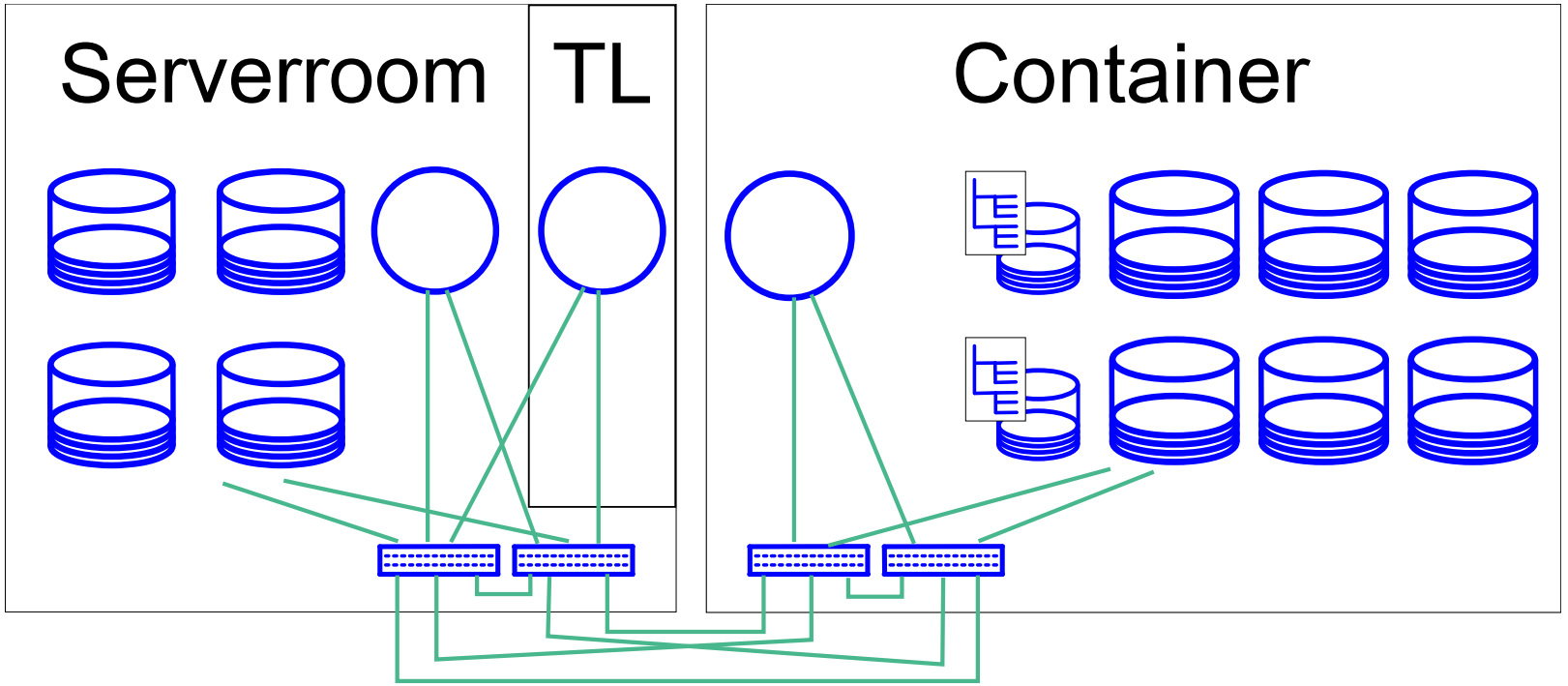
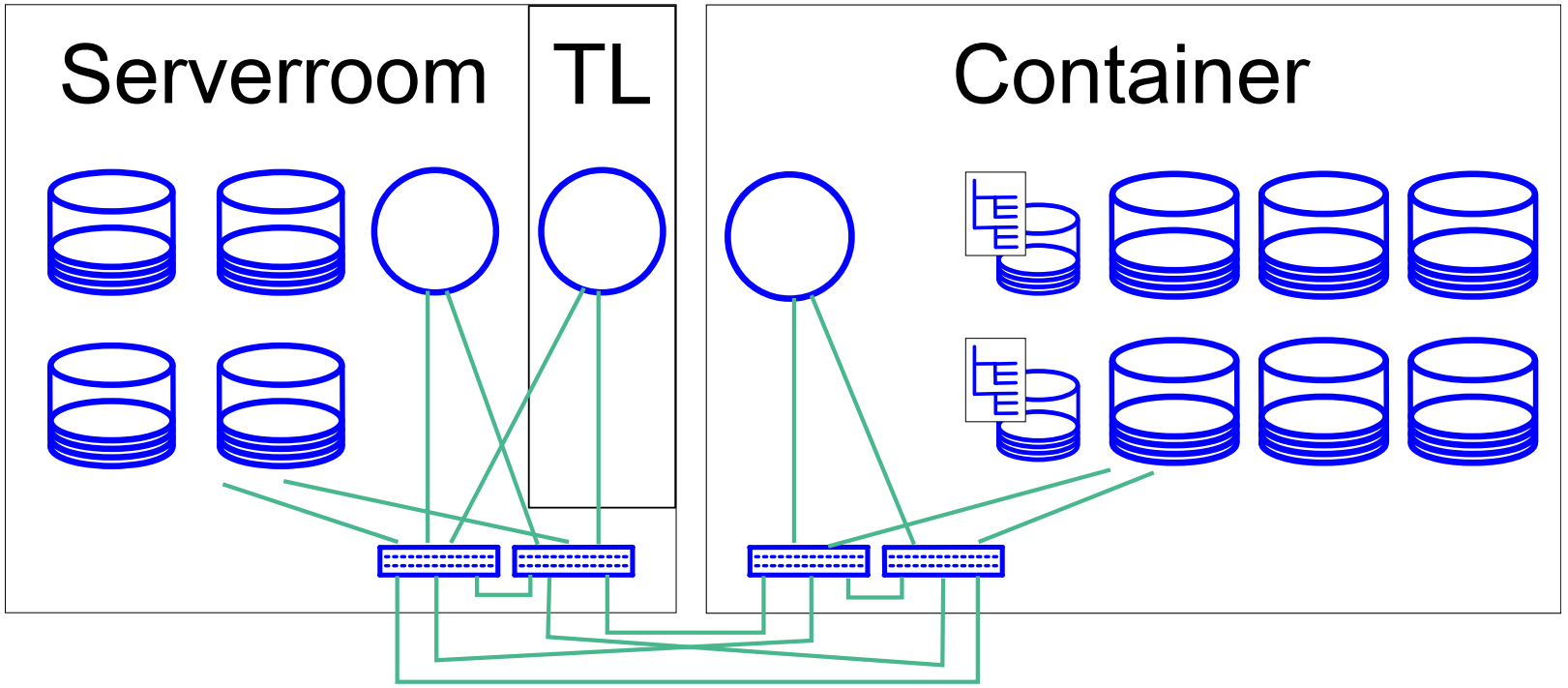
Ceph at DTU Risø - Year One

