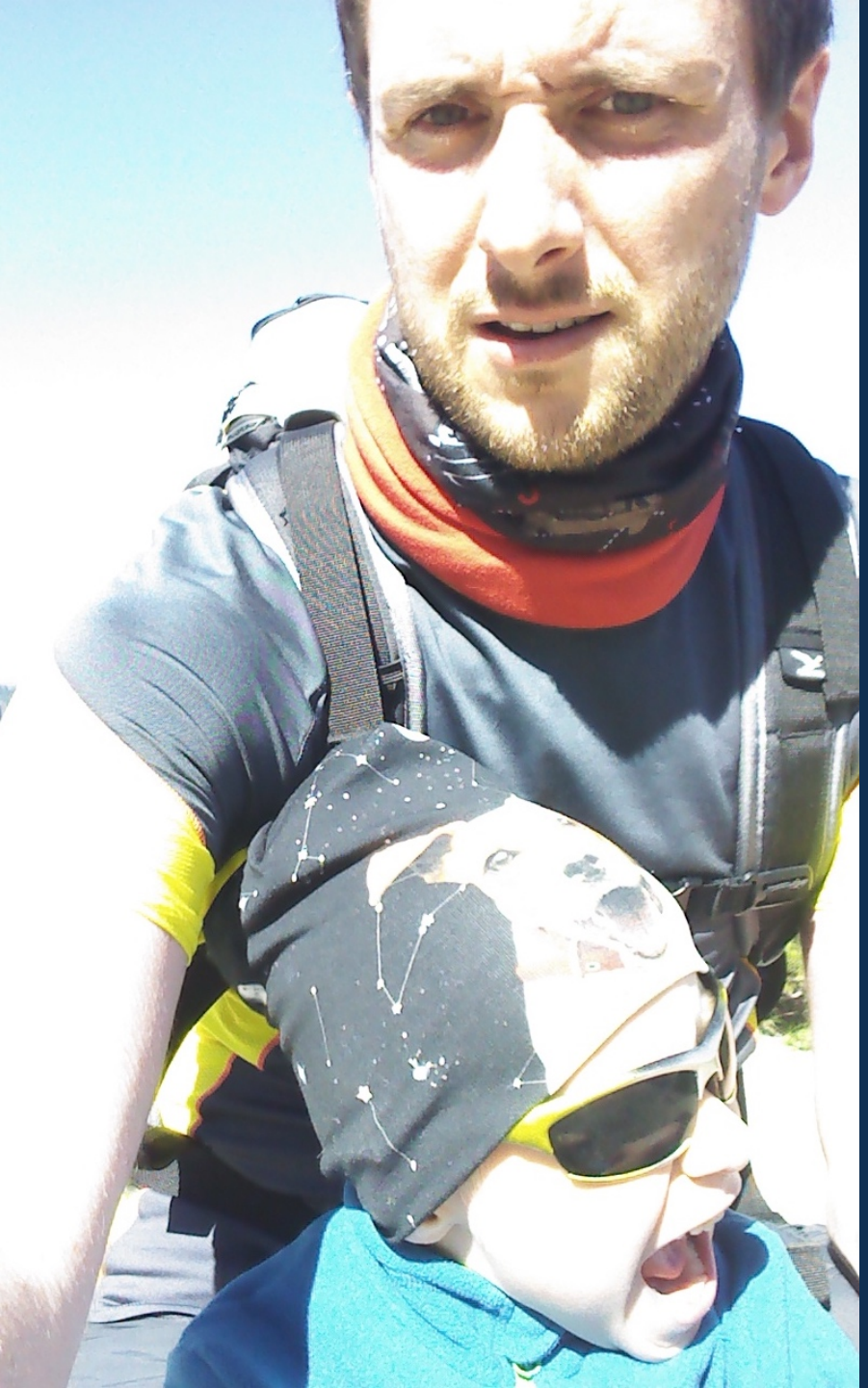
OVH.com

Innovation is Freedom

How to backup Ceph at scale

FOSDEM, Brussels, 2018.02.04



About me

Bartłomiej Święcki

OVH

Wrocław, PL

Current job:

More Ceph awesomeness

OVH.com

Speedlight Ceph intro

- Open-source
- Network storage
- Scalable
- Reliable
- Self-healing
- Fast

ceph

Ceph @ OVH

- Almost 40 PB of raw HDD storage
- 150 clusters
- Mostly RBD images

ceph

OVH.com

Why we need Ceph backup ?

