QCon
全球软件开发大会

美团云
Meituan Open Services

美团云
分布式块存储系统
设计与实现

李慧霸 博士

2015 10 17

上海

个人简介

美团云
Meituan Open Services

- 李慧霸
- 美团云
  - ✤ 分布式块存储项目负责人
- 国防科学技术大学
  - ✤ 并行与分布处理国家重点实验室
  - ✤ 研究员

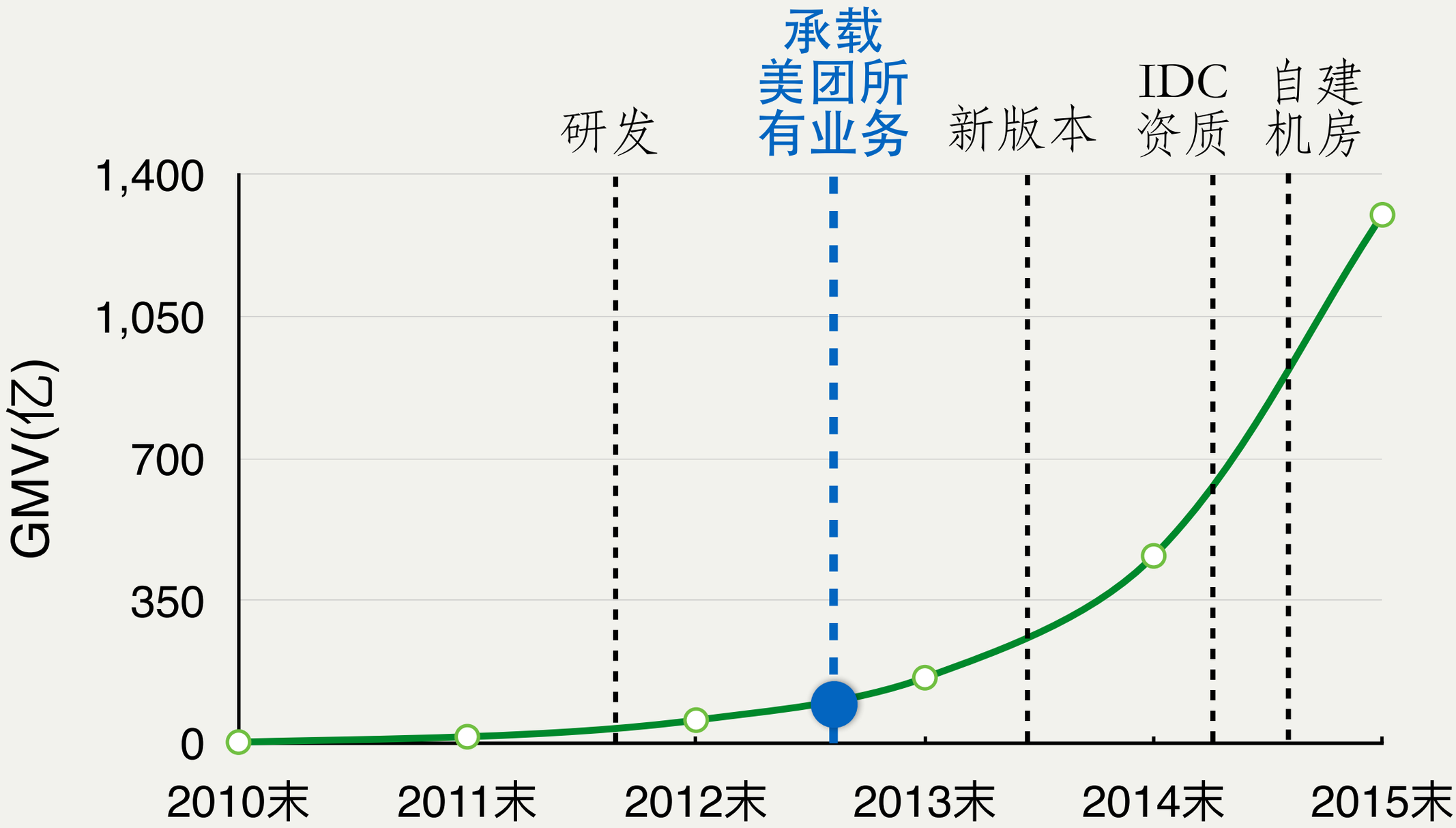
内容提要

美团云
Meituan Open Services

- 个人简介
- 云计算@美团
- 云计算对块存储的需求
- 设计与实现
  - ✤ 设计理念、存储结构、并发模型、
  - ✤ 系统架构、写入策略、错误防护
- 性能测试

美团云

美团云
Meituan Open Services



稳定承载美团所有业务

美团云
Meituan Open Services

云计算承载研发环境

美团云
Meituan Open Services

- 方案验证
- 新技术探索
- 压力测试
- 鲁棒性测试
- “小白鼠”平台
- 配额内自助使用

极高的**交付效率**

极高的**资源利用率**

月**资源“换手”率** > 50%

美团云 - 公有云服务

美团云
Meituan Open Services

Clover
雪球
聪明的投资者都在这里
geekcar
坚果游戏
传送门
chuansong.me
外卖库
送食送
JiaShiSong.com
良仓
生活家的良品分享
科学松鼠会
ivymobi
mds
Your IT Inspiration
全国空气质量指数
快简历
推酷
钛牛科技
TAINIU.CN
糗事百科
創業魔法學院
MAGIC ENTREPRENEUR COLLEGE
博泰
H&G IES
JobDeer
TAL 好未来
eduise 艾德睿智
六人行
彼岸吉他
byguitar.com
华慧微
huarim
华夏航空
CHINA EXPRESS
HAOCHENG
樂童音樂

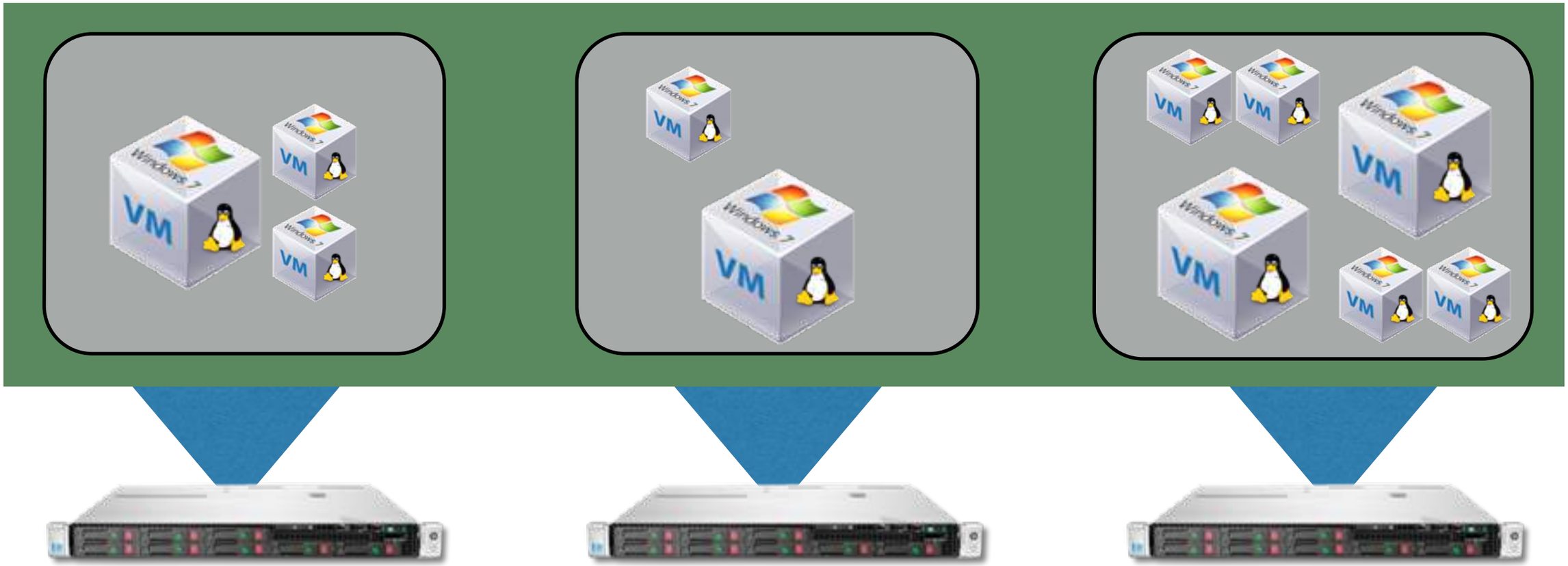
内容提要

美团云
Meituan Open Services

- 个人简介
- 云计算@美团
- 云计算对块存储的需求
- 设计与实现
  - ✤ 设计理念、存储结构、并发模型、
  - ✤ 系统架构、写入策略、错误防护
- 性能测试