Frank Schilder



Ceph at DTU Risø - Year One



- Eash OSD host 4SSDs and 12HDDs.
- Each OSD server split up into 2 failure domains.
- 8+2 EC pool located in container (fs-data).
- 3(2) rep pool located in container (fs-meta).
- 6+2 EC pool located in server room, 6SR+2CON, (rbd-data).
- 3(2) rep pool located in server room (rbd-meta).
- 4(2) rep stretched pool, 2SR+2CON.

Ceph at DTU Risø - Year One

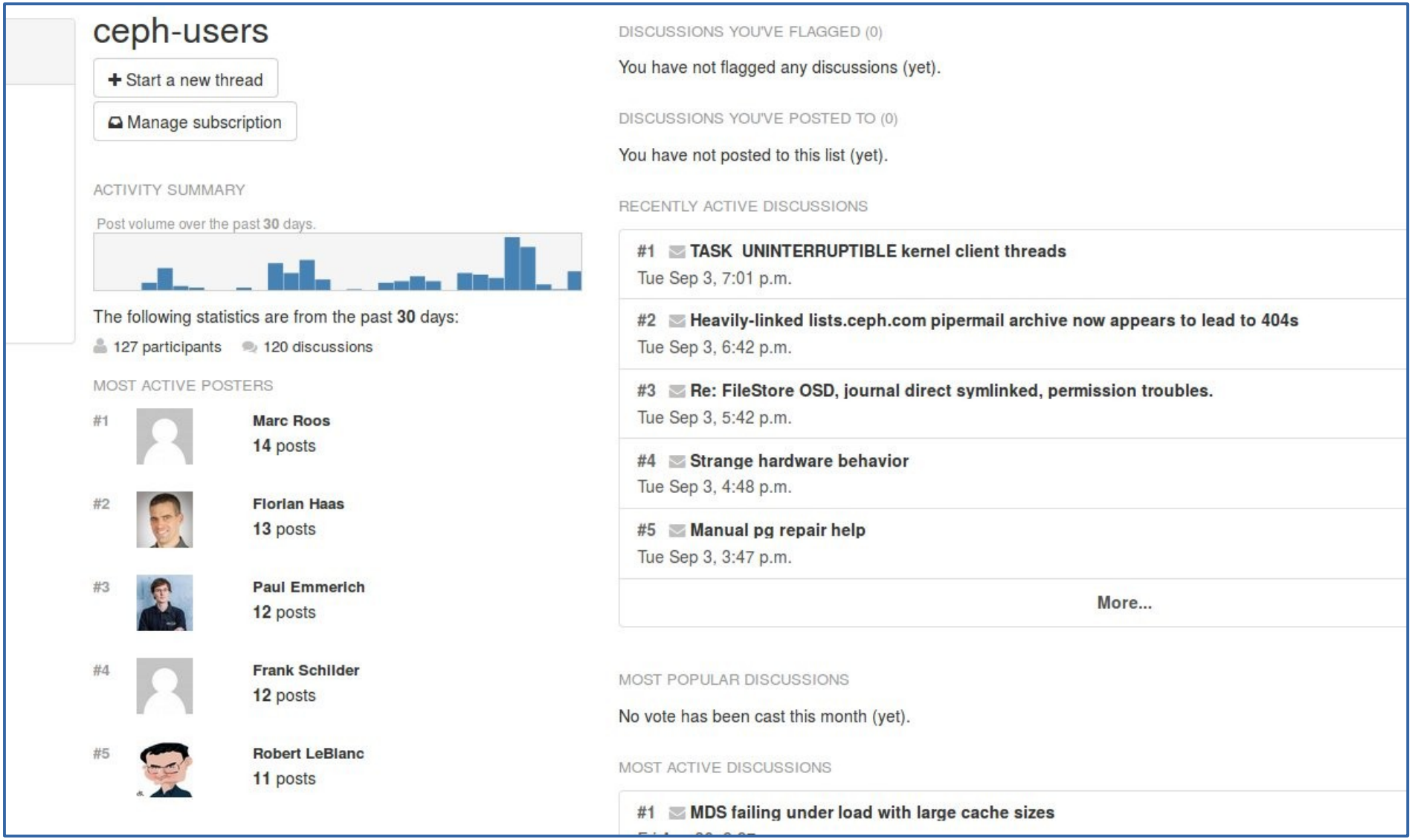
Any problems?

