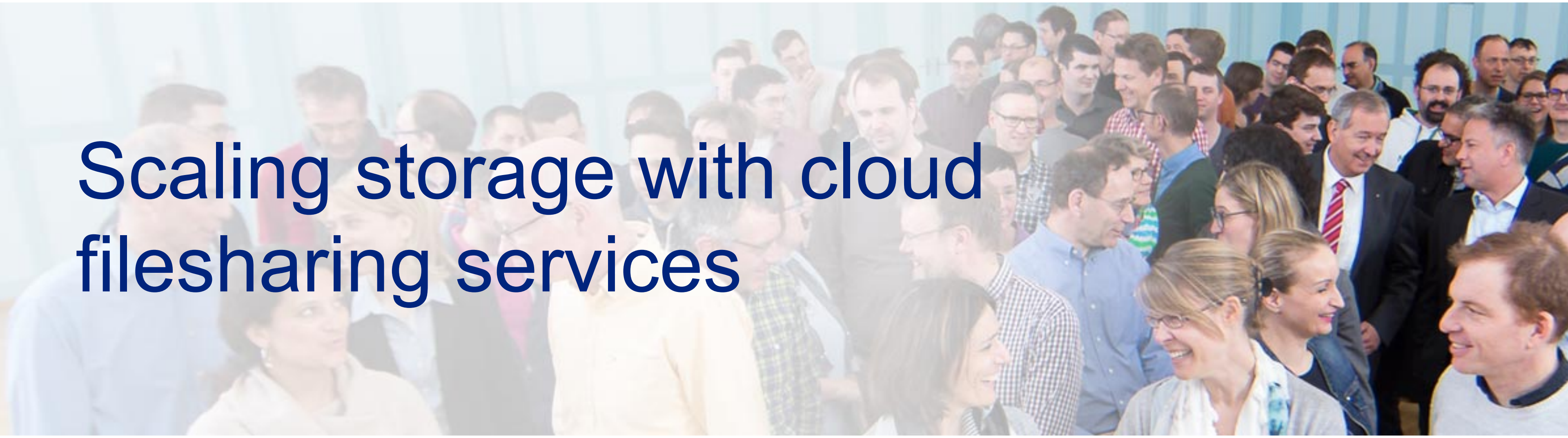


Scaling storage with cloud filesharing services

SWITCH

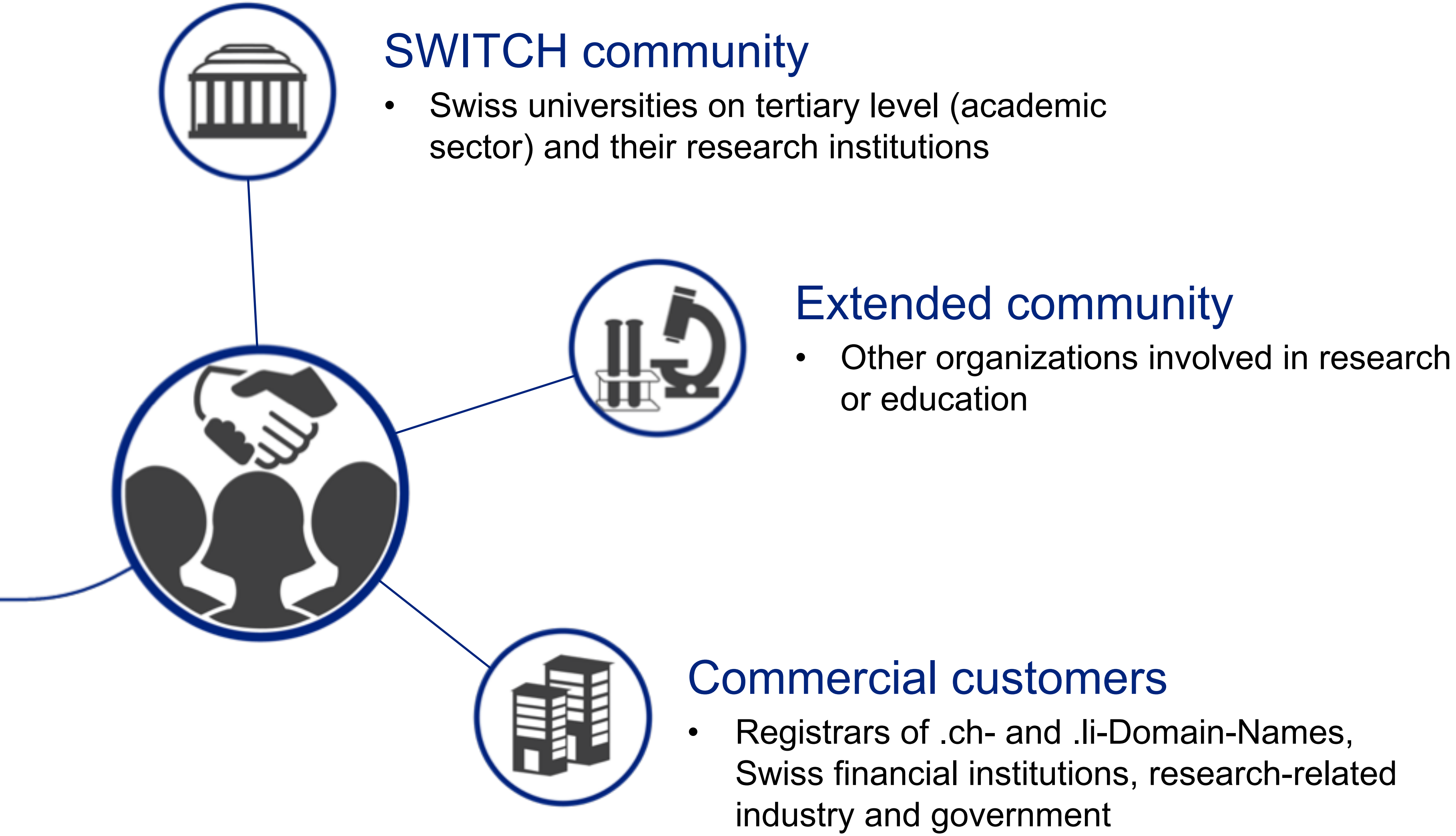
Greg Vernon
greg.vernon@switch.ch

Krakow, 31 January 2018

SWITCH

- We are the Swiss National Research and Education Network.
- We network the Institutes of Higher Education and Research to each other, and the rest of the world.
- We provide additional services such as Federated Authentication, Video, and File Sharing to our Educational customers.
- We manage the Top Level Domains for Switzerland (.ch) and Liechtenstein (.li).
- We provide SWITCH-CERT security service.

Our customers



SWITCH community

- Swiss universities on tertiary level (academic sector) and their research institutions

