ByteDance字节跳动

New Features

字节跳动基础架构部内核与虚拟化团队

祝嘉

实践与探索——New Features

01

Failover

02

EROFS Shared Domain

03

Daemonless

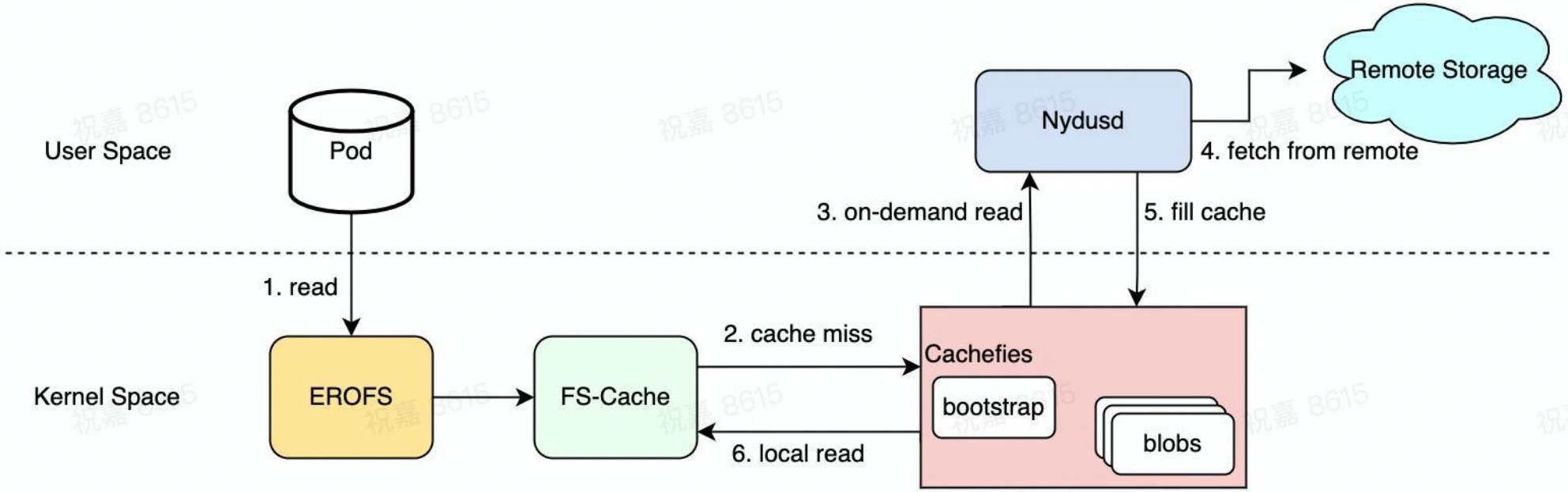
稳定性提升：User Daemon Failover

背景：

- 用户态进程crash后，IO请求将返回错误，影响容器中运行的业务

预期目标：容器不感知用户态进程crash/restart

- In-Flight IO不丢失
- User Daemon crash后 ~ 重启前 产生的IO不丢失
