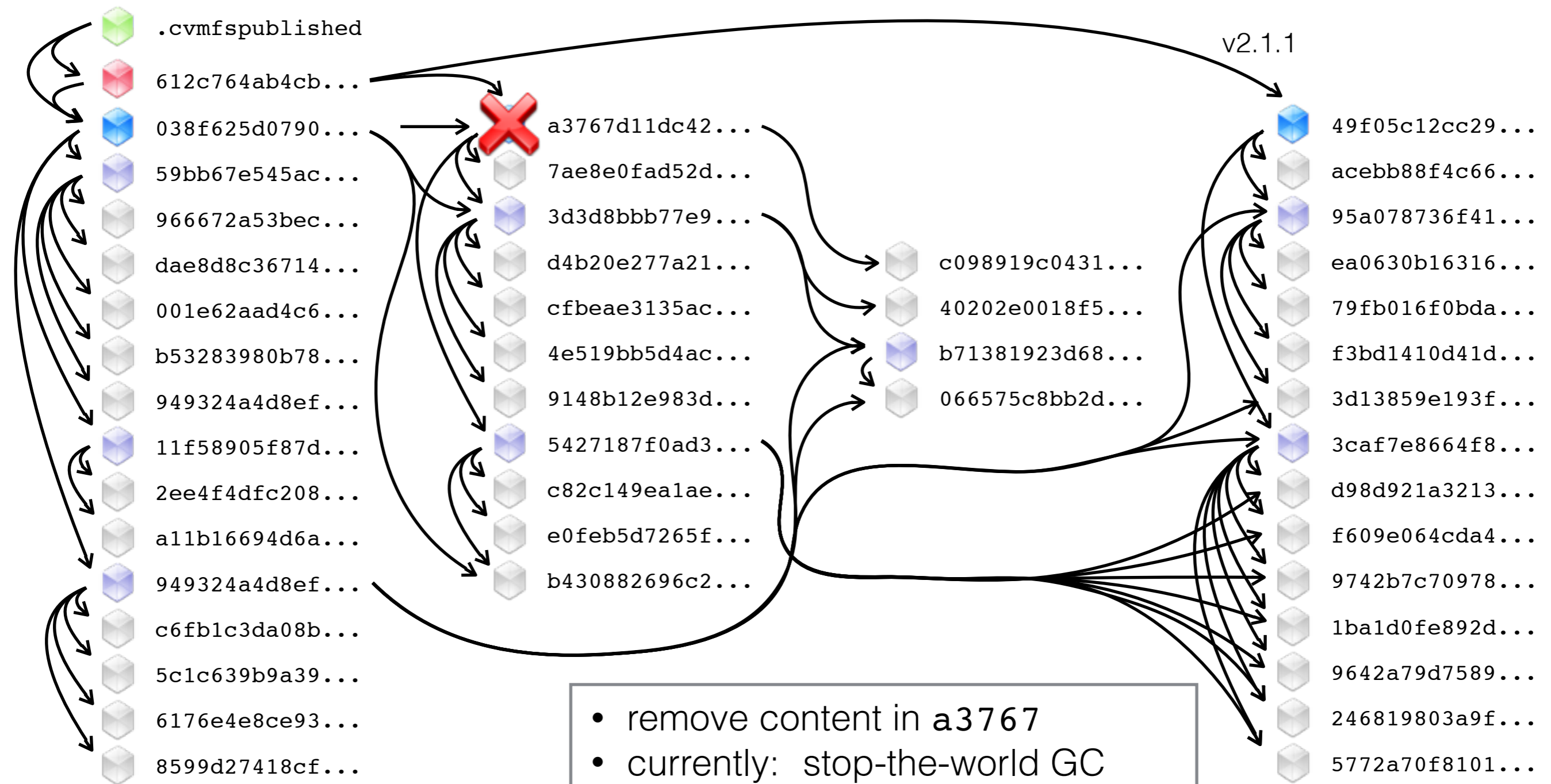


(schematic -not an actual repository)

- remove content in a3767
- currently: stop-the-world GC
no concurrent publish



Garbage Collection



(schematic -not an actual repository)

- remove content in a3767
- currently: stop-the-world GC
no concurrent publish



Garbage Collection



(schematic -not an actual repository)

- remove content in a3767
- currently: stop-the-world GC
no concurrent publish



Garbage Collection



- remove content in a3767
- currently: stop-the-world GC
no concurrent publish

(schematic -not an actual repository)



Garbage Collection



(schematic -not an actual repository)

- remove content in a3767
- currently: stop-the-world GC
no concurrent publish



Garbage Collection



(schematic -not an actual repository)

- remove content in a3767
- currently: stop-the-world GC
no concurrent publish



Garbage Collection



(schematic -not an actual repository)



Garbage Collection

- **Mark-and-Sweep implementation**
- Two-stage approach:
 - Traverse preserved catalogs and log referenced objects
 - Traverse condemned catalogs and match against log
- Full walk of the repository's catalog graph required
- Time consuming task



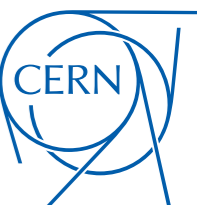
Smart Stratum1 Servers

Automatic Stratum1 Ordering,
Push Replication to Stratum1 Servers



Smart Stratum1 Servers

- Equip Stratum1 servers with **RESTful API**
- **Automatic Stratum1 ordering** on the client side
 - Based on GeoIP database to determine closest replicas
 - (to come in CernVM-FS 2.1.20 - Dave Dykstra - Fermilab/OSG)
- **Triggered (instant) replication** of new revisions
 - Repository's private key for authentication



Wrap up



New Challenges



Large Files (> 200 MiB)

→ File chunking



Long term data preservation

→ Named repository snapshots



Rapidly changing repository content

→ Garbage collection



Increasing configuration distribution effort

→ Central bootstrapping configuration repository



Instant repository update propagation

→ Push replication from Stratum0 to Stratum1



Main New Features

- **Alternative Storage Backends** based on key-value stores through the S3 API (Seppo Heikkila)
- **Named Snapshots and History** for long term software accessibility and error recovery
- **Garbage Collection** for rapidly changing repositories (expected in CernVM-FS 2.1.20)
- **Transactional Repository Updates** for better publishing performance and a robust backend



Other New Features

- **Automatic Ordering of Stratum1 Mirrors** based on geo-IP location (Dave Dykstra - OSG)
- **Configuration Bootstrap Repositories** to facilitate public key and configuration distribution (not yet released)
- **Chunking of Large Files** for better cache exploitation and traffic efficiency
- **Fully Parallel File Processing** to speed up snapshot publishing



