JD.COM 京东

Docker在京东的应用与实践

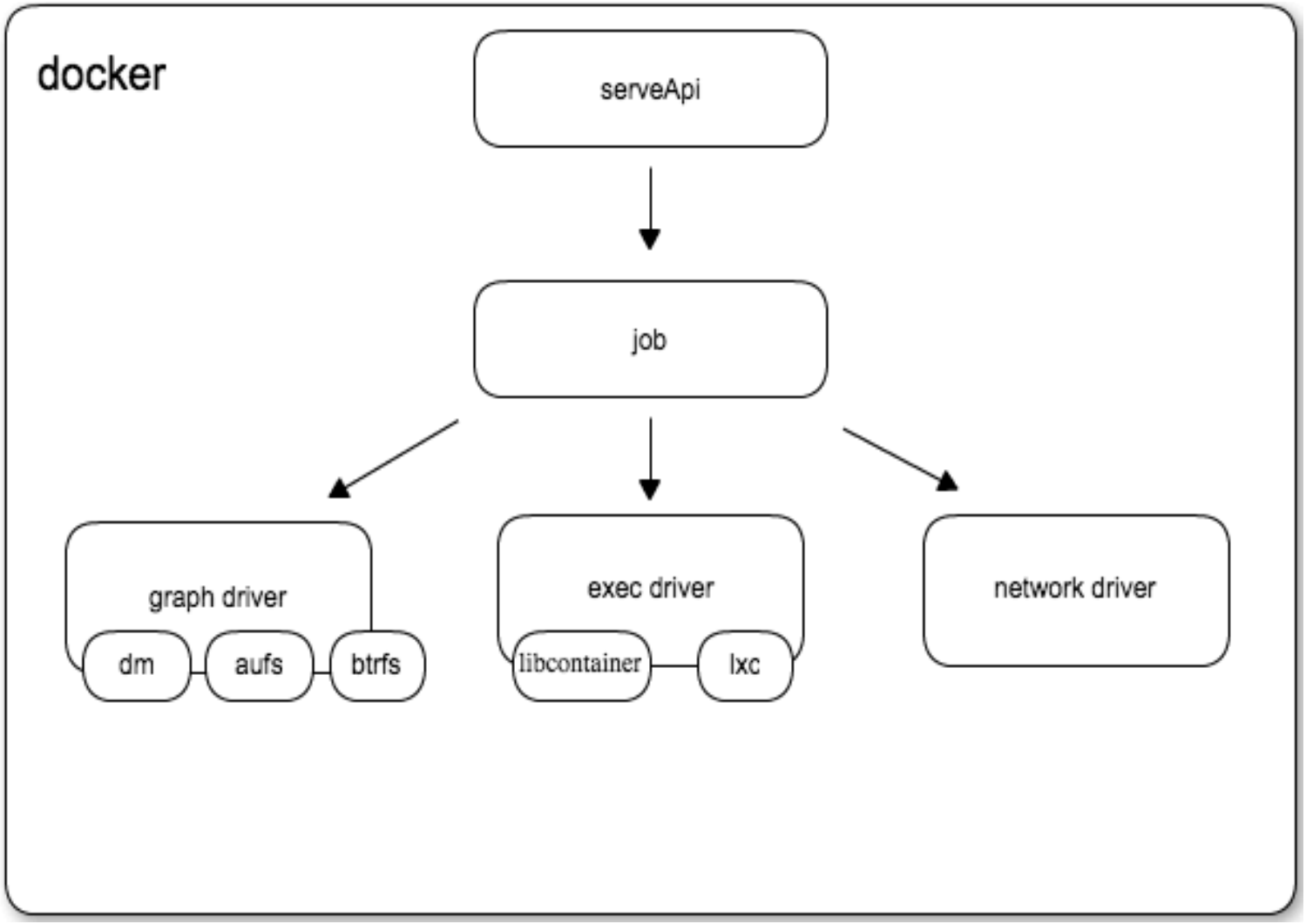
季锡强

JD.COM 京东

- Docker系统整体架构
- Namespace
- CGroup
- Device Mapper
- Start/Stop Container
- Commit Image
- Volume

