JD.COM 京东

docker 原理与应用实践

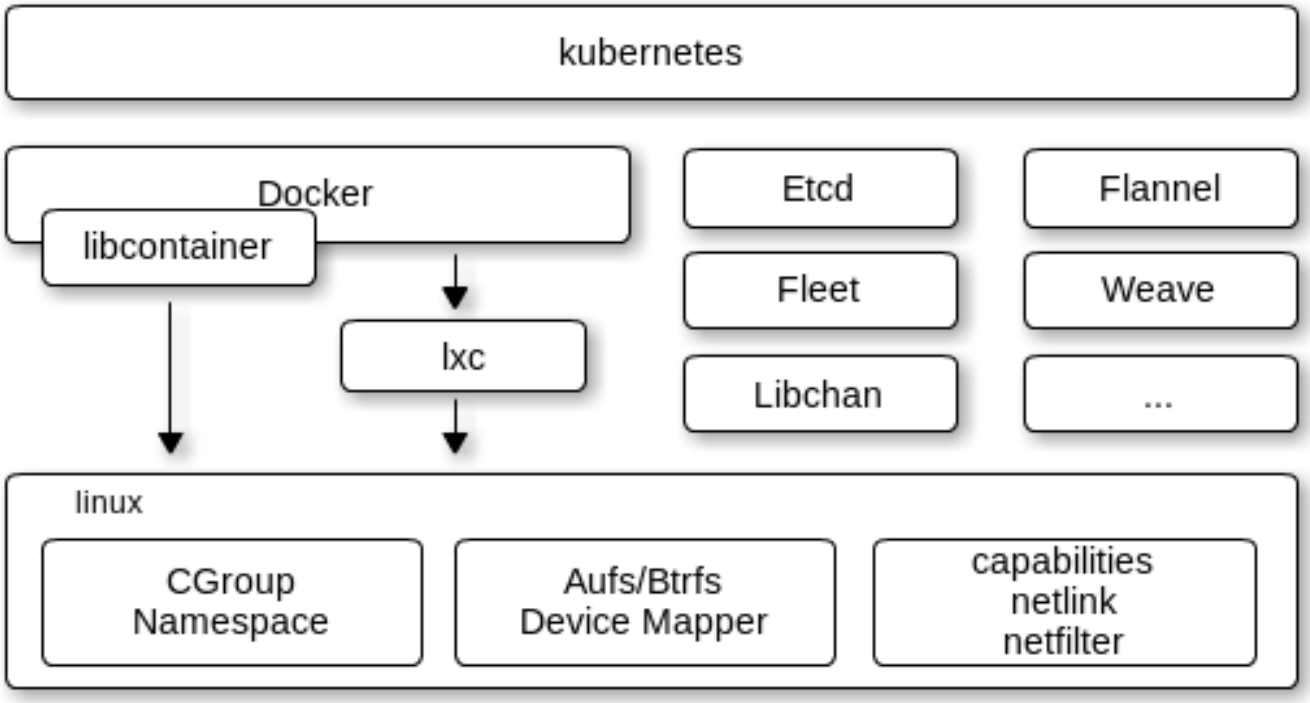
张成远

JD.COM 京东

- 容器系统整体架构
- Namespace
- CGroup
- Device Mapper
- Pull Image
- Start Container
- Stop Container
- Docker Image Storage

