KubeCon | CloudNativeCon

Europe 2022

WELCOME TO VALENCIA

KubeCon CloudNativeCon Europe 2022

Your Manila CephFS Share Backups Belong to S3

Robert Vasek, CERN

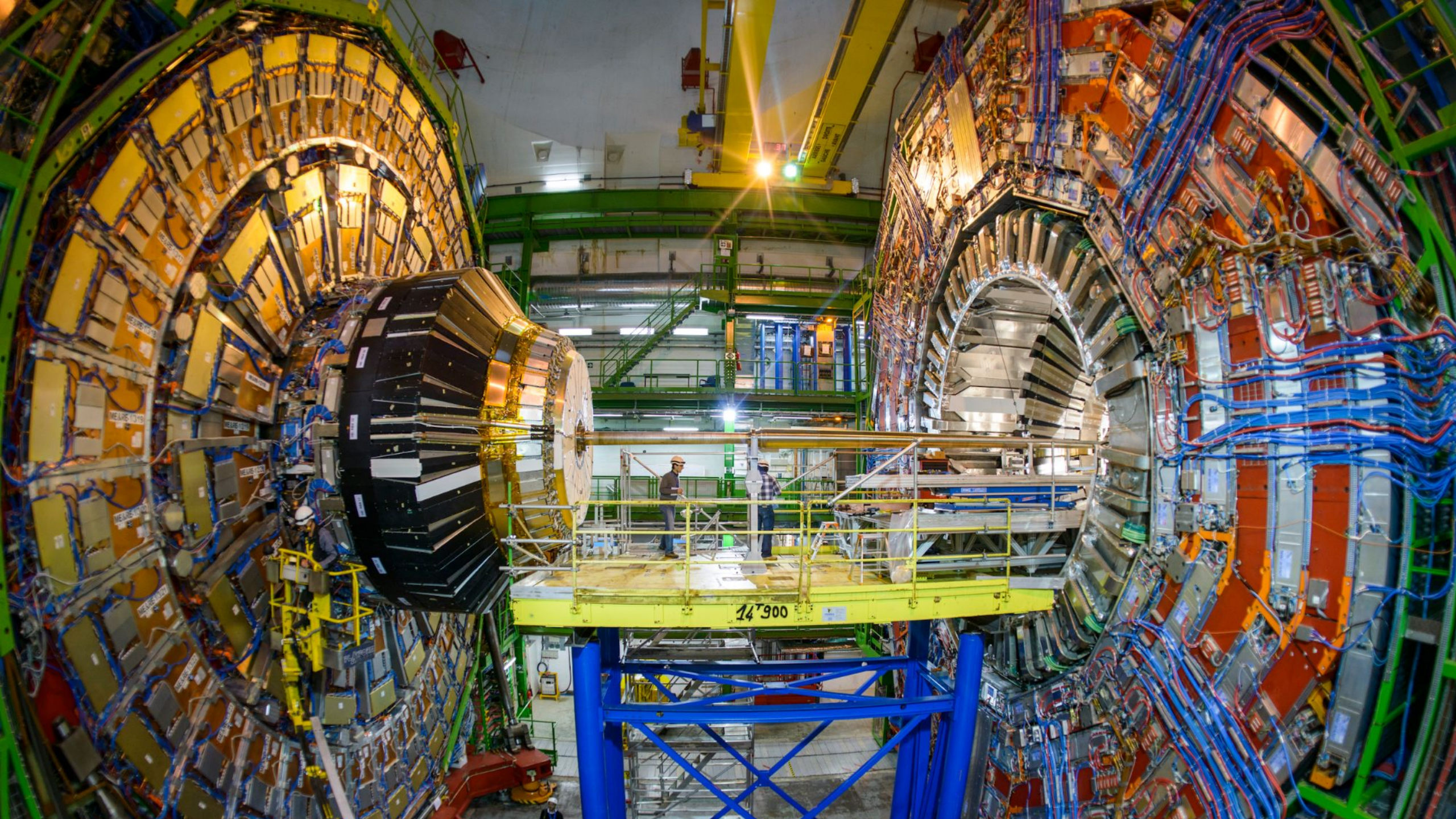
Agenda

- Who needs backups at CERN anyway?
- What is Manila? CephFS?
- How to backup Manila CephFS shares?
- Where to go from here?
- Conclusion!