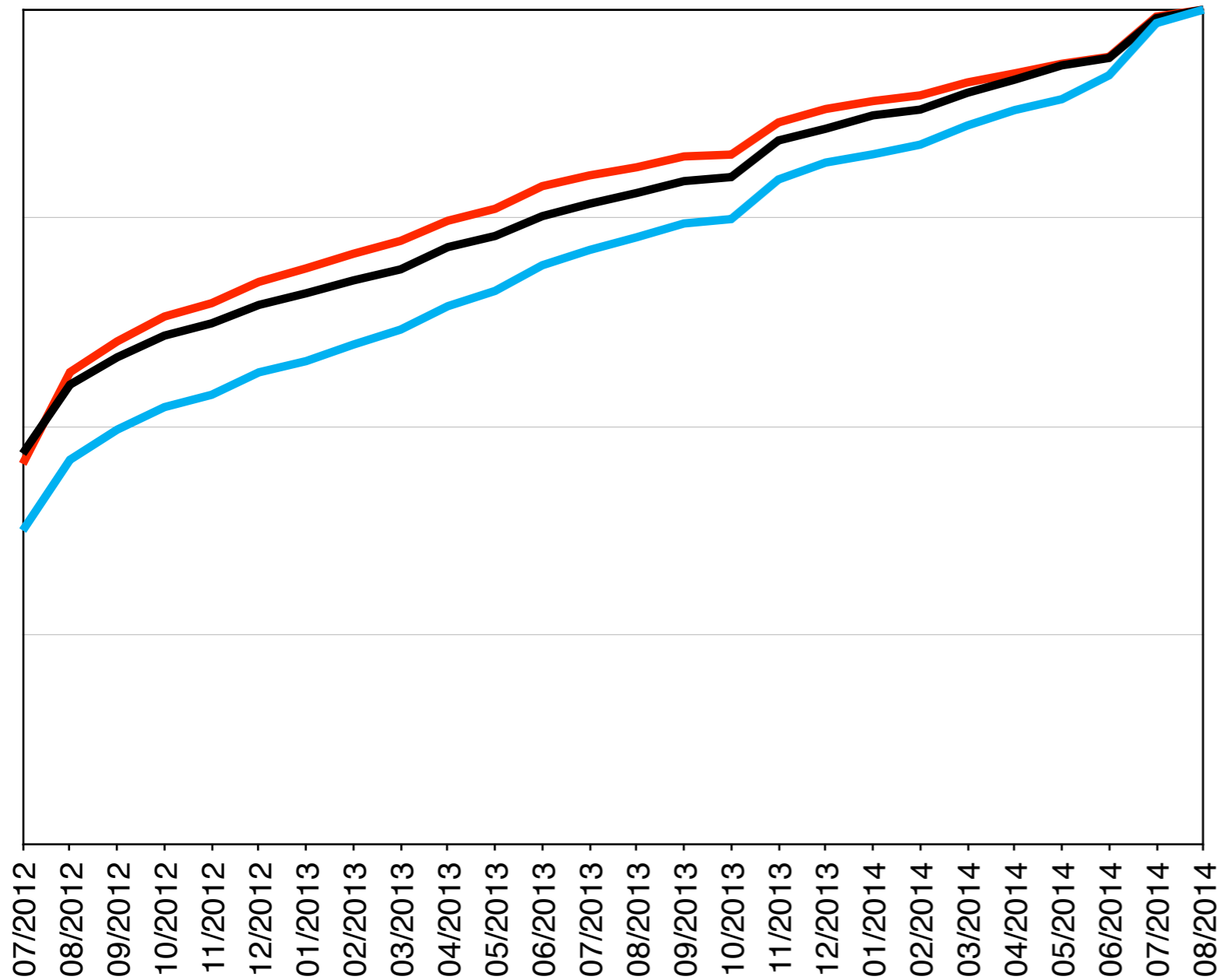


Backup Slides

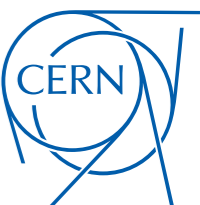


Repository Growth

— Data Volume — Referenced Objects — Directory Entries



- Example Repository: **atlas.cern.ch**
- Size approximately doubled in two years
- Maximal values:
 - Data: 2.1 TiB
 - Entries: 48.0 M
 - Objects: ~3.8 M



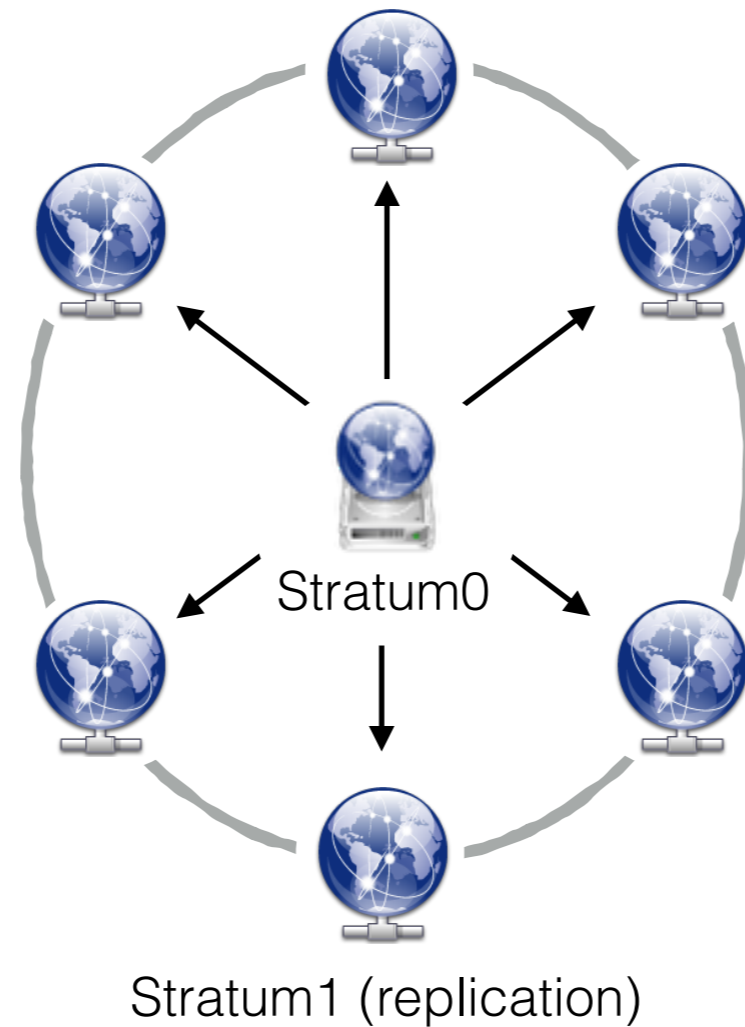
What is CernVM-FS?



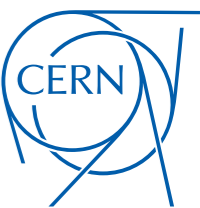
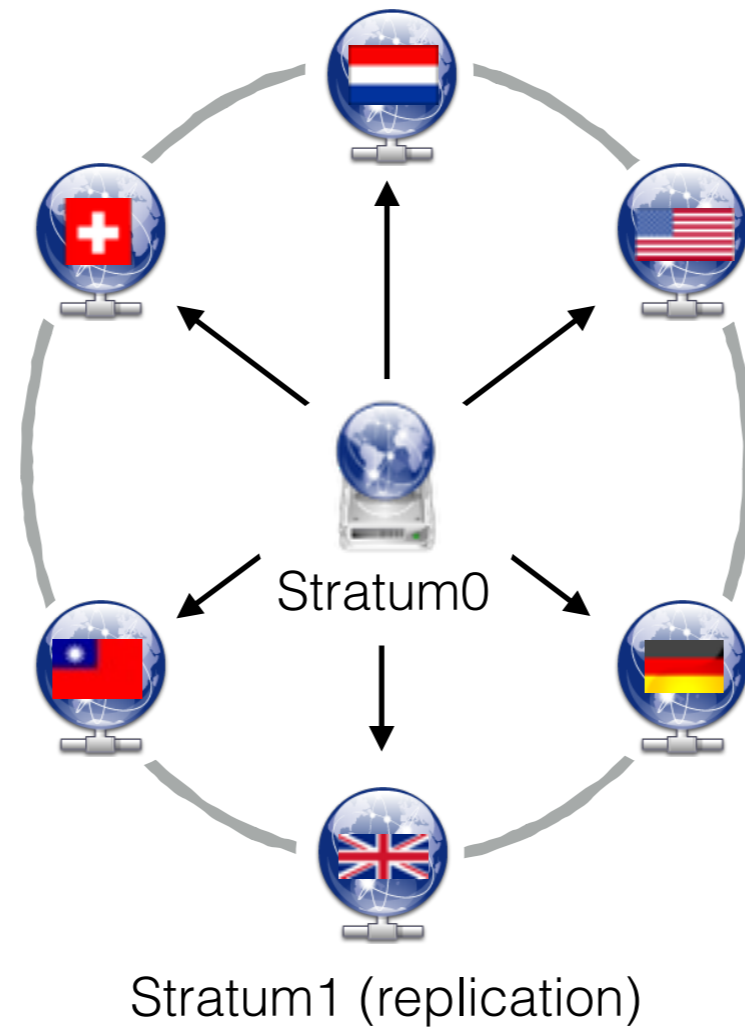
Stratum0