- IO在有限时间内恢复



稳定性提升：User Daemon Failover

方案设计

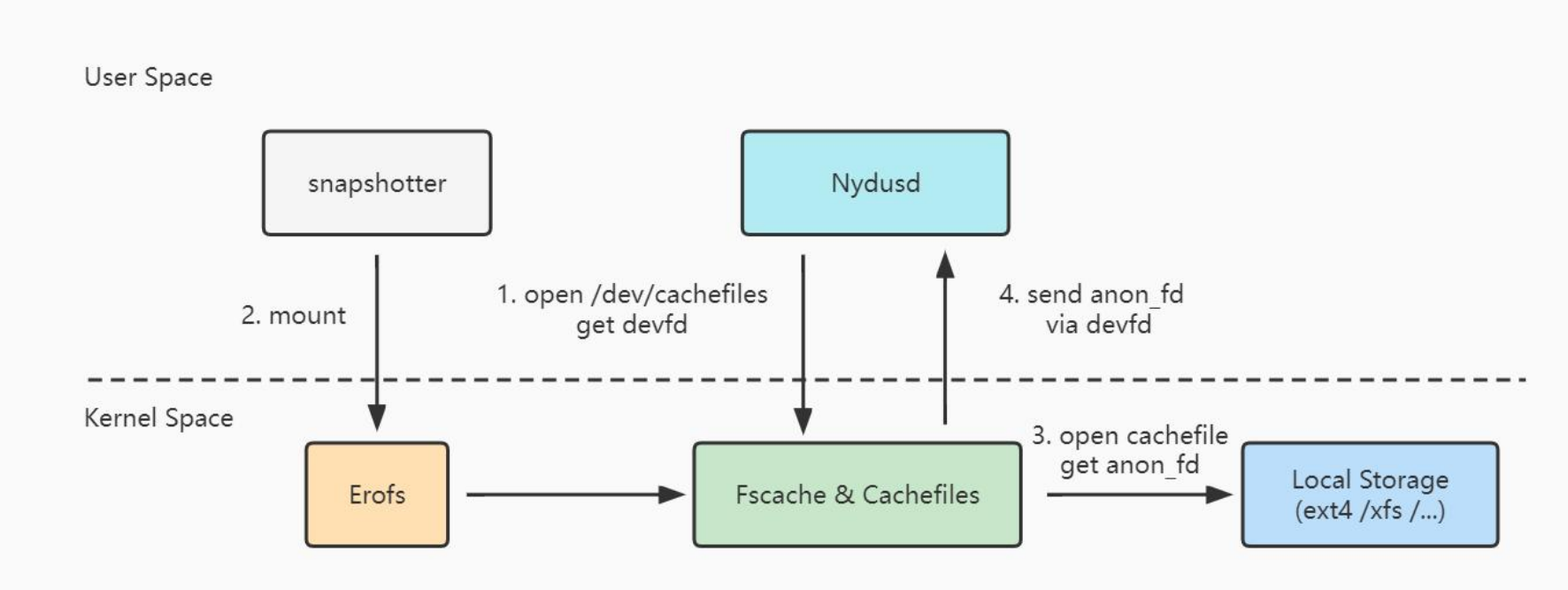
- 异常监控
  - 用户态 supervisor 监控 user daemon 状态，触发故障恢复流程
- 资源保持与恢复
  - 控制面：/dev/cachefiles fd的保持和传递（UDS）
  - 数据面：Reopen mechanism：由按需IO驱动重新发送 anonymous fd
  - cachefiles 'restore' cmd：触发恢复In-Flight IO