What is 96% of the universe made of?

What was the state of matter just after the Big Bang?

Are there more particles to be discovered?


14T900

General / CERN OpenStack overview

Last 60 days

Openstack services statistics

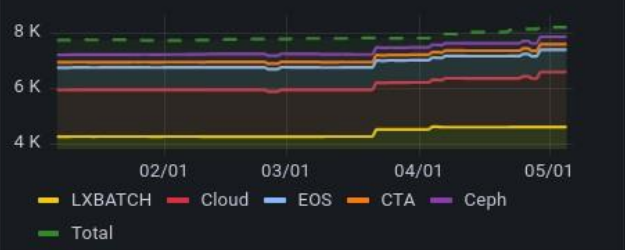
Users	Projects	Kubernetes clusters	Images	Volumes	Volumes size	File Shares	File Shares s...	Object Store ...	Object Store ...
3326	4496	356	3533	7349	3.78 PB	5304	890 TB	452	47.9 TB

Servers				Cores			RAM			Batch		
Physical	Physical in use	Hypervisors	Virtual	Physical	Hypervisors	Virtual	Physical	Hypervisors	Virtual	Servers	Cores	RAM
8658	8226	1996	13299	460 K	57.7 K	87.2 K	1.91 PB	375 TB	206 TB	4885	261651	989 TB