Extended community

- Other organizations involved in research or education

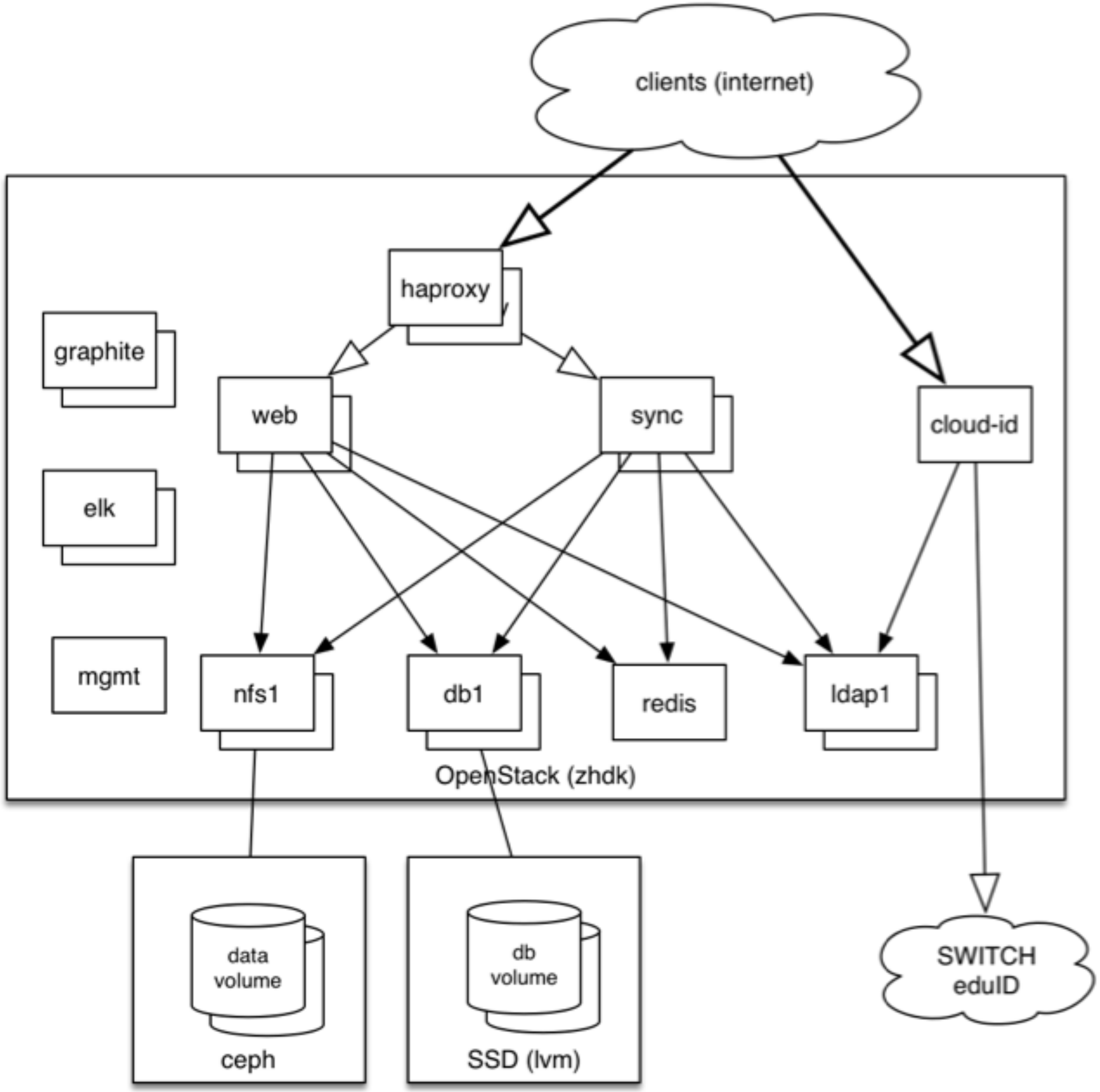
Commercial customers

- Registrars of .ch- and .li-Domain-Names, Swiss financial institutions, research-related industry and government

SWITCHdrive

SWITCHdrive is our branded ownCloud offering. We have the following:

- About 30,000 Users
- 125,000,000 files
- 125,000,000 rows in our oc_filecache table
- 3 Mariadb servers in a Galera cluster
- 9 Apache Servers(4 Sync/4 Web/1 Management)
- Redis
- 3 LDAP Servers
- 5 NFS servers running atop CEPH (130 TB currently)
- 2 HAproxy load balancers
- Monitoring (Graphite, ELK)
- Runs atop SWITCHengines, our OpenStack offering
- Most services are Docker containers


clients (internet)
haproxy
graphite
web
sync
cloud-id
elk
mgmt
nfs1
db1
redis
ldap1
OpenStack (zhdk)
data
volume
ceph
db
volume
SSD (lvm)
SWITCH
eduID

At the beginning...

- There was FileSender on Amazon
  - Storage was s3.
  - Hosting was in Dublin, so the network was the bottleneck, which we saw when we did some benchmarking.
  - It wasn't fast.
  - But it was scalable.

"Don't use the Cloud, be the Cloud"

- SWITCHdrive on our 'Building Cloud Competence' experimental service
  - OpenStack infrastructure
  - Ceph storage
  - Speeds were USB1. ☹

Filesender Redux

- We brought FileSender back as a service, as SWITCHdrive was going through teething pains
  - Storage is now Ceph, but still slow.
  - Expanding volumes take significant time as data must be copied.
  - Snapshots via OpenStack were glacially slow.
  - Used LVM to abstract the data layer, so volumes can be grown without having to take an outage.
  - Over time the Ceph storage has become faster, we rarely see any complaints from users.

So, what exactly is Ceph?