- Protection against software bugs
  - Didn’t see that yet but better safe than sorry
- One more protection against disaster
  - Probability spikes at scale (i.e. HDD failures)
  - XFS (used by Ceph) can easily corrupt during power failures
- Human mistakes – those always happen
  - Ops accidentally removing data
  - Clients removing / corrupting data by mistake
- Geographically separated backups
  - Not easily available in Ceph (yet)

Resource estimation and planning

Software selection

- Compression
- Deduplication
- Encryption
- Speed
- Work with data streams
- Support for OpenStack SWIFT

Software selection

- No perfect match at that time
- Selected duplicity – already used at OVH
- Promising alternatives (i.e. Restic)

Storage, network

- Assumed compression and deduplication – 30% of raw data
- Use existing OVH services – PCA (swift)
- Dynamically scale computing resources with OVH Cloud

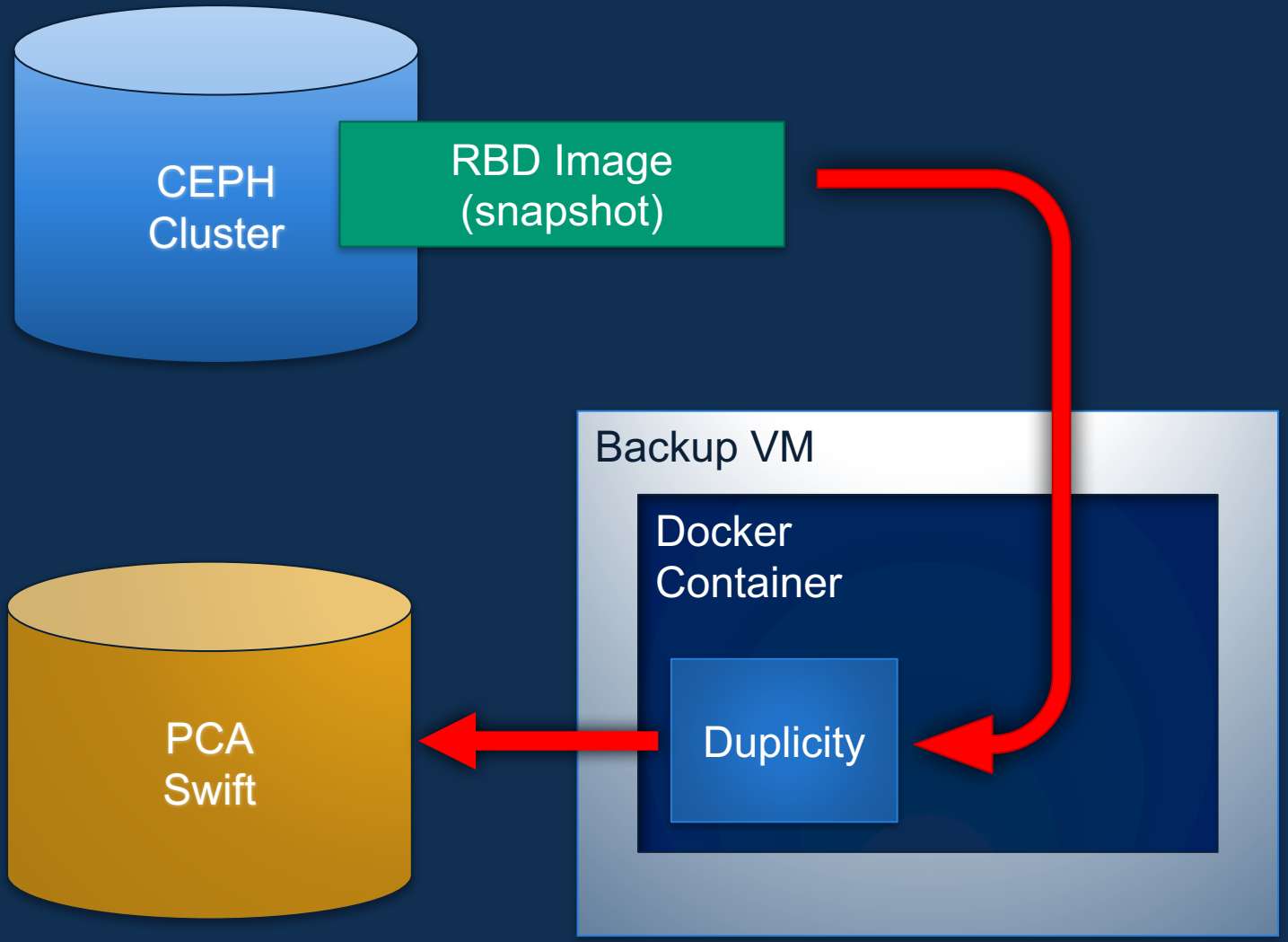
Impact on Ceph infrastructure

20PB raw data: 6.6 PB of data without replicas

For daily backup:

- ~281 GB / h = ~ 4.7 GB /min = ~ 0.078 GB / sec
- 0.63 Gb/sec constant traffic

Backup architecture – idea

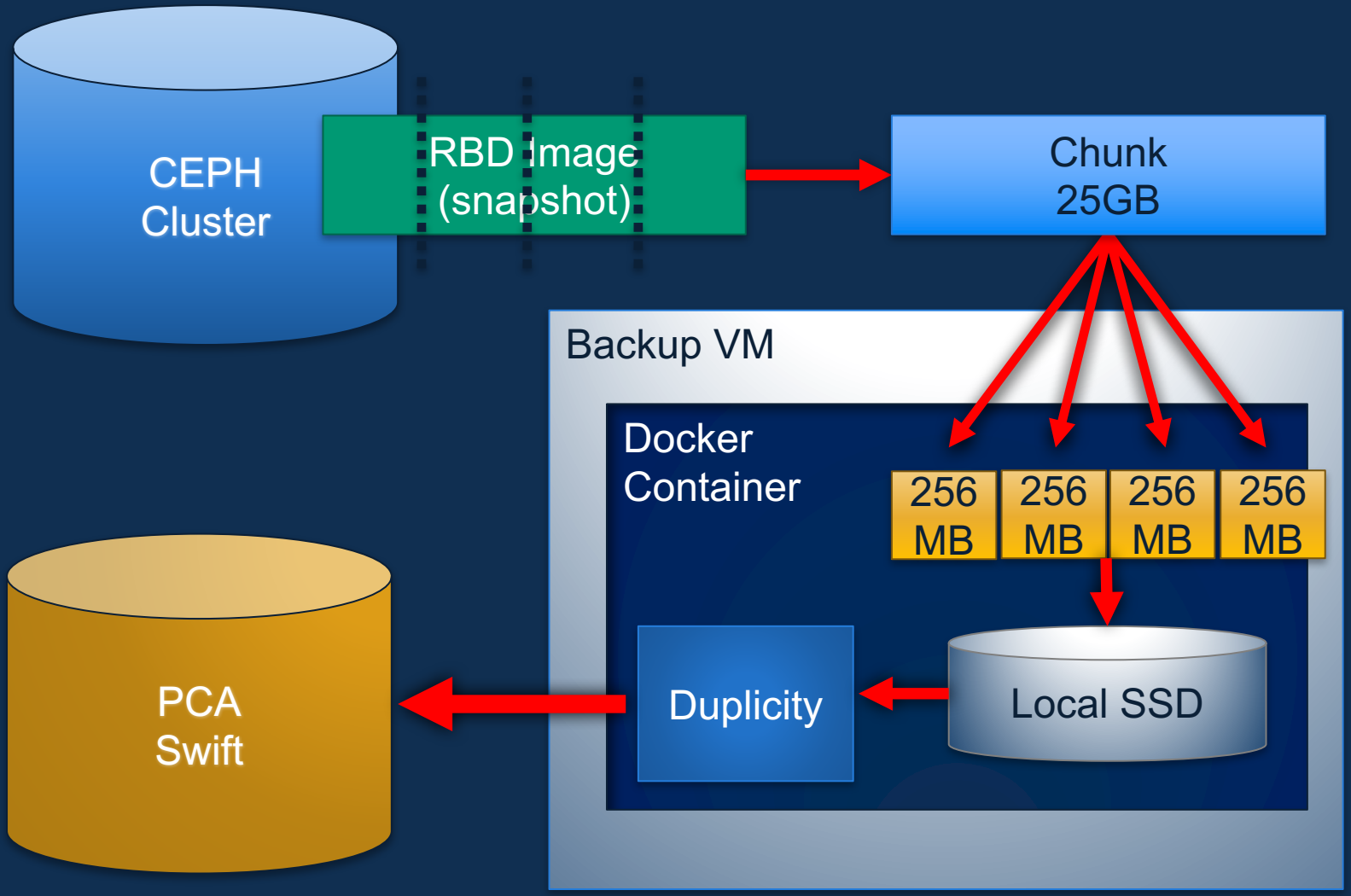


Implementation challenges

Duplicity quirks

- Can backup files only – export rbd image locally need temporary storage
- Files should not be larger than few MB due to librsync limits – rbd image split into files of up to 256MB size
- Can not backup large images (large >= 500GB): not enough local storage, timeouts, interruptions – split image into 25GB chunks and backup separately

Duplicity + SWIFT overview



FUSE to the rescue

- Expose part of image through FUSE
- Can easily work on part of the image
- Can expose image as list of smaller files
- No need for local storage, all can be done in memory
- Restore a bit more problematic but possible

Prod impact

- Throttle number of simultaneous backups
  - Global limit imposed by our compute resources
  - Limits per cluster
  - Limits per backup VM
  - No simultaneous backups of one RBD image
- Used locks and semaphores stored in zookeeper