容器系统整体架构

JD.COM 京东

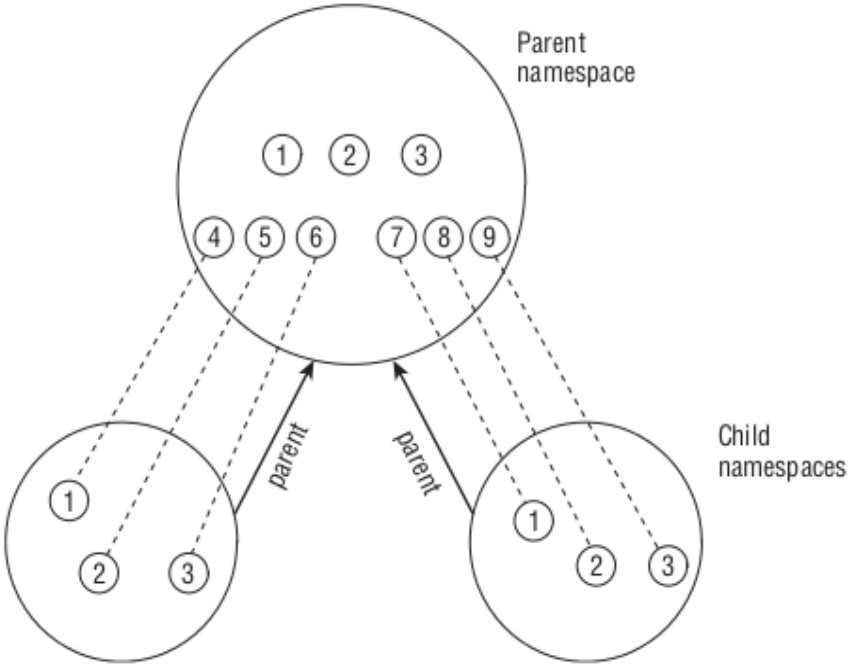


- 提供进程级别的资源隔离
- 为进程提供不同的命名空间视图
- 与虚拟机不同

Namespace 概念及当前支持的种类

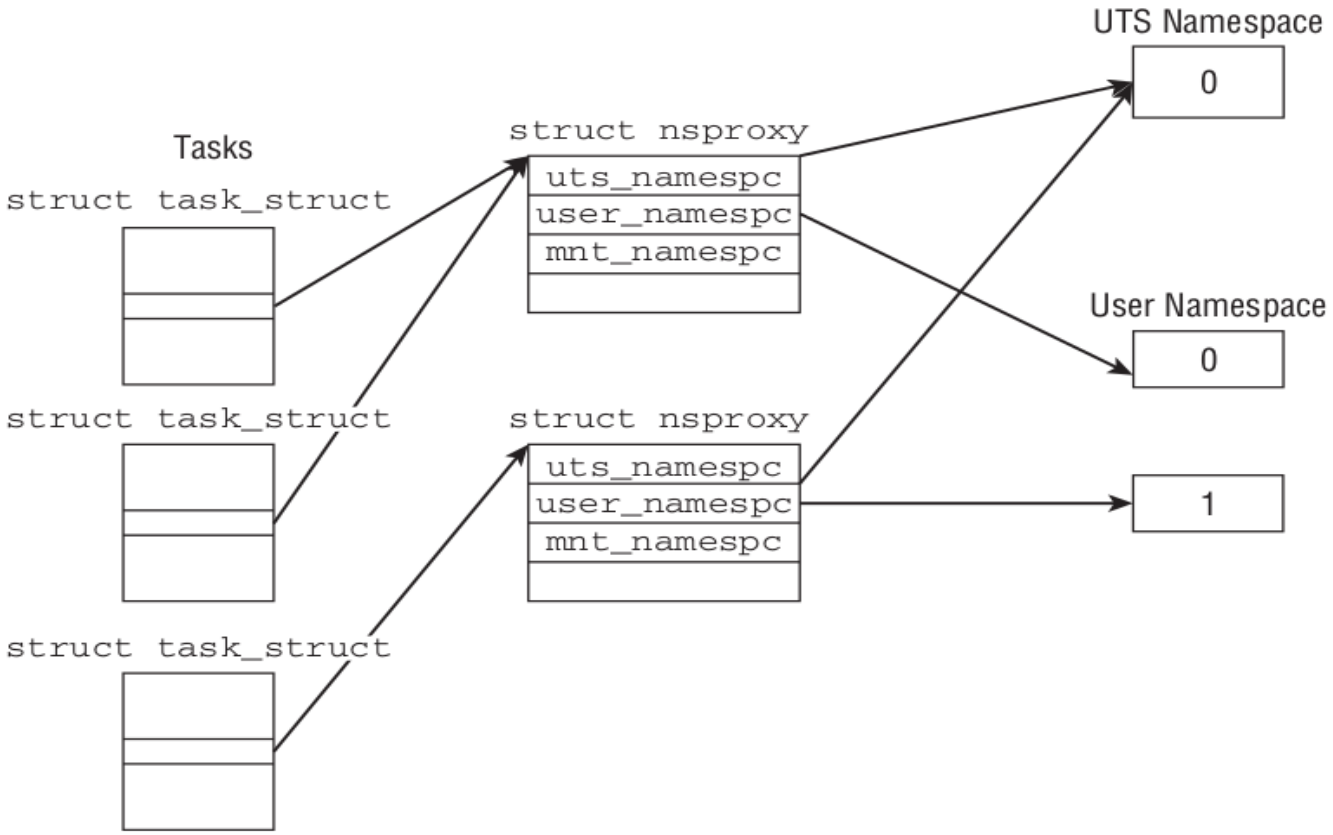
JD.COM 京东

- mnt (Mount points)
- pid (Processes)
- net (Network stack)
- ipc (System V IPC)
- uts (Hostname)
- user (UIDS)



