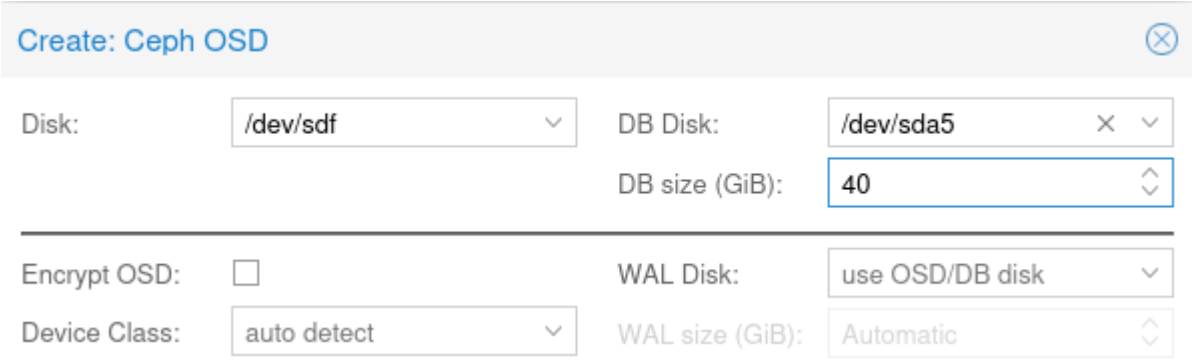
DB adding

NOTE: if partition is formatted as LVM PV, it is possible to choose this partition for DB from Proxmox UI when creating OSD:

```
pvcreate /dev/sda5
vgcreate ceph-db-40gb /dev/sda5
```



Adding journal DB/WAL partition

If an OSD needs to be shutdown for maintenance (i.e. adding new disc), please set `ceph osd set noout` to prevent unnecessary data balance.

Create parition on NVM drive

Reorganize existing NVM/SSD disc to make some free space. Create empty partition on free space.

Cut some space from zpool cache NVM partition

```
# remove cache partition from zpool
zpool list -v
zpool remove rpool /dev/nvme0n1p3
...
reorganize partition
...
blkid
zpool add rpool cache 277455ae-1bfa-41f6-8b89-fd362d35515e
```

Cut some space from zpool

Example how to cut some space from `nvmpool` zpool with spare temporary drive:

- we have 1 spare HDD which will be new Ceph OSD in future

- zpool doesn't support online shrinking.
- move `nvmpool` to spare HDD, to release NVM `nvmpool` partition.

```
zpool replace nvmpool nvme0n1p4 sdc
```

```
# zpool status nvmpool

  pool: nvmpool
 state: ONLINE
status: One or more devices is currently being resilvered.  The pool will
    continue to function, possibly in a degraded state.
action: Wait for the resilver to complete.
  scan: resilver in progress since Thu May  6 14:13:32 2021
    70.2G scanned at 249M/s, 21.0G issued at 74.4M/s, 70.2G total
    21.1G resilvered, 29.91% done, 00:11:17 to go
config:

    NAME            STATE     READ WRITE CKSUM
    nvmpool         ONLINE       0     0     0
      replacing-0   ONLINE       0     0     0
        nvme0n1p4   ONLINE       0     0     0
        sdc         ONLINE       0     0     0  (resilvering)
```

- wait for resilver
- reorganize partitions
- replace disks again

```
zpool replace nvmpool sdc nvme0n1p4
```

Replace OSD

```
ceph osd tree

ceph device ls-by-host pve5
DEVICE                      DEV  DAEMONS  EXPECTED FAILURE
TOSHIBA_HDWD120_30HN40HAS  sdc  osd.2

### Switch OFF OSD. Ceph should rebalance data from replicas when OSD is
switched off directly
ceph osd out X
## or better use lines below:
# this is optional for safety for small clusters instead of using ceph out
osd.2
ceph osd reweight osd.X 0
# wait for data migration away from osd.X
watch 'ceph -s; ceph osd df tree'
# Remove OSD
ceph osd out X

ceph osd safe-to-destroy osd.X
```

```
ceph osd down X
systemctl stop ceph-osd@X.service
ceph osd destroy X

#pveceph osd destroy X


# to remove partition table, boot sector and any OSD leftover:
ceph-volume lvm zap /dev/sdX --destroy

## it is not possible to specify DB partition with pveceph command (read
begining of page):
# pveceph osd create /dev/sdc --db_dev /dev/nvme0n1p3
## it requires whole device as db dev with LVM and will create new LVM on
free space, i.e.
# pveceph osd create /dev/sdc --db_dev /dev/nvme0n1 --db_size 32G
## so direct ceph command will be used:

# Prevent backfilling when new osd will be added
ceph osd set nobackfill

### Create OSD:
ceph-volume lvm create --osd-id X --bluestore --data /dev/sdc --block.db
/dev/nvme0n1p3
# or split above into two step:
ceph-volume lvm prepare --bluestore --data /dev/sdX --block.db
/dev/nvme0n1pX
ceph-volume lvm activate --bluestore X e56ecc53-826d-40b0-a647-xxxxxxxxxxxx
# also possible: ceph-volume lvm activate --all

## DRAFTS:
#ceph-volume lvm create --cluster-fsid 321bdc94-39a5-460a-834f-6e617fdd6c66
--data /dev/sdc --block.db /dev/nvme0n1p3
#ceph-volume lvm activate --bluestore <osd id> <osd fsid>
```

Verify:

```
ls -l /var/lib/ceph/osd/ceph-X/
lrwxrwxrwx 1 ceph ceph  93 sty 28 17:59 block ->
/dev/ceph-16a69325-6fb3-4d09-84ee-c053c01f410f/osd-block-e56ecc53-826d-40b0-
a647-5f1a1fc8800e
lrwxrwxrwx 1 ceph ceph  14 sty 28 17:59 block.db -> /dev/nvme0n1p3

ceph daemon osd.X perf dump | jq '.bluefs'
{
  "gift_bytes": 0,
  "reclaim_bytes": 0,
  "db_total_bytes": 42949664768,    --> 39,99GB  (40GB partition created)
  "db_used_bytes": 1452269568,      -->  1,35GB
  "wal_total_bytes": 0,
  "wal_used_bytes": 0,
```

```
...
# OR
  "db_total_bytes": 4294959104,      --> 3,9GB (4GB partition)
  "db_used_bytes": 66052096,         -->

ceph device ls
```

And restore backfilling:

```
ceph osd unset nobackfill
```

Check benefits:

- Observe better latency on OSD with NVM/SSD:

  ```
  watch ceph osd perf
  ```

- check `iotop` output. Now `[bstore_kv_sync]` should take less time.