ByteDance字节跳动

稳定性提升：User Daemon Failover

fd资源保持与恢复

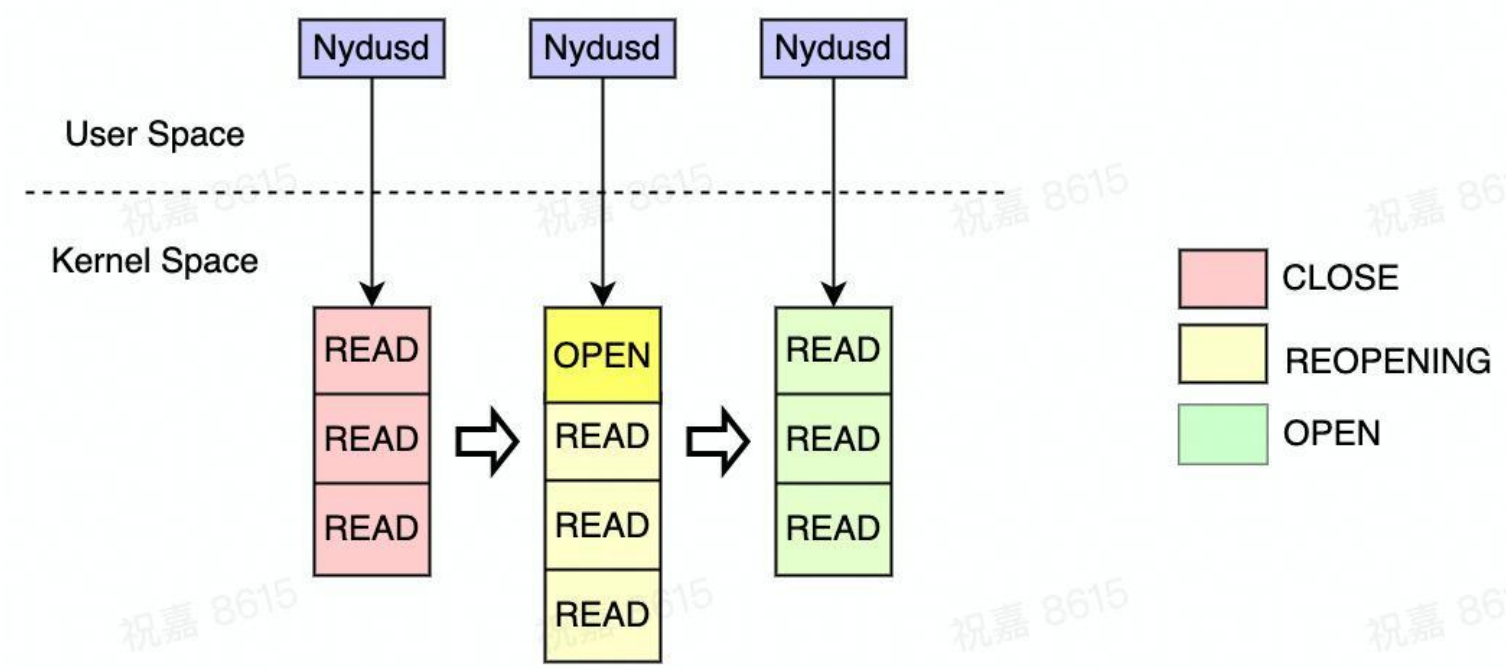
- /dev/cachefiles fd：用户态使用UDS保存和恢复



稳定性提升：User Daemon Failover

fd资源保持与恢复

- 缓存文件匿名fd：由按需IO触发，重新生成OPEN请求打开文件

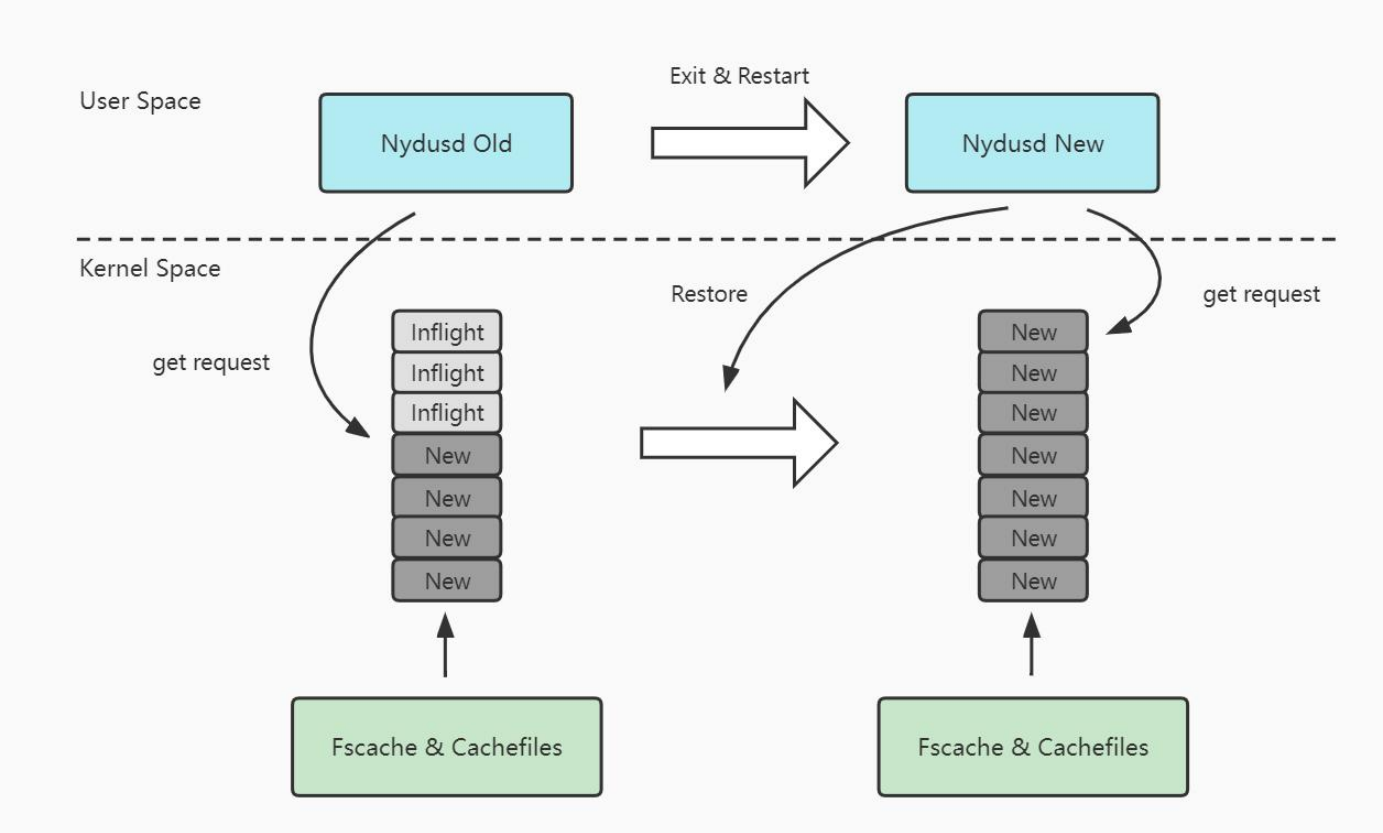


IO request xaray

ByteDance字节跳动