Ceph at DTU Risø - Year One

Any problems?

Well ...

Ceph at DTU Risø - Year One



ceph-users

+ Start a new thread

Manage subscription

ACTIVITY SUMMARY

Post volume over the past **30** days.

The following statistics are from the past **30** days:

127 participants 120 discussions

MOST ACTIVE POSTERS

#1 **Marc Roos** **14** posts

#2 **Florian Haas** **13** posts

#3 **Paul Emmerich** **12** posts

#4 **Frank Schilder** **12** posts

#5 **Robert LeBlanc** **11** posts

DISCUSSIONS YOU'VE FLAGGED (0)

You have not flagged any discussions (yet).

DISCUSSIONS YOU'VE POSTED TO (0)

You have not posted to this list (yet).

RECENTLY ACTIVE DISCUSSIONS

#1 **TASK UNINTERRUPTIBLE kernel client threads**
Tue Sep 3, 7:01 p.m.

#2 **Heavily-linked lists.ceph.com pipermail archive now appears to lead to 404s**
Tue Sep 3, 6:42 p.m.

#3 **Re: FileStore OSD, journal direct symlinked, permission troubles.**
Tue Sep 3, 5:42 p.m.

#4 **Strange hardware behavior**
Tue Sep 3, 4:48 p.m.

#5 **Manual pg repair help**
Tue Sep 3, 3:47 p.m.

More...

MOST POPULAR DISCUSSIONS

No vote has been cast this month (yet).

MOST ACTIVE DISCUSSIONS

#1 **MDS failing under load with large cache sizes**

Ceph at DTU Risø - Year One

[ceph-users] Can't create erasure coded pools with k+m greater than hosts?

(Slightly abbreviated)

Den tors 24 okt. 2019 kl 09:24 skrev Frank Schilder <frans@dtu.dk>:

> What I learned are the following:
>
> 1) Avoid this work-around too few hosts for EC rule at all cost.
>
> 2) Do not use EC 2+1. It does not offer anything interesting for production. Use 4+2 (or 8+2, 8+3 if you have the hosts).
>
> 3) If you have no perspective of getting at least 7 servers in the long run (4+2=6 for EC profile, +1 for fail-over automatic rebuild), do not go for EC.
>
> 4) Before you start thinking about replicating to a second site, you should have a primary site running solid first.
>
> This is collected from my experience. I would do things different now and maybe it helps you with deciding how to proceed. Its basically about what resources can you expect in the foreseeable future and what compromises are you willing to make with regards to sleep and sanity.

Amen to all of those points. We did similar-but-not-same mistakes on an EC cluster here. You are going to produce more tears than I/O if you make these mis-designs mentioned above.

==> Most ceph problems are due to design choices.

Ceph at DTU Risø - Year One

Did we encounter any bugs?

Yes, a small number:

- monitors do not always clean up ops list after dispatch (rm_snap)
  [fixed in new revision]