Scaling issues

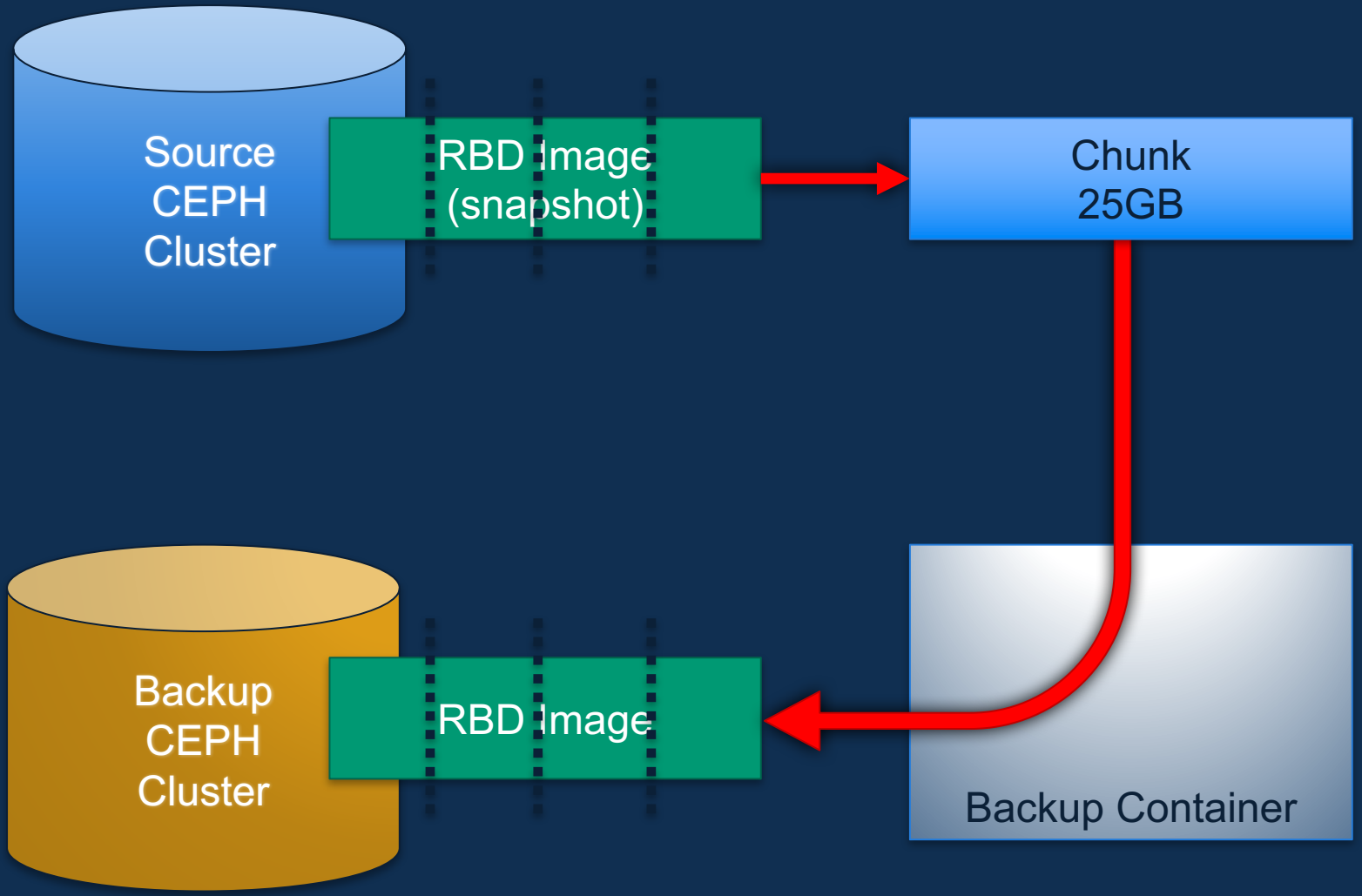
- Zookeeper does not work well with frequently changing data
- Lots of issues with celery workers – memory leaks, ulimit, ping timeouts, rare bugs
- Issues with docker – orphaned network interfaces, local storage not removed
- Duplicity requires lots of CPU to restore backup (restore 4x slower than backup)