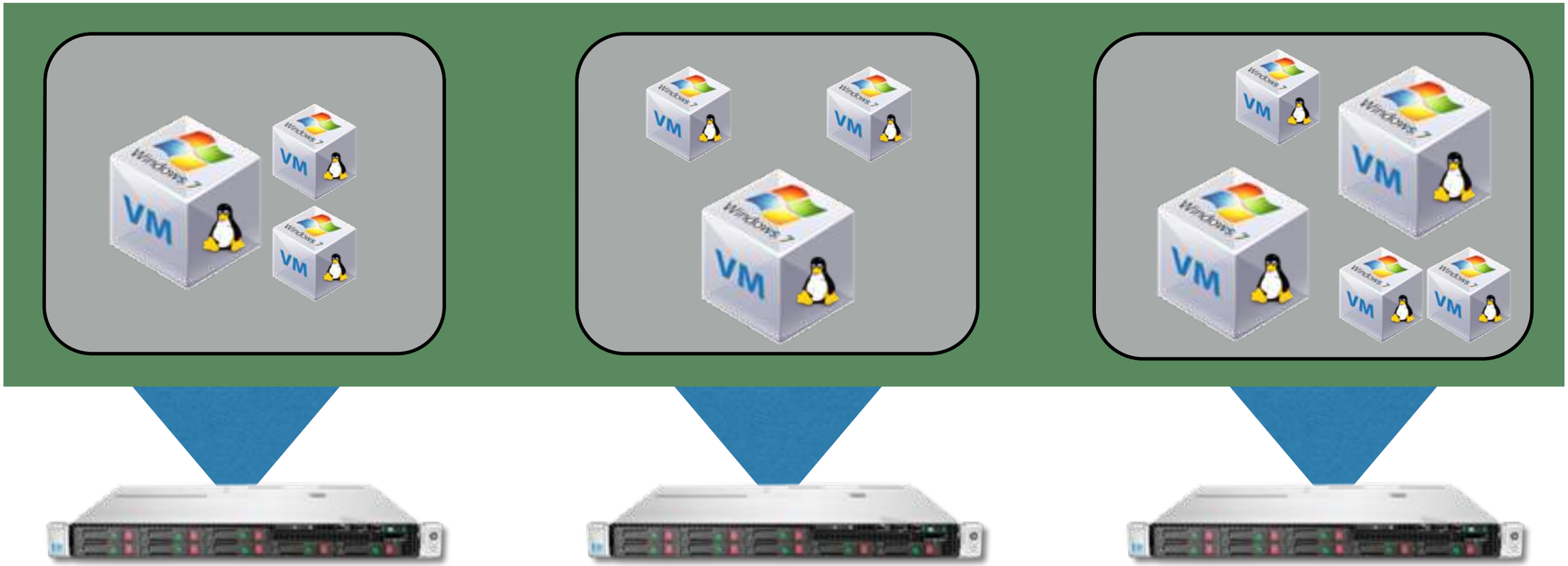
IaaS云计算

美团云
Meituan Open Services



IaaS云计算

美团云
Meituan Open Services



三类存储架构

美团云
Meituan Open Services

- 本地磁盘存储
  - ✤ 技术简单，性能高，成本低
  - ✤ VM迁移困难，可用性低

- 集中共享存储
  - ✤ 技术简单，VM易于迁移
  - ✤ 成本高，可用性低，扩展性低

- 分布共享存储
  - ✤ 可靠性高，可用性高，VM易于迁移
  - ✤ 技术困难

三类存储架构

美团云
Meituan Open Services

- 本地磁盘存储
  - ✤ 技术简单，性能高，成本低
  - ✤ VM迁移困难，可用性低

- 集中共享存储
  - ✤ 技术简单，VM易于迁移
  - ✤ 成本高，可用性低，扩展性低

- 分布共享存储 ✓
  - ✤ 可靠性高，可用性高，VM易于迁移
  - ✤ 技术困难

IaaS中的存储需求

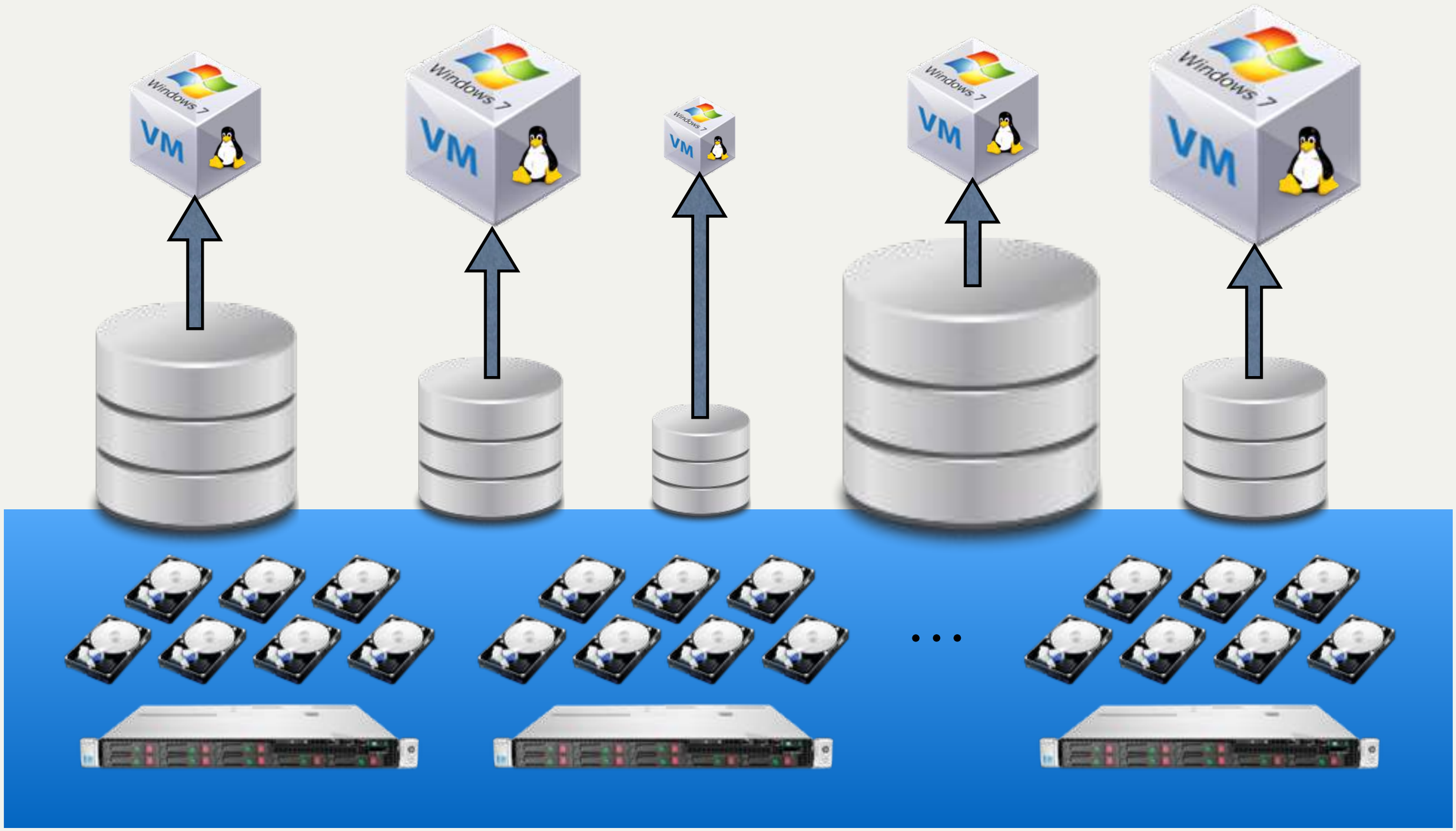
美团云
Meituan Open Services

IaaS中的存储需求

美团云
Meituan Open Services

IaaS中的存储需求

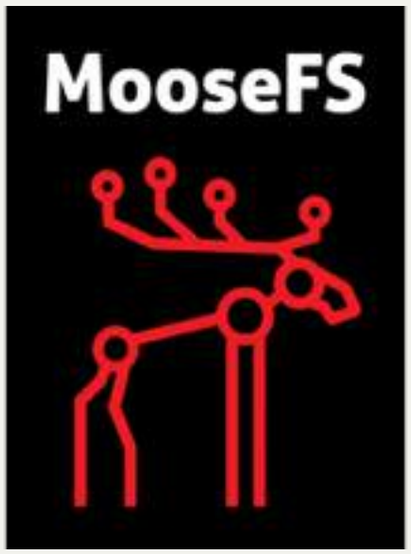
美团云
Meituan Open Services



相关开源项目

美团云
Meituan Open Services

Sheepdog



HDFS/GFS

HLFS

Ceph的效率问题

美团云
Meituan Open Services

服务端：

21282	root	20	0	809m	80m	11m	S	37.4	0.1	6:38.88 ceph-osd
17771	root	20	0	802m	80m	10m	S	27.8	0.1	4:54.46 ceph-osd
19448	root	20	0	839m	79m	10m	S	27.1	0.1	6:42.14 ceph-osd
18580	root	20	0	824m	82m	10m	S	26.5	0.1	5:52.29 ceph-osd
20392	root	20	0	798m	77m	10m	S	13.9	0.1	4:02.15 ceph-osd