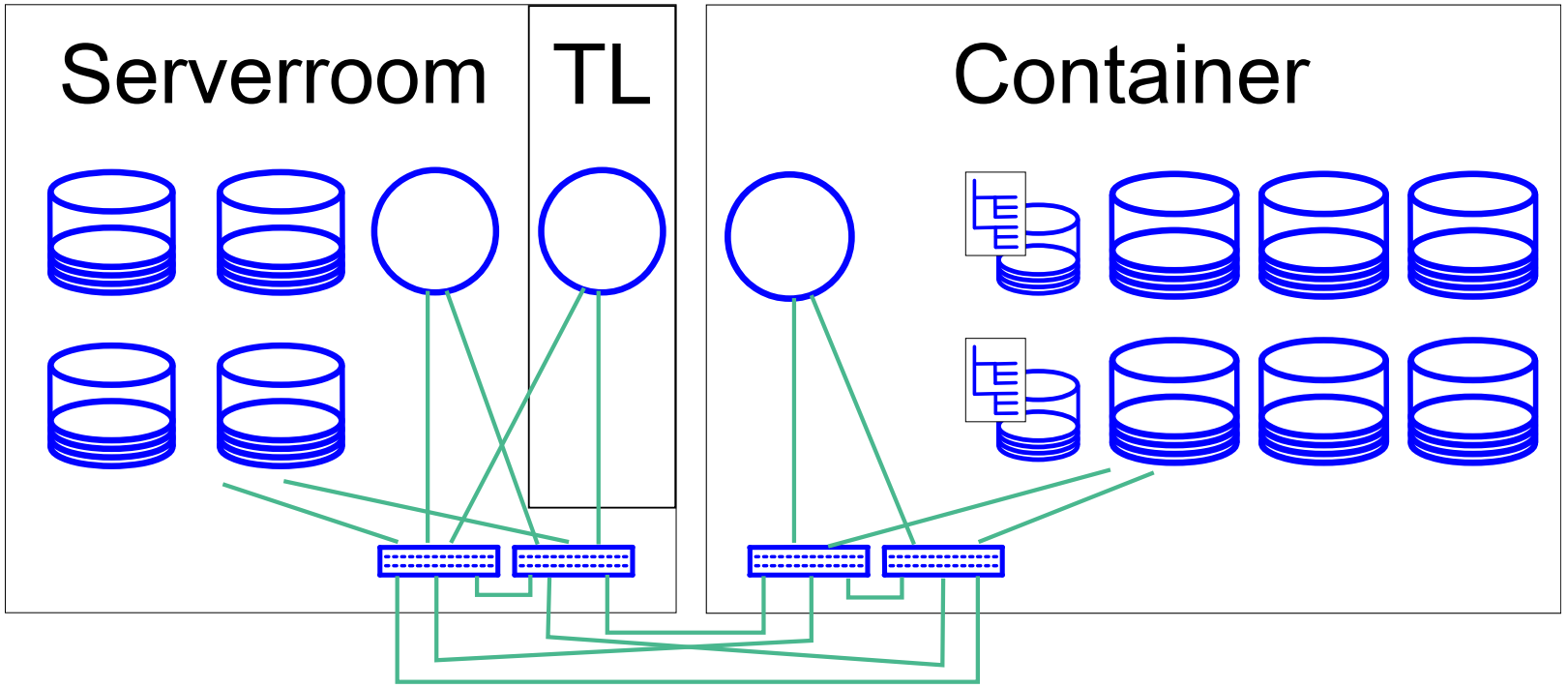
- MDS daemons can get stuck in dirfrag operations
  [open in mimic, almost fixed in nautilus]

- concurrent append write truncates files
  [fixed in 7.5 (kernel-3.10.0-862.33.1.el7)
  and 7.6 (kernel-3.10.0-957.16.1.el7)]

We had to disable file system snapshots and we instructed users how to work around the append write bug.

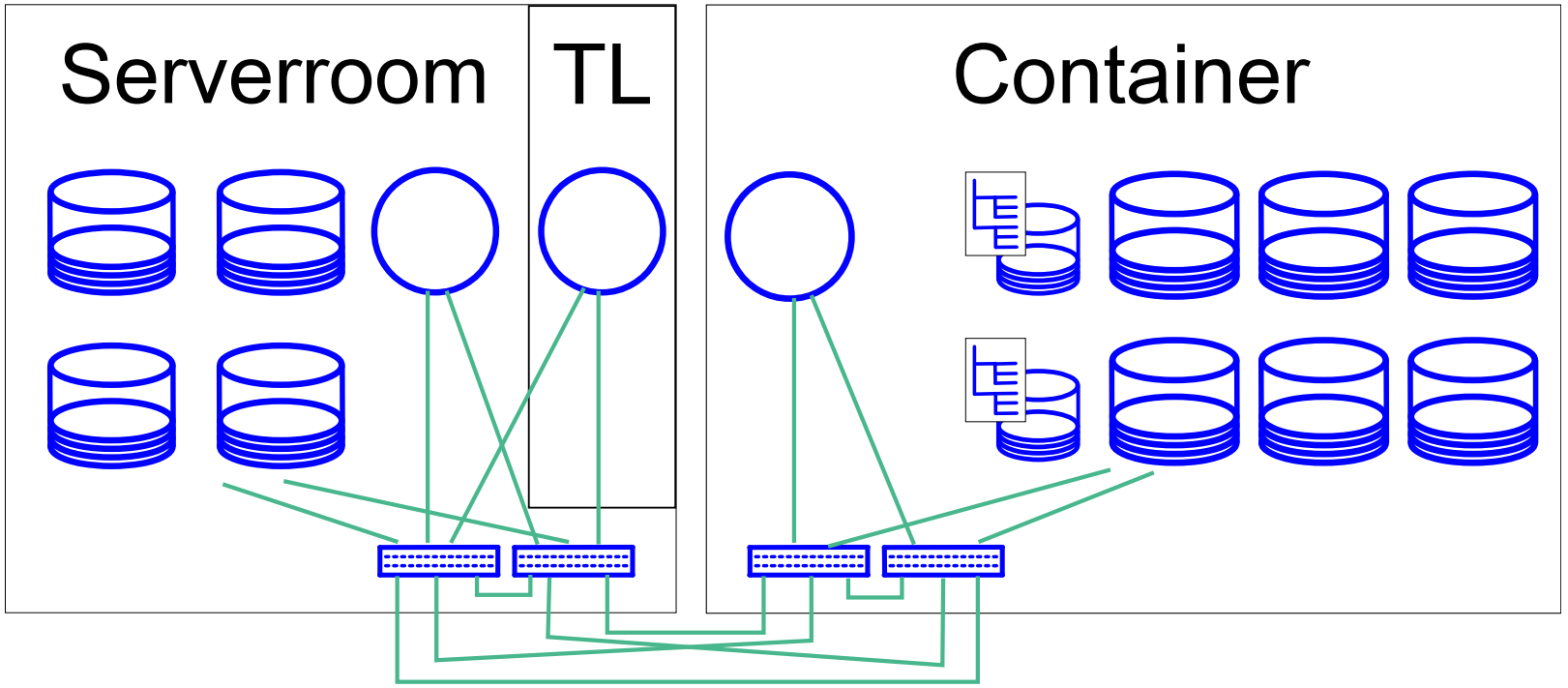
What about design choices?

