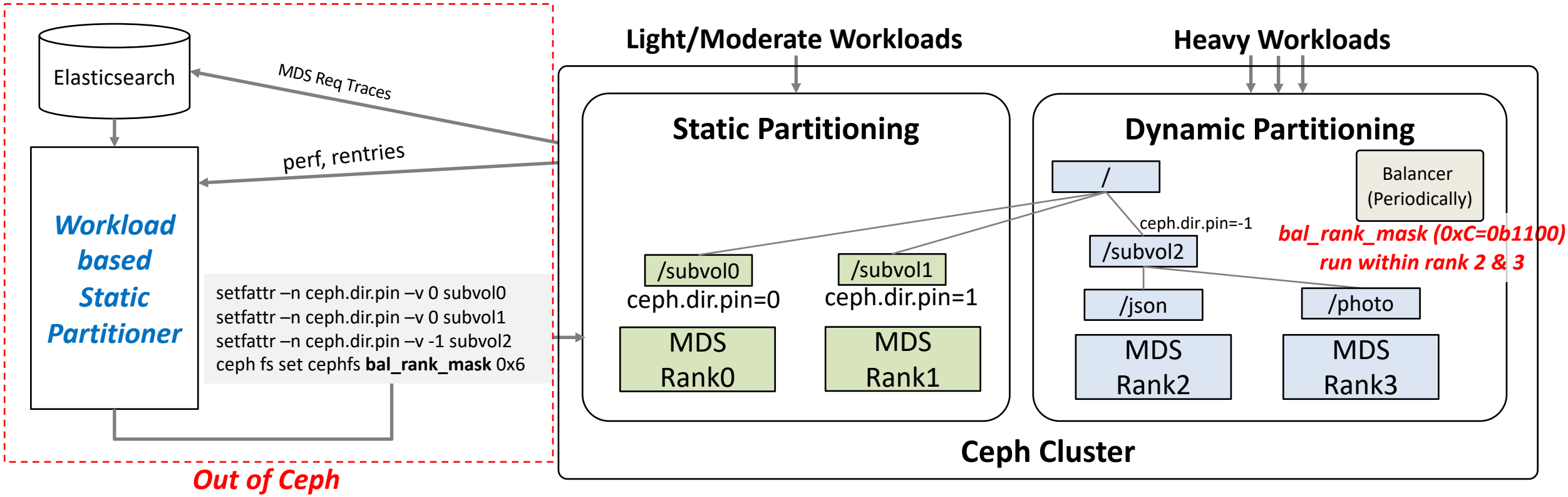
A New MDS Partitioner for CephFS

LINE

Yongseok Oh

Background: Workload Based Static Partitioner with bal_rank_mask



- This idea was presented at **Cephalocon2023**
- ***Workload based static partitioner*** pins subvolumes
  - Workload calculation based on working set, rentries, and performance
  - Rarely or manually conducted if loads are uneven or latencies get higher
  - Make subvolumes involving heavy workloads managed by MDS balancer with bal_rank_mask

Technical Issues with our in-house partitioner

- It is useful for performance
  - It distributes subdirs based on workloads compared to simple pinning
  - However, it is unavailable as open source
  - It needs to be revised and reimplemented for Ceph community
- bal_rank_mask needs to be enhanced
  - It can isolate unpinned large subtrees within certain MDS ranks from pinned subtrees
  - But, migrating large subdirs incur metadata movements
  - per subdir rank mask will be useful
    - e.g., setfattr –n ceph.dir.bal.mask –v 0xf /ceph/home/yongseok

A New MDS Partitioner

- MDS Subtree Partition Module in MGR

```
mds_partitioner enable $fs_name # enable partitioning for $fs_name
mds_partitioner disable $fs_name # disable partitioning for $fs_name
mds_partitioner status $fs_name # show mds_partitioner status

mds_partitioner analyze start $fs_name # analyze client workloads obtained from MDSs
mds_partitioner analyze status $fs_name # report analysis results and recommend optimal the number of MDSs
mds_partitioner analyze list $fs_name # show last N analysis results

mds_partitioner dir_path add $fs_name $dir_path # add a subdir path (e.g., /volumes/_nogroup/*)
mds_partitioner dir_path list $fs_name $dir_path. # show a subdir path list
mds_partitioner dir_path rm $fs_name $dir_path # remove a subdir path

mds_partitioner partition start $fs_name # start partitioning
mds_partitioner partition status $fs_name # show partitioning status
mds_partitioner partition abort $fs_name # abort current partitioning
mds_partitioner partition suspend $fs_name # suspend current partitioning
mds_partitioner partition resume $fs_name # resume current partitioning
```