123.7% @ 4,101读IOPS

客户端： 5xx% @16,407读IOPS

Sheepdog的数据丢失问题

美团云
Meituan Open Services

```
[root@node3-10gtest ~]# collie cluster check
fix vdi test1-3
99.9 % [=============================================================>] 50 GB / 50 GB
fixed replica 3e563000000fca
99.9 % [=============================================================>] 50 GB / 50 GB
fixed replica 3e563000000fec
100.0 % [============================================================>] 50 GB / 50 GB
fixed replica 3e5630000026f5
100.0 % [============================================================>] 50 GB / 50 GB
fixed replica 3e563000002da6
100.0 % [============================================================>] 50 GB / 50 GB
fixed replica 3e563000001e8c
100.0 % [============================================================>] 50 GB / 50 GB
fixed replica 3e563000001530
...
fix vdi test2-9
50.9 % [==================================>                           ] 25 GB / 50 GB
no majority of d781e300000123
51.2 % [===================================>                          ] 26 GB / 50 GB
no majority of d781e30000018a
53.2 % [====================================>                         ] 27 GB / 50 GB
...
```

Sheepdog的数据丢失问题

美团云
Meituan Open Services

```
[root@node3-10gtest ~]# collie cluster check
fix vdi test1-3
99.9 % [=============================================================>] 50 GB / 50 GB
fixed replica 3e563000000fca
99.9 % [=============================================================>] 50 GB / 50 GB
fixed replica 3e563000000fec
100.0 % [============================================================>] 50 GB / 50 GB
fixed replica 3e5630000026f5
100.0 % [============================================================>] 50 GB / 50 GB
fixed replica 3e563000002da6
100.0 % [============================================================>] 50 GB / 50 GB
fixed replica 3e563000001e8c
100.0 % [============================================================>] 50 GB / 50 GB
fixed replica 3e563000001530
...
fix vdi test2-9
50.9 % [================================>                             ] 25 GB / 50 GB
no majority of d781e300000123
51.2 % [=================================>                            ] 26 GB / 50 GB
no majority of d781e30000018a
53.2 % [==================================>                           ] 27 GB / 50 GB
...
```

Sheepdog的数据丢失问题

美团云
Meituan Open Services

```
[root@node3-10gtest ~]# collie cluster check
fix vdi test1-3
99.9 % [=============================================================>] 50 GB / 50 GB
fixed replica 3e563000000fca
99.9 % [=============================================================>] 50 GB / 50 GB
fixed replica 3e563000000fec
100.0 % [============================================================>] 50 GB / 50 GB
fixed replica 3e5630000026f5
100.0 % [============================================================>] 50 GB / 50 GB
fixed replica 3e563000002da6
100.0 % [============================================================>] 50 GB / 50 GB
fixed replica 3e563000001e8c
100.0 % [============================================================>] 50 GB / 50 GB
fixed replica 3e563000001530
...
fix vdi test2-9
50.9 % [================================>                             ] 25 GB / 50 GB
no majority of d781e300000123
51.2 % [=================================>                            ] 26 GB / 50 GB
no majority of d781e30000018a
53.2 % [==================================>                           ] 27 GB / 50 GB
...
```

内容提要

美团云
Meituan Open Services

- 个人简介
- 云计算@美团
- 云计算对块存储的需求
- 设计与实现
  - ✤ 设计理念、存储结构、并发模型、
  - ✤ 系统架构、写入策略、错误防护
- 性能测试

设计理念

美团云
Meituan Open Services

- 稳定可靠
  - ★ 在做trade off时，稳定压倒一起
- 高性能
  - ★ 能发挥万兆网和SSD的潜力
- 勤俭
  - ★ 低人力物力投入
    - ＊ 能使用单路低端CPU
    - ＊ 尽可能重用已有服务

设计理念

- CAP取舍
  - 舍C，取AP
- 服务无状态
  - stateless, timeless or functional
- 简化并发模型
- 模块化、层次化

系统架构

美团云
Meituan Open Services

- 有中心（元数据服务器）
  - ✤ HDFS/GFS, MooseFS, …
- 无中心（元数据服务器）
  - ✤ ceph, gluster, sheepdog, …

