PostgreSQL中文社区

2021 PostgreSQL China Conference

主办：PostgreSQL 中文社区

第 11 届 PostgreSQL 中国技术大会

开源论道 × 数据驱动 × 共建数字化未来

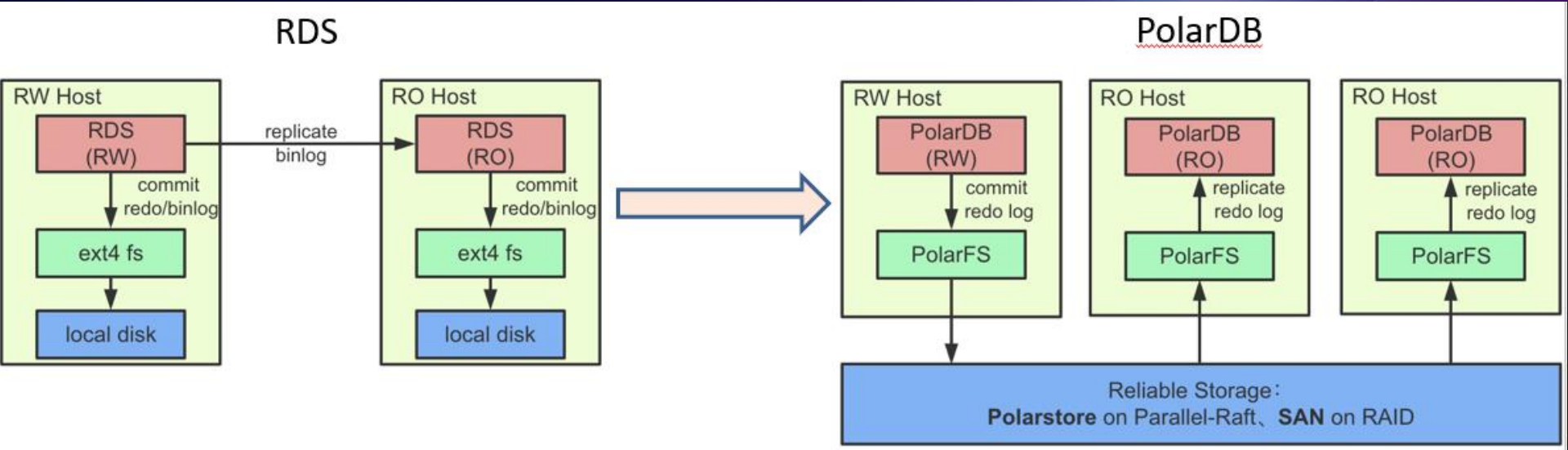
PolarDB存储原理与实践

阿里云数据库产品事业部-PolarDB基础设施
朱元(圆珠)

分享内容

- PolarDB存储原理简介
  - PolarDB存储的基本工作原理
- PolarDB存储实践
  - PolarFS的部署
  - 基于SAN存储的部署
  - 基于NBD存储的部署
  - 基于阿里云共享存储的部署