Namespace 实现

JD.COM 京东



JD.COM 京东

- 创建新进程及 namespace

```
int clone(int (*fn)(void *), void *child_stack,
          int flags, void *arg, ...
          /* pid_t *ptid, struct user_desc *tls, pid_t *ctid */ );
```

- 加入当前进程到新建 namespace 中

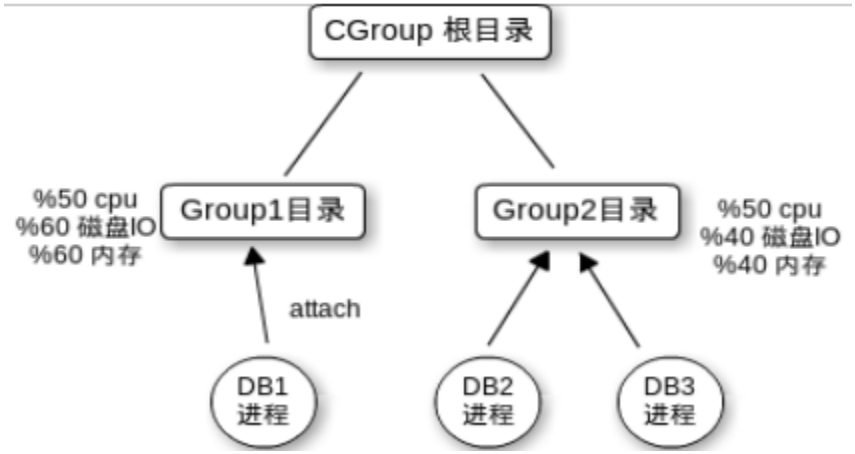
```
int unshare(int flags);
```

- 改变当前进程的 namespace

```
int setns(int fd, int nstype);
```

JD.COM 京东

- 提供进程的资源管理功能
- 资源管理主要涉及内存 ,CPU,IO 等
- 不依赖于 Namespace ，可单独使用
- 管理功能通过 VFS 接口暴露
- CGroups 提供通用框架，各子系统负责实现