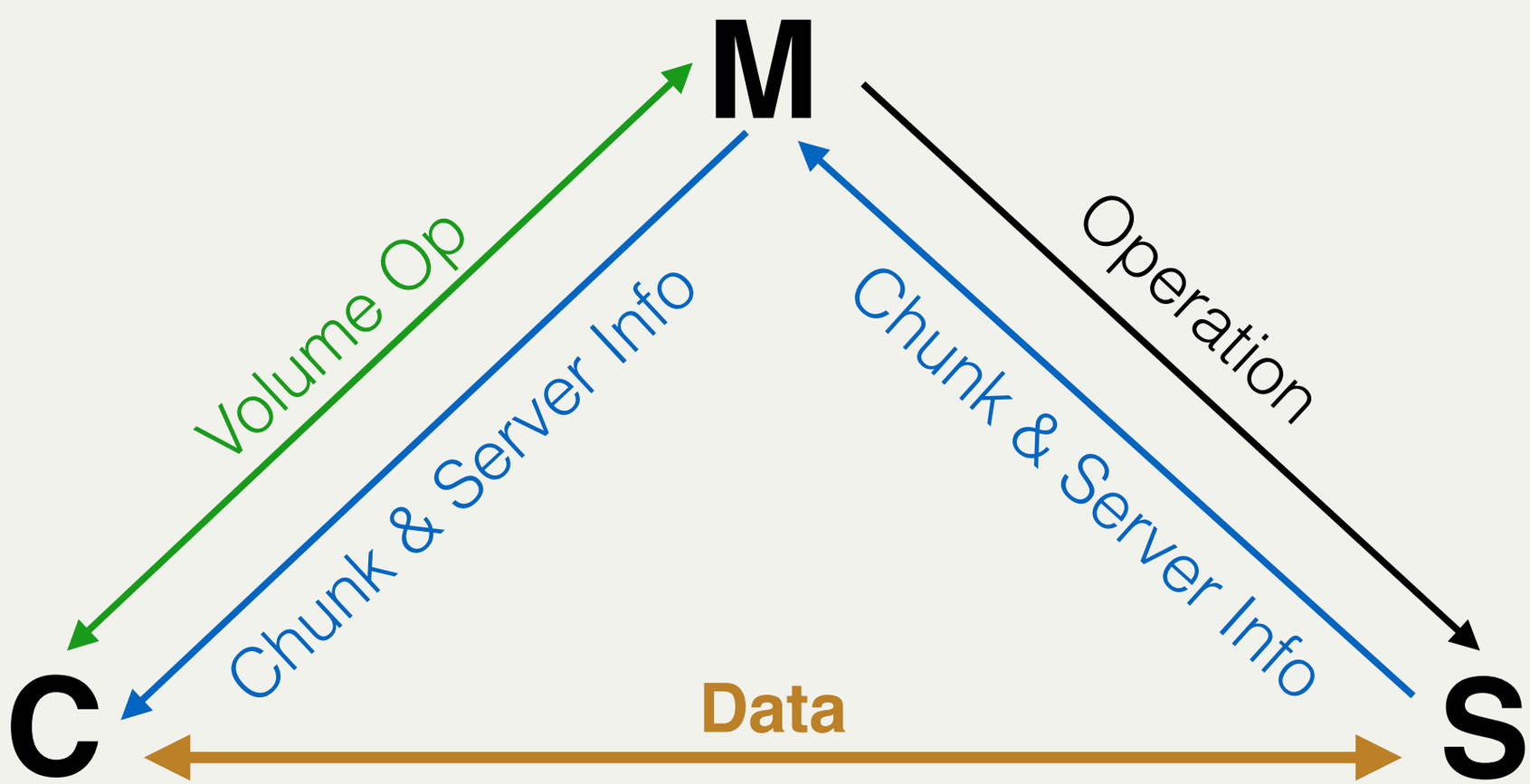
系统架构

美团云
Meituan Open Services

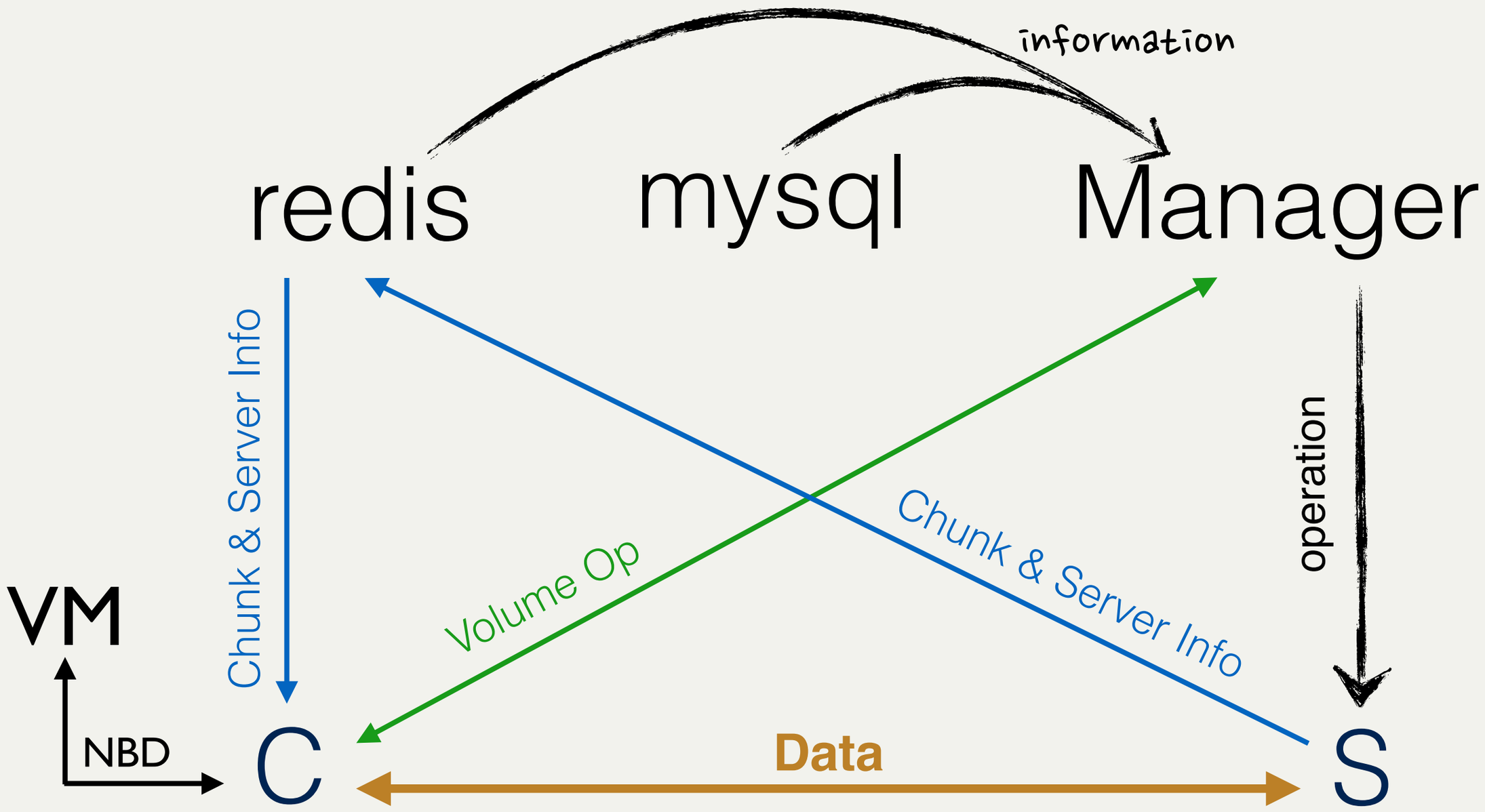
- 有中心（元数据服务器）
  - ✤ HDFS/GFS, MooseFS, …

- 无中心（元数据服务器）
  - ✤ ceph, gluster, sheepdog, …

系统架构


M
Volume Op
Chunk & Server Info
Chunk & Server Info
Operation
C
Data
S

系统架构


information
redis
mysql
Manager
Chunk & Server Info
operation
Volume Op
Chunk & Server Info
VM
NBD
C
Data
S

并发模型

- 事件驱动
  - ✤ select(), poll(), epoll(), ...
- 多线程
- 多进程
  - ✤ 每块硬盘对应一个进程
- 多协程

并发模型

美团云
Meituan Open Services

- 事件驱动 ✓
  - ✤ select(), poll(), epoll(), ...
- 多线程
- 多进程
  - ✤ 每块硬盘对应一个进程
- 多协程

并发模型

美团云
Meituan Open Services

- 事件驱动 ✓
  - ✤ select(), poll(), epoll(), ...
- 多线程
- 多进程 ✓
  - ✤ 每块硬盘对应一个进程
- 多协程

并发模型

美团云
Meituan Open Services

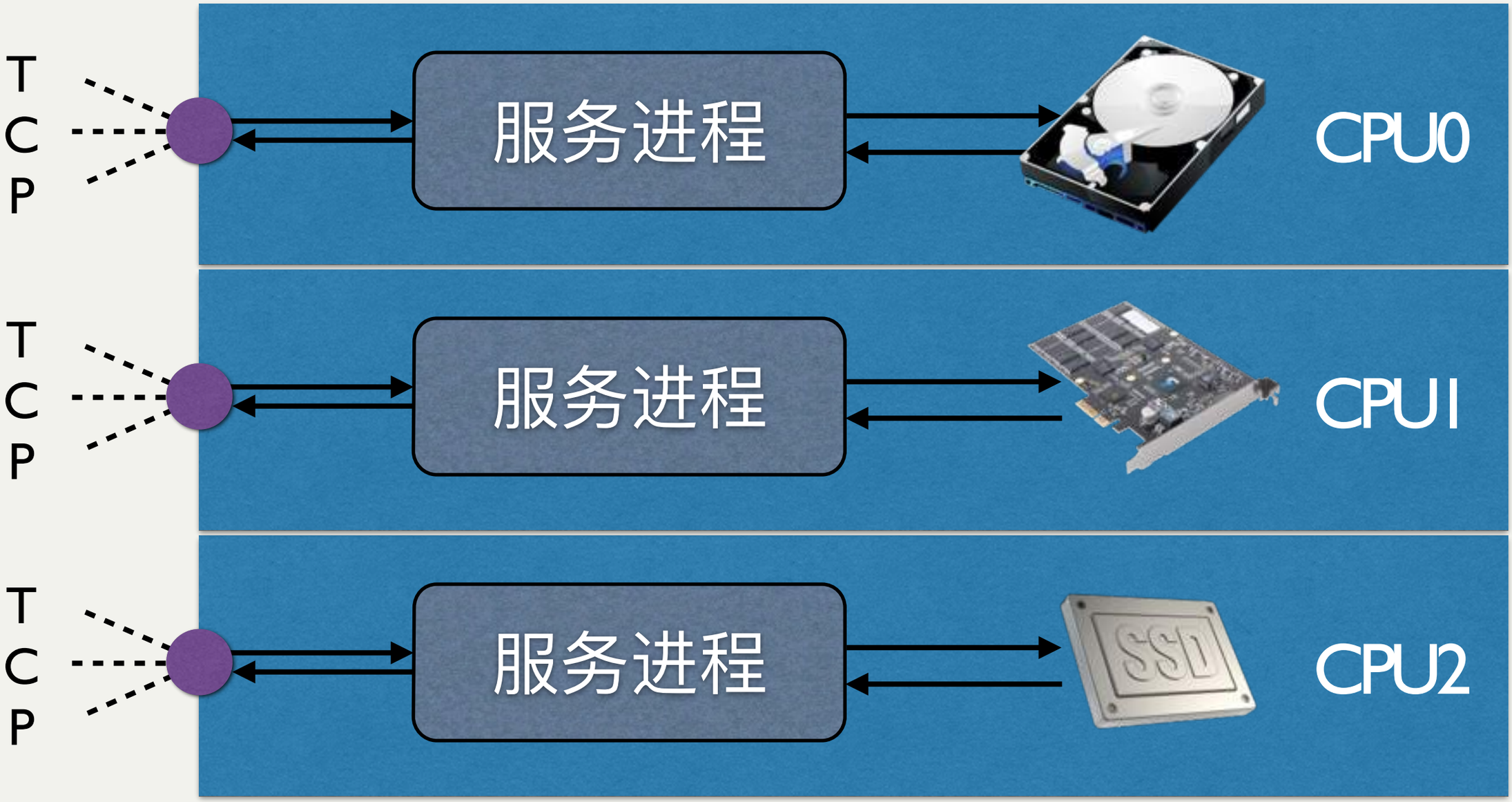
- 事件驱动 ✓
  - ✤ select(), poll(), epoll(), ...
- 多线程
  - ✤ 每块硬盘对应一个进程
- 多进程 ✓
  - ✤ 多协程 ✓

并发模型

- 事件驱动 ✓
  - ✤ select(), poll(), epoll(), ...
- 多线程
- 多进程 ✓
  - ✤ 每块硬盘对应一个进程
- 多协程 ✓

