



WELCOME

Olivier Chaze

14th October 2026, Zurich

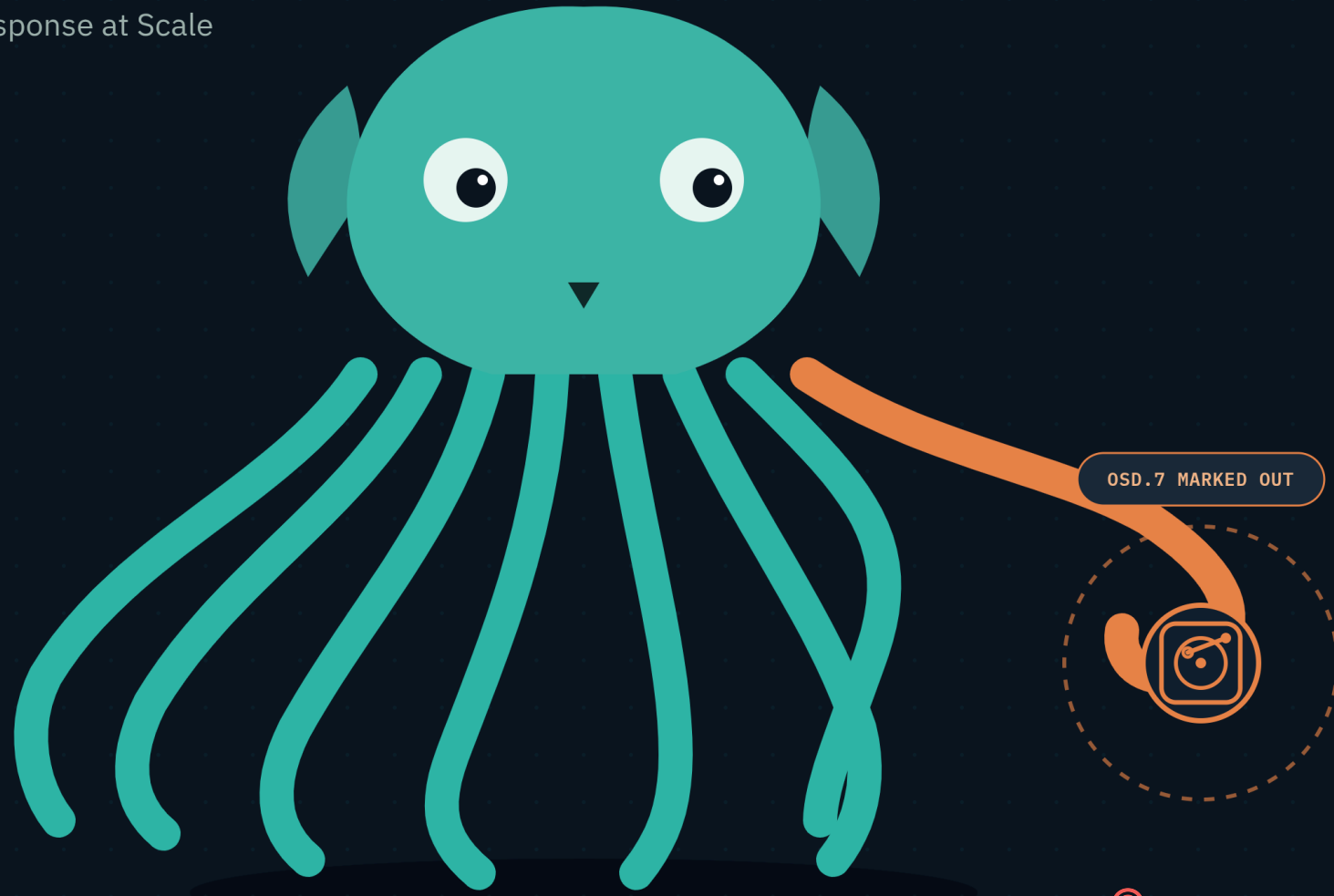


One Slow Disk, Zero Customer IOPS

Automating Ceph Failure Response at Scale

infomaniak The Ethical Cloud⁺

olivier.chaze@infomaniak.com



Ceph Days - Zürich - October 2026

Infomaniak & Ceph - Context

Technology	Purpose
OpenStack	Since 2012. Most of our infrastructure is based on OpenStack We build and maintain the packages for Debian (Thomas Goirand)
OpenStack Swift	Since 2012. (Backup Solutions · S3 Storage · Web Hosting · Collaborative Tools · Drive · VOD)
Ceph RBD	Since 2019. (Public Cloud, DBaaS, KaaS)
Ceph RGW	Since 2024.