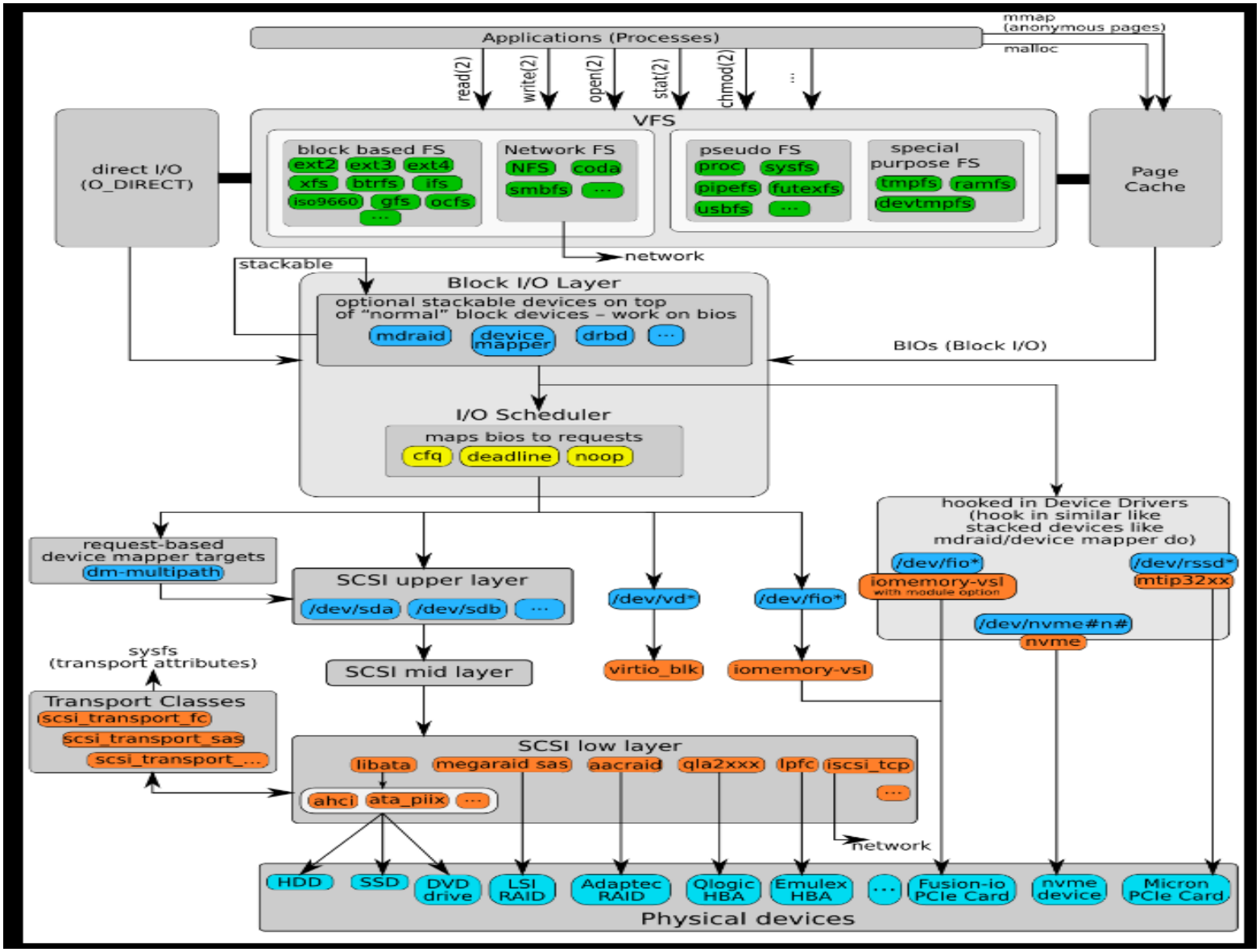


- blkio — 块设备 I/O 限制
- cpu — CPU 限制
- cpuacct — 自动生成 CPU 使用报告
- cpuset — 限定所使用的核
- memory — 限制内存
- devices— 控制任务访问设备
- freezer— 挂起 / 恢复任务

- DM 框架为上层应用提供了丰富的设备映射及 IO 策略方面的支持
- Docker 存储端实现之一使用 DM - thin provision
- 上层通过 dmsetup 工具或 libdevmapper 库使用

Device Mapper

JD.COM 京东



- Docker 支持 Aufs,Btrfs,DM 等
- 由于 DM 基于设备层，对上层文件系统 layer Diff 无法直接支持， Docker 手工比对文件实现
- 启动 docker 如果未指定 stroage driver ，依据 os 依次选择 aufs 、 btrfs 、 devicemapper

JD.COM 京东

- many virtual devices to be stored on the same volume
- an arbitrary depth of recursive snapshots
- metadata is stored on a seperate device from data

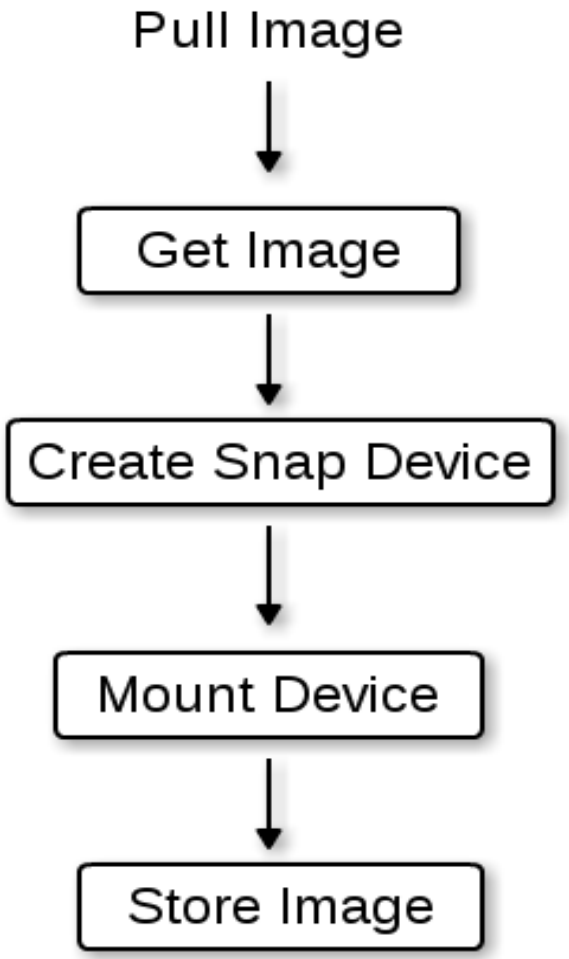
JD.COM 京东

- dd if=/dev/zero of=metadata bs=1024k count=128
- dd if=/dev/zero of=data bs=1024k count=1024
- losetup /dev/loop7 metadata
- losetup /dev/loop6 data
- dmsetup create pool --table "0 20971520 thin-pool /dev/loop7 /dev/loop6 128 512"