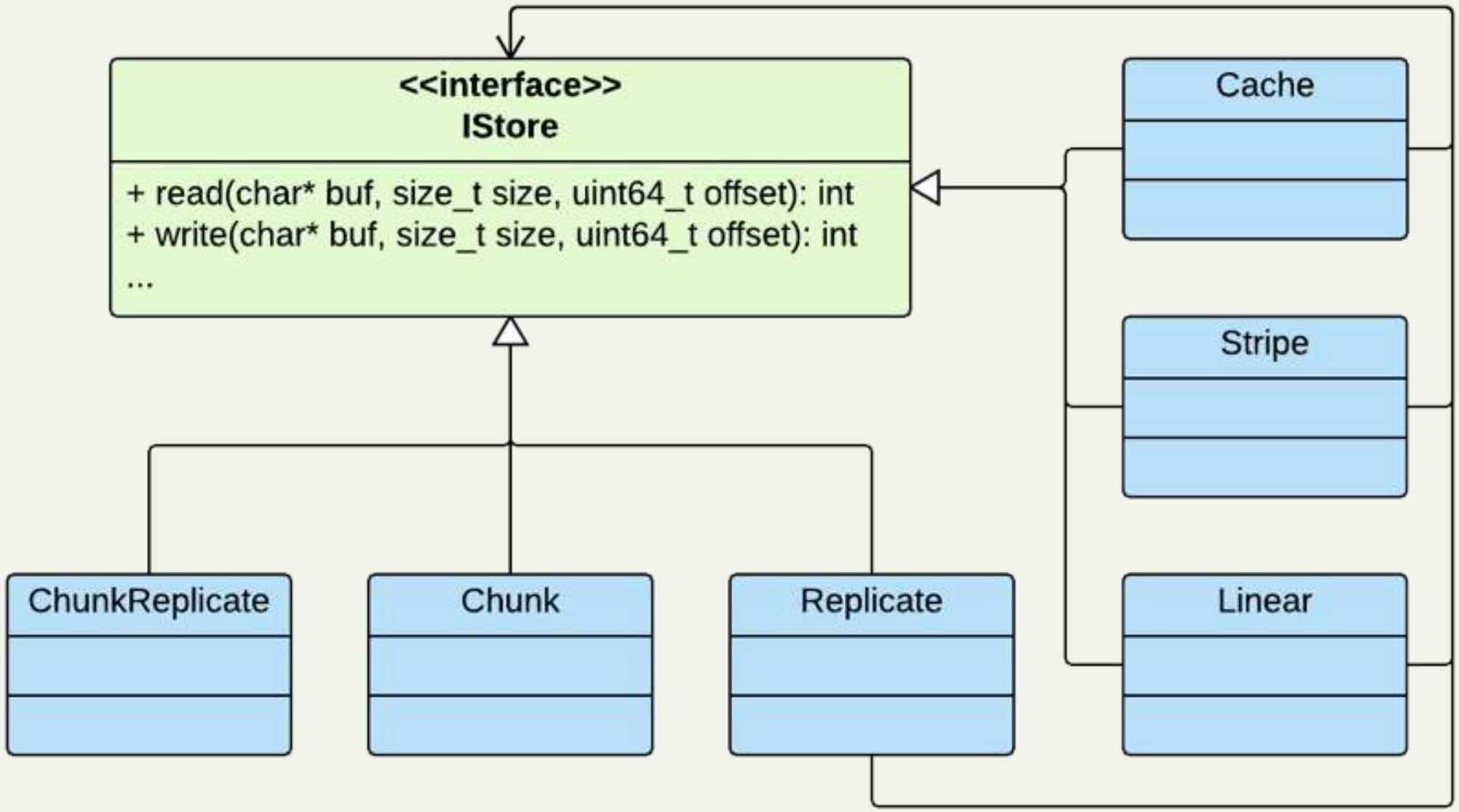
同步化

CPU核心绑定



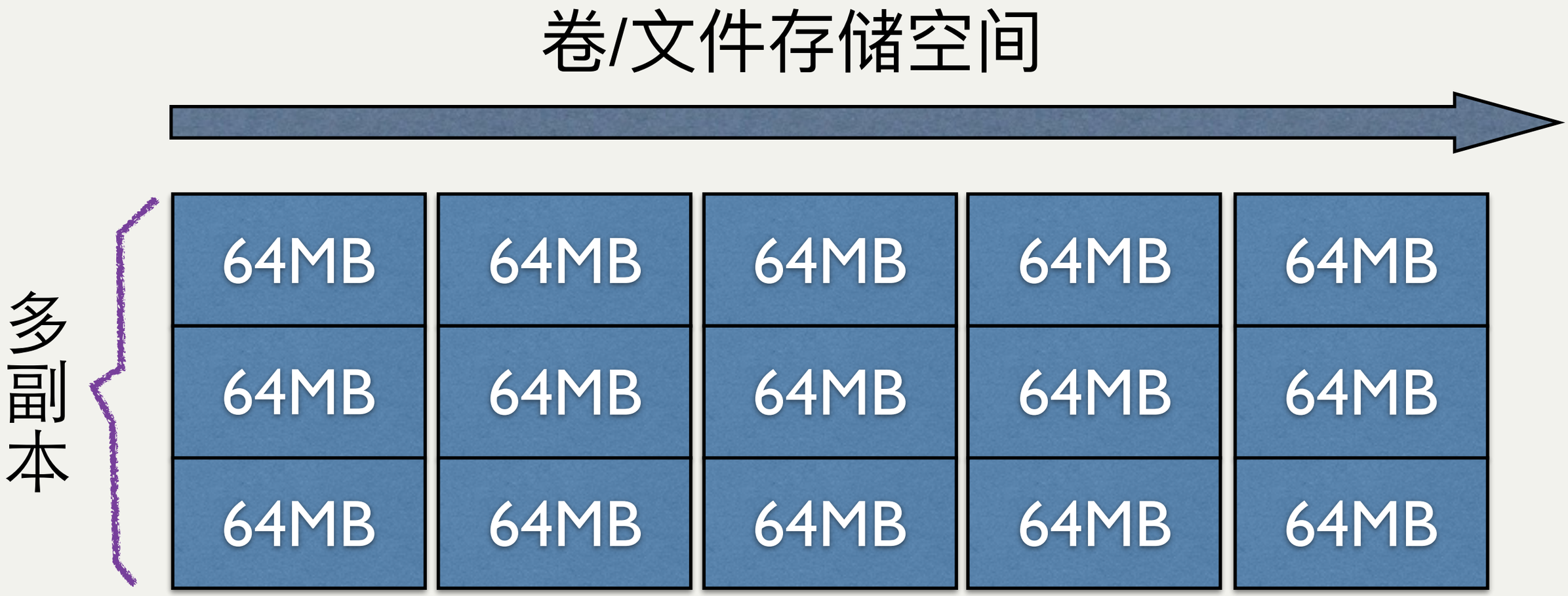
可组合的模块结构

美团云
Meituan Open Services



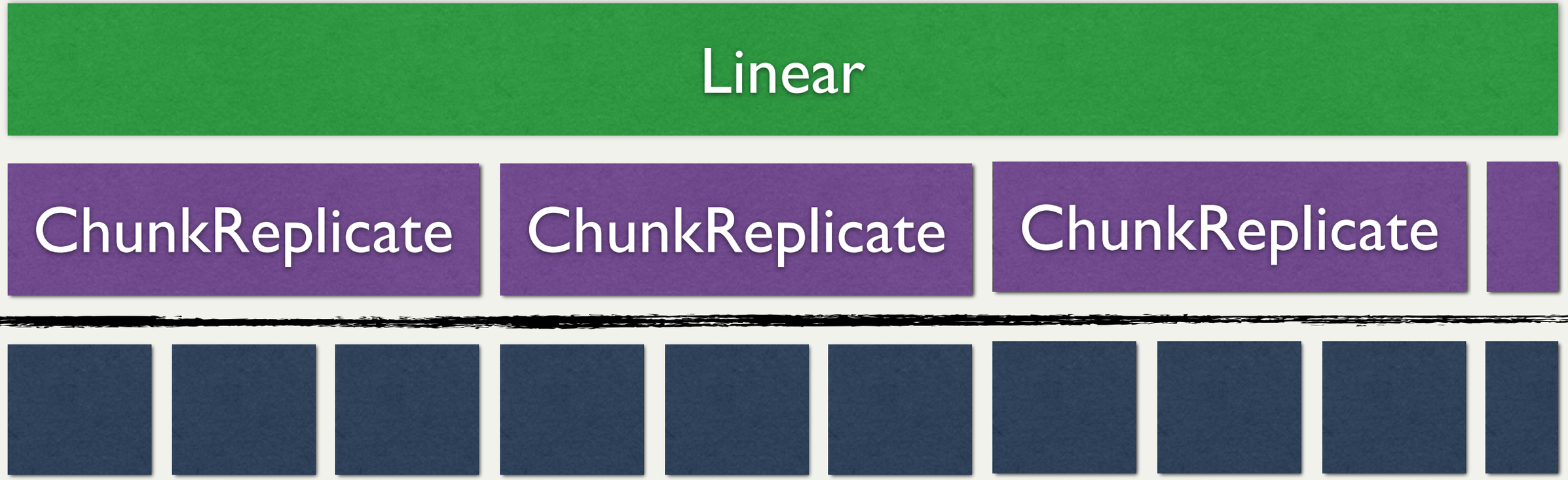
存储结构

美团云
Meituan Open Services



存储结构

卷/文件存储空间



存储结构

卷/文件存储空间

Linear

Stripe

Stripe

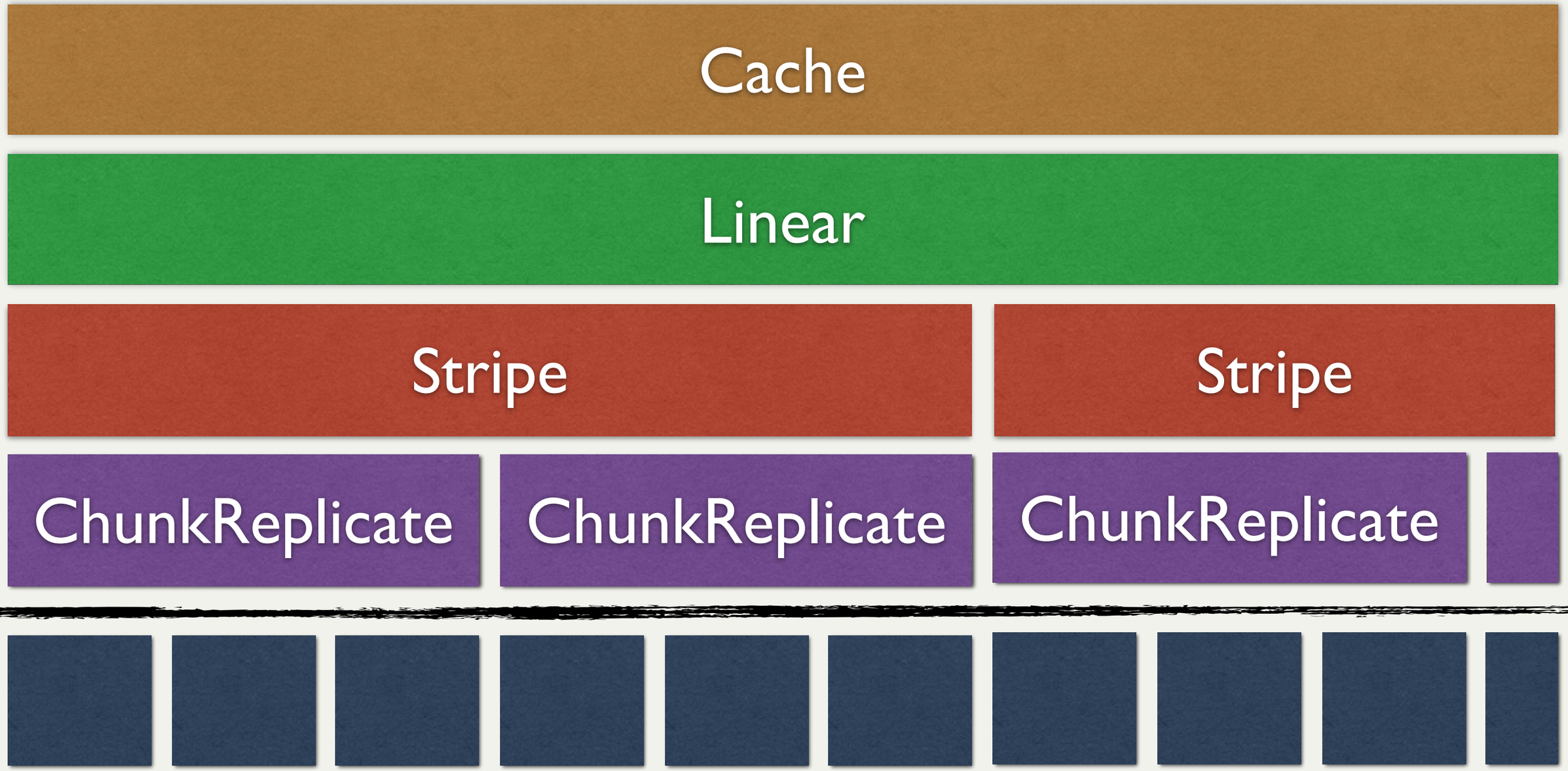
ChunkReplicate

ChunkReplicate