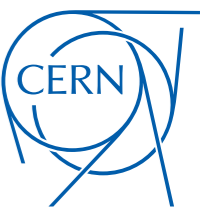
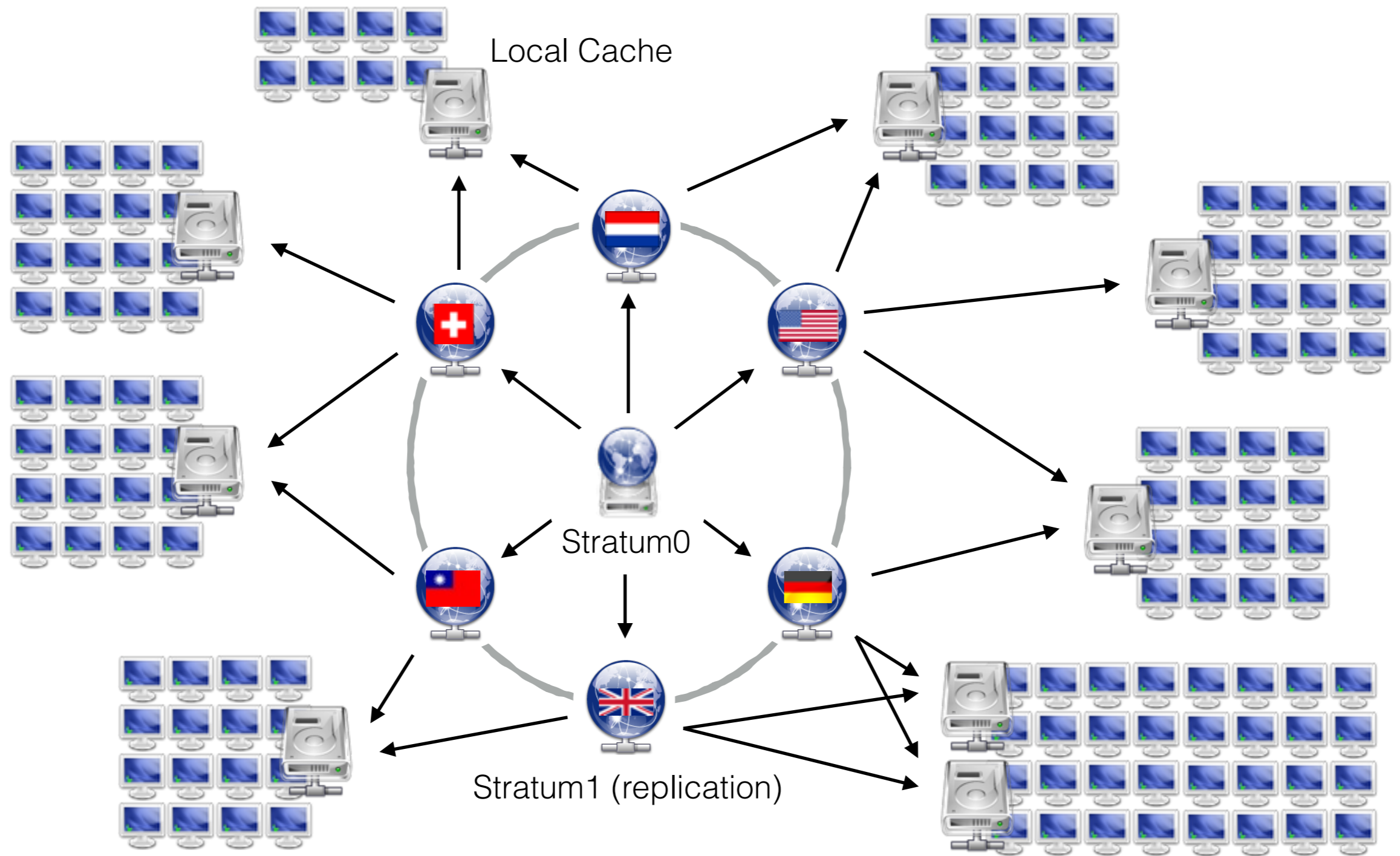
What is CernVM-FS?



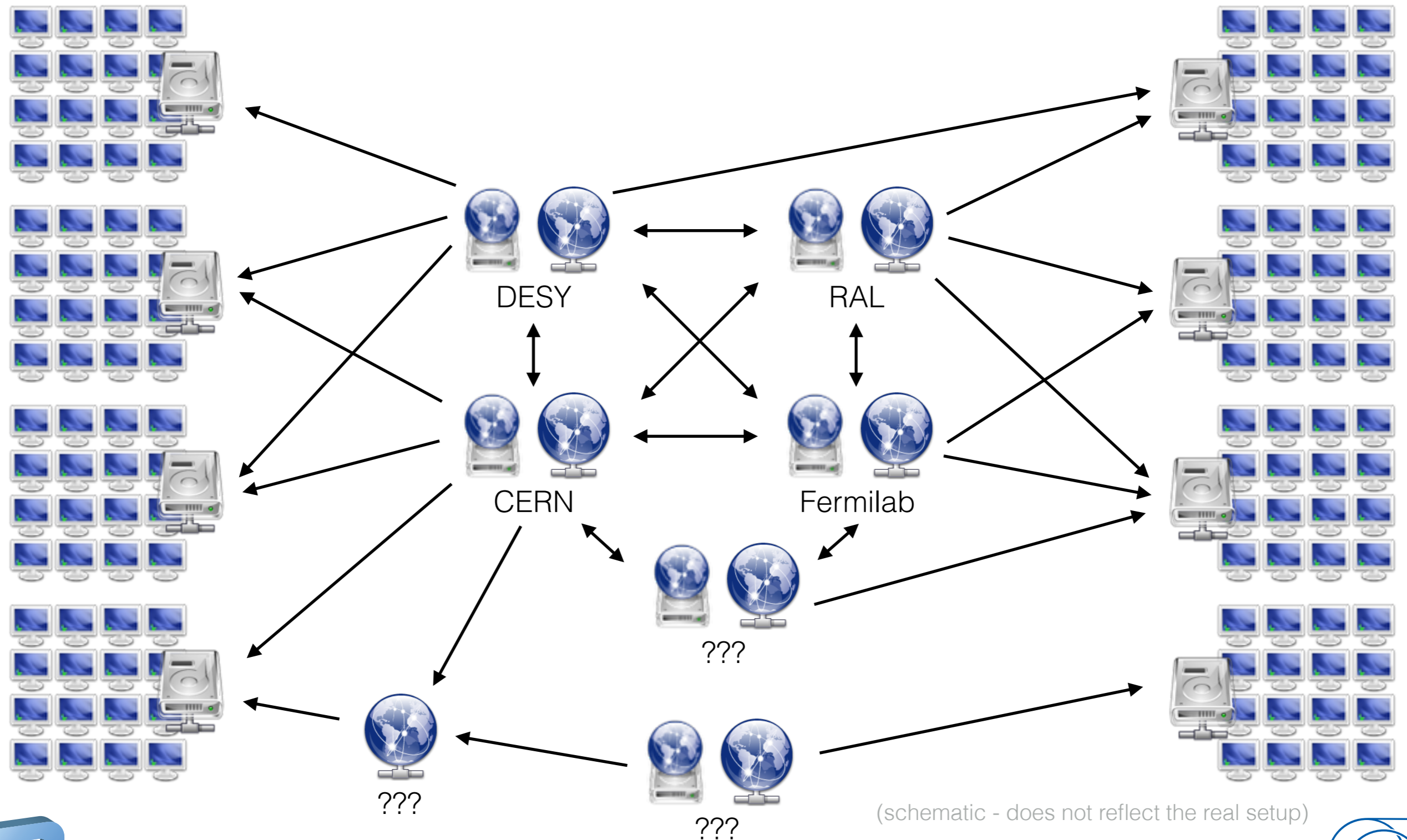
What is CernVM-FS?

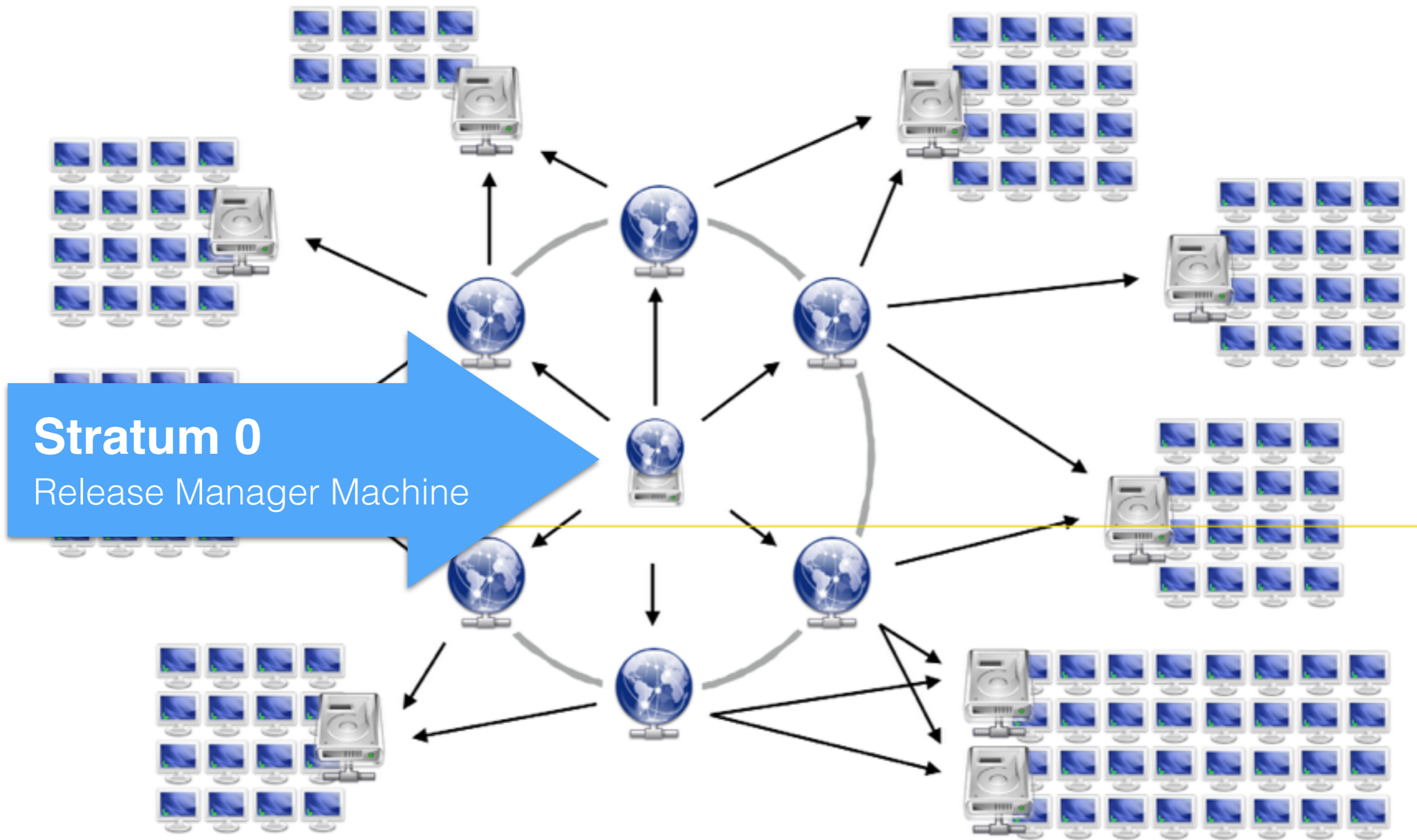


What is CernVM-FS?