Docker系统整体架构

JD.COM 京东

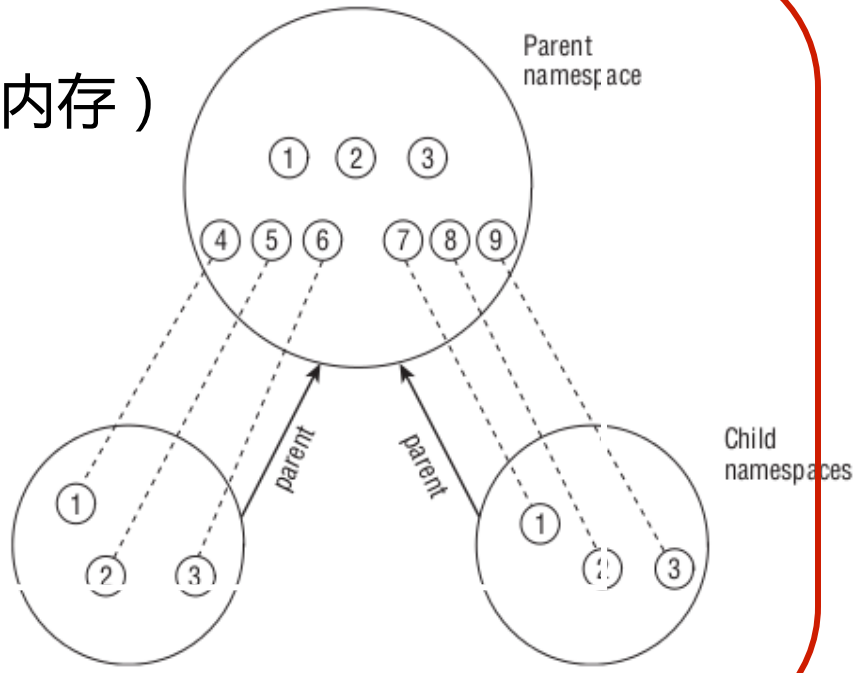


- 提供进程级别的资源隔离
- 为进程提供不同的命名空间视图
- 与虚拟机不同

Namespace概念及当前支持的种类

JD.COM 京东

- ipc (信号量、消息队列和共享内存)
- mnt (挂载点)
- pid (进程编号)
- net (网络设备、网络栈等)
- uts (主机名与域名)
- user (用户和用户组)



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

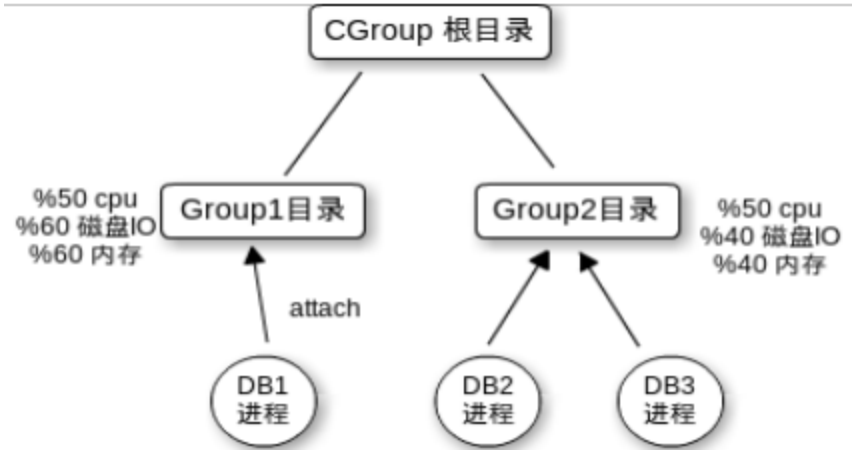
```
int unshare(int flags);
```

- 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

- 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

- 允许多个虚拟设备存储在相同的数据卷中
- 支持任意深度的快照
- 支持元数据存储到单独的设备上

DM-thin provision example

JD.COM 京东

- dd if=/dev/zero of=metadata bs=1024k count=4096
- dd if=/dev/zero of=data bs=1024k count=20480
- losetup /dev/loop7 metadata
- losetup /dev/loop6 data
- dmsetup create pool --table "0 20971520 thin-pool /dev/loop7 /dev/loop6 128 0"