稳定性提升：User Daemon Failover

- 恢复IO请求



稳定性提升：User Daemon Failover

- 进展：正在推动进入upstream

- https://lore.kernel.org/all/20221014080559.42108-1-zhujia.zj@bytedance.com/

实践与探索——New Features

01

Failover

02

EROFS Shared Domain

03

Daemonless

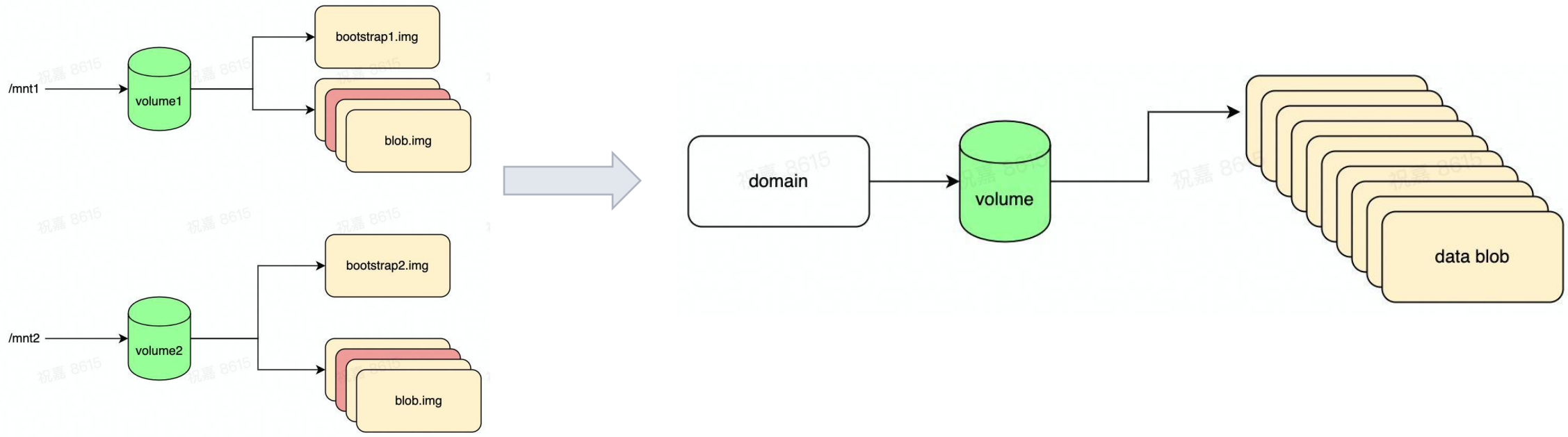