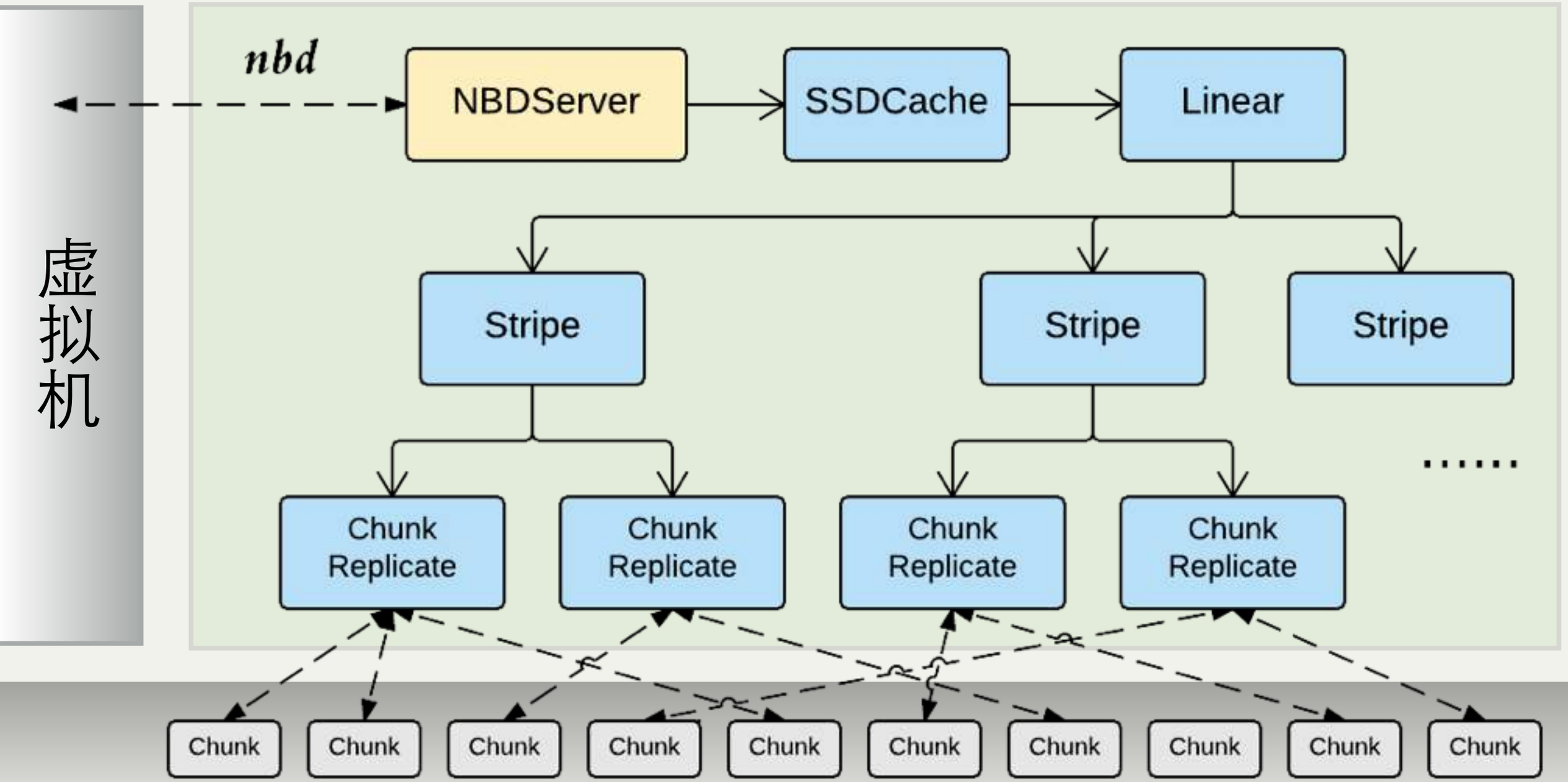
ChunkReplicate

存储结构

美团云
Meituan Open Services

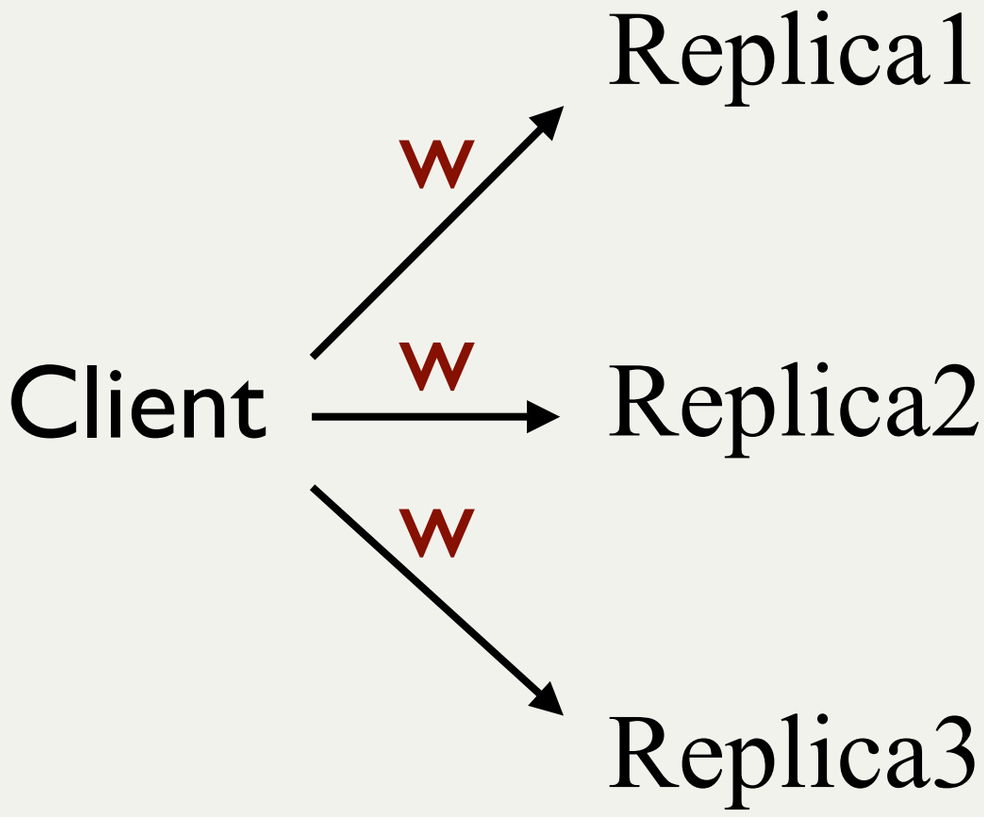


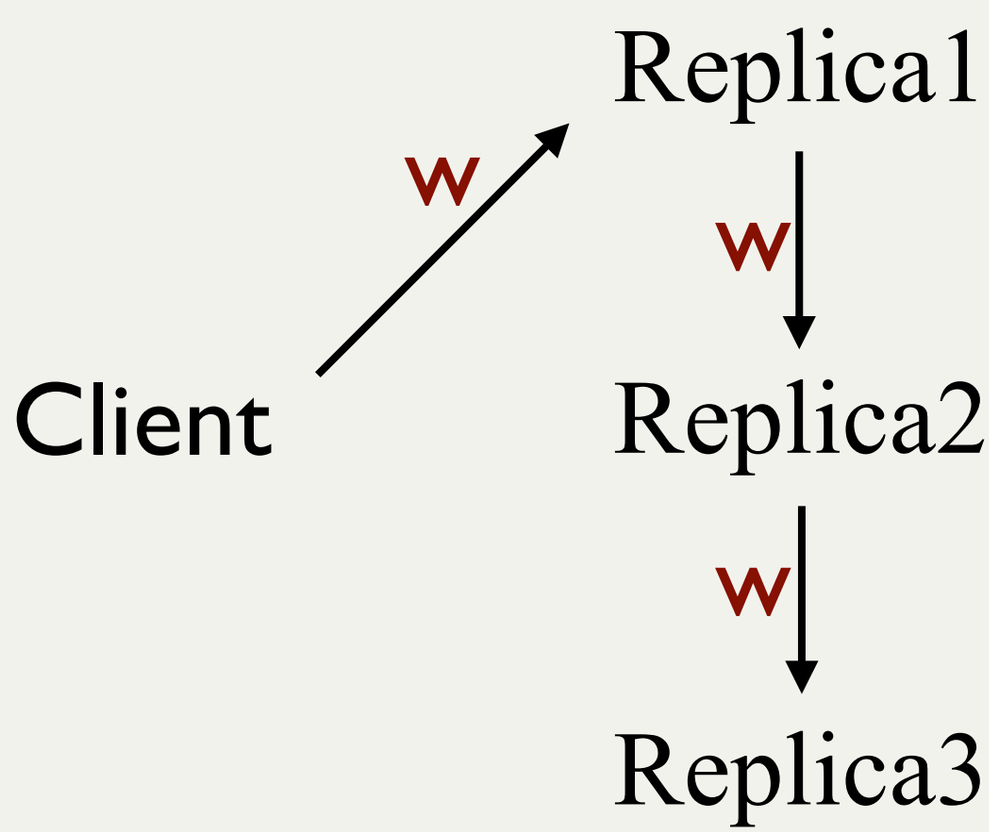
运行时模块结构


虚拟机
nbd
NBDServer
SSDCache
Linear
Stripe
Stripe
Stripe
Chunk Replicate
Chunk Replicate
Chunk Replicate
Chunk Replicate
.....
Chunk
Chunk
Chunk
Chunk
Chunk
Chunk
Chunk
Chunk
Chunk
Chunk