Ceph at DTU Risø - Year One



- Eash OSD host 4SSDs and 12HDDs.
- Each OSD server split up into 2 failure domains.
- 8+2 EC pool located in container (fs-data).
- 3(2) rep pool located in container (fs-meta).
- 6+2 EC pool located in server room, 6SR+2CON, (rbd-data).
- 3(2) rep pool located in server room (rbd-meta).
- 4(2) rep stretched pool, 2SR+2CON.

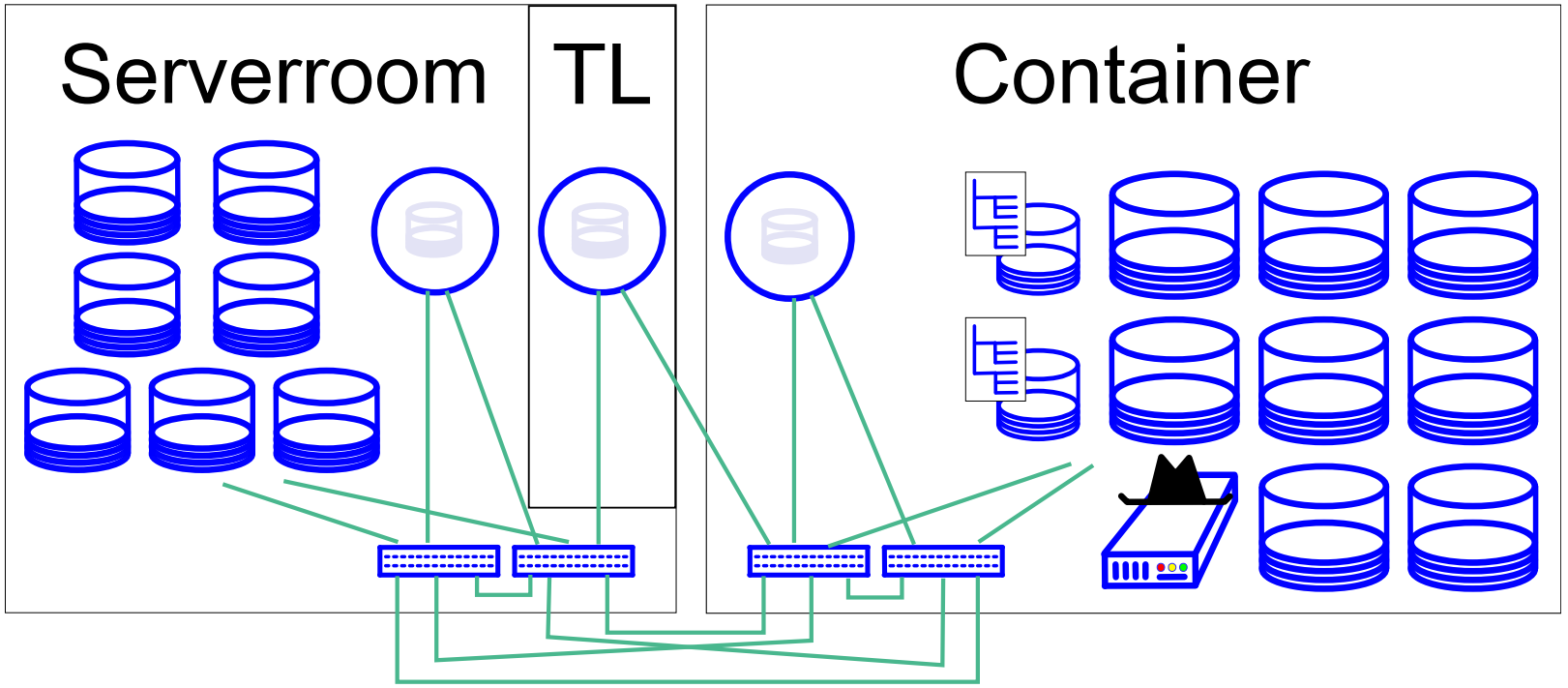
Ceph at DTU Risø - Year One



- Zoning of OSD hosts dramatically increases admin effort.
- Overlapping pools for different tier levels does not work as expected.
- Dimensioning pool I/O profile for RBD and cephfs are vastly different.