Numbers

Ceph RGW: 150+ PB

Standard RGW Cluster Setup: 6PB to 29PB

- 24x HDDs (24+ Seagate TB) per server + 3 NVMEs (4TB)
- EC 8+3
- Active sync to another DC
- Adding capacity in bunches of 12 servers
- Max sync speed between clusters: 120 Gbps

Ceph RBD: 15+ PB

Standard RBD Cluster Setup

- 10 to 18 NVMEs per server
- Replica 3

Other info

- Cephadm
- Debian Trixie
- BGP to the host
- Ansible (provisioning) / Salt (Configuration)
- Dedicated haproxy servers in front of RGW
- mclock custom profile
- Clusters always in WARN

Standard RGW cluster

--- RAW STORAGE ---

CLASS	SIZE	AVAIL	USED	RAW USED	%RAW USED
hdd	27 PiB	11 PiB	17 PiB	17 PiB	61.44
ssd	279 TiB	275 TiB	4.7 TiB	4.7 TiB	1.67
TOTAL	28 PiB	11 PiB	17 PiB	17 PiB	60.85

--- POOLS ---

POOL	ID	PGS	STORED	OBJECTS	USED	%USED	MAX AVAIL
.mgr	1	1	167 MiB	43	669 MiB	0	65 TiB
.rgw.root	2	32	220 KiB	59	1.6 MiB	0	65 TiB
kdrive04a.rgw.control	3	32	0 B	9	0 B	0	1.2 PiB
kdrive04a.rgw.meta	4	32	1.0 GiB	2.79M	34 GiB	0	1.2 PiB
kdrive04a.rgw.log	5	32	8.6 GiB	13.63M	51 GiB	0.02	65 TiB
kdrive04a.rgw.buckets.index	6	32	826 GiB	12.65M	3.2 TiB	0.06	1.2 PiB
kdrive04a.rgw.buckets.non-ec	7	32	77 KiB	1	323 KiB	0	1.2 PiB
kdrive04a.rgw.buckets.data	8	8192	12 PiB	7.27G	16 PiB	76.53	3.5 PiB