DM-thin provision example

JD.COM 京东

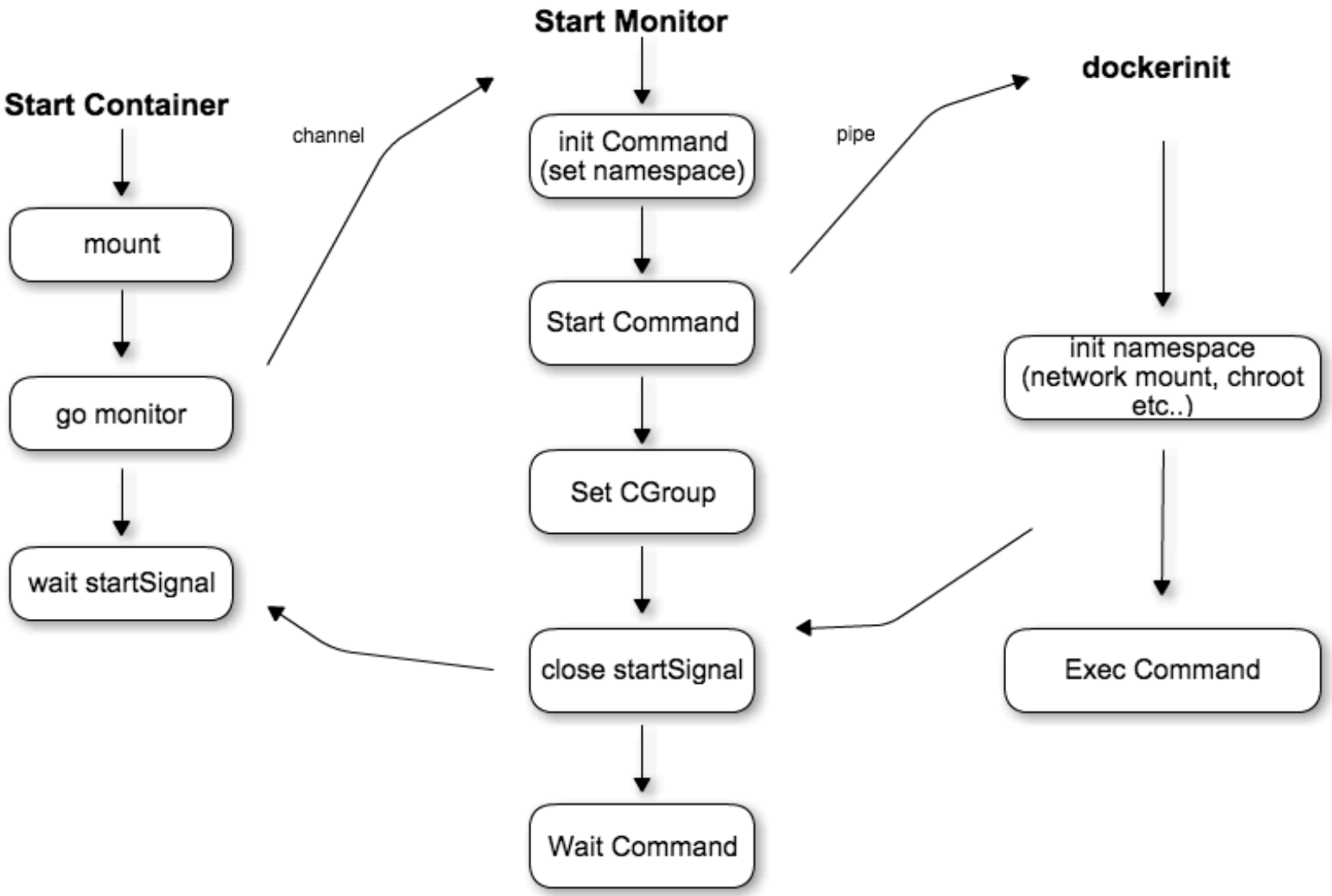
- dmsetup message /dev/mapper/pool 0 "create_thin 0"
- dmsetup create thin --table "0 2097152 thin /dev/mapper/pool 0"
- mkfs.ext4 /dev/mapper/thin
- mount /dev/mapper/thin /export