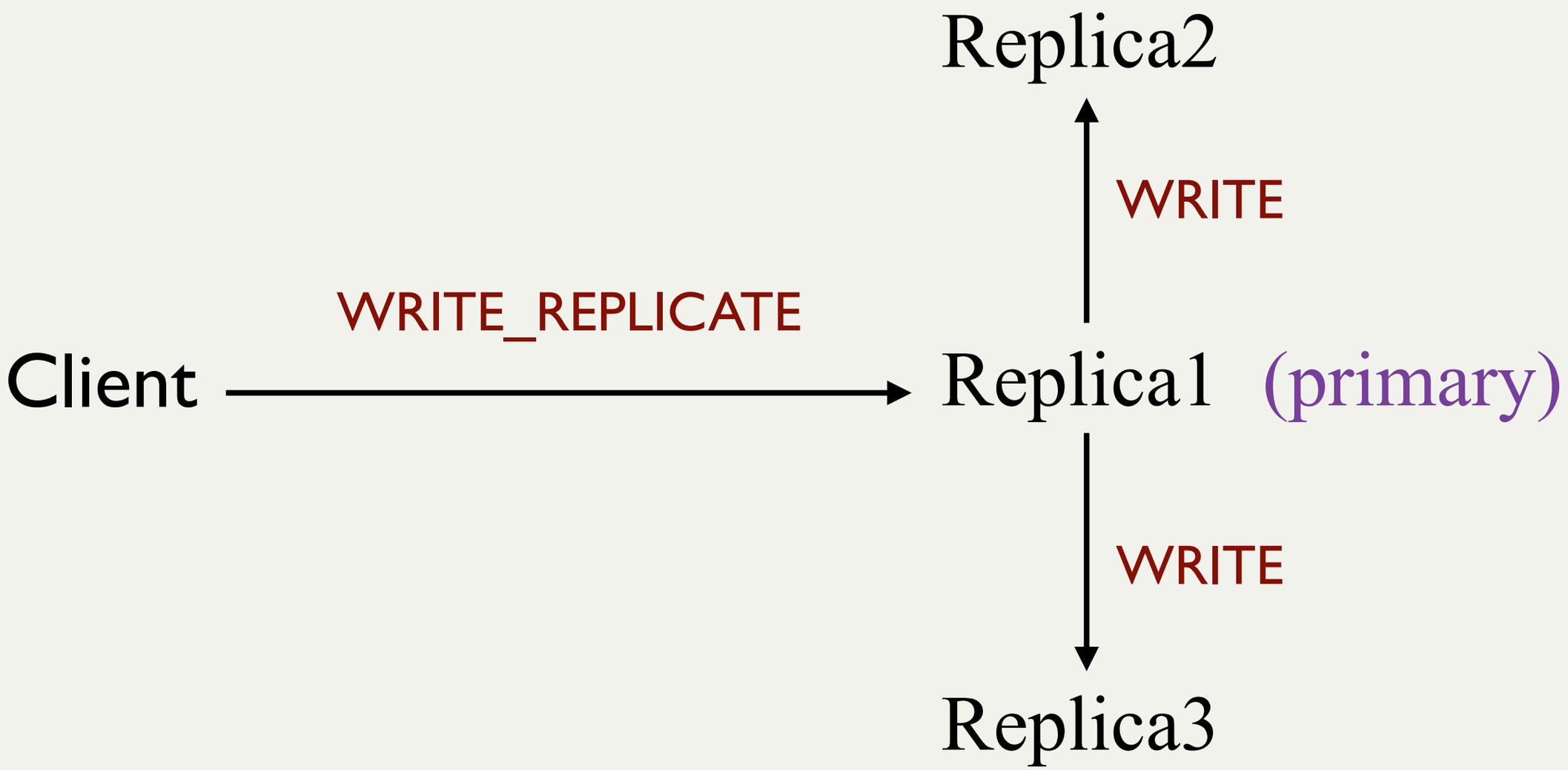
写入策略

美团云
Meituan Open Services

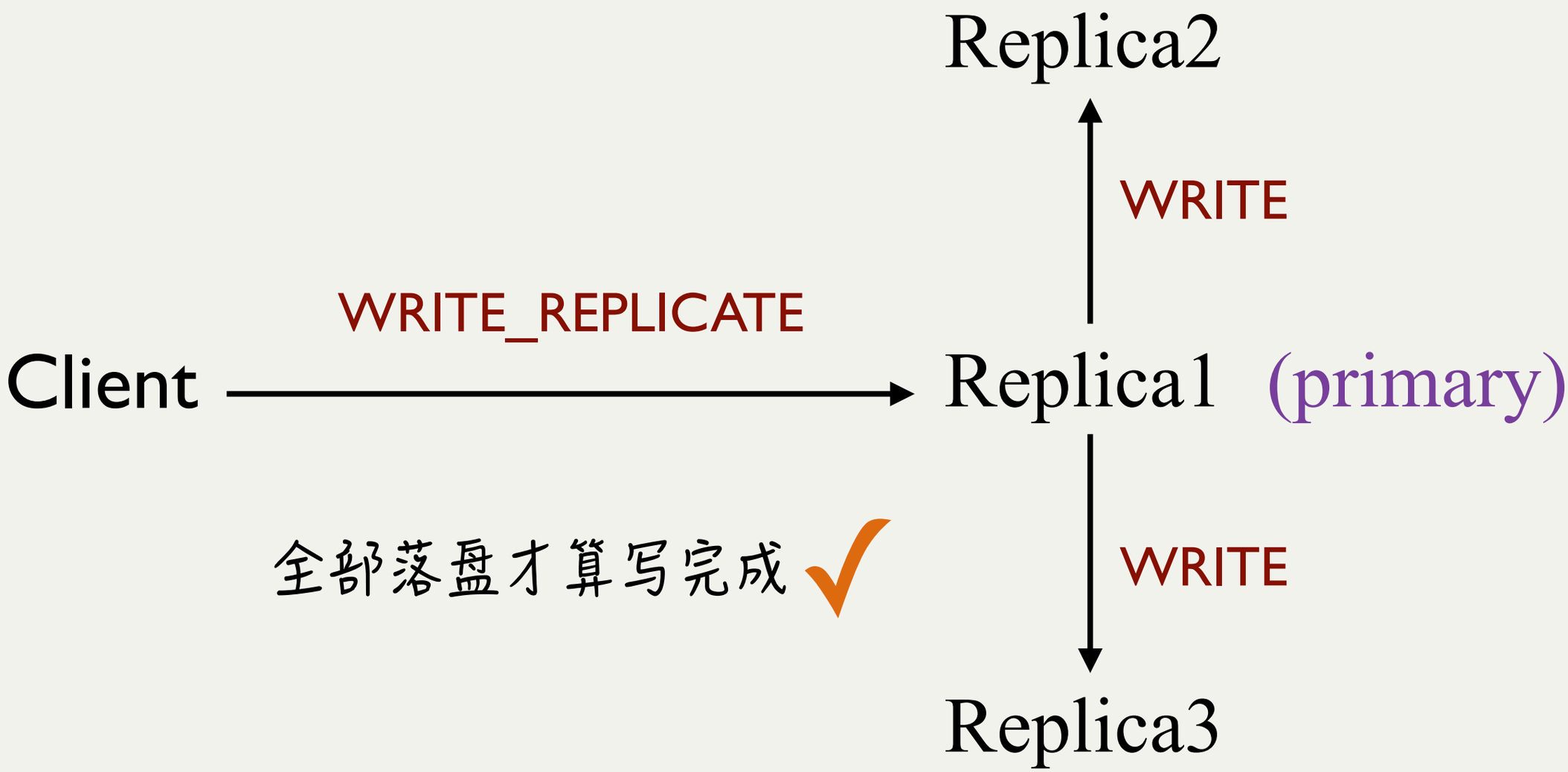




写入策略


Replica2
WRITE
WRITE_REPLICATE
Client
Replica1 (primary)
WRITE
Replica3

写入策略


Replica2
WRITE
WRITE_REPLICATE
Client
Replica1 (primary)
全部落盘才算写完成
WRITE
Replica3