2s

0 to 100 km/h

Yes

Ceph vs Swift - 2 philosophies



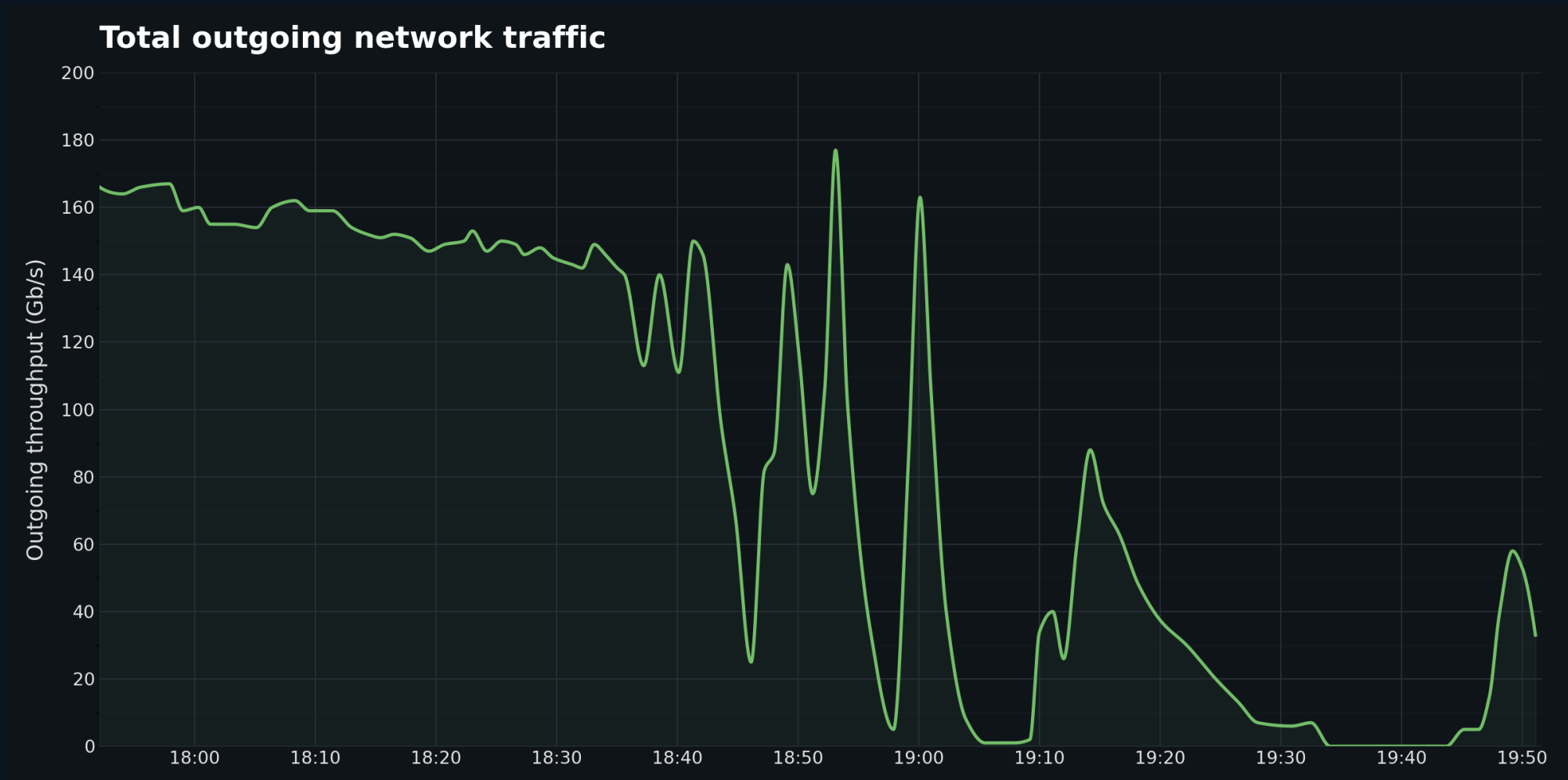
~~2s~~
2ms
Synchronous
No
Moody

~~0 to 100 km/h~~
Data Consistent
Consistency model
Latency-tolerant
When things go sideways