JD.COM 京东

- 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

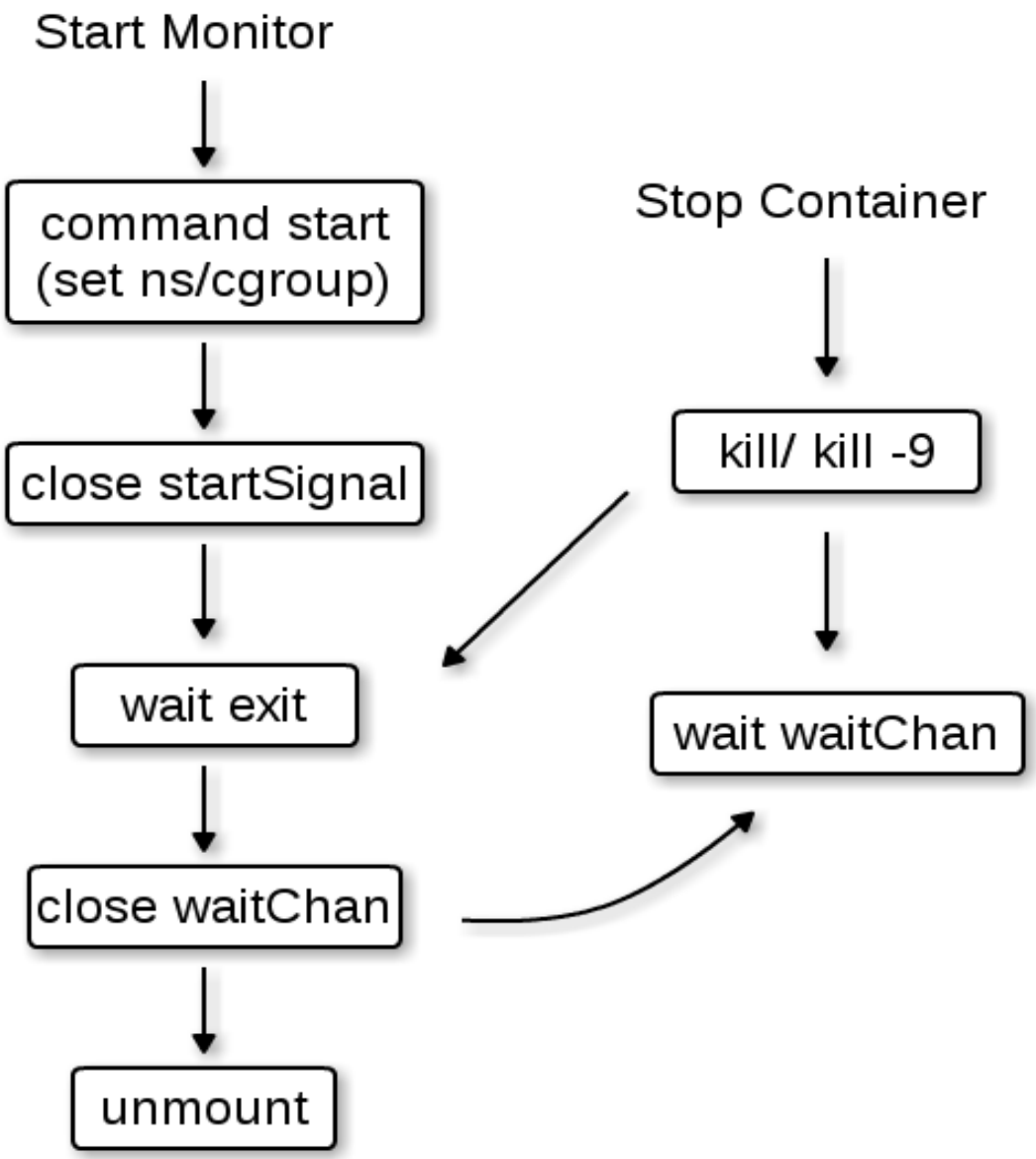
Start Container

JD.COM 京东



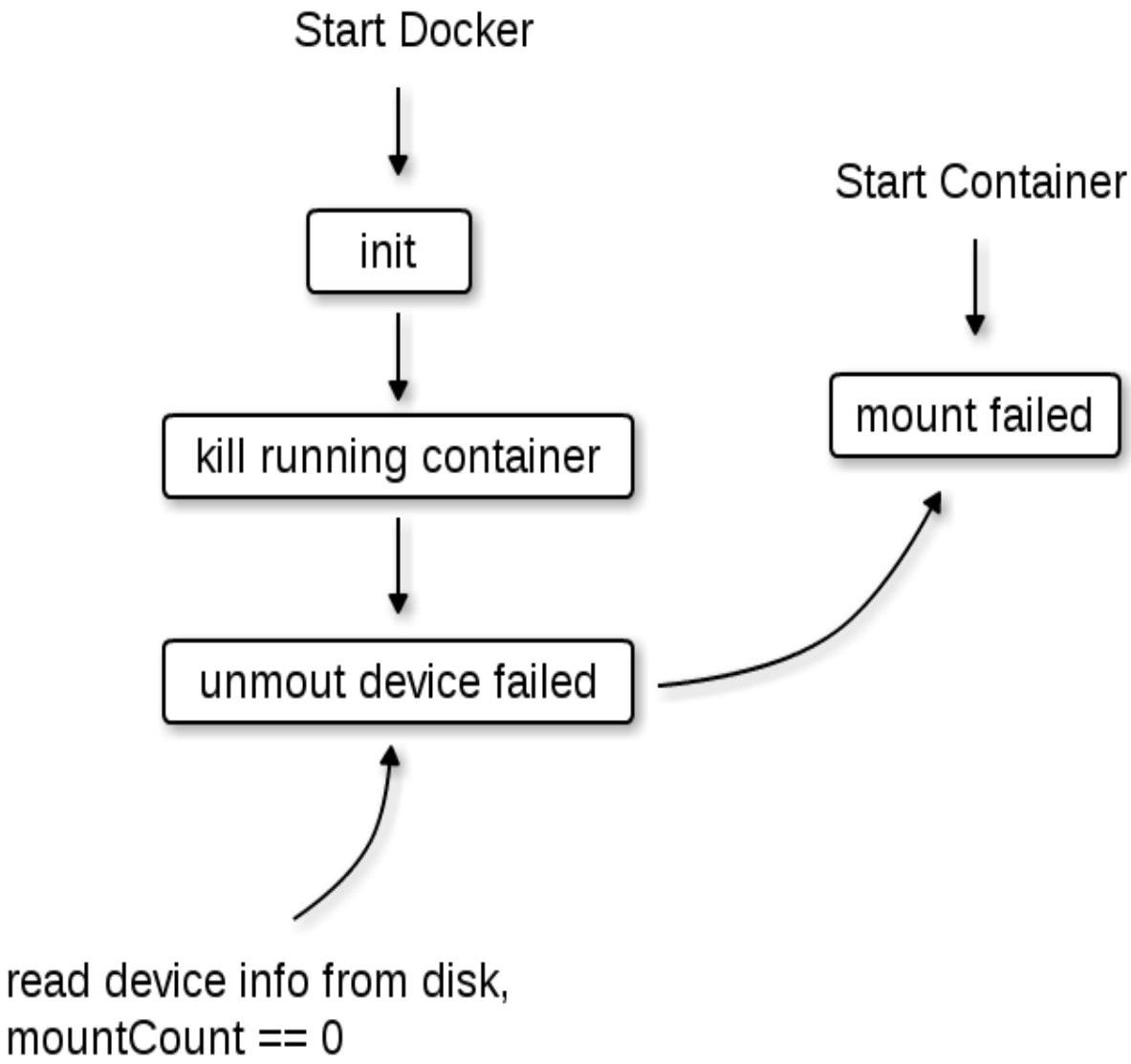
Stop Container