Hot / cold backup strategy

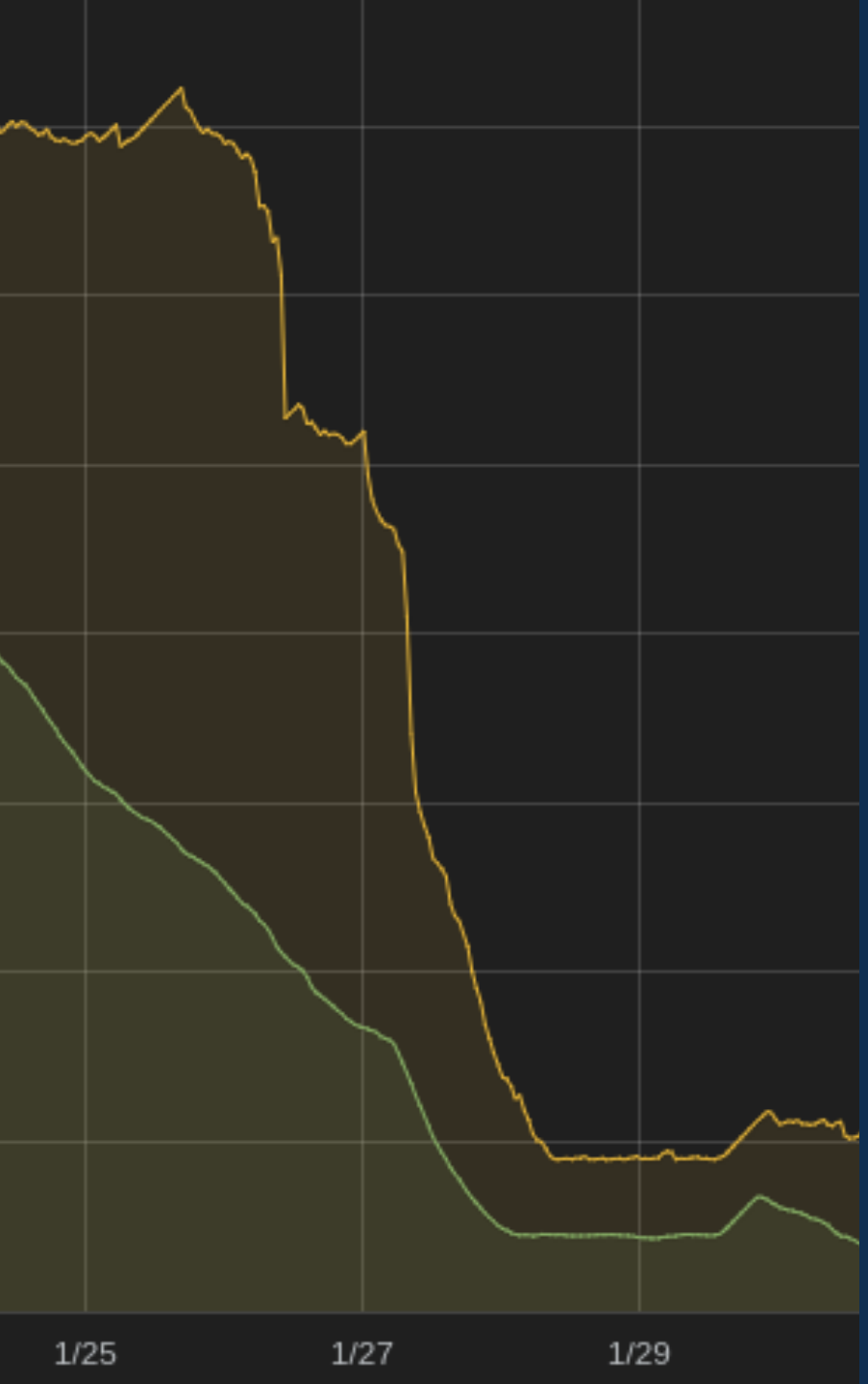
Backup to Ceph

- Separate Ceph cluster with copy of data
- Export / import diff a huge advantage
- Can use backup cluster as a hot-swap replacement
- Reuse previous backup architecture
- Can backup spare cluster as before – cold backup