JD.COM 京东

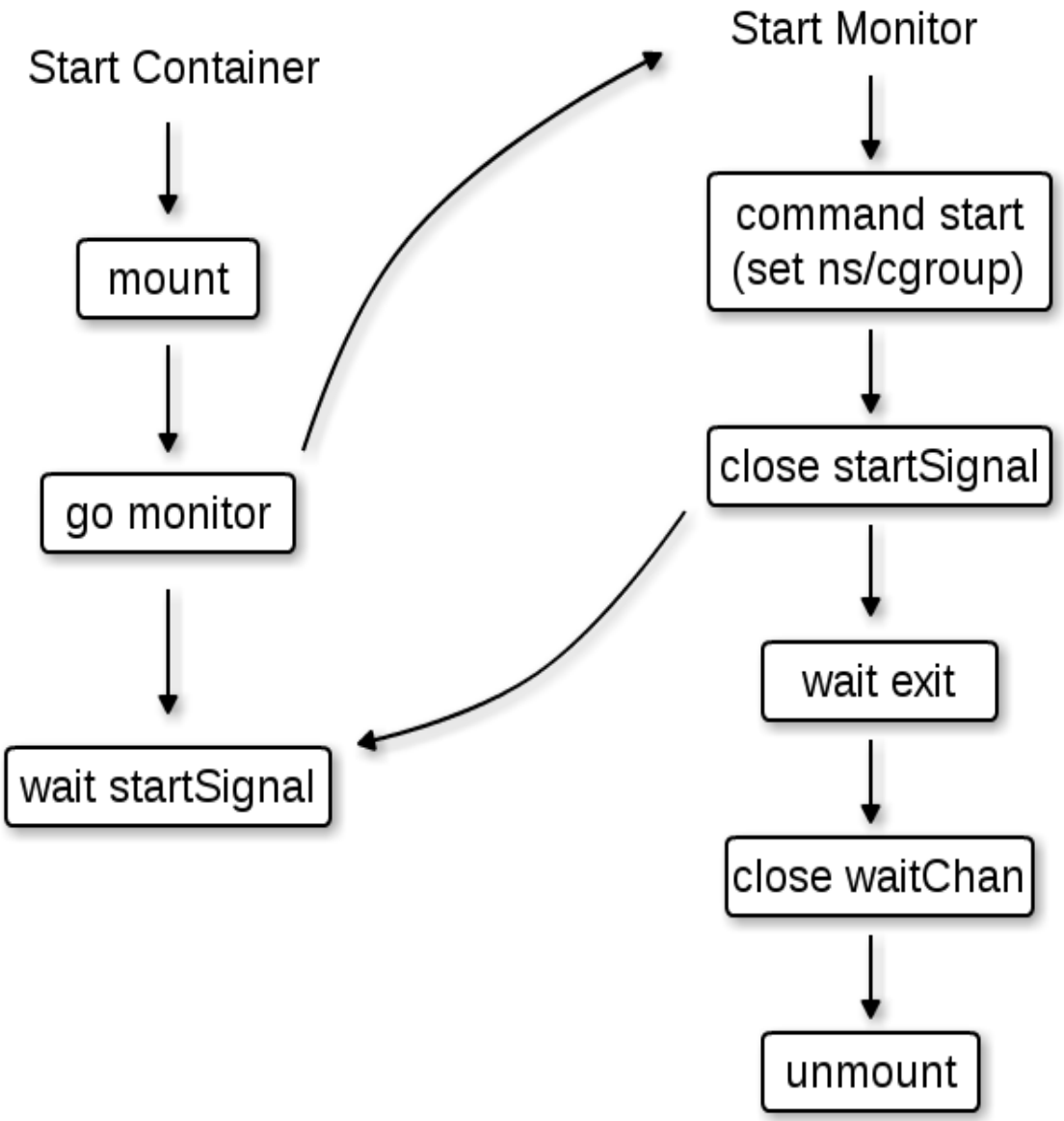
- dmsetup message /dev/mapper/pool 0 "create_thin 0"
- dmsetup create thin --table "0 2097152 thin /dev/mapper/pool 0"
- mkfs.ext4 /dev/mapper/thin
- mount /dev/mapper/thin /export