数据去重：EROFS Shared Domain

背景：

- 不同EROFS挂载点间，相同的镜像无法共享

预期目标：

- 实现镜像layer粒度的共享



ByteDance字节跳动

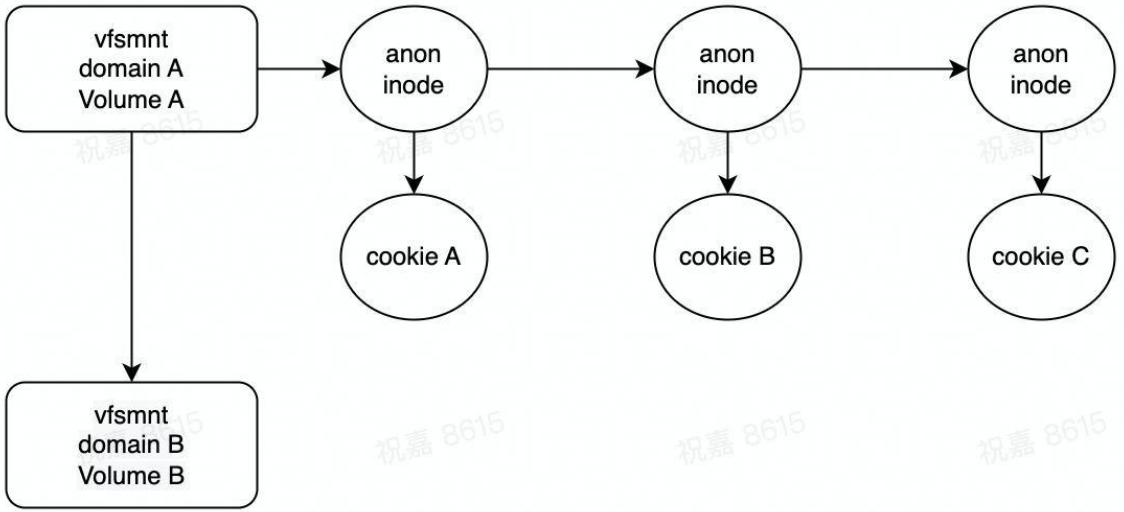
数据去重：EROFS Shared Domain

设计

- EROFS侧抽象Shared Domain模型，对应fscache volume
- pseudo挂载点
  - 管理domain生命周期
  - anonymous inodes 管理镜像文件生命周期，对应fscache cookie（blob）