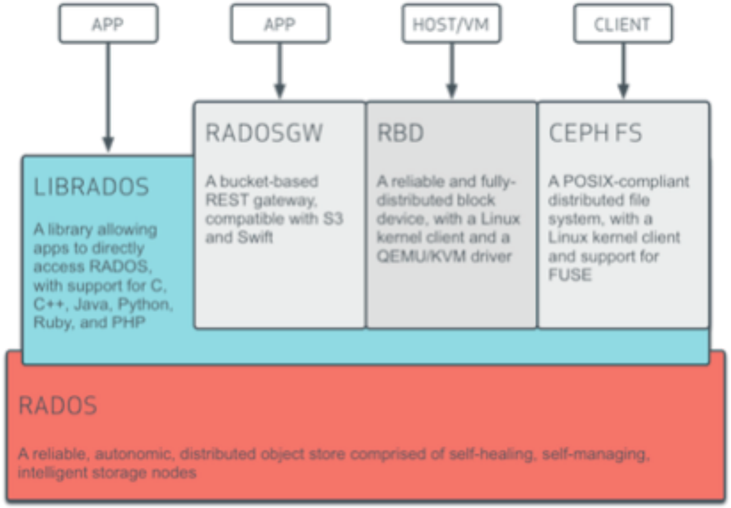
- Ceph is a horizontally distributed storage system
  - block storage (which we always use)
  - S3-like object storage (which is interesting)
  - posix file storage (CephFS)
- Very little overhead between storage and clients.
- There are separate nodes that allow for redundant storage devices (in our case, mostly spinning rust).
- It's popular among OpenStack operators, including SWITCHengines, which is what we use for our SWITCHdrive system.
- Distributed using CRUSH algorithm that hashes, a client will know exactly which device it will talk to.
- For replicated storage, the client can contact any of the replicas and will get the exact result.
- OSDs are the atomic units, and there are three for every Ceph volume.

What is Ceph? (continued)

- OSDs do journaling to (we use SSDs here), master writes request in journal, then syncs with other two, and then in background writes to spinning rust. Basically like a DB write-ahead log. The SSDs have advanced power-loss protection, and the SSDs cheat by using their RAM buffer, and we're battery-backed (with capacitors, enough to keep it with the ability to write to flash)
- You have a HUGE number of small 4MB objects that _should_ be distributed among many servers (more servers == faster) you hope this is random to get the distribution nicely sorted.

Ceph Architecture

(Source: http://docs.ceph.com/docs/giant/architecture/)



APP

APP

HOST/VM

CLIENT

LIBRADOS

A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP

RADOSGW

A bucket-based REST gateway, compatible with S3 and Swift

RBD

A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

CEPH FS

A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE

RADOS

A reliable, autonomic, distributed object store comprised of self-healing, self-managing, intelligent storage nodes

Why use Ceph?

- It's relatively cheap
- It's relatively safe
- It works well, and is well supported, with OpenStack
- It has very smooth scaling
- It is flexible, you can mismatch hardware without problems, except that your slowest devices will define your speeds

Is it slow?

- NO! Ceph is not slow any longer. We did some testing with fio, and found that it is reasonably fast for network storage. Not lighting quick, but (fast|cheap|safe), and we're cheap and safe.
  - fio: https://github.com/axboe/fio
- Ceph + NFS + XFS is slow, but we don't know exactly why, but we strongly suspect that there is an issue with large numbers of small files in single directories.
- Ceph snapshots are cheap to make, but really expensive to remove.
- Our current limitations might have more to do with our architecture than Ceph itself.

SWITCHdrive Today

- Stable service.
- Some Ceph Turbulance from deleting snapshots.
- Some issues with directories with many small files.
- But we think it could be faster.

SWITCHdrive in the future

- Ceph stays.
- Ceph Object storage (with ownCloud 10).
- zfs replacing xfs, and replacing Ceph snapshotting
- using zfs for offsite disaster recovery.
- Using a zfs pool to replace all of the nfs volumes, but we still have the large number of Ceph volumes.

Questions

Email: greg.vernon@switch.ch

30 Years

SWITCH – an integral part of the Swiss academic community since 1987.

www.switch.ch/30years