Time series

Physical servers

Last 4 months



Virtual Machines

Virtual Machines created/deleted

Last 7 days

Shared cells availability

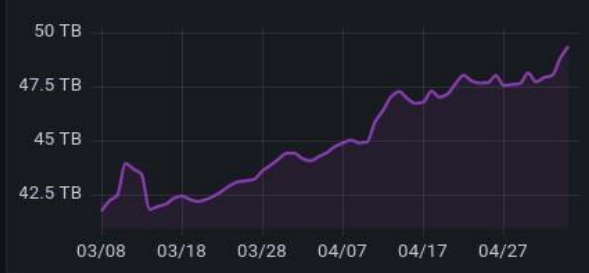
Last 30 days

Kubernetes clusters

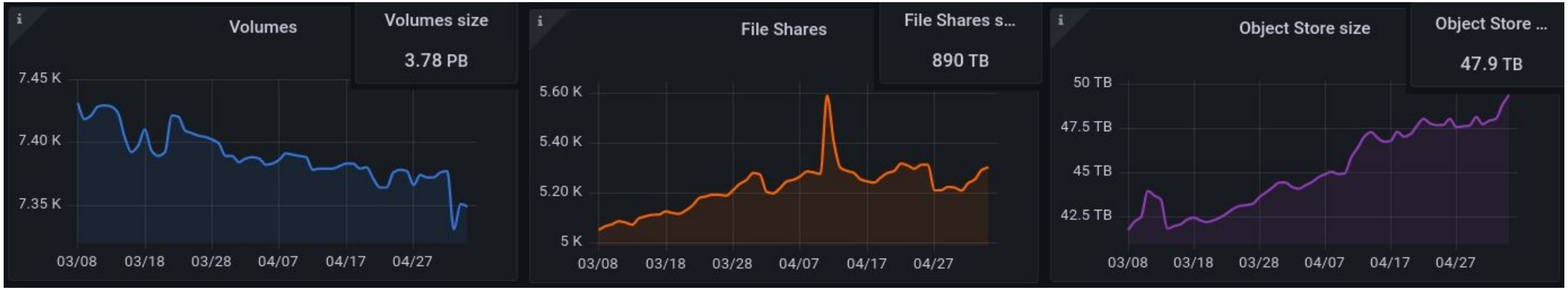
Volumes

File Shares

Object Store size



Storage used at CERN Cloud



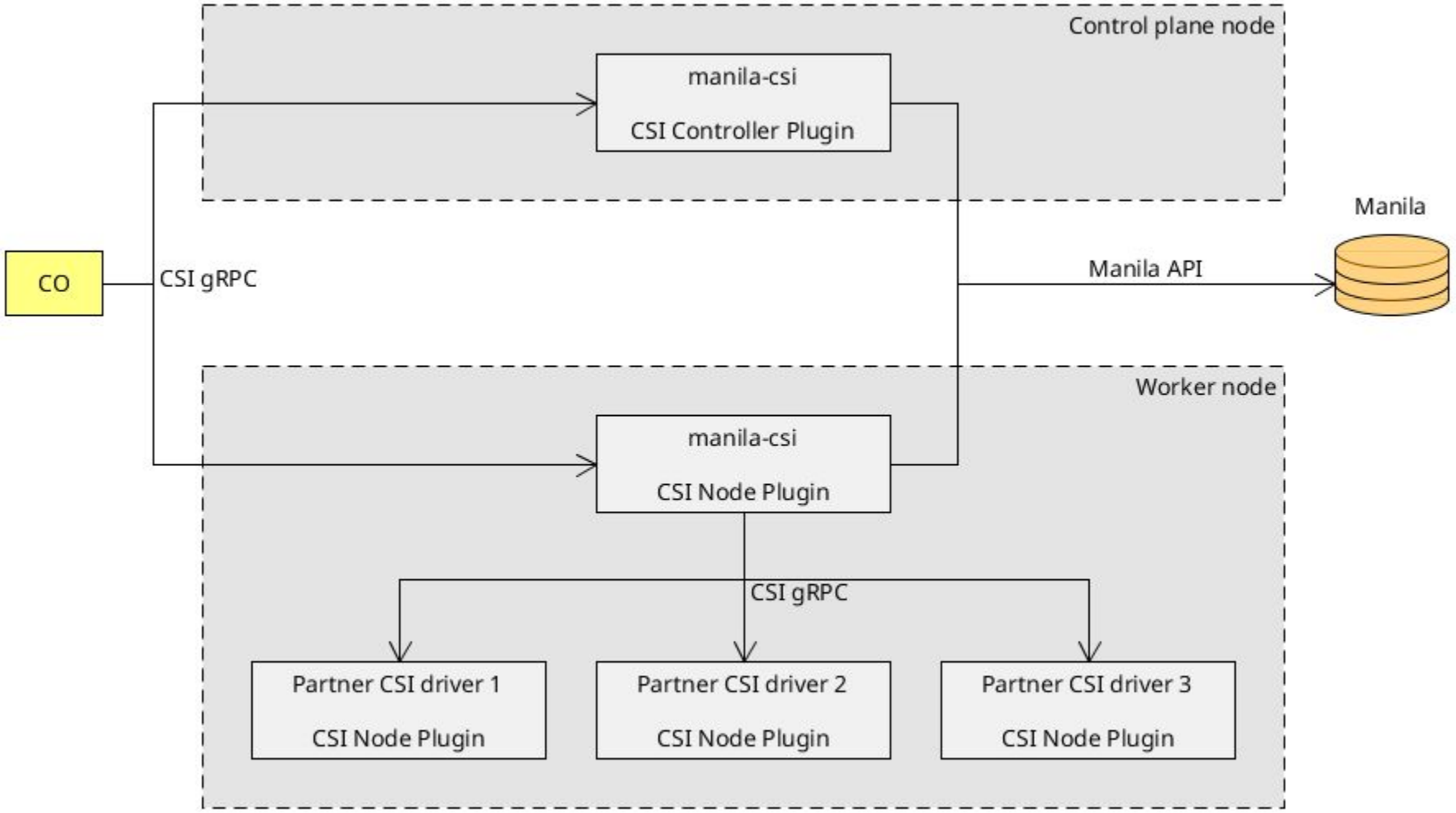
Target users for Manila backups at CERN

- 65 projects
- 159 production clusters
- 74 TiB total capacity reserved by provisioned shares