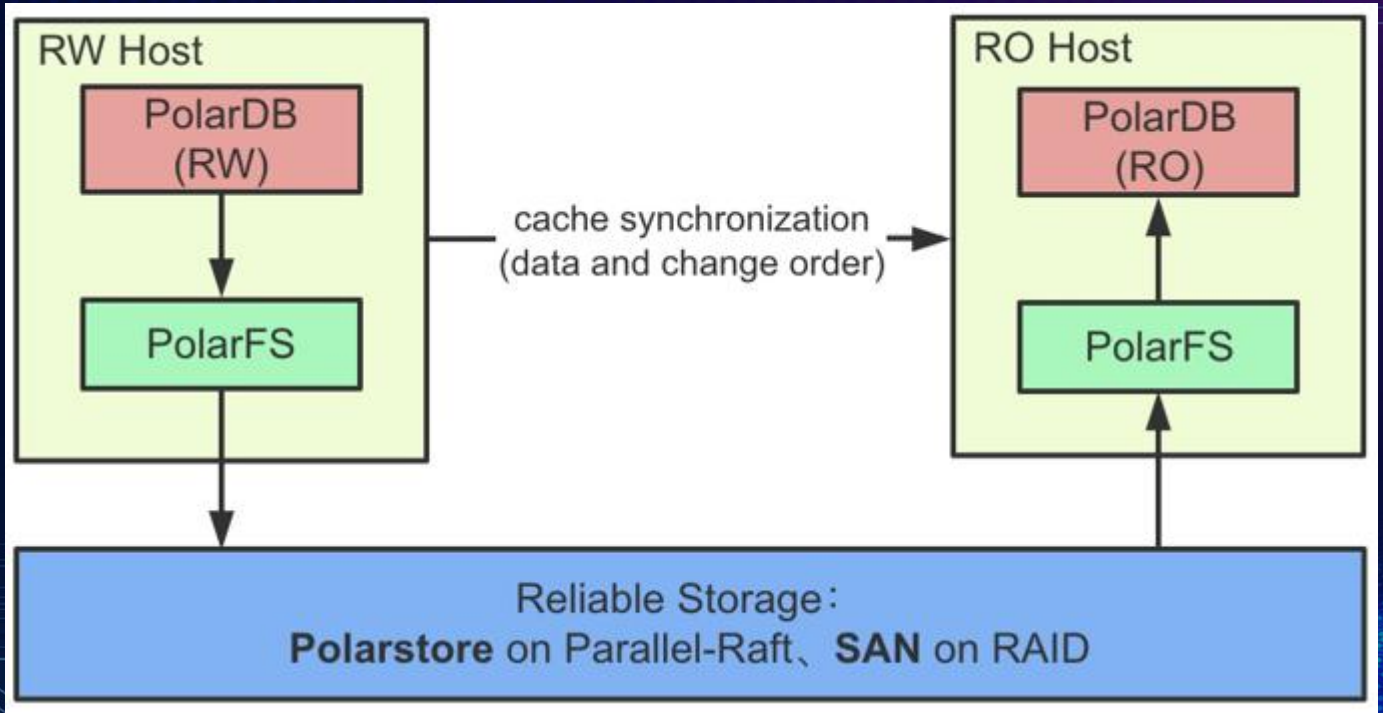
基于共享存储的PolarDB



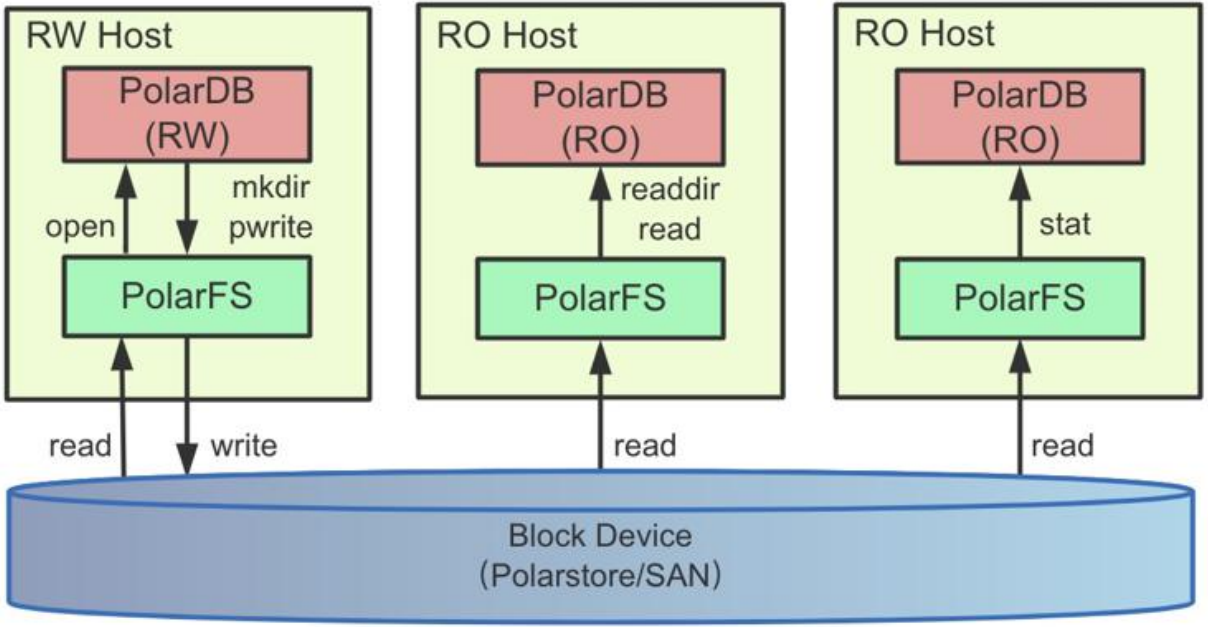
- 存储成本正比于节点数
- 存储预分配，节点独占本地存储

- 存储和计算层分离，存储成本和节点数无关
- 存储分布式共享，无需预分配可动态扩容
- 存储层独立实现存储可靠性和复制功能（副本冗余、快照）

共享存储架构下的计算层修改：缓存同步



PolarFS：PolarDB的文件系统



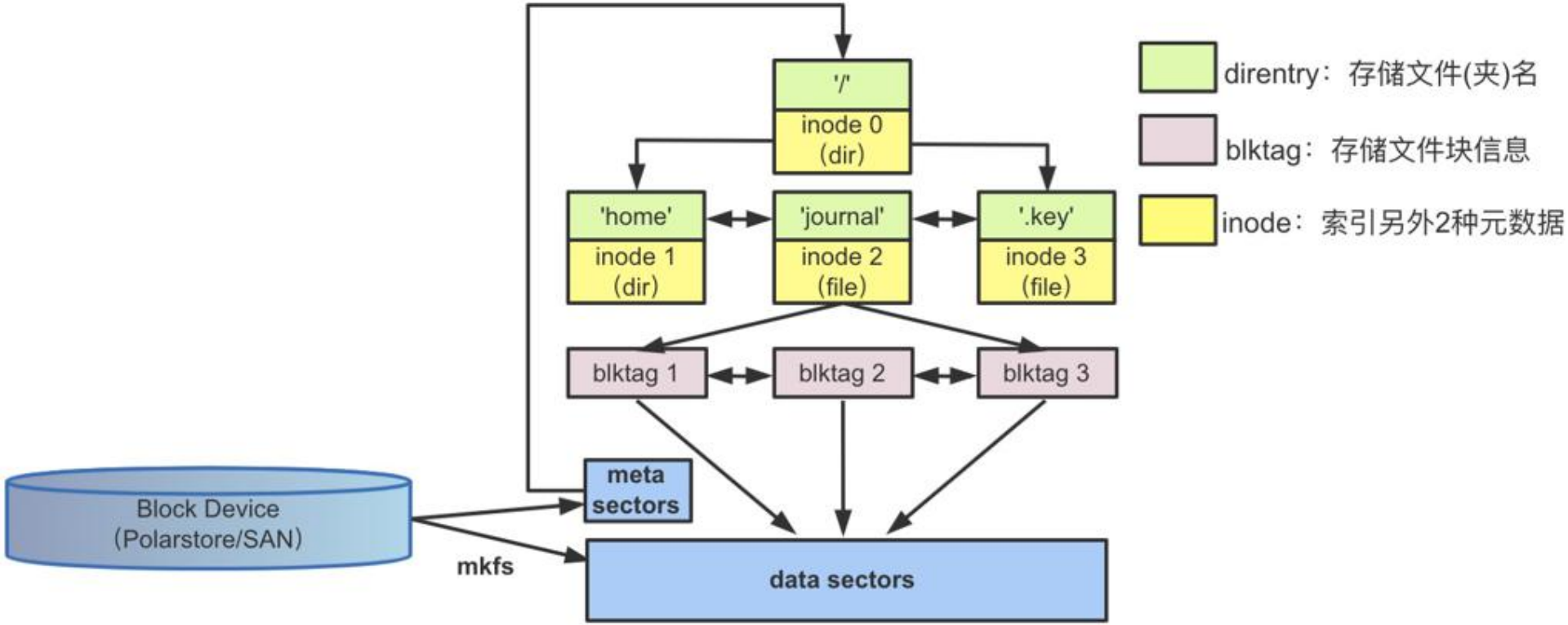