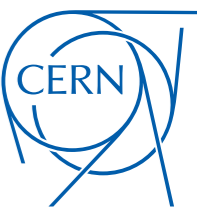
What is CernVM-FS?





Stratum 0

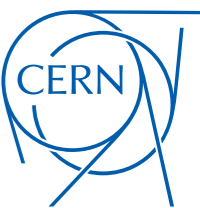
Release Manager Machine



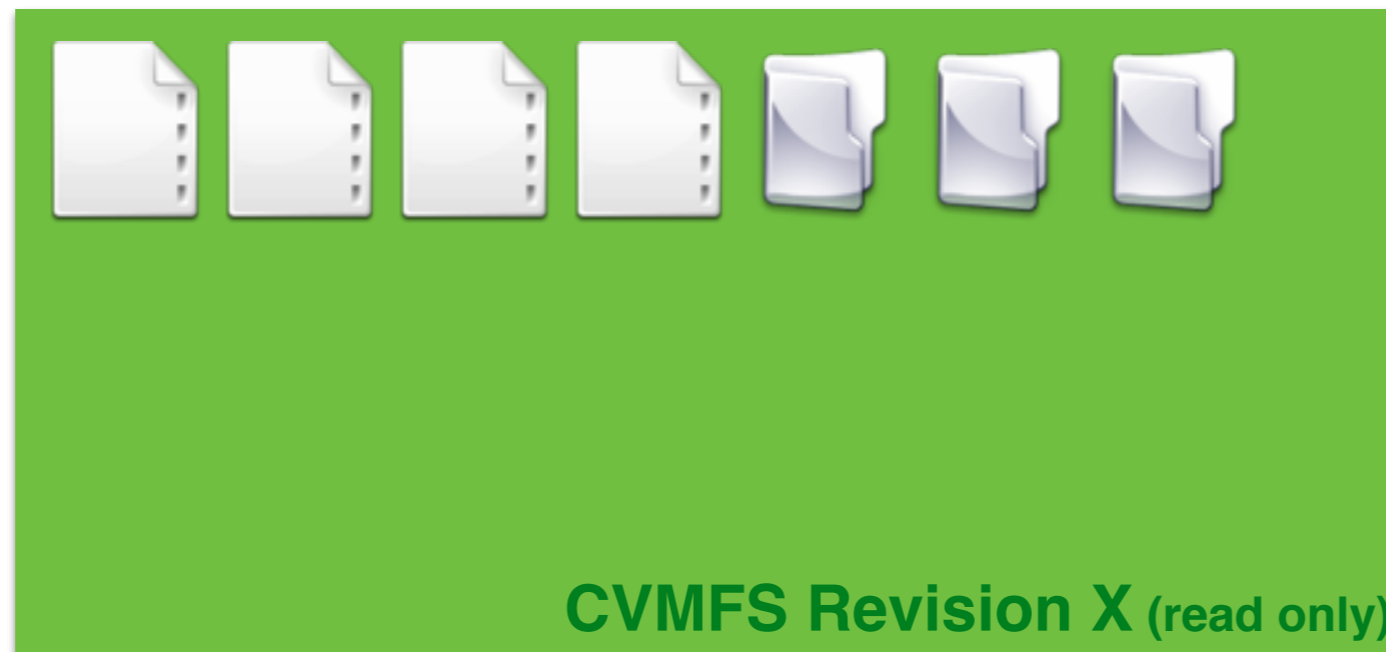
Updating a Repository



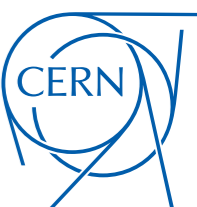
Stratum0
(backend storage)



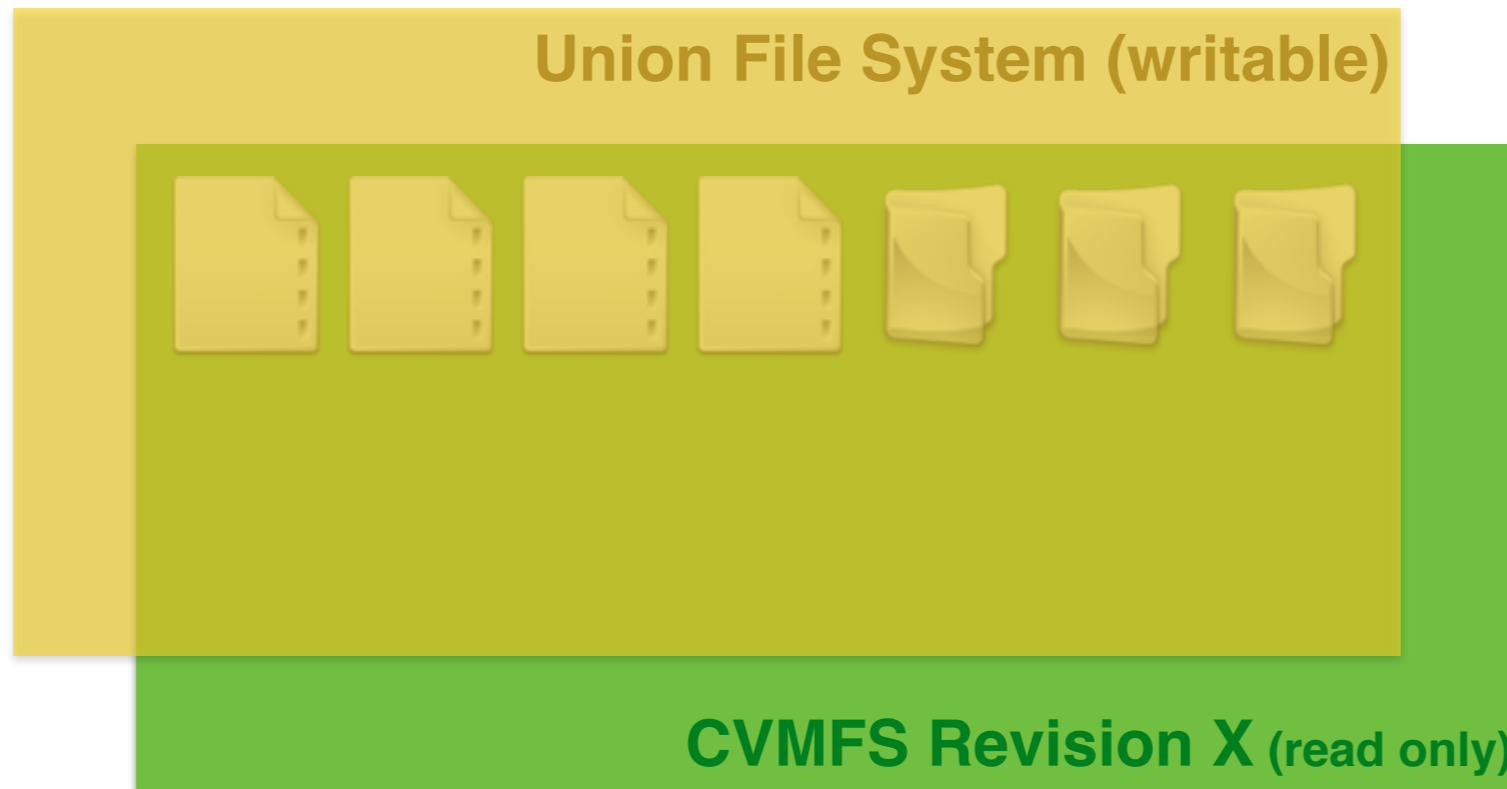
Updating a Repository



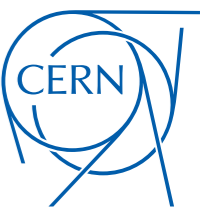
Stratum0
(backend storage)