使用

- EROFS挂载选项'domain_id'
- 创建/加共享域：同一共享域内所有挂载点共享镜像
- 不使用domain_id：不与任何挂载点共享镜像数据



ByteDance字节跳动

数据去重：EROFS Shared Domain

```
CACHE    REF   VOLS  OBJS  ACCES S NAME
======== ===== ===== ===== ===== = ===============
00000009     3     2    11     1 A CacheFiles
VOLUME   REF   nCOOK ACC FL CACHE           KEY
======== ===== ===== === == =============== ================
00000009     7     6   1 00 CacheFiles      erofs,7ea193b433cad62d1b909ecd6a948201fb17ff5d2aca3284154bad090109b186
0000000a     6     5   1 00 CacheFiles      erofs,32e72ef854d48d4850f6d677b5a085dc25cd69250129636ae0cd8322f7722201
COOKIE   VOLUME   REF ACT ACC S FL DEF
======== ======== === === === = == ================
00000022 00000009   2   1   0 A 6134 376561313933623433336361643632643162393039656364366139343832303166623137666635643261636133323834313534626164303930313039623138 36
00000023 00000009   2   1   0 A 6134 353132653066646465646131306533663337636135633730313732616539313261346166643231333663373932383734336538613466363035663731386562 64
00000024 00000009   2   1   0 A 4134 633733333861363438623163326363623533646137373561666434623234633466636661333538393537353834623035663263393438333634633063613566 62
00000025 00000009   2   1   0 A 4134 636238343138346136336439623362343339366339616262666561313230666436613565373531326639336264313263393934646637613430633039376235 35
00000026 00000009   2   1   0 A 6134 353937333165323362653337616163353363366465643138393434646634343135366161313431313262343436396436323432316616646463363939373435 32
00000027 00000009   2   1   0 A 6134 323634323461383236333564313537616233323766613562363737646331316465363535636661326663383038396566303237356433333464366434306232 37
00000028 0000000a   2   1   0 A 6134 333265373265663835346434386434383530663664363737623561303835646332356364636932353031323936333661653063643833323266373732323031 31
00000029 0000000a   2   1   0 A 6134 353132653066646465646131306533663337636135633730313732616539313261346166643231333663373932383734336538613466363035663731386562 64
0000002a 0000000a   2   1   0 A 4134 633733333861363438623163326363623533646137373561666434623234633466636661333538393537353834623035663263393438333634633063613566 62
0000002b 0000000a   2   1   0 A 4134 636238343138346136336439623362343339366339616262666561313230666436613565373531326639336264313263393934646637613430633039376235 35
0000002c 0000000a   2   1   0 A 6134 353937333165323362653337616163353363366465643138393434646634343135366161313431313262343436396436323432316616646463363939373435 32
```

```
CACHE    REF   VOLS  OBJS  ACCES S NAME
======== ===== ===== ===== ===== = ===============
00000007     2     1     7     1 A CacheFiles
VOLUME   REF   nCOOK ACC FL CACHE           KEY
======== ===== ===== === == =============== ================
00000007     8     7   1 00 CacheFiles      erofs,ABigDomain
COOKIE   VOLUME   REF ACT ACC S FL DEF
======== ======== === === === = == ================
0000001a 00000007   2   1   0 A 6134 376561313933623433336361643632643162393039656364366139343832303166623137666635643261636133323834313534626164303930313039623138 36
0000001b 00000007   2   1   0 A 6134 353132653066646465646131306533663337636135633730313732616539313261346166643231333663373932383734336538613466363035663731386562 64
0000001c 00000007   2   1   0 A 4134 633733333861363438623163326363623533646137373561666434623234633466636661333538393537353834623035663263393438333634633063613566 62
0000001d 00000007   2   1   0 A 4134 636238343138346136336439623362343339366339616262666561313230666436613565373531326639336264313263393934646637613430633039376235 35
0000001e 00000007   2   1   0 A 6134 353937333165323362653337616163353363366465643138393434646634343135366161313431313262343436396436323432316616646463363939373435 32
0000001f 00000007   2   1   0 A 6134 323634323461383236333564313537616233323766613562363737646331316465363535636661326663383038396566303237356433333464366434306232 37
00000020 00000007   2   1   0 A 6134 333265373265663835346434386434383530663664363737623561303835646332356364636932353031323936333661653063643833323266373732323031 31
```

Shared Domain

数据去重：EROFS Shared Domain

- 进展
  - 已进入upstream v6.1-rc1
  - https://lore.kernel.org/all/20220916085940.89392-1-zhujia.zj@bytedance.com/