职责：

- 提供文件夹、文件语义类Posix接口(mkdir/readdir/stat/read ...等)
- 支持分布式共享块设备一写多读

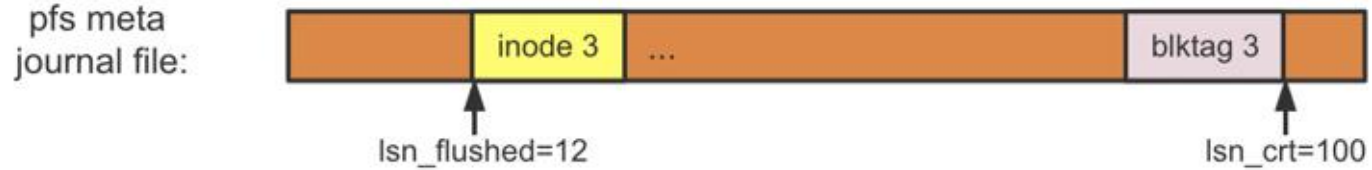
定位：

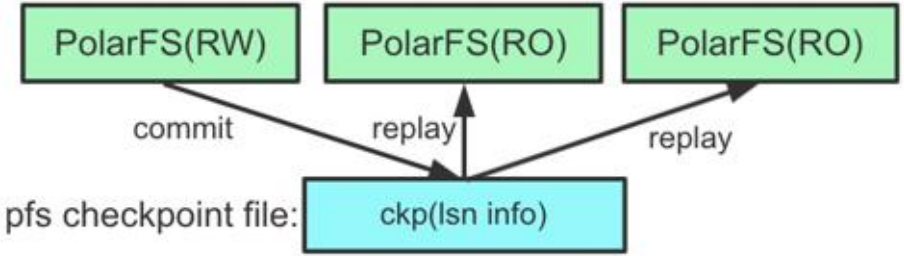
- 用户态、高性能

PolarFS的元数据结构



PolarFS的元数据分布式同步





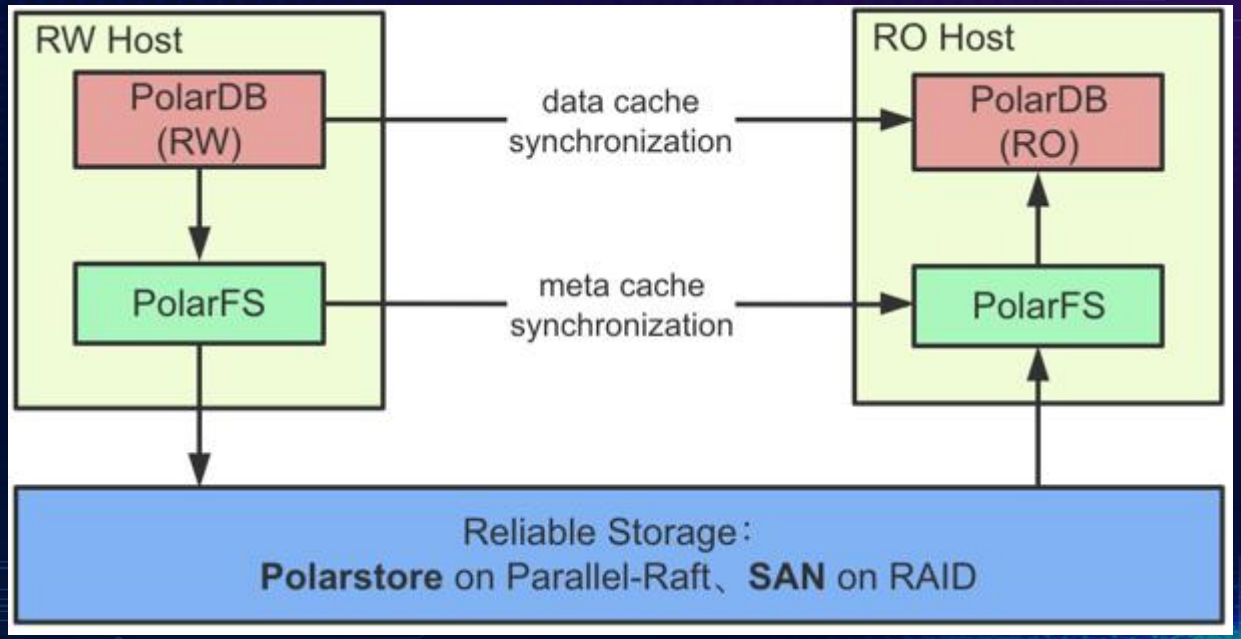
RW:

WAL+checkpoint

RO:

每次api调用都读checkpoint文件，尝试replay元信息修改

PolarDB和PolarFS的缓存同步分工



存储实践

- PolarDB存储原理简介
  - PolarDB存储的基本工作原理

- PolarDB存储实践
  - **PolarFS的部署**
  - SAN存储的部署
  - NBD存储的部署
  - **基于阿里云共享存储的部署**

PolarFS的编译和安装

1.编译

- https://github.com/ApsaraDB/polardb-file-system 下载源码
- 准备需要的第三方库（libaio, libzlog）
- ./autobuild.sh

2.安装

- sudo ./install.sh 会生成二进制工具pfs（支持以类似busy-box的形态执行gnu-util的部分文件系统命令）和文件系统服务pfs_daemon
- sudo pfs mkfs 进行磁盘格式化
- 格式化后可以使用ls， mkdir等命令进行操作（不支持相对路径）

PolarFS bash工具使用示例

```
$lsblk |grep nvme|head -8
nvme10n1 259:5    0   1.8T  0 disk
nvme11n1 259:4    0   1.8T  0 disk
nvme0n1  259:8    0 349.3G  0 disk
nvme1n1  259:7    0 349.3G  0 disk
nvme2n1  259:9    0   1.8T  0 disk
nvme3n1  259:10   0   1.8T  0 disk
nvme4n1  259:6    0   1.8T  0 disk
nvme5n1  259:1    0   1.8T  0 disk

[yuanzhu.zy@e03g04233.eu6sqa /home/yuanzhu.zy]
$sudo pfs -C disk mkfs -f nvme5n1  1>/dev/null 2>/dev/null

[yuanzhu.zy@e03g04233.eu6sqa /home/yuanzhu.zy]
$sudo pfs -C disk ls /nvme5n1/
  File  1     4194304           Tue Oct 26 16:16:36 2021  .pfs-paxos
  File  1     1073741824        Tue Oct 26 16:16:37 2021  .pfs-journal
total 2105344 (unit: 512Bytes)

[yuanzhu.zy@e03g04233.eu6sqa /home/yuanzhu.zy]
$sudo pfs -C disk mkdir /nvme5n1/test

[yuanzhu.zy@e03g04233.eu6sqa /home/yuanzhu.zy]
$sudo pfs -C disk ls /nvme5n1/
  File  1     4194304           Tue Oct 26 16:16:36 2021  .pfs-paxos
  File  1     1073741824        Tue Oct 26 16:16:37 2021  .pfs-journal
   Dir  1     0                 Tue Oct 26 16:16:55 2021  test
total 2105344 (unit: 512Bytes)
```

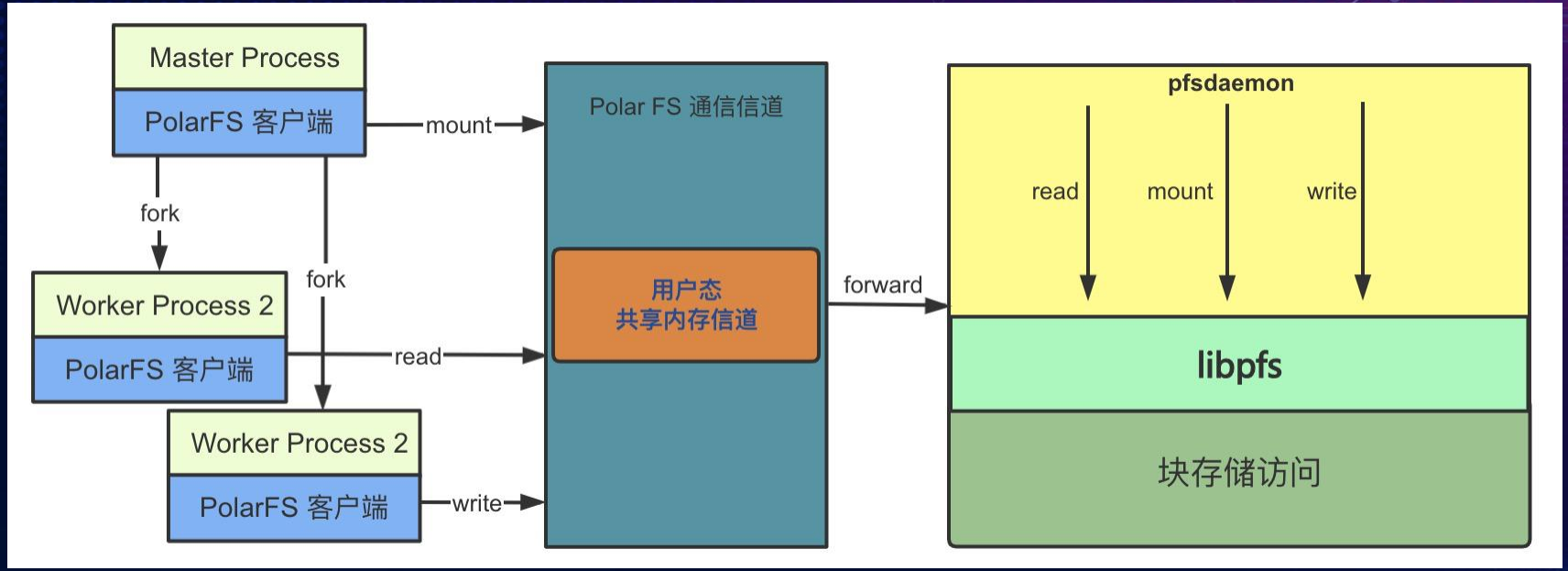
PolarDB postgresql 实际文件示例

```
[root@r03.dbm-01 ~]$pfs -C disk ls /mapper_360050767088080a2680000000000684f/data/
   Dir  1     640               Tue Oct 19 11:52:50 2021  base
   Dir  1     9344              Tue Oct 19 11:52:51 2021  global
   Dir  1     0                 Tue Oct 19 11:52:51 2021  pg_tblspc
   Dir  1     640               Sat Oct 23 13:10:32 2021  pg_wal
   Dir  1     640               Tue Oct 26 05:45:11 2021  pg_logindex
   Dir  1     0                 Tue Oct 19 11:52:54 2021  pg_twophase
   Dir  1     896               Tue Oct 26 07:35:13 2021  pg_xact
   Dir  1     0                 Tue Oct 19 11:52:55 2021  pg_commit_ts
   Dir  1     256               Tue Oct 19 11:52:55 2021  pg_multixact
   Dir  1     256               Tue Oct 26 14:25:18 2021  pg_csnlog
   Dir  1     256               Tue Oct 19 11:52:55 2021  polar_dma
   Dir  1     128               Tue Oct 19 11:53:01 2021  polar_fullpage
  File  1     32                Tue Oct 19 11:52:59 2021  RWID
   Dir  1     128               Tue Oct 19 11:53:11 2021  pg_replslot
total 8192 (unit: 512Bytes)
[root@r03.dbm-01 ~]$pfs -C disk ls /mapper_360050767088080a2680000000000684f/data/base
   Dir  1     83840             Mon Oct 25 20:26:53 2021  16328
   Dir  1     57344             Tue Oct 19 11:52:48 2021  16327
   Dir  1     57856             Tue Oct 19 11:56:56 2021  16330
   Dir  1     57344             Tue Oct 19 11:52:50 2021  1
   Dir  1     57344             Tue Oct 19 11:52:51 2021  16329
total 0 (unit: 512Bytes)
[root@r03.dbm-01 ~]$pfs -C disk ls /mapper_360050767088080a2680000000000684f/data/base/16327
  File  1     16384             Tue Oct 19 11:52:47 2021  2656
  File  1     8192              Tue Oct 19 11:52:47 2021  13824_vm
  File  1     16384             Tue Oct 19 11:52:47 2021  2699
  File  1     16384             Tue Oct 19 11:52:47 2021  2661
  File  1     24576             Tue Oct 19 11:52:47 2021  2616
  File  1     24576             Tue Oct 19 11:52:47 2021  3603_fsm
  File  1     8192              Tue Oct 19 11:52:47 2021  2600_vm
  File  1     8192              Tue Oct 19 11:52:47 2021  2753_vm
  File  1     8192              Tue Oct 19 11:52:47 2021  8895_vm
  File  1     8192              Tue Oct 19 11:52:47 2021  15724
  File  1     8192              Tue Oct 19 11:52:47 2021  16099
  File  1     65536             Tue Oct 19 11:52:47 2021  2704
  File  1     32768             Tue Oct 19 11:52:47 2021  2662
  File  1     0                 Tue Oct 19 11:52:47 2021  8899
  File  1     0                 Tue Oct 19 11:52:47 2021  15725
  File  1     32768             Tue Oct 19 11:52:47 2021  2757
```