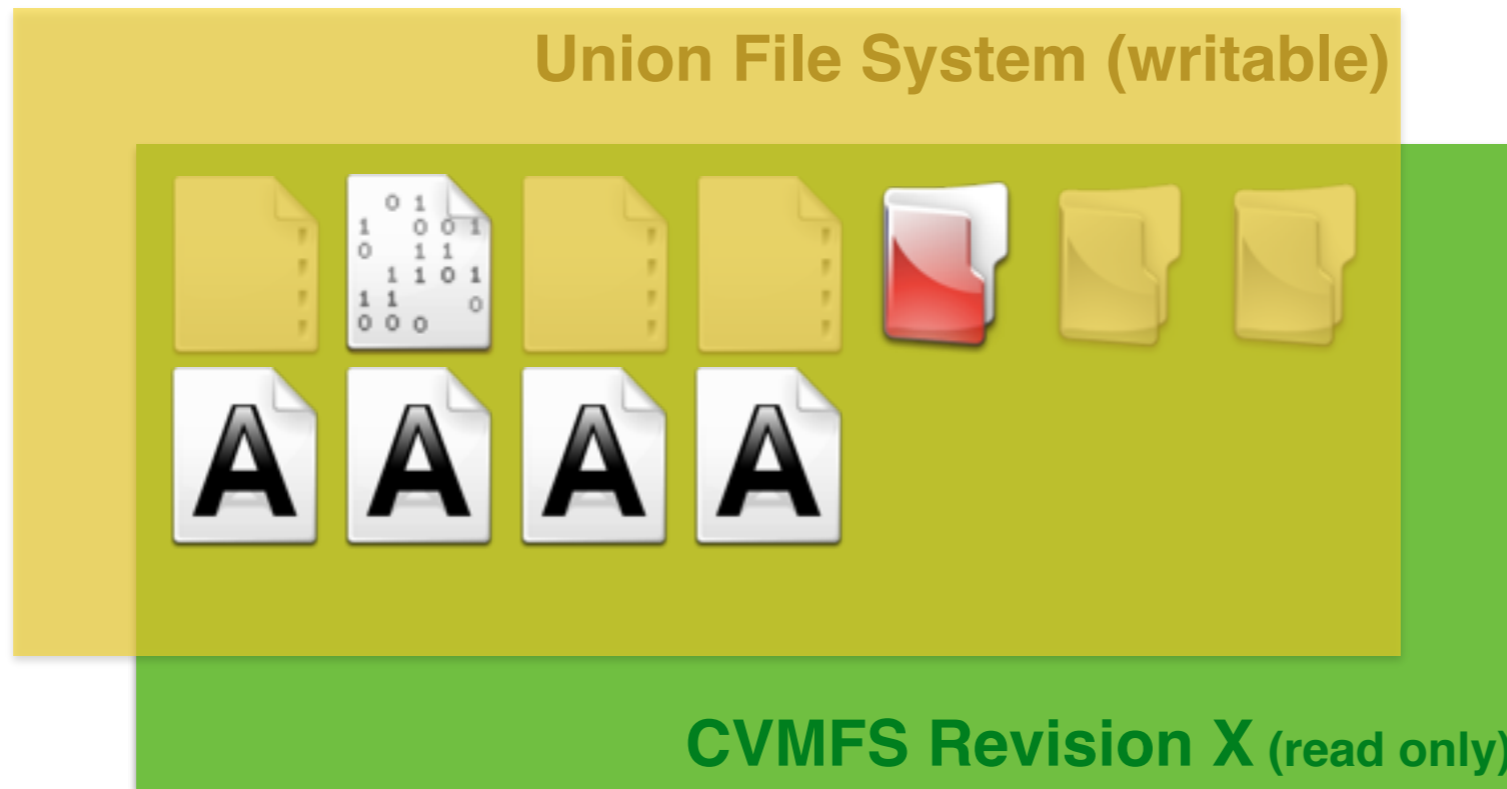
Updating a Repository



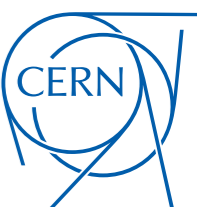
Stratum0
(backend storage)



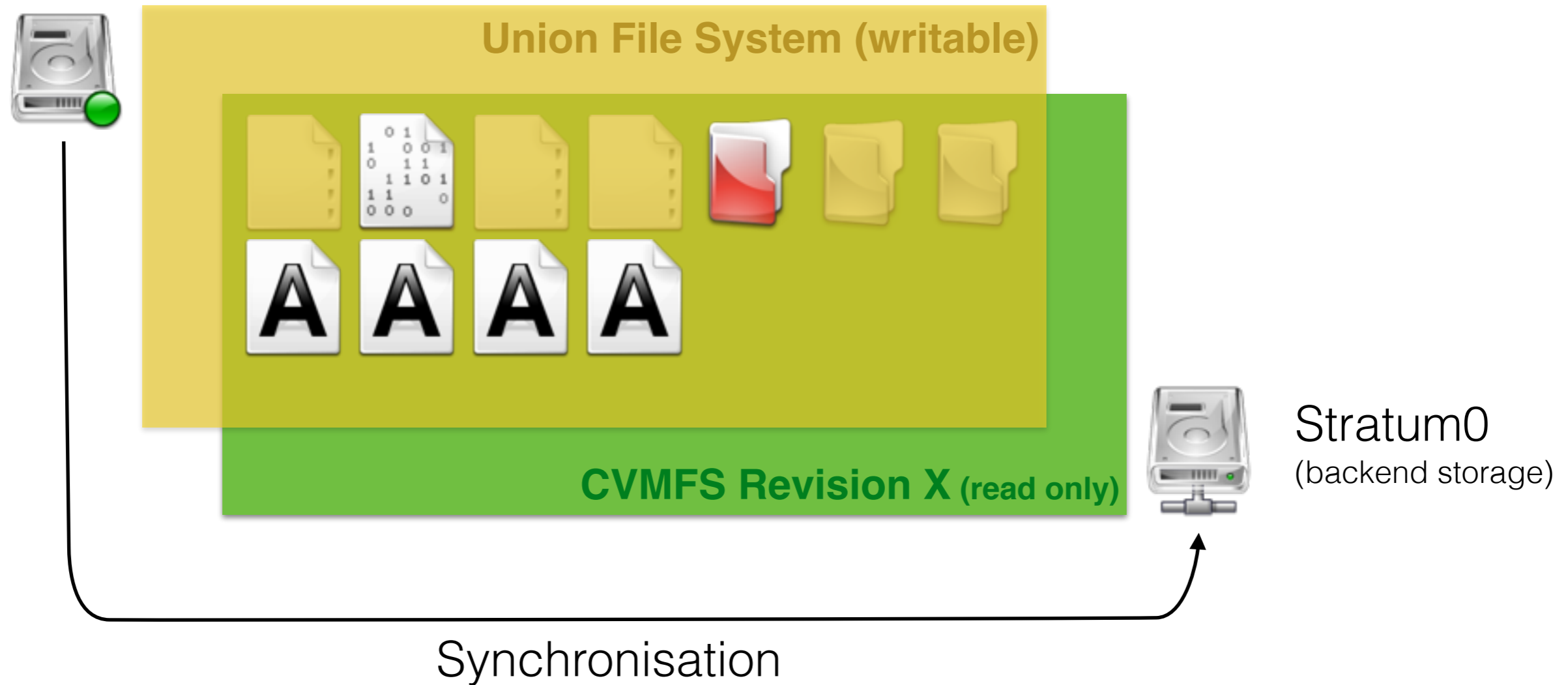
Updating a Repository



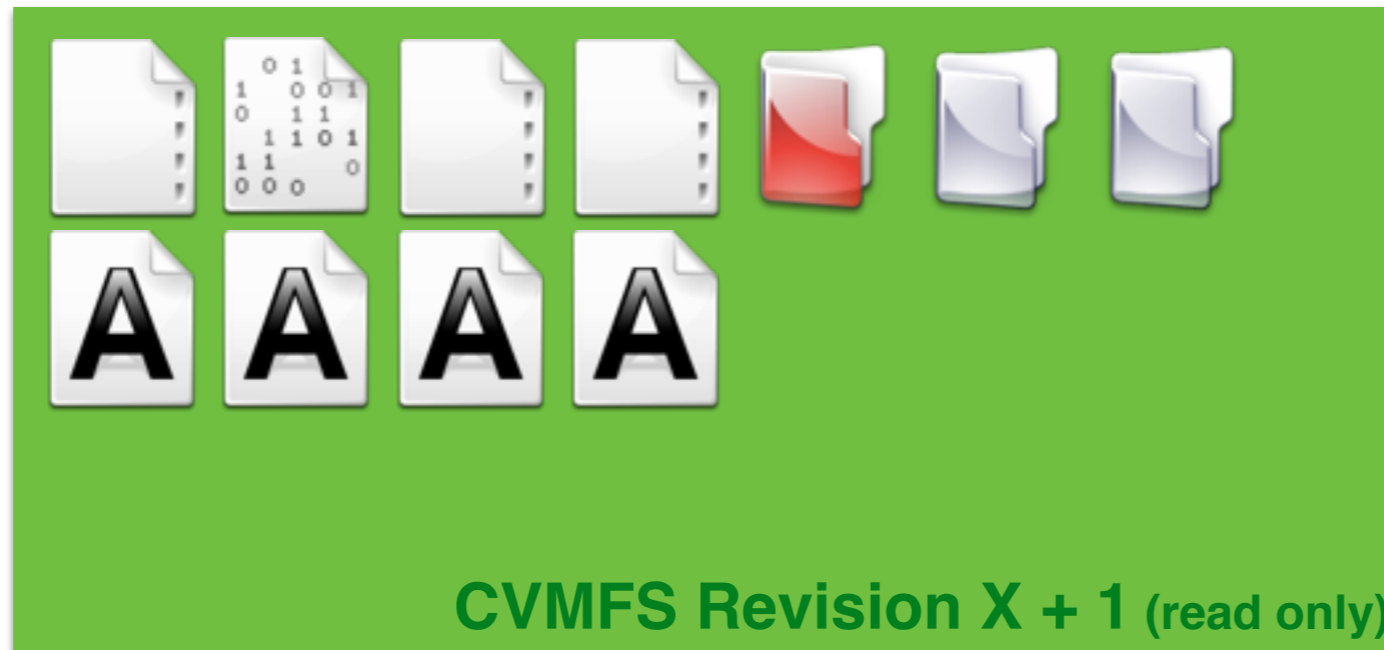
Stratum0
(backend storage)