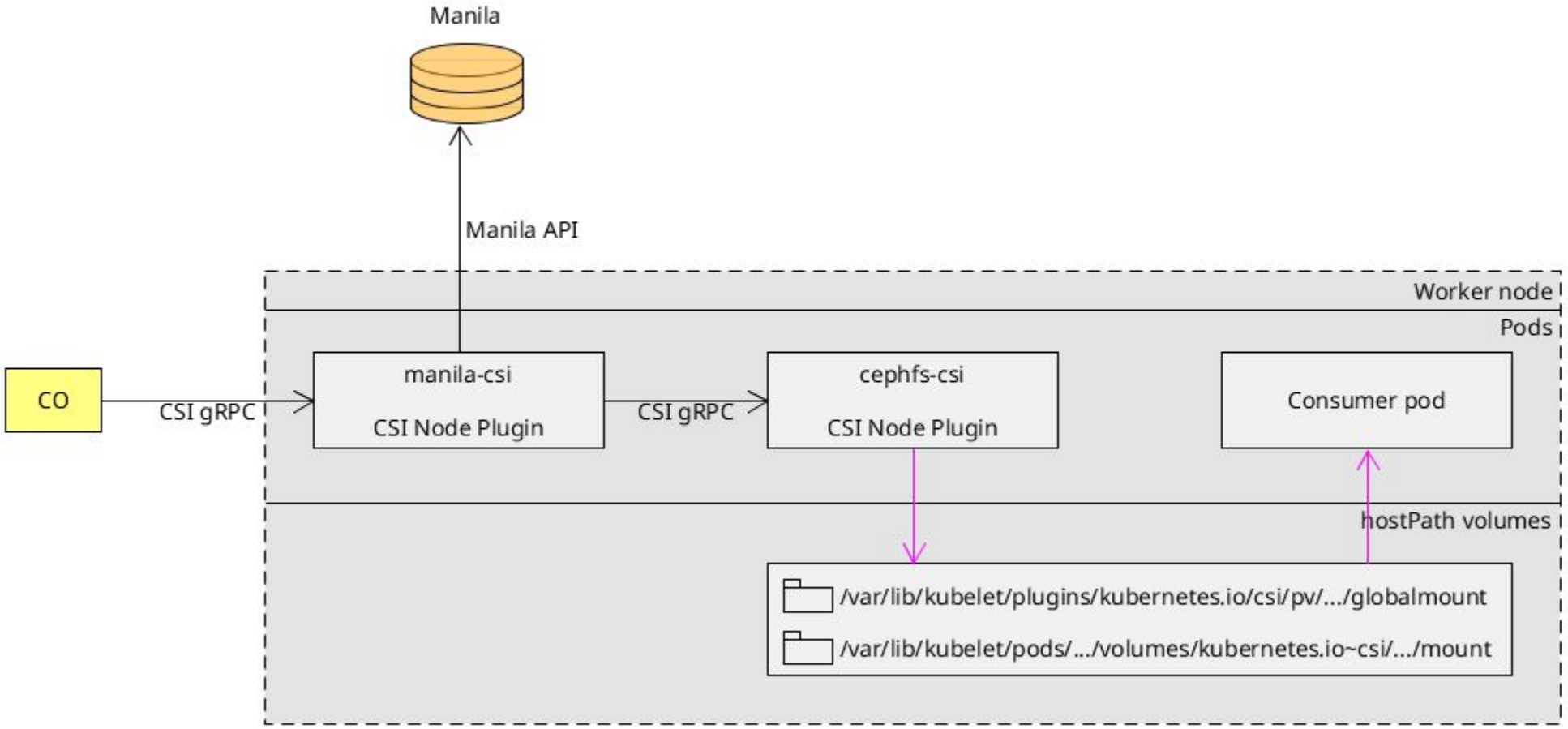
Vocabulary

- Ceph
  - Scalable distributed storage system
  - 3 in 1
    - object store (RADOS)
    - block devices (RBD)
    - file-systems (CephFS)
- OpenStack Manila
  - Shared file-systems service
- CSI (container storage interface)
  - Industry standard for writing storage plugins for container orchestrators
  - Many CSI drivers provided by storage vendors
    - Including ceph-csi-cephfs, openstack-manila-csi

openstack-manila-csi



openstack-manila-csi & ceph-csi-cephfs



Backup and restore workflow

Backup

1. Quiesce application
2. Create snapshot ***s*** of volume v_o
3. Unquiesce application
4. Create volume v_s from snapshot ***s***
5. Backup volume v_s
6. Remove v_s and ***s***

Restore

1. Restore v_s to volume v_o
2. Run application

CephFS Manila backups & recovery in general

- Provide CephFS and Manila CSI drivers with capabilities that fit backup & recovery workflow
- Facilitate users with means of data protection regardless of the backup solution they decide to use
- Reliance on snapshots
- Cheap “PVC from snapshot source” is a must

Snapshot-backed volumes in cephfs-csi

- Snapshots available directly in volumes under *.snap/*
- Exposing them to workloads as separate read-only volumes
- Volume creation is O(1)
- Using RADOS objects for bookkeeping
  - …to make sure the backing snapshot is not deleted before its dependent volumes are.