- data 和 metadata 需要两个块设备
- truncate 生成文件， loop 设备
- dm.loopdatasize=100G
- dm.basesize=10G
- dm.datadev 指定 pool 使用的设备
- dm.metadatadev 指定 metadata 使用的设备
- dd if=/dev/zero of=$metadata_dev bs=4096 count=1


Pull Image
Get Image
Create Snap Device
Mount Device
Store Image

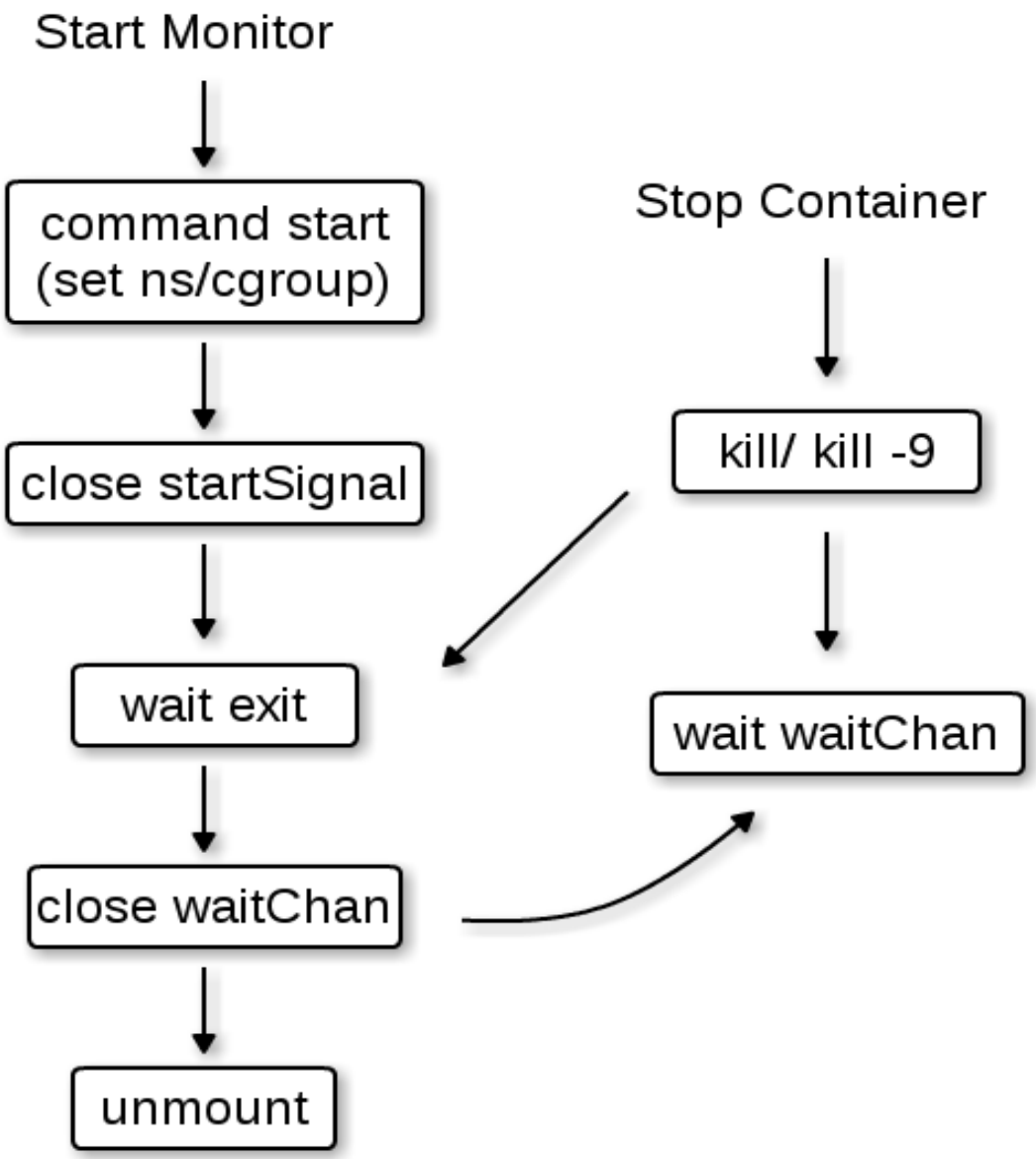
Start Container

JD.COM 京东



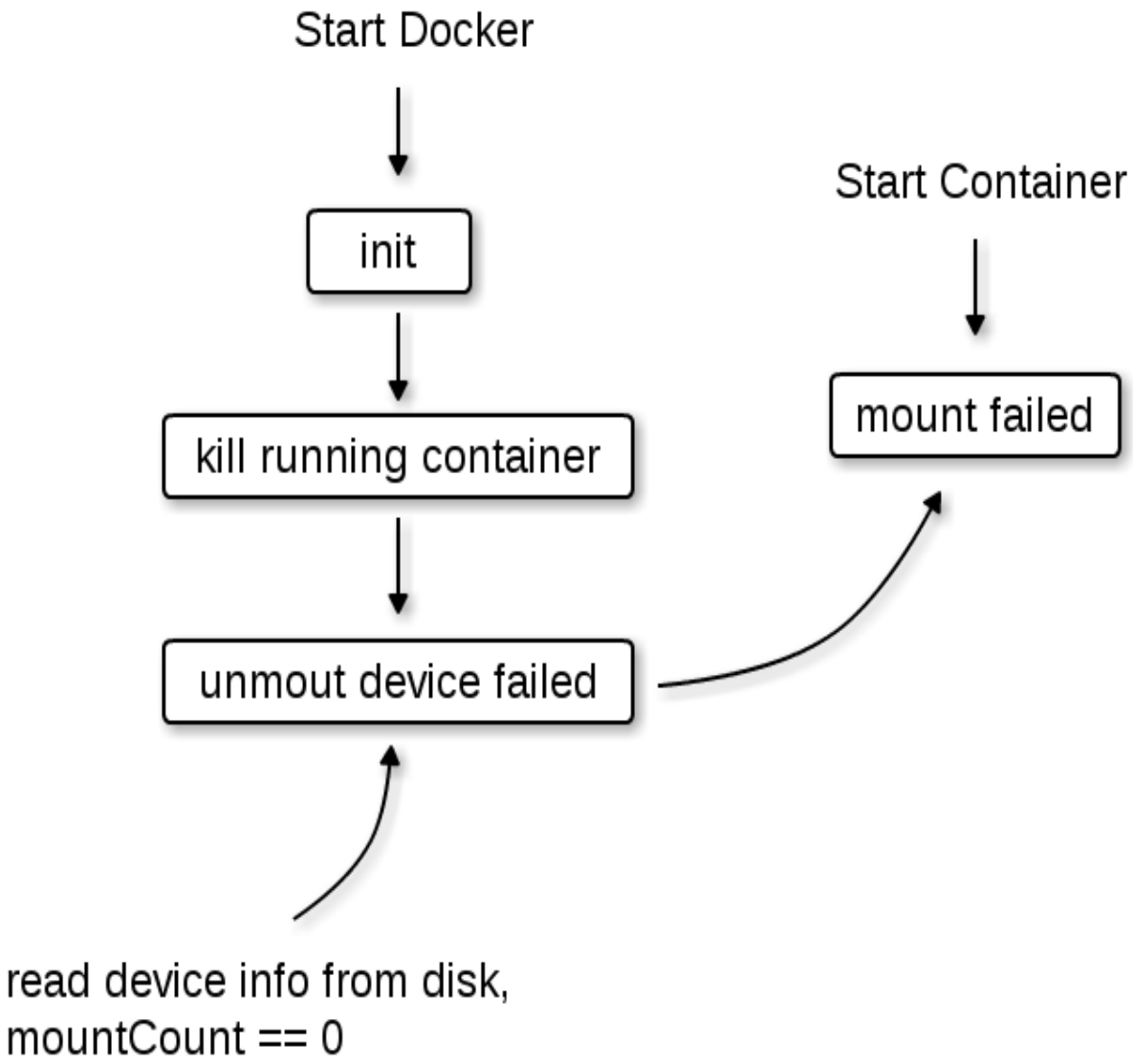
Stop Container