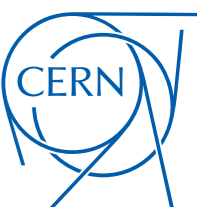
Updating a Repository



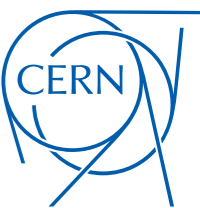
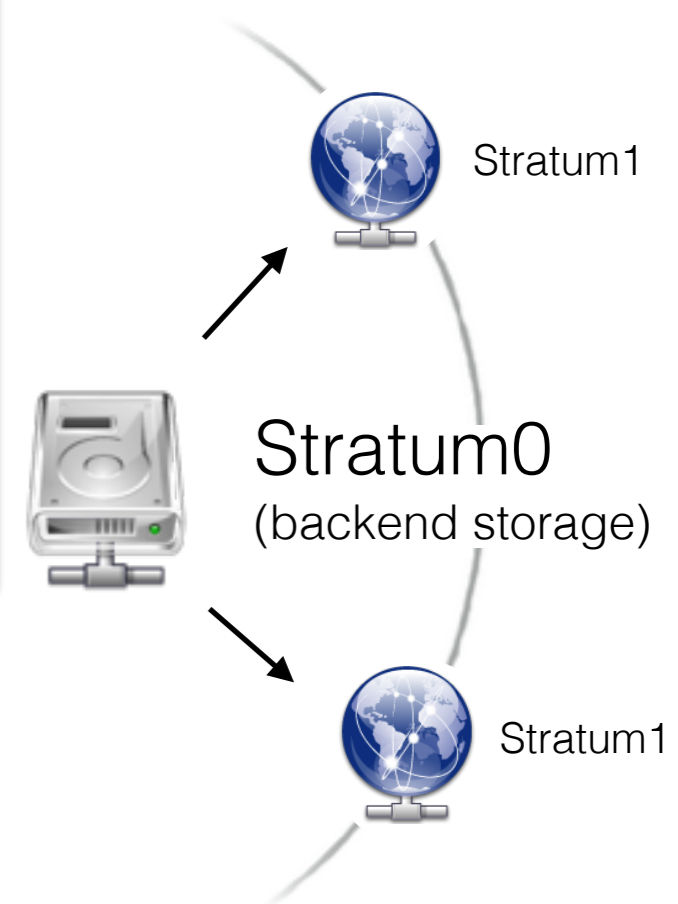
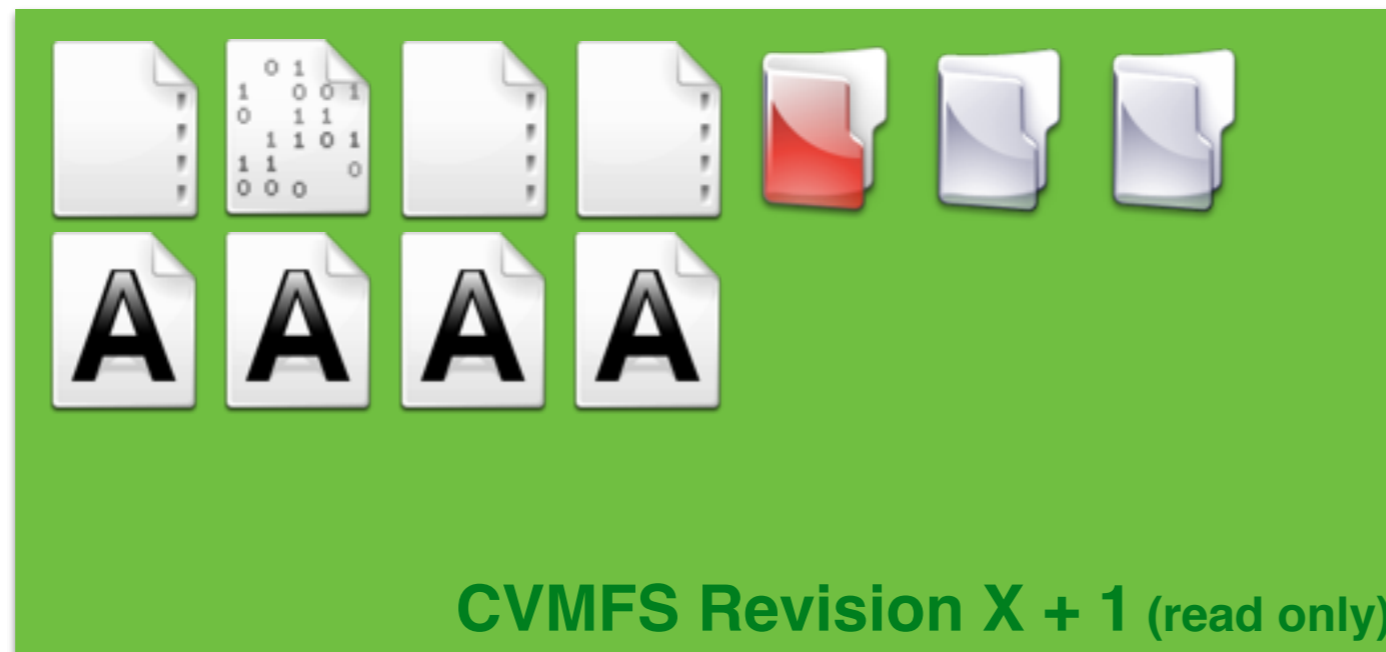
Updating a Repository