- Planned for cephfs-csi v3.7.0

Mounting CephFS snapshots with manila-csi

- Manila proper offers capability to mount snapshots
  - Not for CephFS though, manila-csi will fill in the gap
- Most of the heavy-lifting is implemented in cephfs-csi

- Challenges:
  - Bookkeeping using share metadata?
  - Issue with long snapshot names
    - ceph/ceph#45192
    - ceph/ceph#45312
    - manila/+bug/1967760

- No ETA yet, blocker issues must be resolved first

Mounting CephFS snapshots with manila-csi

```
# ls -l .snap
ls: cannot access '.snap/_8afe40e3-b3dd-4c99-acdb-673be49cc7d1_cf335dad-c57f-4a41-879a-a4cb5a65d56f_1099': No such file or directory
total 0
d????????? ? ?    ?    ?            ? _8afe40e3-b3dd-4c99-acdb-673be49cc7d1_cf335dad-c57f-4a41-879a-a4cb5a65d56f_1099
drwxrwxrwx. 2 root root 5 May  1 16:15 _csi-snap-f61c8b6f-b1d6-11ec-8fbb-0242ac110003_1099511628283
#
```

```
82  std::string_view SnapInfo::get_long_name() const
83  {
84    if (long_name.empty() ||
85        long_name.compare(1, name.size(), name) ||
86        long_name.find_last_of("_") != name.size() + 1) {
87      char nm[80];
88      snprintf(nm, sizeof(nm), "_%s_%llu", name.c_str(), (unsigned long long)ino);
89      long_name = nm;
90    }
91    return long_name;
92  }
```

Backup & restoration solutions

Trilio, Kasten, Stash, Velero, Kanister,

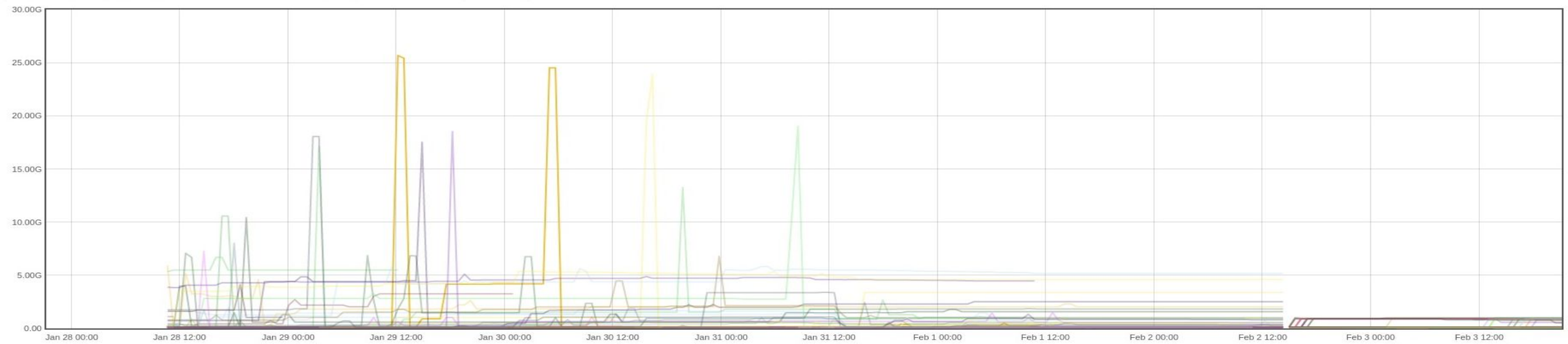
in-Kubernetes solution (WG Data Protection @ SIG Storage/Apps), …

- Evaluating Velero as optional offering to our users
- “Velero is an open source tool to safely backup and restore, perform disaster recovery, and migrate Kubernetes cluster resources and persistent volumes.”
- Scheduled backups, pre- and post-backup hooks, data retention