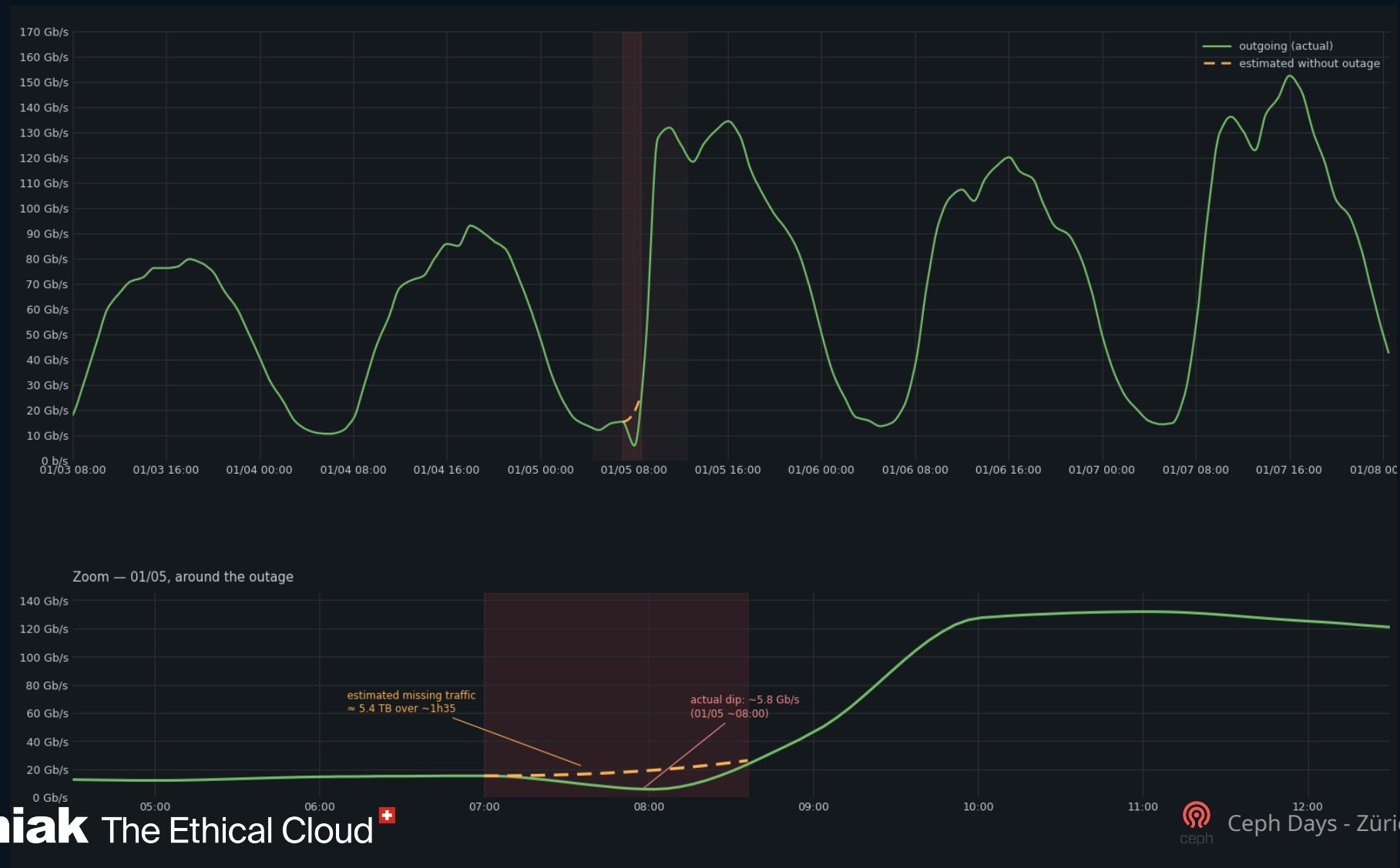


~~Yes~~
Eventually
Asynchronous
Yes
Never gonna give you up

Impact of a failing HDD on a RadosGW cluster



Impact of a failing HDD on a RadosGW cluster



"A disk is failing" isn't an incident. It's a normal day.

- Thousands of spinning disks → **one failing every week**
- Spotting it isn't the hard part
- The hard part is the **decision** to take out the disk

Should I take out the disk?

→ **That should be a decision made by code, not left to human judgment.**

ceph-companion - One Go Binary, 3 usages (operator, analyze, get)

```
every OSD host ────────────  
systemd.timer → `ceph-companion operator`  
├─ actions via: cephadm shell ceph ... / cephadm unit stop  
├─ reads /etc/ceph-companion/rules/  
├─ queries Prometheus/VictoriaMetrics  
└─ local_checks: journalctl, smartctl, lsblk
```

```
/etc/ceph-companion/  
├─ config.yaml  
├─ patterns # OSDs metric filtering and analysis  
│   ├── bluestore_slow_ops.yaml  
│   ├── HighWriteLatencyFinding.yaml  
│   ├── list_clients.yaml  
│   ├── list_events.yaml  
│   ├── osd_scrubs.yaml  
│   ├── scrub_blocked_ops.yaml  
│   ├── slow_ops_in_flight.yaml  
│   ├── slow_osd_op_op_applied.yaml  
│   └─ ...  
└─ rules # conditions to trigger actions  
    ├── ...  
    ├── stop_osd_bluestore_slow_committed_kv_count.yaml  
    ├── stop_osd_with_hw_error.yaml  
    └─ stop_sluggish_osd_service.yaml
```

```
# /etc/ceph-companion/config.yaml  
prometheus:  
  url: "https://vm.example.com/prometheus" # Prometheus or VictoriaMetrics  
  cluster_id: "ae7sdlk-17dd-xxxxx-8320-xxxxx" # fsid  
notifications:  
  webhook: "https://xxxxx.kchat.infomaniak.com/hooks/..." # Optional
```

Built-in function example:

```
$ ceph-companion get osd-performance-drop
```

```
Performance (5m vs 1h), positive diff = traffic increased:  
IOPS: 1172.40 (5m) vs 1107.39 (1h) → diff: +5.87%  
BW: 3.28 GB/s (5m) vs 3.06 GB/s (1h) → diff: +7.24%
```