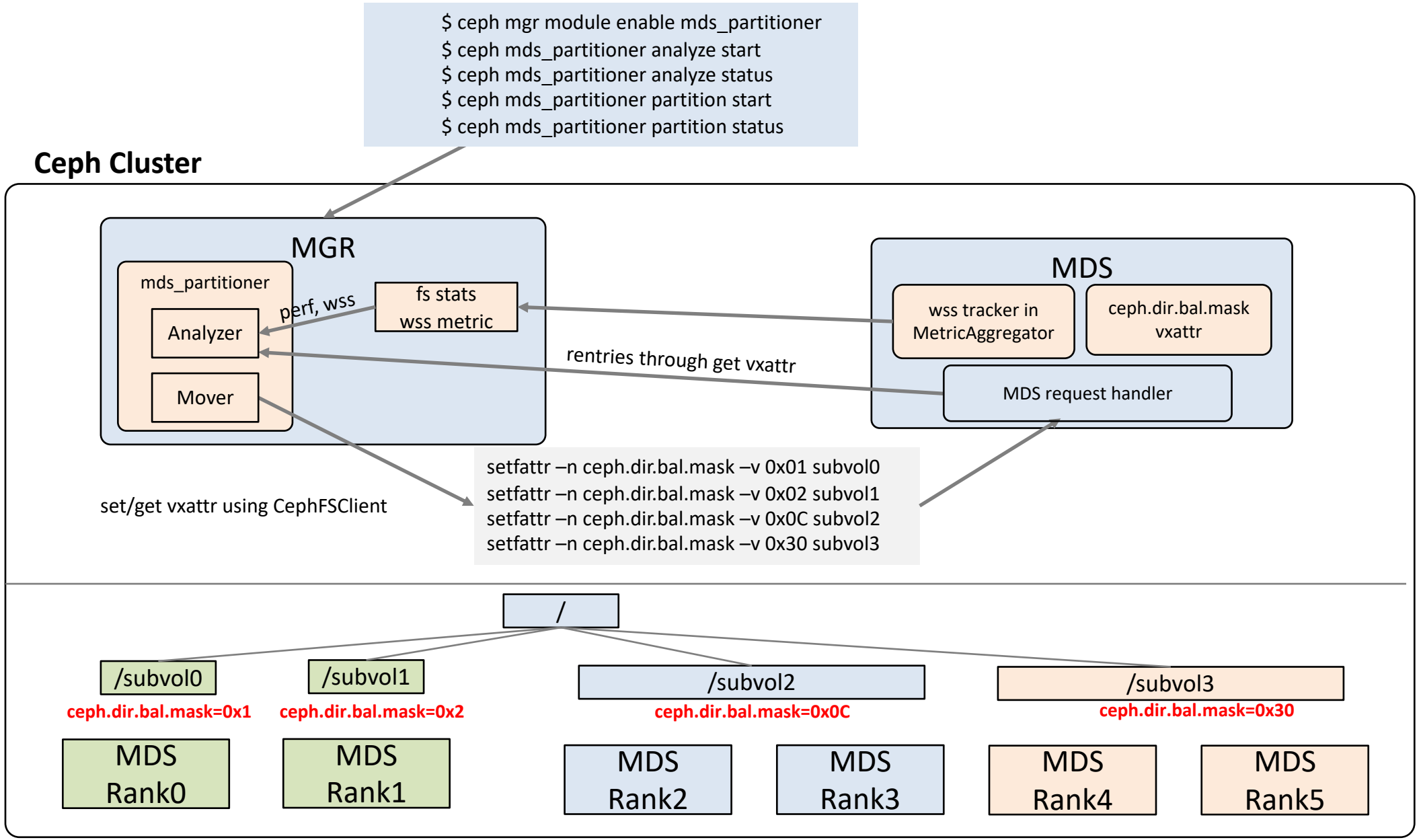
MDS Modifications

- ceph.dir.bal.mask vxattr
  - Distribute a subdir within certain ranks based on ceph.dir.bal.mask
    PR: https://github.com/ceph/ceph/pull/52373/files

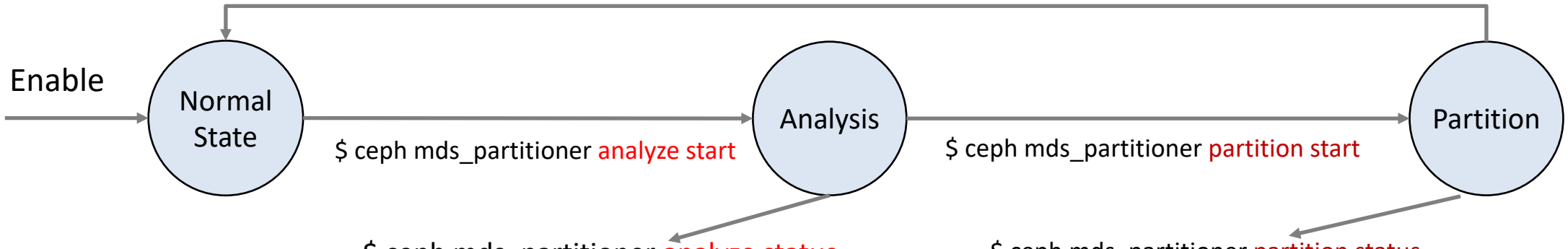
```
setfattr –n ceph.dir.bal.mask –v 0x3 /cephfs/home/yongseok
```

- Working Set Size (WSS) tracker in mds/MetricAggregator
  - How many metadata are accessed
  - WSS can be reported through `ceph fs perf stats`
- Minimize MDS slow requests
  - Exporting metadata must be controlled when occurring MDS slow requests
  - ceph fs set $fs_name noexport true/false
    - Similar to `ceph osd set nobackfill`

Overall Architecture



Example of Operation Flow



Name	wss	reqs	rentries	workload	Current Ranks	New Ranks	Migration Progress	Migration Status
Subvol1	1,000,000	1,203,030	50,000,000	339	0	0,1	20%	In-progress
Subvol2	700,000	500,000	1,000,000	137	1	2	0%	Ready
subvol3	100,000	5,000	5,000,000	21	2	2	100%	Done
subvol4	3,000	20,000	70,000	3	2	2	100%	Done
Total	1,803,000	1,728,030	56,070,000	500				

wss: working set size
reqs: requests
rentries: files + dirs

workload = (working_set_size / total_working_set *2
+ requests / total_requests * 2
+ rentries / total_rentries) * 100