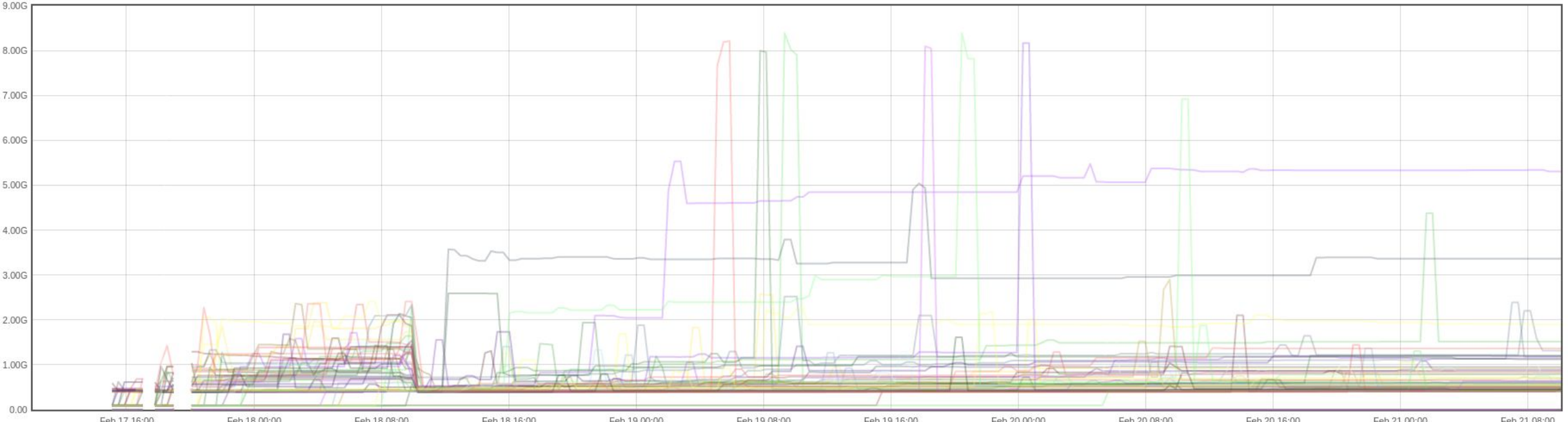
Our experience using Velero

- Works well in general
- Support for Amazon EBS, Google Persistent Disks, Azure Managed Disks
- Implies CSI snapshots are durable
- Others need to copy data with Restic
  - Velero's integration with Restic is advertised as beta quality

- Issues:
  - Large memory consumption with Restic
  - Failed backups stay failed, no retries
  - Scaling issues with many PVCs



- Memory peaks at 25GiB
- As of Velero v1.7.1 it’s significantly better, ~8GiB



Where to go from here?

- Velero plans for the future:
  - Better support for CSI snapshots
  - Adding support for alternatives to Restic, e.g. Kopia
  - Adding abort capability
  - And others…

- Trying out Kanister.io
  - “An extensible open-source framework for application-level data management on Kubernetes”
  - I.e. define your own data moving workflow, or use premade ones
  - Supports Kopia

Restic and Kopia comparison

- Backup and restore ~1.5mil files (uncompressed copies of the same Linux kernel)
- The same Ceph cluster for volumes and backup location (S3)
- Velero v1.8.1, velero-plugin-for-aws v1.4.1, Kanister v0.76.0

backup	**Restic (Velero)**	**Kopia (Kanister)**
Elapsed time [minutes]	54:55	19:27 minutes
Max. memory consumed [MiB]	4997	244
S3 bucket size [GiB]	1.86 (477 objects)	1.17 (71 objects)

restore	**Restic (Velero)**	**Kopia (Kanister)**
Elapsed time [minutes]	63:01	26:47
Max. memory consumed [MiB]	2304	1525

Conclusion

- Our users want consistent backups, point-in-time snapshots are needed
- Verdict: Your Manila CephFS Share Backups Belong to S3…?
  - Depends if you need point-in-time snapshots
  - If yes: you need to wait a bit
  - If no: go and backup your data right now (carefully)
- Fixes coming to Ceph and Manila
- Scaling issues need more attention

Robert VASEK, CERN <rvasek01@gmail.com>

Acknowledgements

Thanks for providing extremely helpful feedback on Velero goes to:

CERN Drupal Infrastructure Team:

- Vineet REDDY RAJULA
- Francisco BARROS
- Konstantinos SAMARAS-TSAKIRIS