PolarFS bash工具命令支持一览

```
[root@e03g04233.eu6sqa /home/yuanzhu.zy]
#pfs
Usage: pfs [-H hostid] [-C|--cluster=clustername] [-t pfsd_timeout] <command> [options] pbdpaths
pfs has following commands
  help        show help info
  tree        list all files in this dir and its subdirs
  ls          list all direntries in this directory
  rmdir       remove an empty directory
  truncate    truncate file
  tail        read file tail incessantly
  fallocate   allocate block for file
  write       write file
  read        read file
  stat        show file info
  touch       create file
  chunk       chunk operations
  cp          copy file or dir
  du          display disk usage statistics
  dumpfs      dump pbd info or data
  dumple      dump log entries
  flushlog    flush log to pbd
  fscp        transfer pfs pbd data
  fstrim      trim filesystem
  info        show pfs meta info
  map         dump a file's block index
  mkdir       create dir
  growfs      grow filesystem
  mkfs        make filesystem
  rename      rename file
  rm          remove file or dir
  usedinfo    dump pbd info or data
```

PolarFS 部署形态




```
root     111110 111093 26 Oct19 ?        1-22:13:34 /usr/local/polarstore/pfsd/bin
/../bin/pfsdaemon -f -w 8 -s 20 -i 8192 -f -p mapper_360050767088080a2680000000000
684f -e 1617 -c /usr/local/polarstore/pfsd/bin/../conf/pfsd_logger.conf
```

期待您的参与和建议

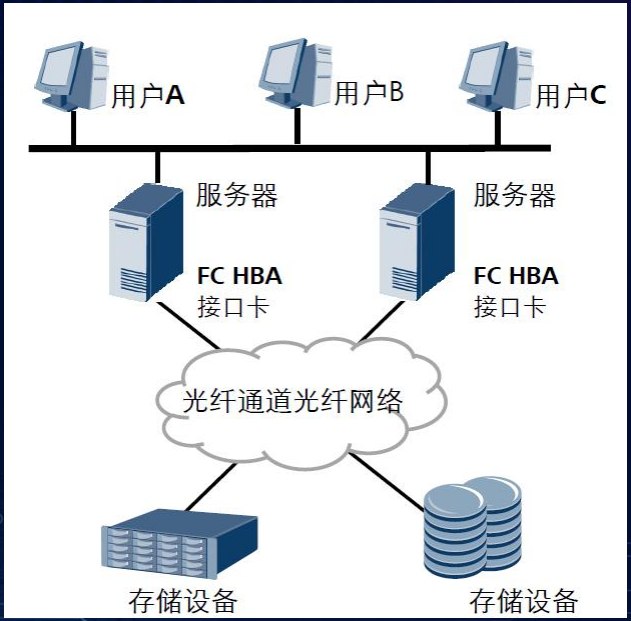
- https://github.com/ApsaraDB/polardb-file-system 开源项目地址
- https://github.com/ApsaraDB/polardb-file-system/releases/download/pfsd4pg-release-1.2.41-20211018/t-pfsd-opensource-1.2.41-1.el7.x86_64.rpm 开源项目预编译rpm包
- https://github.com/ApsaraDB/polardb-file-system/blob/master/Readme-CN.md 安装部署中文文档