JD.COM 京东



Start Docker& Start Container

JD.COM 京东



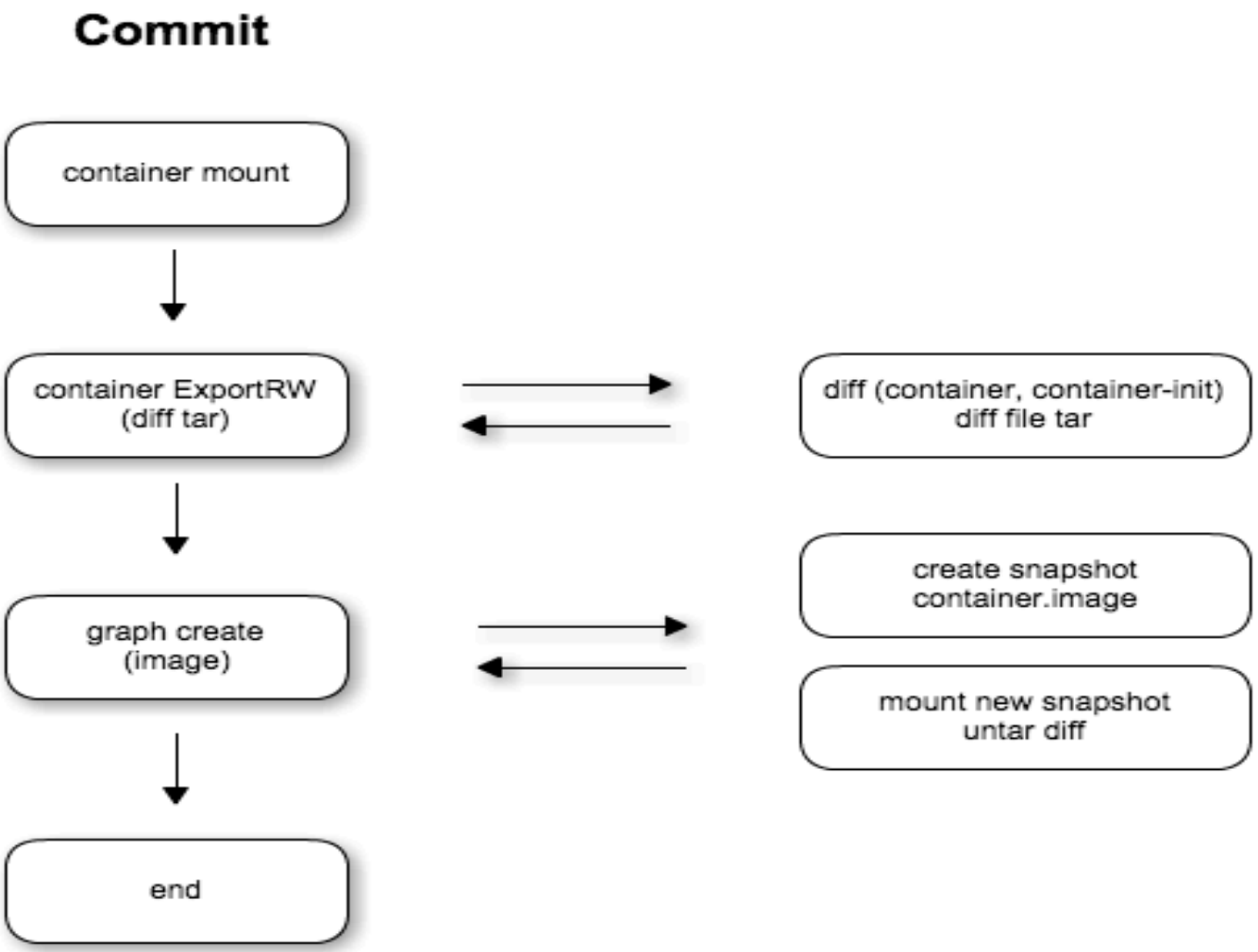
Start/Stop Container注意点

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

Commit Image

JD.COM 京东



- Docker Volume 主要使用mount –bind操作来实现
- Commit 操作是不会将Volume添加到新镜像中
- Volume-from 实现原理

小结

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
- 机器磁盘I/O很高时，会导致commit 操作相当耗时

谢谢

微博：@龙芯o

JD.COM 京东