Ceph on Ceph overview



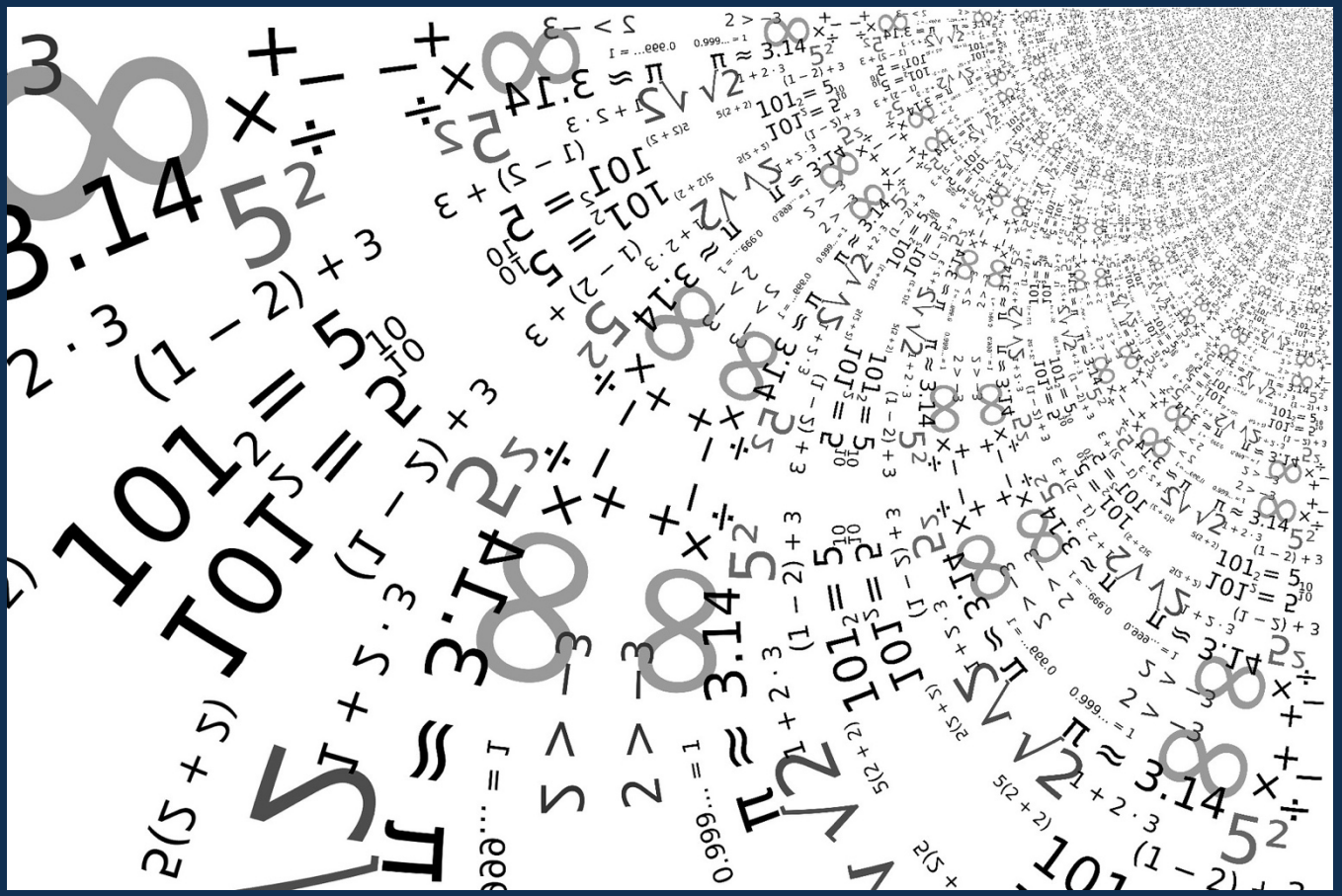
OVH.com



Advantages

- Can backup large cluster in less than 24h
- Greatly reduced compute power needed
- Can recover in minutes, not hours / days

OVH Ceph Backups - numbers



OVH.com

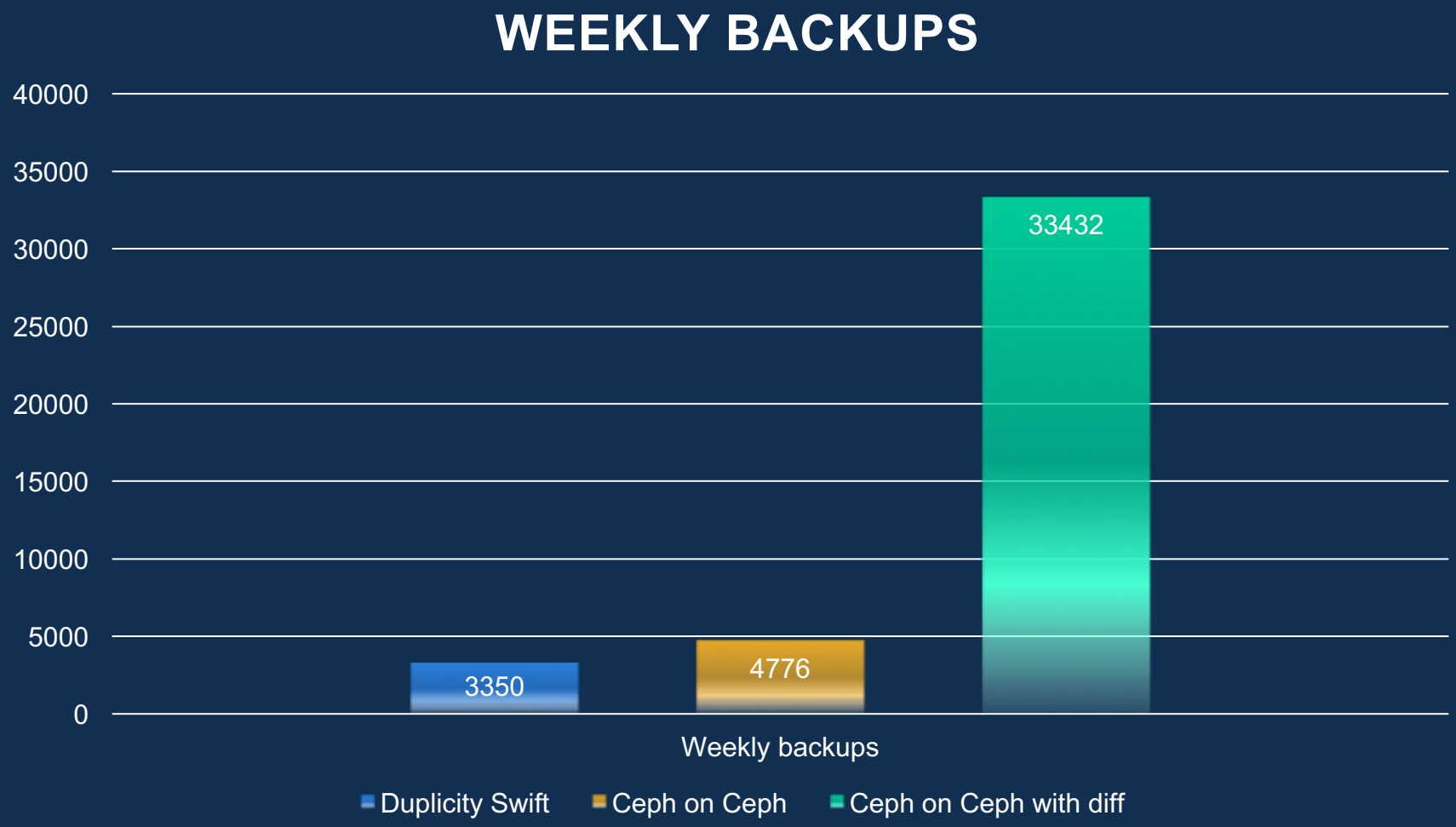
Global info:

34 Clusters with active backup

~9000 backups finished daily

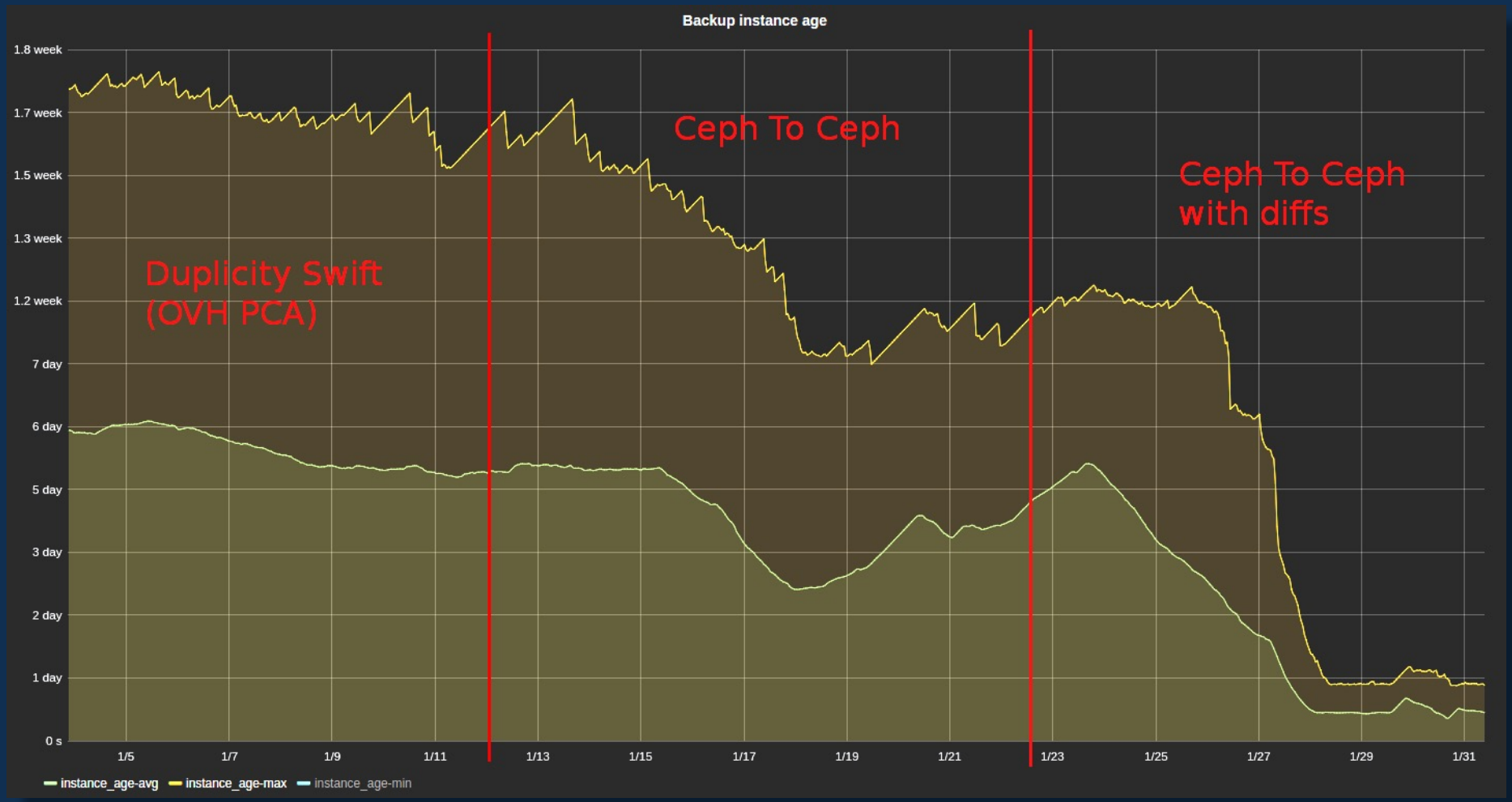
~0.6 PB of data exported daily

Large cluster case study:

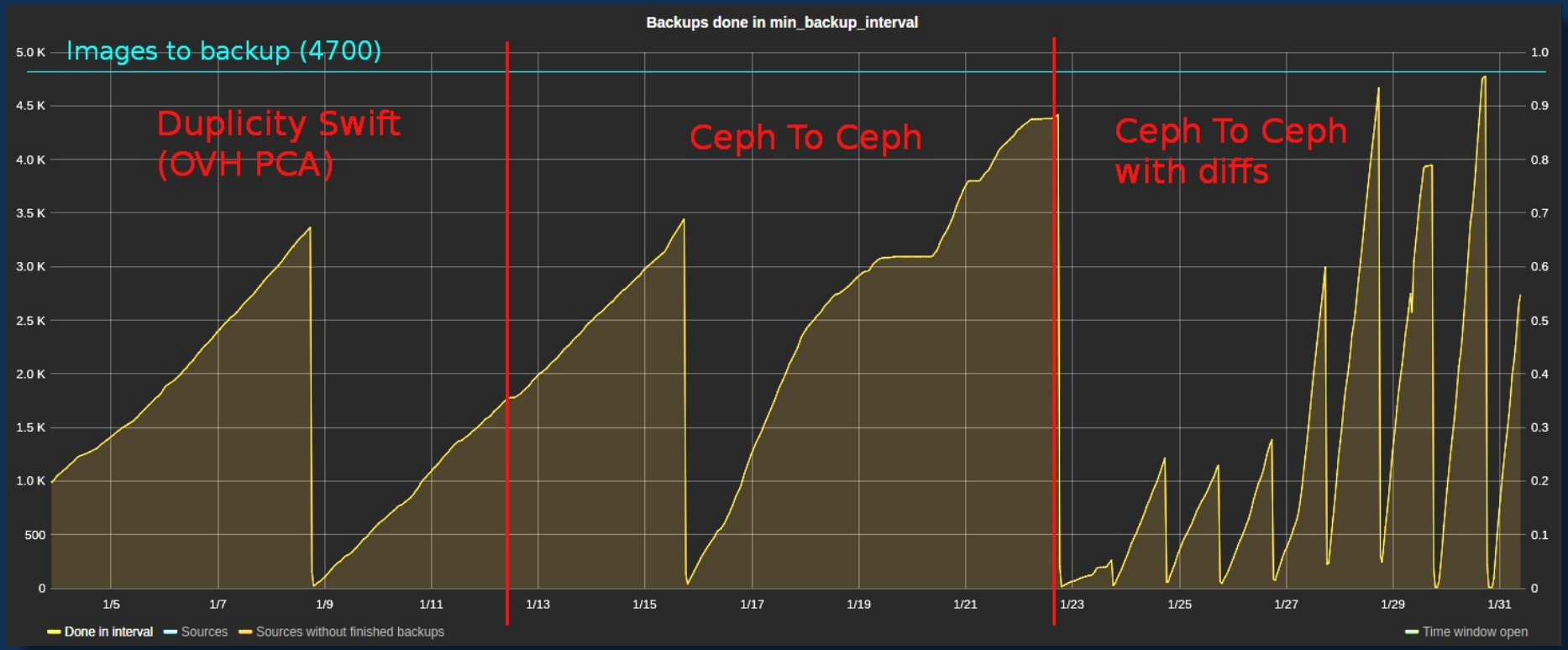


OVH.com

Large cluster case study:



Large cluster case study:



To sum up…

- Backups at scale definitely possible…
- … but better start with Ceph-on-Ceph
- You can get down to 24h backup window on highly utilized clusters
- Alternative storage to Ceph can give even better protection but will be slow
- Ceph-on-Ceph as a first line, alternative storage as a second line backup

Image sources

http://alphastockimages.com/

https://www.flickr.com/photos/soldiersmediacenter/4473414070

https://commons.wikimedia.org/wiki/File:Open_Floodgates_-_Beaver_Lake_Dam_-_Northwest_Arkansas,_U.S._-_21_May_2011.jpg

https://commons.wikimedia.org/wiki/File:Hot_Cold_mug.jpg

Questions?

bartlomiej.swiecki@corp.ovh.com