Part 1 - ceph-companion operator - Automatically taking out a slow HDD

```
/etc/ceph-companion/rules/stop OSD bluestore slow committed kv count.yaml
```

```
conditions:
- command: prometheus_query
  query: "topk(5, (increase(ceph_bluestore_slow_committed_kv_count{cluster_selector}[1h]) > 0))"
  operator: ">"
  threshold: 2
  entity_type: osd

- command: osd_performance_drop
  threshold: 20.0

- command: smart_media_errors
  error_limit: 3

- command: osd_ok_to_stop          # refuse unless Ceph says it's safe

action:
type: ceph_cli
command: |
  cephadm shell ceph osd ok-to-stop osd.{id} && \
  cephadm unit stop --name osd.{id} && \
  cephadm shell ceph osd out osd.{id} && \
  cephadm shell ceph health mute PG_DEGRADED 7d --sticky
```

ceph-companion operator at work - Keep It Simple, Stupid

```
Sep 21 10:00:00 kdrive04a-cephosd-23 systemd[1]: Starting ceph-companion.service
Sep 21 10:00:01 kdrive04a-cephosd-23 ceph-companion[2994395]: Done: no action needed.

Sep 21 10:01:00 kdrive04a-cephosd-23 systemd[1]: Starting ceph-companion.service
Sep 21 10:01:01 kdrive04a-cephosd-23 ceph-companion[2995069]: Done: no action needed.

Sep 21 10:02:00 kdrive04a-cephosd-23 systemd[1]: Starting ceph-companion.service
# condition1
query [kdrive04a]: topk(5, (increase(ceph_bluestore_slow_committed_kv_count{cluster='ae76dh5a-17dd-11f1-8320-6c83754eee8a'}[1h]) > 0))
    osd osd.500: 46
# condition2
Cluster: kdrive04a
    Performance (5m vs 1h), positive diff = traffic increased:
        IOPS: 6077.50 (5m) vs 6631.78 (1h) diff: -8.36%
        BW: 3.10 GB/s (5m) vs 4.41 GB/s (1h) diff: -29.73%
    Significant drop detected!
# condition3
    osd.500: /dev/sdl (SEAGATE ST24000NM007H): 44 media error(s) - SCSI: error-log=1 grown-defects=44 pending-defects=8
    Found 1 OSD(s) with SMART media errors
# condition4
    osd.500: safe

# All conditions passed
[Exec] cephadm shell ceph osd ok-to-stop osd.500 && \
cephadm unit stop --name osd.500 && \
cephadm shell ceph osd out osd.500 && \
logger -t ceph-companion -p kern.warning "OSD osd.500 stopped due to performance drop detected by ceph-operator" && \
cephadm shell ceph health mute PG_DEGRADED 7d --sticky

Done: 1 action(s) executed.
```

Part 2 — per-OSD analysis `ceph-companion analyze`

Data sources:

`osd_historic_ops` · `osd_perf_dump` · `osd_ops_in_flight` · `osd_counter_dump` · `prometheus`

- Can set thresholds and conditions

```
/etc/ceph-companion/  
├── config.yaml  
├── patterns  
│   ├── bluestore_slow_ops.yaml  
│   ├── HighWriteLatencyFinding.yaml  
│   ├── list_clients.yaml  
│   ├── list_events.yaml  
│   ├── osd_scrubs.yaml  
│   ├── slow_osd_op_op_applied.yaml  
│   └── ...  
└── rules  
    └── ...
```

Analysis, as YAML

```
# patterns/slow_HDD.yaml
name: slow_HDD
source: osd_historic_ops
patterns:
  - "event_duration('queued_for_pg', 'MOSDECSubOpWrite') > 2000"
  - "event_duration('txc_committed_kv', 'MOSDECSubOpWrite') > 2000"
  - "event_duration('osd_op', '_do_read') > 1000"
  - "event_duration('osd_op', 'kv_sync') > 1000"
  # ...
findings:
  - "High latency could indicate PG queueing / HDD I/O latency"
  - "If using HDDs, consider putting the OSDs down as an ultimate mitigation"
```