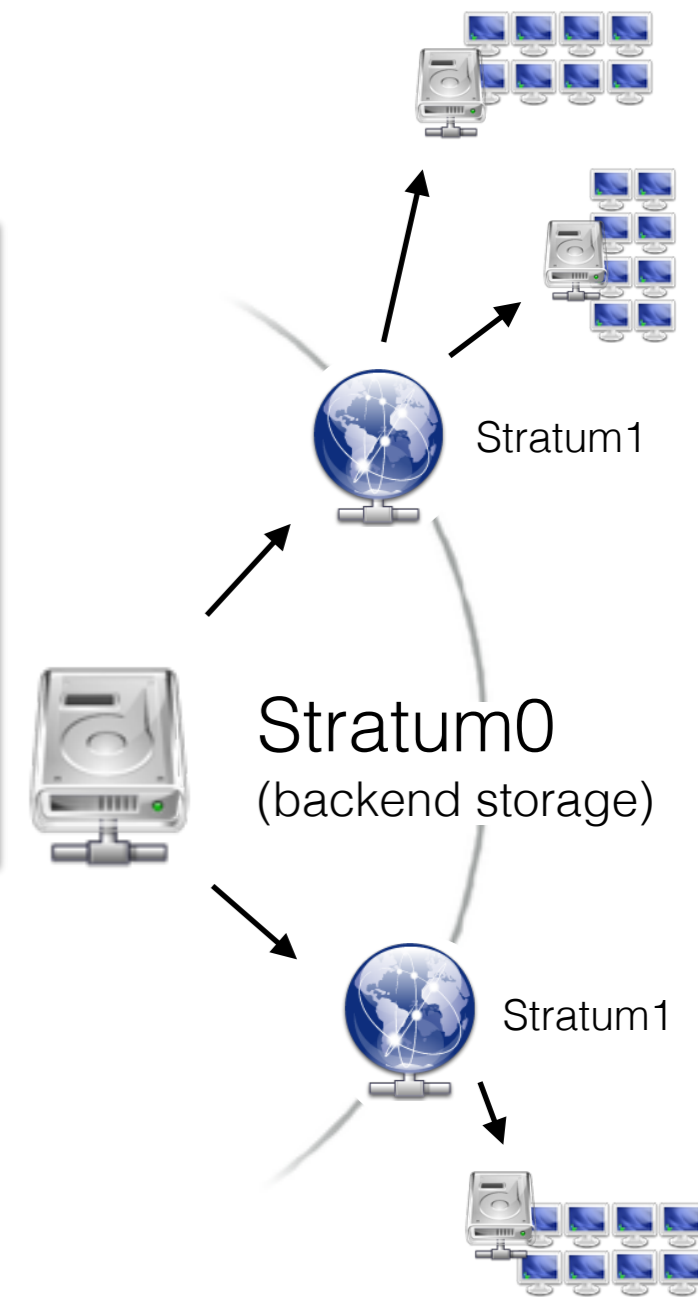
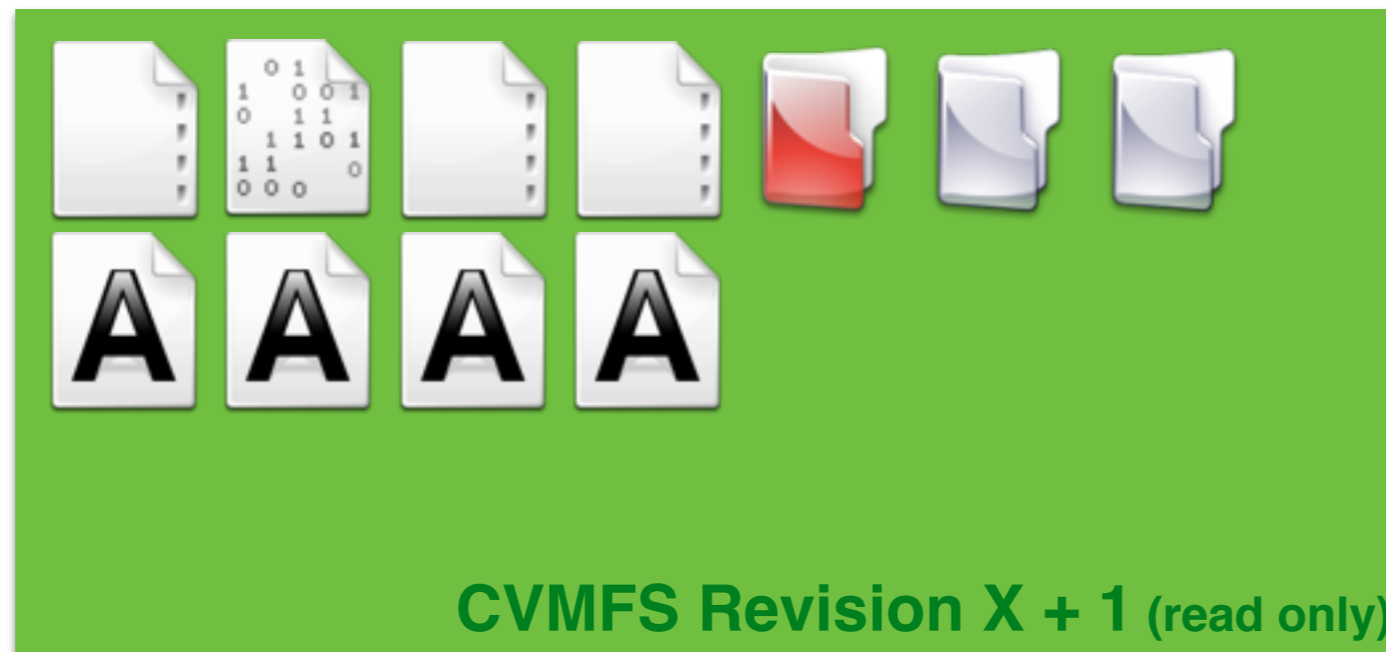
Stratum0
(backend storage)