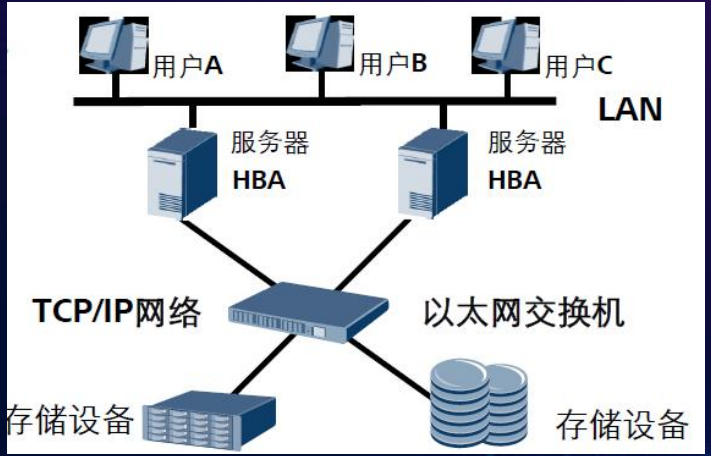
存储实践（二）

- PolarDB存储原理简介
  - PolarDB存储的基本工作原理

- PolarDB存储实践
  - PolarFS的部署
  - SAN存储的部署
  - NBD存储的部署
  - 基于阿里云共享存储的部署

基于光纤或以太交换网络的SAN





SAN on linux 部署

1. SAN网络初始化

- fc san由hba卡自动完成
- ip san需要手工建立网络连接

```
iscsiadm -m discovery -p ${target_ip} -t st
iscsiadm -m node -l -p ${target_ip}
```

2. 在计算主机上扫描linux块设备

```
echo '- - -' > /sys/class/fc_host/$host/scan
```

块设备注册与发现完成

```
brw-rw---- 1 root disk 129, 224 Sep 10 14:40 /dev/sdfc
brw-rw---- 1 root disk 129, 240 Sep 10 14:40 /dev/sdfd
brw-rw---- 1 root disk 130,   0 Sep 10 14:40 /dev/sdfe
brw-rw---- 1 root disk 130,  16 Sep 10 14:40 /dev/sdff
brw-rw---- 1 root disk 130,  32 Oct 22 10:08 /dev/sdfg
brw-rw---- 1 root disk 130,  48 Oct 22 10:08 /dev/sdfh
brw-rw---- 1 root disk 130,  64 Oct 22 10:09 /dev/sdfi
brw-rw---- 1 root disk 130,  80 Oct 22 10:09 /dev/sdfj
brw-rw---- 1 root disk 130,  96 Sep 10 14:40 /dev/sdfk
brw-rw---- 1 root disk 130, 112 Sep 10 14:40 /dev/sdfl
brw-rw---- 1 root disk 130, 128 Sep 10 14:40 /dev/sdfm
brw-rw---- 1 root disk 130, 144 Sep 10 14:40 /dev/sdfn
brw-rw---- 1 root disk 130, 160 Sep 10 14:40 /dev/sdfo
brw-rw---- 1 root disk 130, 176 Sep 10 14:40 /dev/sdfp
brw-rw---- 1 root disk 130, 192 Sep 10 14:40 /dev/sdfq
brw-rw---- 1 root disk 130, 208 Sep 10 14:40 /dev/sdfr
brw-rw---- 1 root disk 130, 224 Sep 10 14:40 /dev/sdfs
brw-rw---- 1 root disk 130, 240 Sep 10 14:40 /dev/sdft
```

管理存储访问

1. 安装

yum –y install device-mapper device-mapper-multipath

2. 可以通过dmsetup 把多块san物理盘合并成一个linux逻辑块设备来使用。Polarfs支持以10GB为单位(chunk)管理设备，不能被10GB整除的剩余空间部分无法使用。

3. 可以在/etc/multipath.conf中配置通过存储网络访问磁盘的路径负载均衡。

存储示例

1. 磁盘合并

```
ls: cannot access /dev/dm-: No such file or directory
[root@r03.dbm-01 ~]$ll /dev/mapper/lvid-test234
lrwxrwxrwx 1 root root 8 Oct 22 11:40 /dev/mapper/lvid-test234 -> ../dm-25
[root@r03.dbm-01 ~]$dmsetup table  | grep linear
lvid-test234: 0 503316480 linear 253:21 0
lvid-test234: 503316480 545259520 linear 253:7 0
[root@r03.dbm-01 ~]$multipath -ll |more
```

2. 访问磁盘路径负载均衡

```
360050767088080a2680000000000684f dm-7 ALIBABA ,MCS
size=260G features='1 queue_if_no_path' hwhandler='0' wp=rw
|-+- policy='round-robin 0' prio=50 status=active
| |- 14:0:2:1  sdaf    65:240  active ready running
| |- 15:0:5:1  sdev    129:112 active ready running
| |- 14:0:7:1  sdey    129:160 active ready running
| `- 15:0:1:1  sdax    67:16   active ready running
`-+- policy='round-robin 0' prio=10 status=enabled
  |- 14:0:3:1  sdbu    68:128  active ready running
  |- 15:0:3:1  sdcx    70:80   active ready running
  |- 14:0:6:1  sddy    128:0   active ready running
  `- 15:0:7:1  sdgh    131:208 active ready running
```