- Re-design pool locality completely disjoint for different tier levels.
- Change to all-flash OSDs for RBD workloads.

Ceph at DTU Risø - Year One



- Added 3 OSD hosts to server room and 2 OSD hosts and a head- and NFS gateway node in the container.
- Reorganize disks, preference: HDDs in container, SSDs in server room (see ceph df next slide). Reduce zoning as much as possible.
- Three completely disjoint sets of OSD hosts for RBD (server room), ceph fs (container) and multi-site replicated RBD (server room, tape library, container; located together with MONs for maximum availability).

Ceph at DTU Risø - Year One

```
[root@ceph-01 ~]# ceph df
GLOBAL:
    SIZE        AVAIL       RAW USED     %RAW USED
    1.8 PiB     1.7 PiB       43 TiB          2.40
POOLS:
    NAME                    ID     USED        %USED     MAX AVAIL     OBJECTS
    sr-rbd-meta-one         1      6.2 GiB      0.07       9.0 TiB        2869
    sr-rbd-data-one         2      2.9 TiB      8.66        30 TiB     1023569
    sr-rbd-one-stretch      3       93 GiB      0.10        89 TiB       25953
    con-fs-meta             4      313 MiB      0.03       1.1 TiB      552525
    con-fs-data             5       33 TiB      3.28       967 TiB    15888856
    con-rbd-meta-hpc-one    7        429 B         0       1.1 TiB          18
    con-rbd-data-hpc-one    8      593 MiB         0       967 TiB         204
    sr-rbd-data-one-hdd     11         0 B         0       289 TiB           0
```

- ceph fs is extremely economic with meta data on SSD.
- The main bottleneck for IO is the number of HDDs, which cannot keep up with the IOPs of the SSD meta data pool. Slow OPS are only observed on HDD.

- RBD is extremely IOPs hungry.
- RBD also requires substantially more meta data per data unit than ceph fs.

Ceph at DTU Risø - Year One

Main challenge: design pool for RBD with sufficient IOPs.

- Data access to ceph fs is structured. Aggregated I/O performance is relevant and OSD bluestore cache allocation is per OSD.

- Data access to RBD images is unstructured. Relative I/O performance is relevant and bluestore cache allocation is per TB, not per OSD.

Why relative I/O performance?

VM images typically have a fixed size and require a certain constant minimum amount of IO per VM. Hence, the number of VMs typically scales with the size of an RBD pool and with it the total IO requirements.

In other words, the ratio of required IOPs / TB is constant, which means that simply growing an RBD pool will not lead to improved performance as in the case of ceph fs.

IOPs Calculation

OS	Av. Im. Size (GB)	Typ. IOPs RBD only	Peak IOPs RBD only
Windows	50	10-25	50
Linux	10	1-5	20

IOPs/VM write	VMs/TB	EC rep fac (=k+m)	Bluestore IOPs/client OP read	Bluestore IOPs/client OP write	IOPs/TB read	IOPs/TB write	IOPs/TB tot
50	10	8	1	2	4000	8000	12000
50	20	8	1	2	8000	16000	24000

IOPs/VM write	EC rep fac (=k+m)	Bluestore IOPs/client OP read	Bluestore IOPs/client OP write	IOPs/VM read	IOPs/VM write	IOPs/VM tot
5	10	1	2	50	100	150
10	10	1	2	100	200	300