JD.COM 京东



Start Docker& Start Container

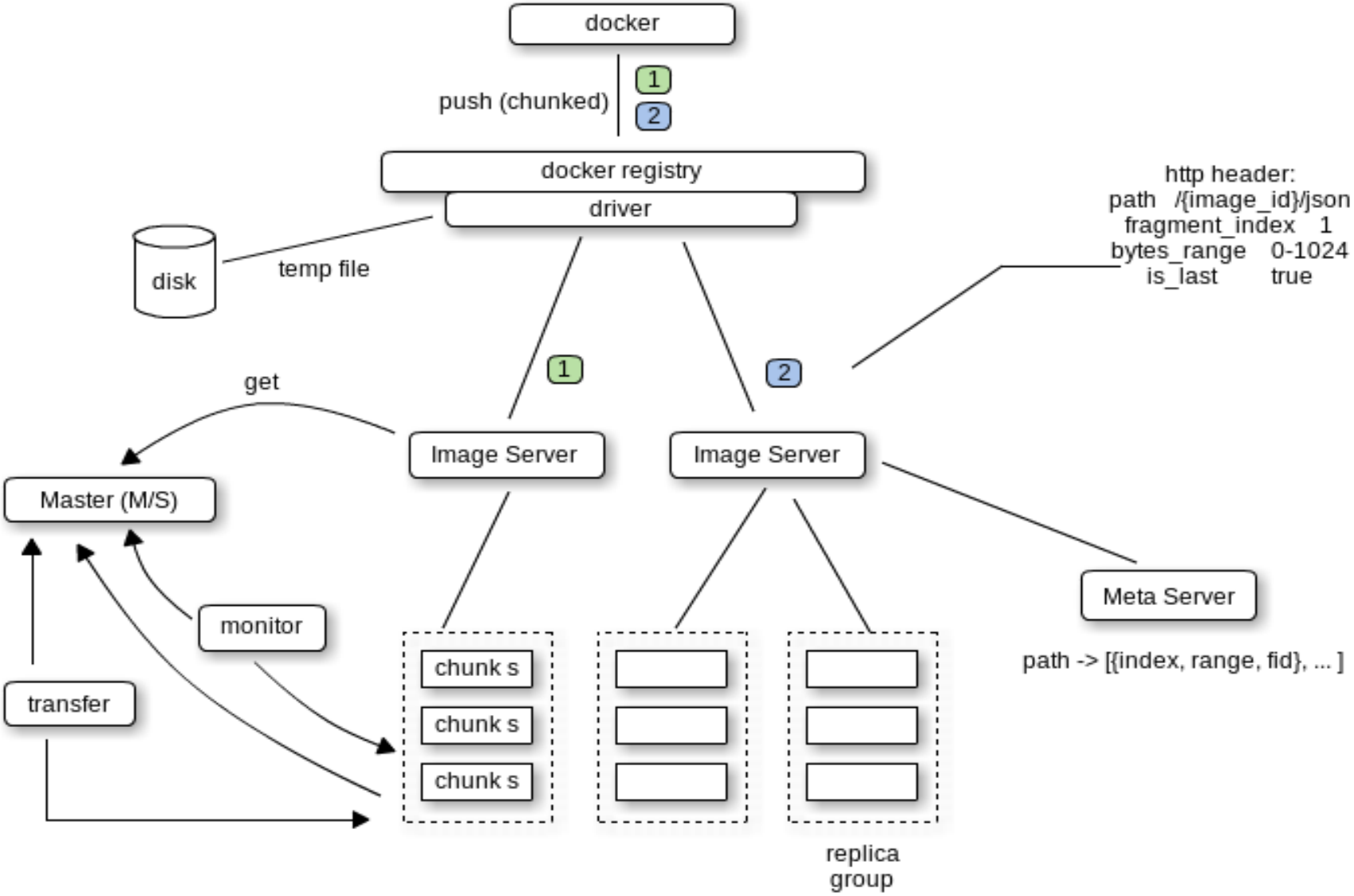
JD.COM 京东



- docker 退出时会依次 stop 各个 container
- Start container 时 mount 相关设备
- Stop container 时， monitor 会做清理工作
- docker 异常停止时，各个 container 对应的 namespace 等相关信息依然存在
- docker 启动时候会将之前的 container stop，但 umount 失败， start container 会失败

JD image storage system

JD.COM 京东



- 当前 Namespace 功能仍不完善，需要更多的隔离
- Docker 主要用到 CGroup 的一部分子系统
- Docker 存储端仍需做一些选择或工作
- 选择 DM thin-provision 时需要注意 data 及 metadata 的设置
- docker start/stop 之间的交互及扫尾工作
- image 本地存储与 dm 之间的关系
- Docker Registry Storage 需要选择开源或定制开发

MariaDB 原理与实现

张金鹏 张成远 季锡强 编著

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

JD.COM 京东

Docker

RDS

Linux kernel

Golang

SDN/Network

联系方式：

- zhanchengyuan@jd.com

JD.COM 京东

谢谢

微博： @NEU_ 寒水

JD.COM 京东