扩容流程

1. 在SAN存储上利用厂商软件扩容，或者把新增的盘合并到旧盘

2. 每台主机上的块设备上感知扩容结果
echo 1 > /sys/block/sdx/device/rescan

3. 文件系统格式化扩容区域
pfs -C disk growfs -o 1 -n 3 sdx 在RW主机上执行

4. 用户态文件系统感知扩容区域
在每一个数据库中调用 pfsd_growfs(“sdx”)， 先RO最后RW。

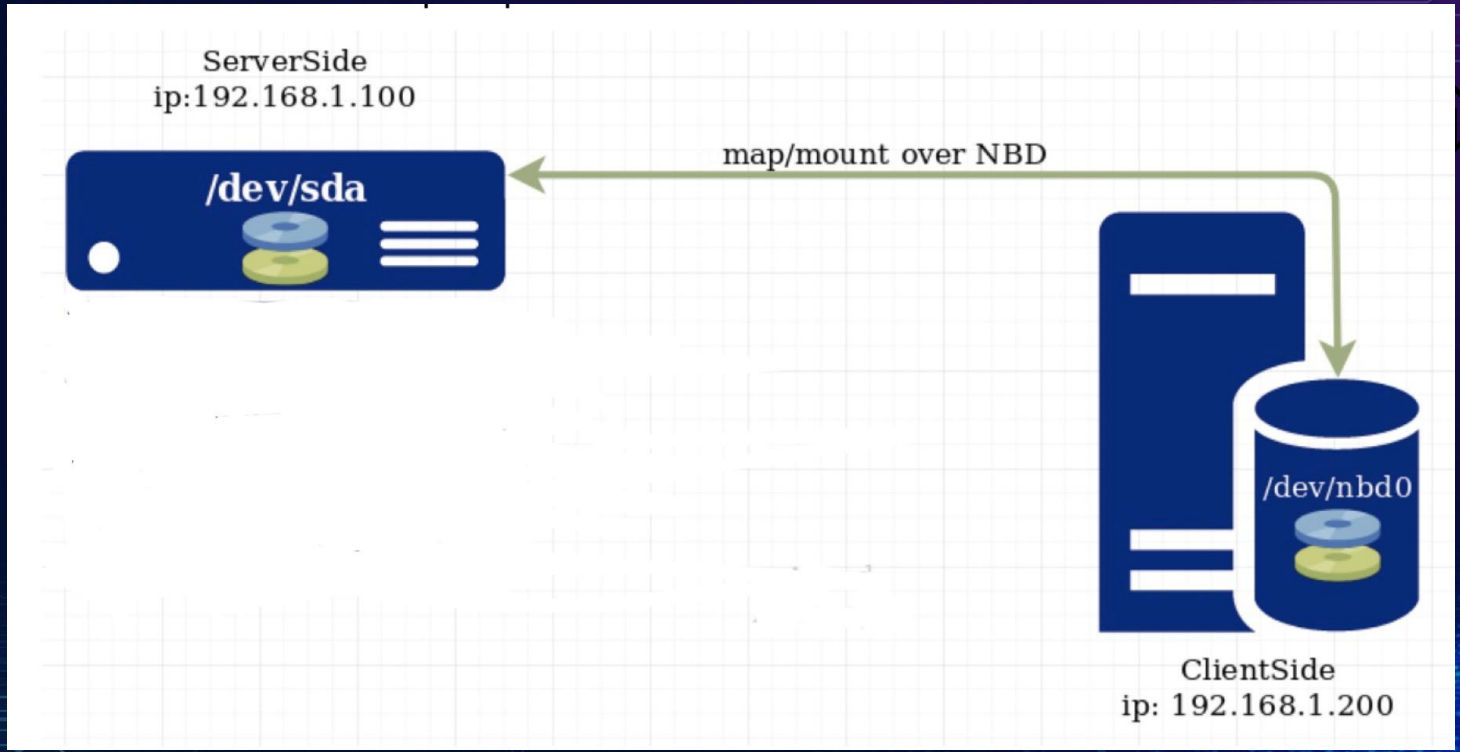
已有部分用户案例

- 中国人寿保险公司
- 深圳边检
- …

存储实践（三）

- PolarDB存储原理简介
  - PolarDB存储的基本工作原理
- PolarDB存储实践
  - PolarFS的部署
  - SAN存储的部署
  - NBD存储的部署
  - 基于阿里云共享存储的部署

NBD的概念


ServerSide
ip:192.168.1.100
/dev/sda
map/mount over NBD
/dev/nbd0
ClientSide
ip: 192.168.1.200

NBD的服务端部署

yum install nbd

服务端部署：

拉起nbd服务即可，按照同步方式(sync/flush=true)配置在某个端口(1921)上监听对某个块设备(vdb)的访问。

```
root     15018 13754  0 10月15 ?      00:00:00 nbd-server -C /root/nbd.conf
[root@iZbp1eo3op9s5gxnvc7aokZ ~]# cat /root/nbd.conf
# This is a comment
[generic]
    # The [generic] section is required, even if nothing is specified
    # there.
    # When either of these options are specified, nbd-server drops
    # privileges to the given user and group after opening ports, but
    # _before_ opening files.
    #user = nbd
    #group = nbd
    listenaddr = 0.0.0.0
    port = 1921
[export1]
    exportname = /dev/vdb
    readonly = false
    multifile = false
    copyonwrite = false
    flush = true
    fua = true
    sync = true
[export2]
```