Aggregated IOPs for spinning disk cluster

#disks	IOPs/disk	IOPs total	#VMs min.	
150	100	15000	50	100

SSD Performance Tests

Vendor Specs

	Dell Toshiba PX05SMB040Y / 5VHHG 400GB 12Gbps SAS eMLC Enterprise Endurance SSD	Micron 5210 ION Enterprise SATA QLC SSD	Micron 5210 ION Enterprise SATA QLC SSD	Micron 5210 ION Enterprise SATA QLC SSD	Micron 5200 PRO Enterprise SATA SSD	Micron 5200 PRO Enterprise SATA SSD	Kingston Data Centre DC500M SSD	Kingston Data Centre DC500M SSD	Kingston Data Centre DC500M SSD
Capacity (GB)	400	1920	3840	7680	1920	3840	960	1920	3840
Seq. Read (MB/s)	1992	540	540	540	540	540	555	555	555
Seq. Write (MB/s)	1100	260	350	360	520	520	520	520	520
4K Rand Read	270000	70000	83000	90000	95000	95000	98000	98000	98000
4K Rand Write	105000	13000	6500	4500	32000	24500	70000	75000	75000
DWPD	10	0.2	0.09	0.05	1.7	2.5	1.3	1.3	1.3
MTW (4K rand w)	7300	701	631	701	5950	17600	2278	4555	9110
Power loss prot.		yes	yes	yes	yes	yes	yes	yes	yes
SMART	yes	yes	yes	yes	yes	yes	yes	yes	yes
Price approx.	7000	1400	2670	5130	2170	7600	1730	3160	4760
Price/TB	17500	729	695	668	1130	1979	1802	1646	1240
IOPs/TB Read	675000	36458	21615	11719	49479	24740	102083	51042	25521
IOPs/TB Write	262500	6771	1693	586	16667	6380	72917	39063	19531
BWL IO size read (KB)	7.55	7.90	6.66	6.14	5.82	5.82	5.80	5.80	5.80
BWL IO size write (KB)	10.73	20.48	55.14	81.92	16.64	21.73	7.61	7.10	7.10

Benchmarks (write cache flag = 0)

iodepth — IOP/s w (randwrite 4K) `fio -ioengine=libaio -name=test -sync=1 -direct=1 -bs=4k -rw=randwrite -runtime=60 -filename=/dev/sdb -iodepth=1`

1	27981		29825		32343		37463		
2	51261		26430		51498		67460		
4	103567		26460		63924		66846		

IOP/s w/TB

1	69952	0	7767	0	16845	0	39024	0	0
2	128152	0	6883	0	26822	0	70271	0	0
4	258918	0	6891	0	33294	0	69631	0	0

IOP/s r/w (randrw 4K, 50% read) `fio -ioengine=libaio -name=test -sync=1 -direct=1 -bs=4k -rw=randrw -runtime=60 -filename=/dev/sdb -iodept`

1	7783		3791		6381		6841		
	7773		3792		6369		6831		
2	16267		6306		10859		11484		
	16247		6295		10852		11470		
4	30059		8772		13742		17371		
	30025		8762		13726		17347		

IOP/s (r/w)/TB, 50% read

1	19458	0	987	0	3323	0	7126	0	0
	19432	0	987	0	3317	0	7115	0	0
2	40668	0	1642	0	5656	0	11963	0	0
	40618	0	1639	0	5652	0	11948	0	0
4	75148	0	2284	0	7158	0	18095	0	0
	75062	0	2282	0	7149	0	18070	0	0

IOP/s r/w (randrw 4K, 80% read) `fio -ioengine=libaio -name=test -sync=1 -direct=1 -bs=4k -rw=randrw -rwmixread=80 -runtime=60 -filename=/d`

1	9904		4964		7727		8282		
	2485		1246		1940		2080		
2	21304		8073		13915		13828		
	5329		2028		3486		3465		
4	40860		10765		19925		19814		
	10217		2702		4986		4960		

IOP/s (r/w)/TB, 80% read

1	24760	0	1293	0	4024	0	8627	0	0
	6212	0	325	0	1011	0	2166	0	0
2	53260	0	2102	0	7247	0	14404	0	0
	13324	0	528	0	1816	0	3610	0	0
4	102150	0	2803	0	10378	0	20639	0	0
	25542	0	704	0	2597	0	5167	0	0

IO size (KB) — BW (w/IOPs) (MB/s, randwrite xK) `fio -ioengine=libaio -name=test -sync=1 -direct=1 -rw=randwrite -runtime=60 -iodepth=1 -filename=/dev/sd`

blue

16	324		254		288		315		
	20745		16247		18444		20187		
32	494		289		365		389		
	15809		9255		11683		12447		
64	669		312		421		441		
	10699		4987		6739		7055		
128	778		322		461		468		
	6226		2579		3687		3744		

BW (w/IOPs)/TB

16	810	0	66	0	150	0	329	0	0
	51862	0	4231	0	9606	0	21028	0	0
32	1235	0	75	0	190	0	405	0	0
	39522	0	2410	0	6085	0	12965	0	0
64	1672	0	81	0	219	0	459	0	0
	26749	0	1299	0	3510	0	7349	0	0
128	1946	0	84	0	240	0	488	0	0
	15566	0	672	0	1920	0	3900	0	0

iodepth — IOP/s r/w (randrw 16K, 50% read) `fio -ioengine=libaio -name=test -sync=1 -direct=1 -bs=16k -rw=randrw -runtime=60 -filename=/dev/sdb -iod`

1	6054		3157		3022		3711		
	6047		3153		3015		3711		

IOP/s (r/w)/TB, 50% read

1	15136	0	822	0	1574	0	3866	0	0
	15118	0	821	0	1570	0	3865	0	0

Ceph at DTU Risø - Year One

Very good article about SSD performance:

`https://yourcmc.ru/wiki/Ceph_performance`

With current SSD prices, EC all-flash storage already beats replicated HDD pools, which could also provide acceptable RBD performance per TB.

This development also makes cache tiering less interesting.