A support engineer's tool.

```

ceph-companion analyze --test osd_scrubs --osd-id 100
=====
Running tests on OSD.100
=====
[Pattern] osd[0].counters.successful_scrubs_ec_elapsed.avgtime > 7200 → ● True
Pattern evaluated successfully
Found osd[0].counters.successful_scrubs_ec_elapsed.avgtime = 74767.26850706

[Pattern] osd[0].counters.num_scrubs_started_ec > osd[0].counters.successful_scrubs_ec → ● True
Pattern evaluated successfully
Found osd[0].counters.num_scrubs_started_ec = 81
Found osd[0].counters.successful_scrubs_ec = 74

[Pattern] osd_scrub_dp_ec[0].counters.write_blocked_by_scrub > 1 → ✓ False
Pattern evaluated successfully
Found osd_scrub_dp_ec[0].counters.write_blocked_by_scrub = 0

[Pattern] osd_scrub_sh_ec[0].counters.write_blocked_by_scrub > 1 → ✓ False
Pattern evaluated successfully
Found osd_scrub_sh_ec[0].counters.write_blocked_by_scrub = 0

[Pattern] osd_scrub_dp_ec[0].counters.preemptions > 1 → ● True
Pattern evaluated successfully
Found osd_scrub_dp_ec[0].counters.preemptions = 6

[Pattern] osd_scrub_sh_ec[0].counters.preemptions > 1 → ● True
Pattern evaluated successfully
Found osd_scrub_sh_ec[0].counters.preemptions = 37

↓ Matched Patterns Triggered Findings:
[Finding] osd scrub (EC) average elapsed time per successful scrub: 74767.26850706 s
[Finding] osd scrub (replicated) average elapsed time per successful scrub: no data found s
[Finding] num_scrubs_started > successful_scrubs means scrubs are ongoing or unfinished, impacting latency
[Finding] scrub chunk counters (write_blocked_by_scrub, preemptions) show scrub impact on regular IO

```

Thank you!

ceph-companion is on GitHub: <https://github.com/Infomaniak/ceph-companion>

Our current concerns
- Latency: Ceph could be more latency-tolerant (e.g., for cross-DC deployments).
- Multi-site sync: Many orphans are left behind; compression breaks synchronization of multi-TB objects.
- Cluster expansion: Prevent the endless throughput race between new data ingress and PG remapping.
- Data integrity check: Scrubs consume too much I/O.

Backup slides

Side note

`diskprediction_local` mgr module

- Relies on `devicehealth` which doesn't work with our Seagate HDDs due to a `smartmontools` bug (<https://tracker.ceph.com/issues/74698>)
- Predicts only, no remediation.

Detecting a slow HDD - example 2

```
/etc/ceph-companion/rules/stop_sluggish_OSD_service.yaml
```

```
conditions:
- command: prometheus_query
  query: "topk(5, (increase(ceph_bluestore_slow_committed_kv_count{cluster_selector}[1h]) > 0))"
  operator: ">"
  threshold: 1
  entity_type: osd

- command: osd_performance_drop
  threshold: 20.0

- command: prometheus_query
  query: "topk(5, sort_desc(ceph_osd_apply_latency_ms{cluster_selector} + ceph_osd_commit_latency_ms{cluster_selector}))"
  operator: ">"
  threshold: 1500.0
  entity_type: osd

- command: osd_ok_to_stop      # refuse unless Ceph says it's safe

action:
...
```

ceph-companion operator - Notify if too many write_blocked_by_scrub

```
name: scrub_write_blocked_rate
description: |
  Notify when client writes blocked by scrub exceed a per-OSD rate threshold for the past 10 minutes.

enabled: true

conditions:
- command: prometheus_query
  query: "sort_desc(topk(20, sum by (ceph_daemon) (increase(ceph_osd_scrub_dp_ec_write_blocked_by_scrub{cluster_selector}[10m])
or increase(ceph_osd_scrub_sh_ec_write_blocked_by_scrub{cluster_selector}[10m])
or increase(ceph_osd_scrub_dp_repl_write_blocked_by_scrub{cluster_selector}[10m])
or increase(ceph_osd_scrub_sh_repl_write_blocked_by_scrub{cluster_selector}[10m]))))"
  operator: ">"
  threshold: 50
  entity_type: osd
  entity_labels: ["ceph_daemon"]

action:
  type: ceph_cli
  command: "true"
  dry_run_text: "WOULD NOTIFY: {value} writes blocked by scrub in the last 1h on {full_name}"
  max_items_per_run: 1
  send-notif: true
  notif-message: |
    **Scrub blocking client writes**

    **Cluster:** {cluster}
    **OSD:** {items}
    **Blocked writes (1h):** {value}
```




BLOCK 2 - Wrap Up: **Lunch Break & Networking**

→ See you back at 2 p.m.

14th October 2026, Zurich