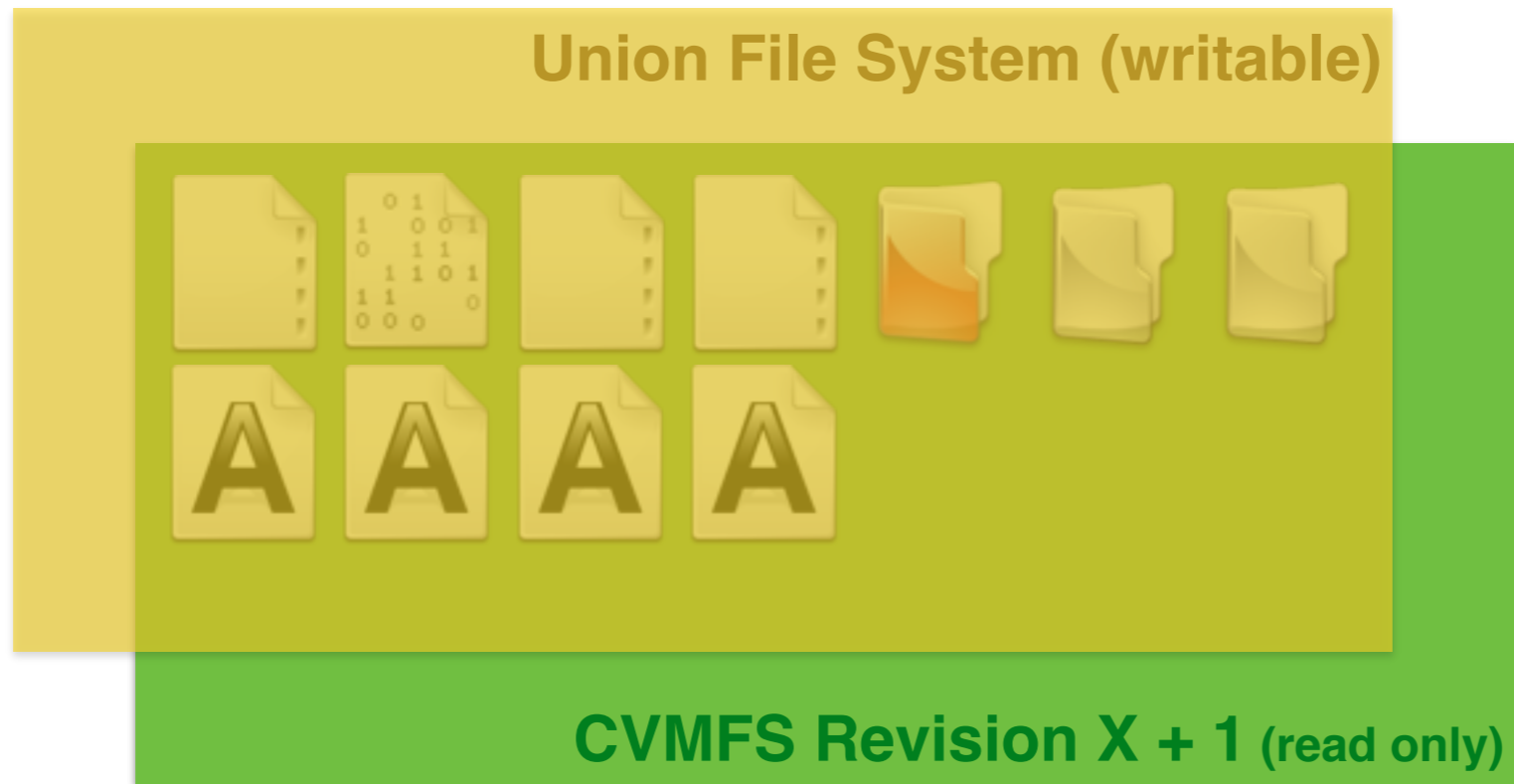
Updating a Repository



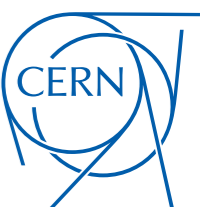
Updating a Repository



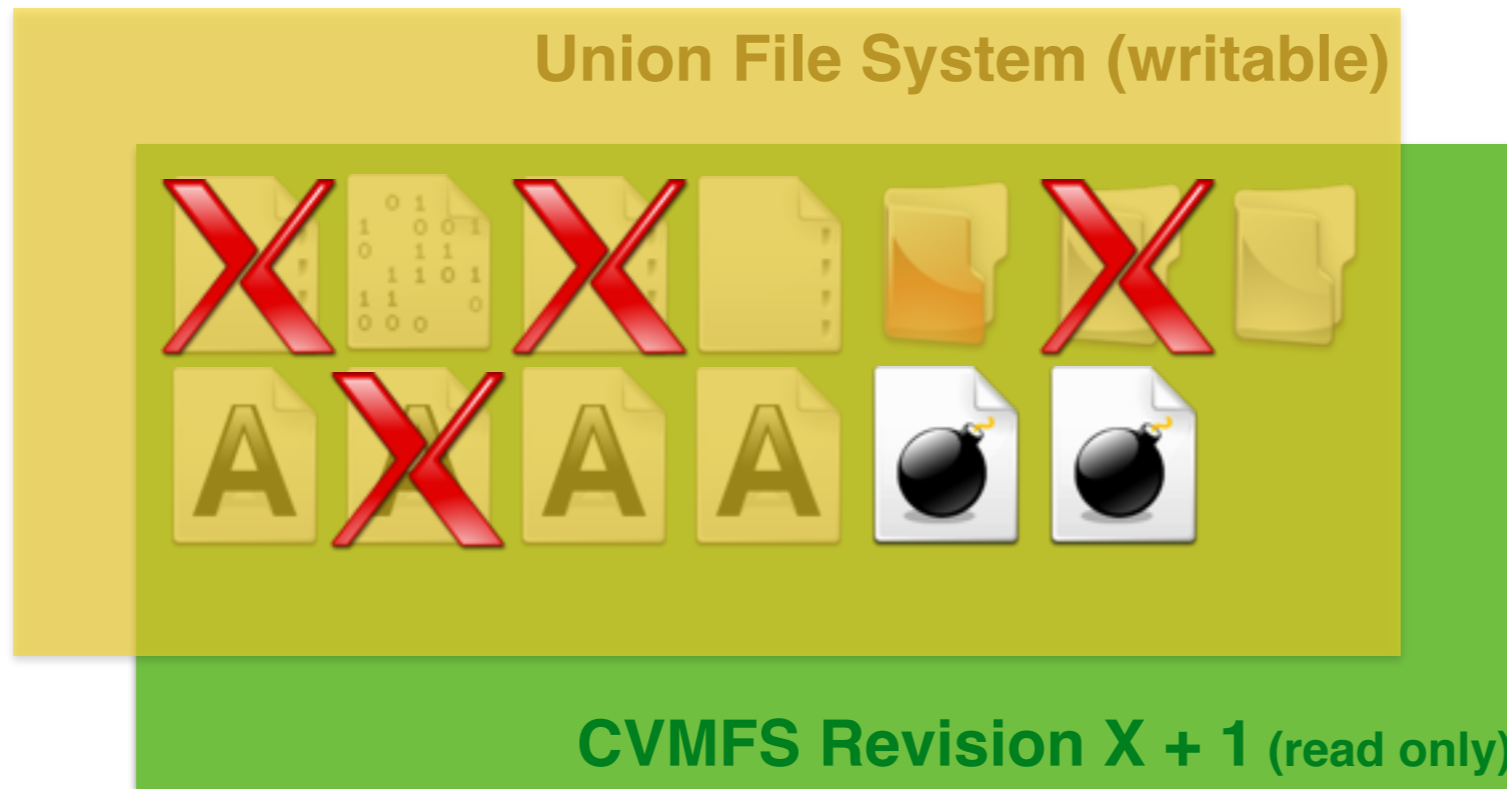
Updating a Repository



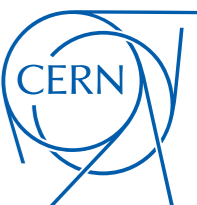
Stratum0
(backend storage)



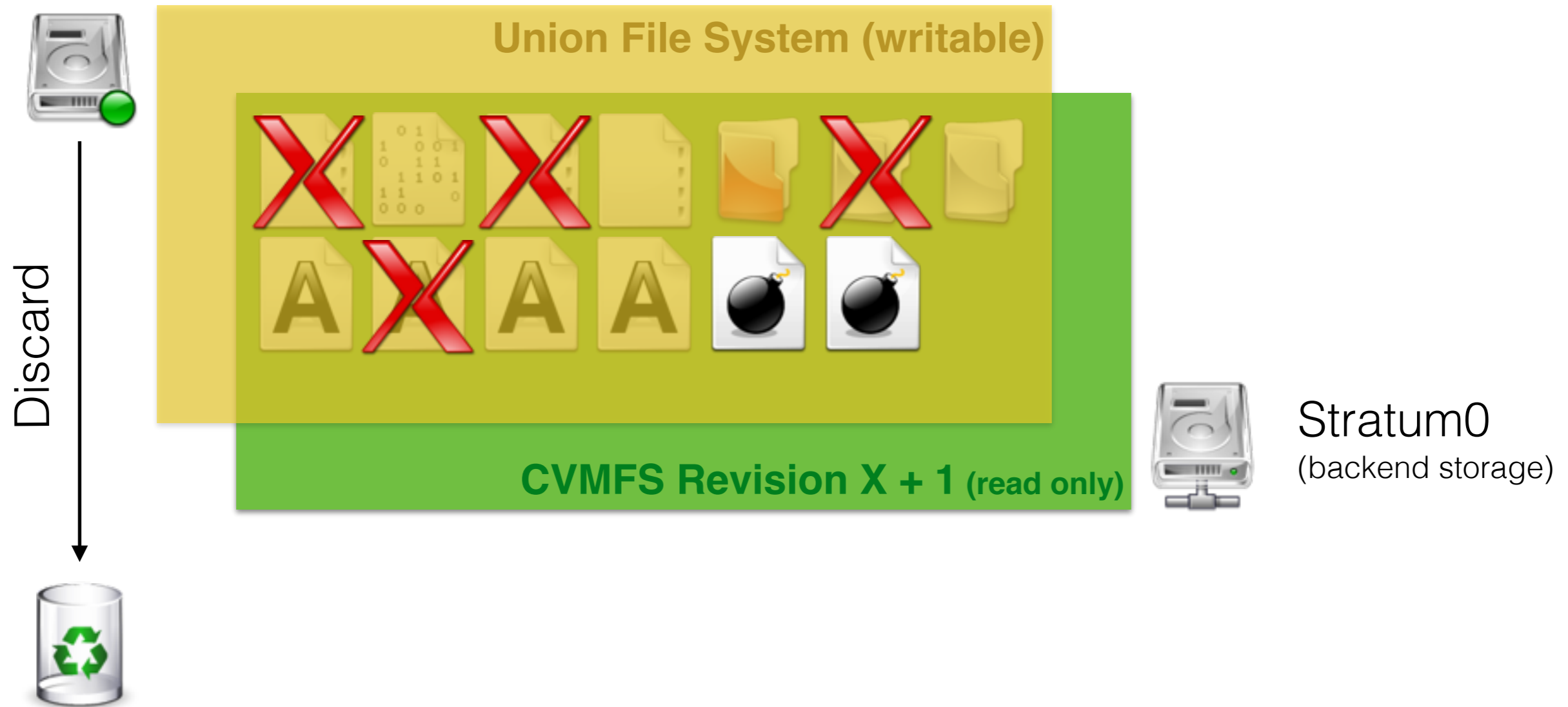
Updating a Repository



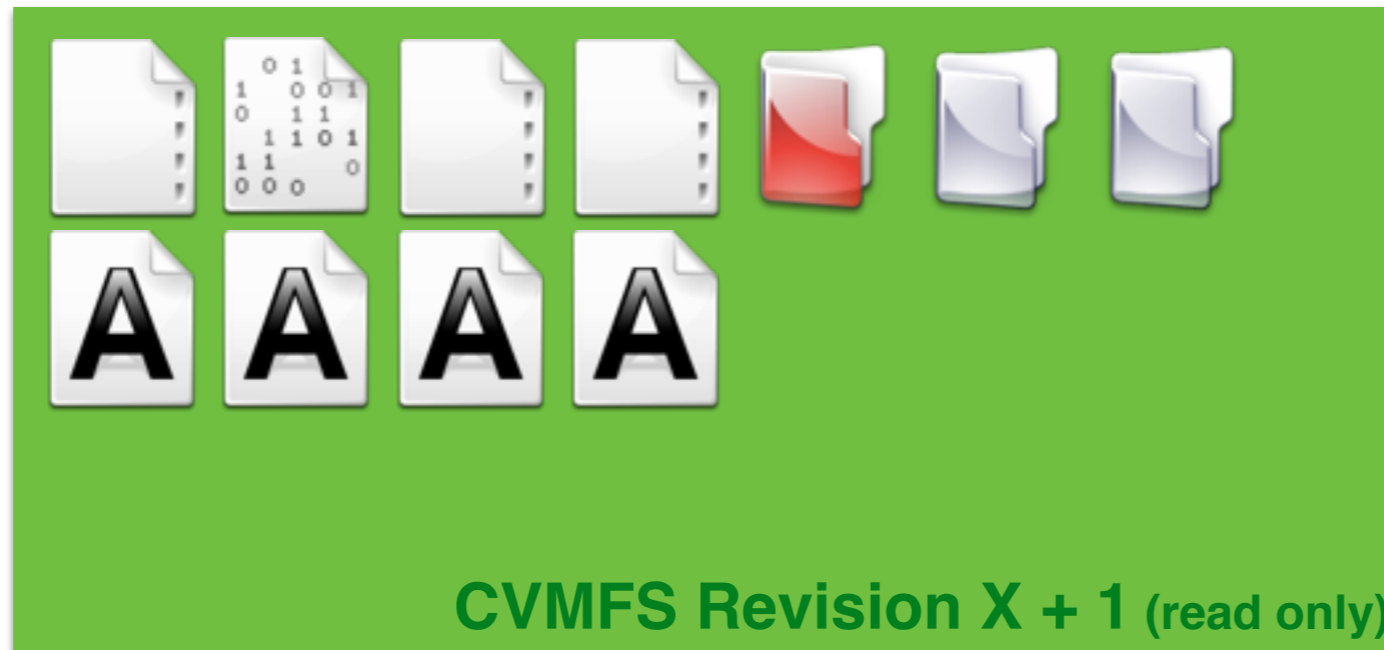
Stratum0
(backend storage)



Updating a Repository



Updating a Repository



Stratum0
(backend storage)