Ceph at DTU Risø - Year One

Other lessons learned:

1) Set

osd_op_queue = wpq
osd_op_queue_cut_off = high

on OSD and MDS. This will reduce the impact of backfill as well as prevent daemons missing heart beats under heavy load.

Ceph at DTU Risø - Year One

Other lessons learned:

2) Procedure for adding OSDs:

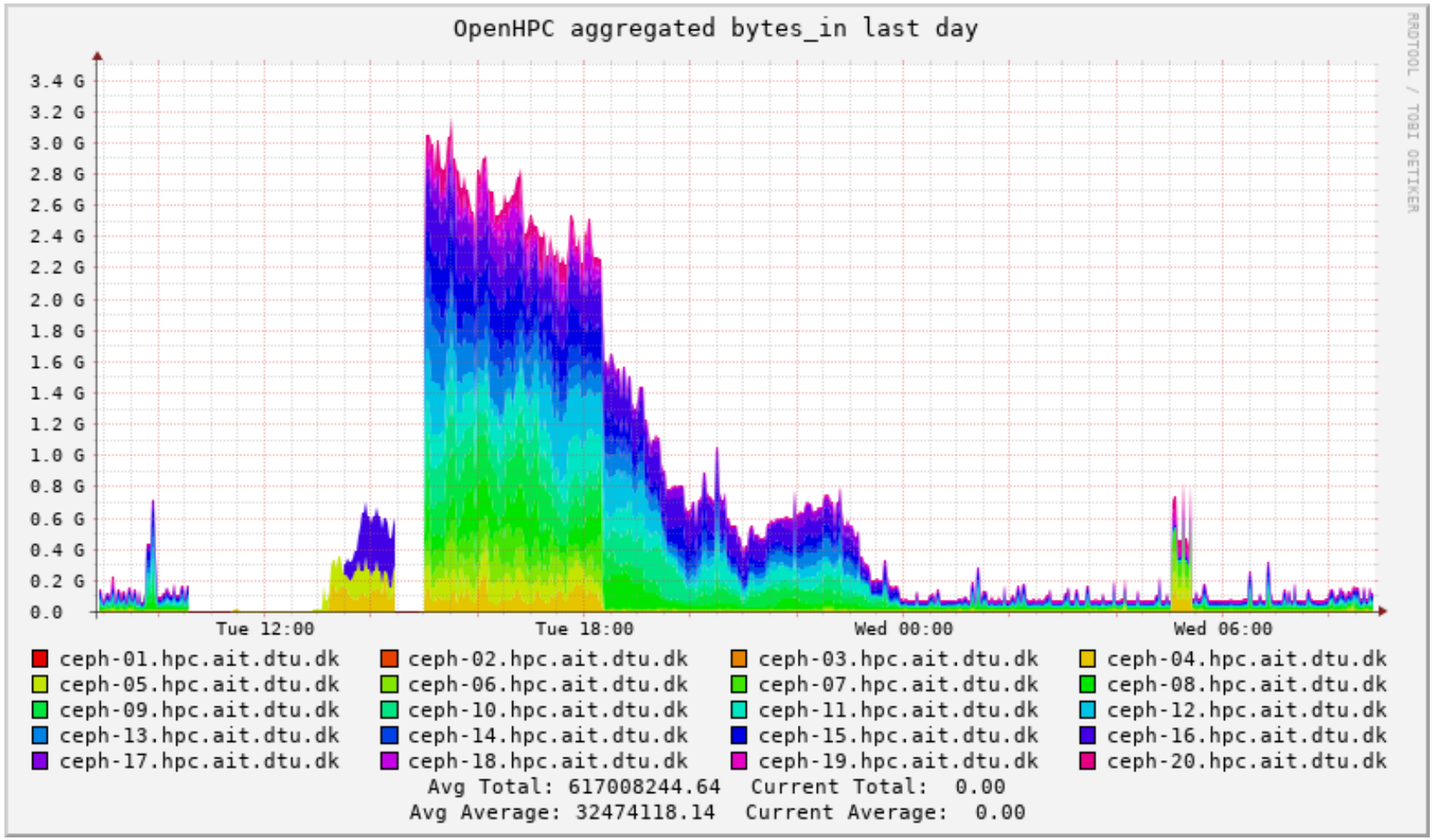
- Make sure cluster is HEALTH_OK
- ceph osd set norebalance
- ceph osd set nobackfill
- Add all the OSDs
- Wait for all PGs to peer. All PGs must be active before continuing.
- ceph osd unset nobackfill
- ceph osd unset norebalance
- Wait for HEALTH_OK

This procedure allows adding an arbitrary number of OSDs with minimal amount of data movement. I did not observe any client IO problems with the setting of lesson (1) and

osd_max_backfills = 3
osd_recovery_max_active = 8
osd_recovery_sleep = 0.05

Ceph at DTU Risø - Year One

Other lessons learned:



Ceph at DTU Risø - Year One

Other lessons learned:

3) When moving/adding OSDs, always move/add while up, not with (implied) change of crush location (of down OSDs) and restart.