NBD的客户端部署（一）

源码在 drivers/block/**nbd**.c

编译内核依赖和组件：

```
make menuconfig # Device Driver -> Block devices -> Set “M” On “Network block device support”
make prepare && make modules_prepare && make scripts
make CONFIG_BLK_DEV_NBD=m M=drivers/block
```

看一下是否正常生成了驱动

```
modinfo drivers/block/nbd.ko
```

拷贝，生成依赖并插入内核

```
cp drivers/block/nbd.ko /lib/modules/$(uname -r)/kernel/drivers/block
depmod -a
modprobe nbd
```

此时在/dev/下会生成/dev/nbdxx设备，根据服务端的配置把远程的块设备映射到本地的某个nbd设备

NBD的客户端部署（二）

```
[root@iZbp1eo3op9s5gxnvc7aokZ ~]# ll /dev/nbd*
brw-rw---- 1 root disk 43,  0 10月  16 10:00 /dev/nbd0
brw-rw---- 1 root disk 43,  1 10月  16 10:00 /dev/nbd1
brw-rw---- 1 root disk 43, 10 10月  16 10:00 /dev/nbd10
brw-rw---- 1 root disk 43, 11 10月  16 10:00 /dev/nbd11
brw-rw---- 1 root disk 43, 12 10月  16 10:00 /dev/nbd12
brw-rw---- 1 root disk 43, 13 10月  16 10:00 /dev/nbd13
brw-rw---- 1 root disk 43, 14 10月  16 10:00 /dev/nbd14
brw-rw---- 1 root disk 43, 15 10月  16 10:00 /dev/nbd15
brw-rw---- 1 root disk 43,  2 10月  16 10:00 /dev/nbd2
brw-rw---- 1 root disk 43,  3 10月  16 10:00 /dev/nbd3
brw-rw---- 1 root disk 43,  4 10月  16 10:00 /dev/nbd4
brw-rw---- 1 root disk 43,  5 10月  16 10:00 /dev/nbd5
brw-rw---- 1 root disk 43,  6 10月  16 10:00 /dev/nbd6
brw-rw---- 1 root disk 43,  7 10月  16 10:00 /dev/nbd7
```

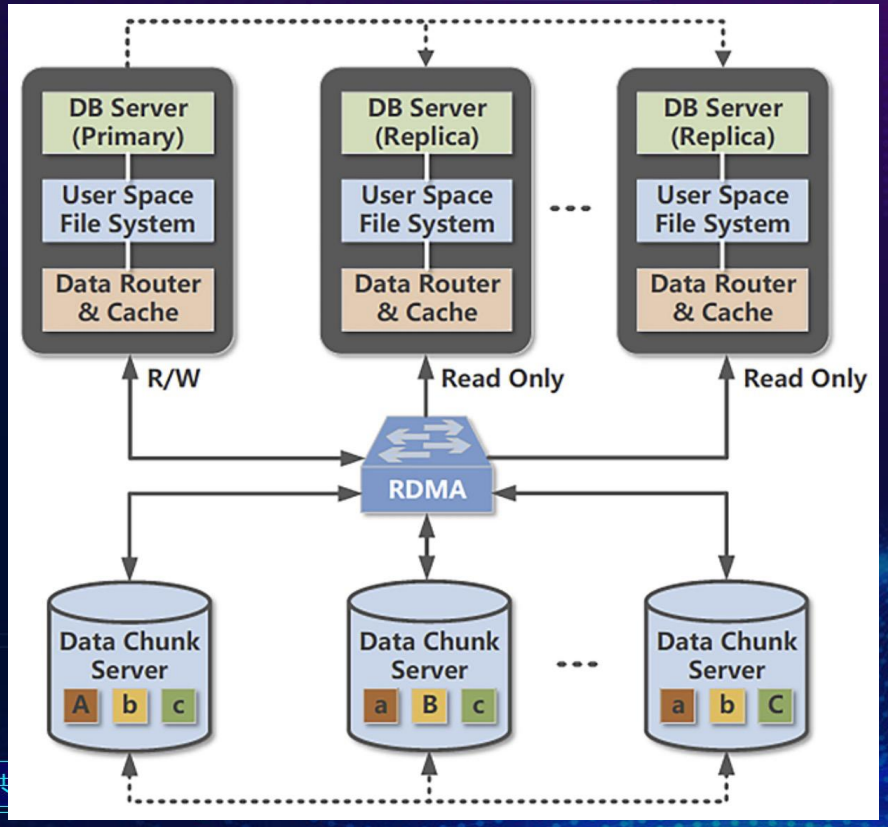
```
root     29130     1  0 10月15 ?      00:02:14 nbd-client 172.17.164.66 1921 -N ex
port1 /dev/nbd0
root     29132     2  0 10月15 ?      00:02:32 [nbd0]
root     29148     1  0 10月15 ?      00:00:00 nbd-client 172.17.164.66 1921 -N ex
port2 /dev/nbd1
root     29150     2  0 10月15 ?      00:00:00 [nbd1]
[root@iZbp1eo3op9s5gxnvc7aolZ ~]# lsblk
NAME   MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
vda    253:0    0  100G  0 disk
└─vda1 253:1    0  100G  0 part /
nbd0    43:0    0  100G  0 disk
nbd1    43:1    0  100G  0 disk
```

存储实践（三）

- PolarDB存储原理简介
  - PolarDB存储的基本工作原理
- PolarDB存储实践
  - PolarFS的部署
  - SAN存储的部署
  - NBD存储的部署
  - 基于阿里云共享存储的部署

云上共享存储：Polarstore

阿里云控制台直接购买PolarDB：
PolarDB-M(5.6/5.7/8.0)
PolarDB-PG(11)
PolarDB-Oracle兼容



THANKS

谢谢观看