- 未来工作
  - 发挥内核文件系统优势，探索page cache sharing

实践与探索——New Features

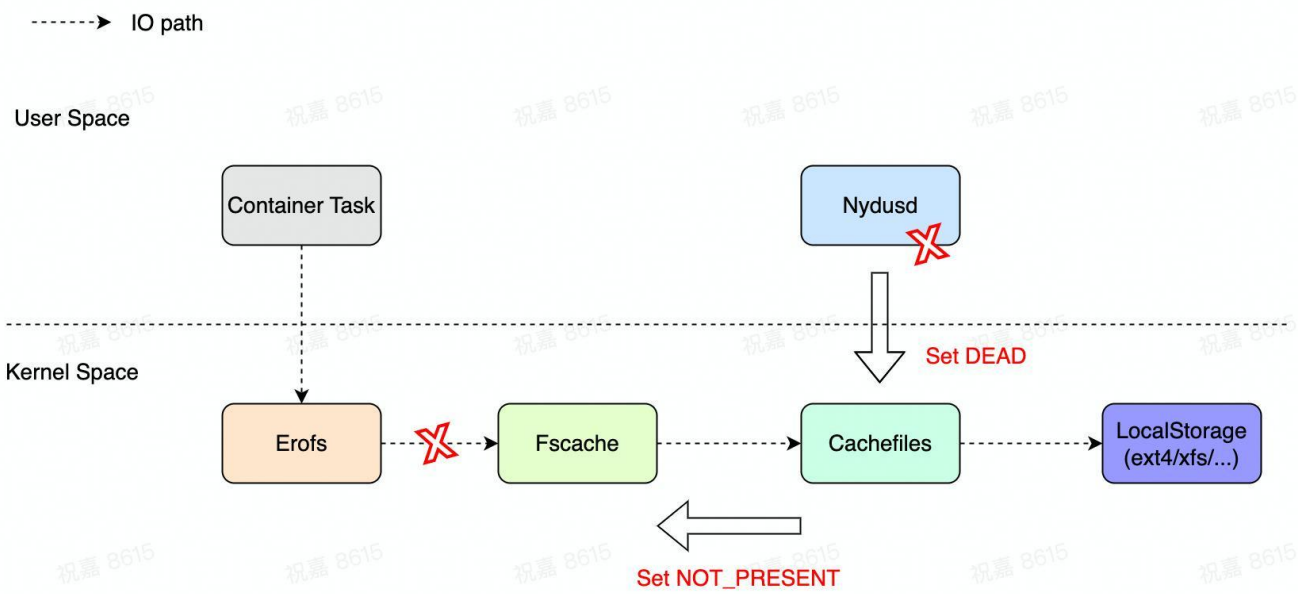
01 Failover

02 EROFS Shared Domain

03 **Daemonless**

0开销：Daemonless

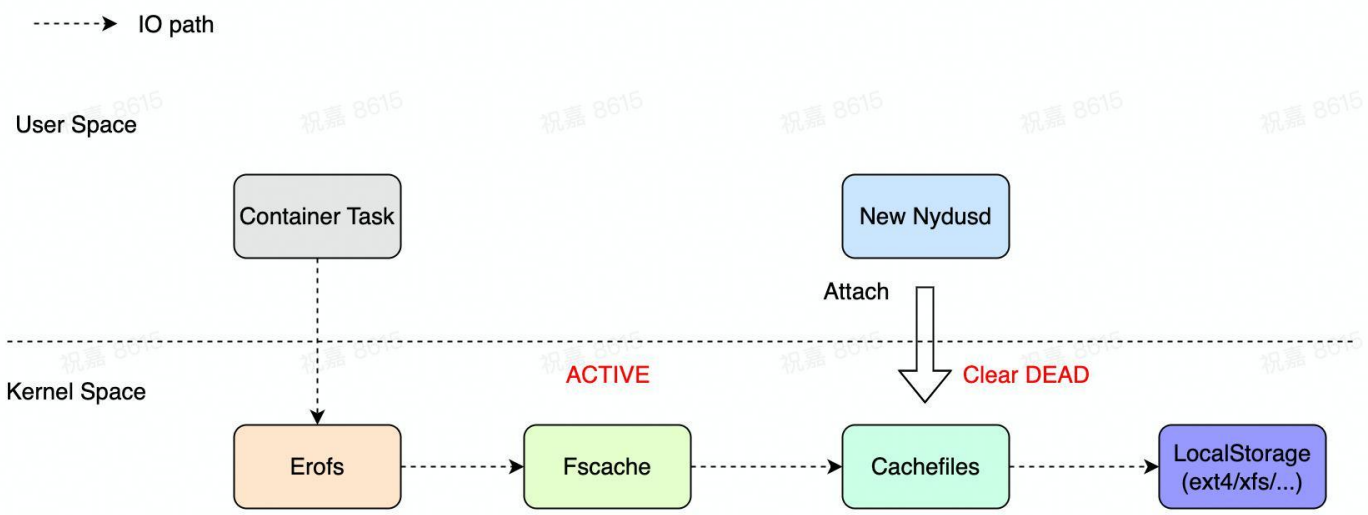
- 背景：按需加载完成后，IO转为纯内核路径，不再需要用户进程参与，可使其退出，节省资源
- 面临的挑战：用户进程退出后，fscache相关内核结构被置为DEAD状态，EROFS无法从fscache中读取数据
- 设计目标：cachefiles实现新运行模式，解除本地数据访问对user daemon的依赖



0开销：Daemonless

设计目标

- daemon 退出后，内核不释放相关结构，EROFS依然可以访问本地缓存文件
- daemon启动，可重新接管上述本地缓存文件，继续进行cache维护管理。



0开销：Daemonless

进展

代码已验证完成，即将提交内核社区

ByteDance字节跳动

Thanks