一些“黑科技” 

美团云
Meituan Open Services

一些“黑科技” 👽

Nether Swap
移形换位

一些“黑科技” 👽

美团云
Meituan Open Services

Nether Swap
移形换位

Vanguard
先锋盾

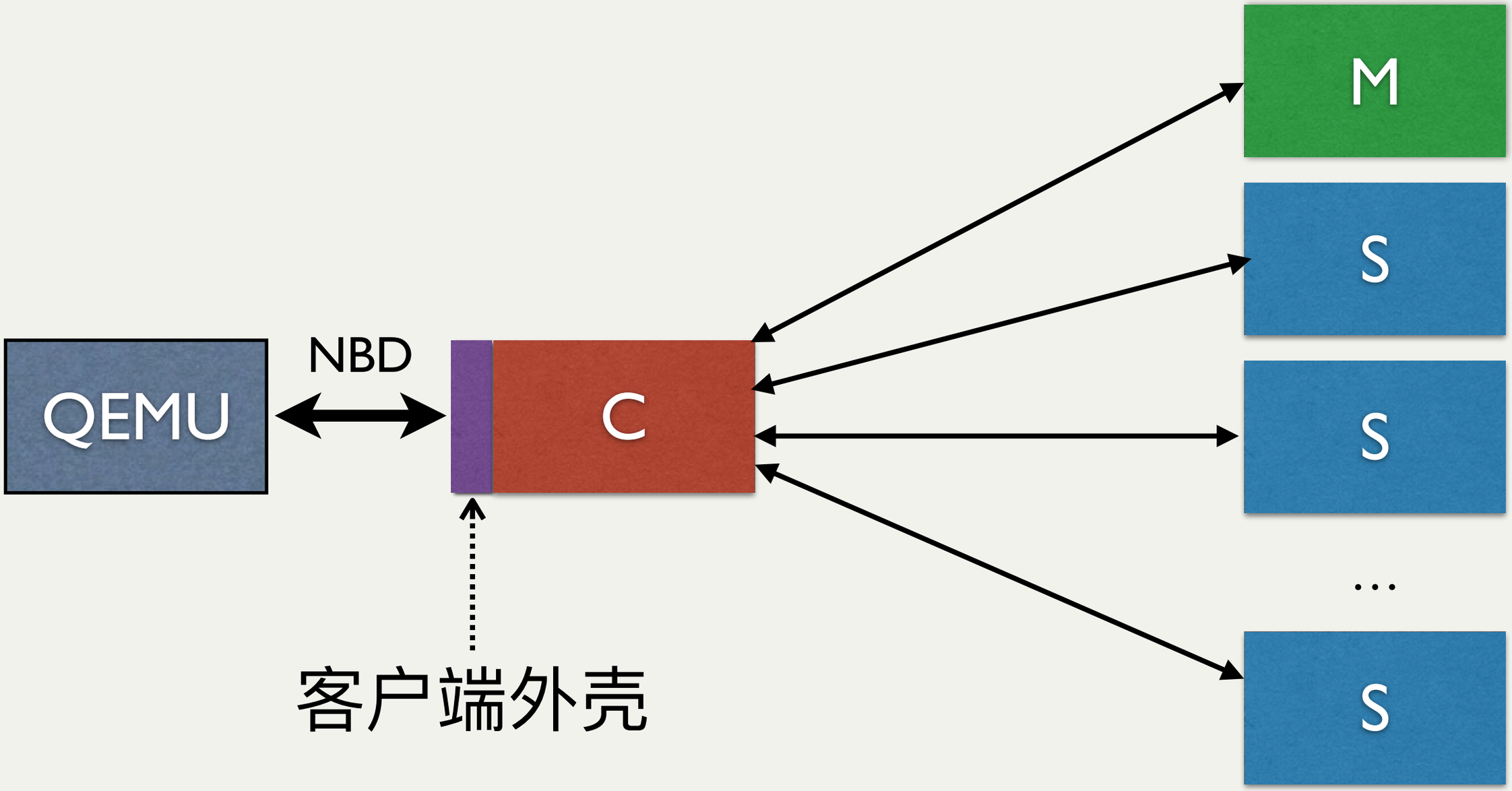
一些“黑科技” 👽

Nether Swap
移形换位

Aegis of the Immortal
不朽之盾

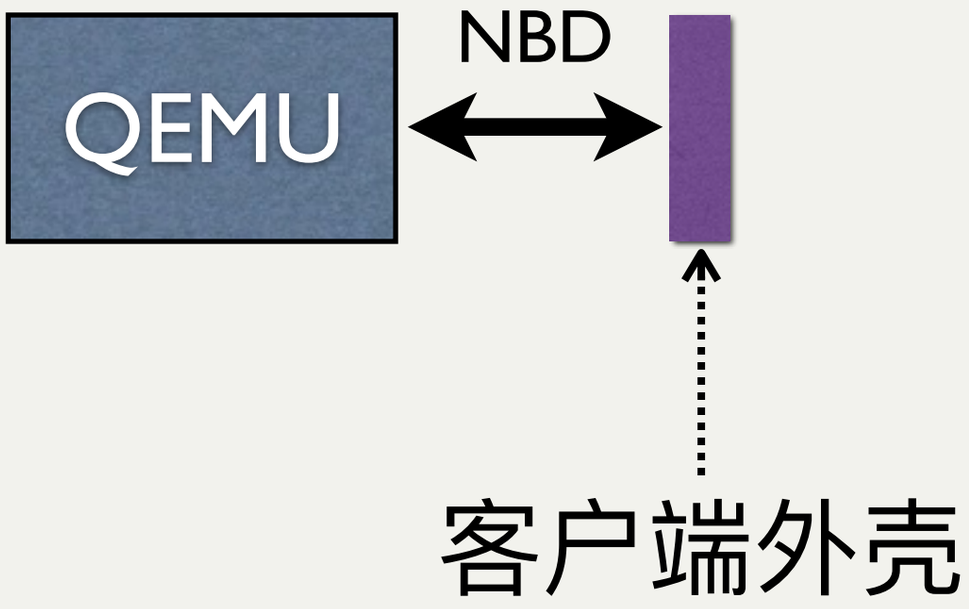
Vanguard
先锋盾

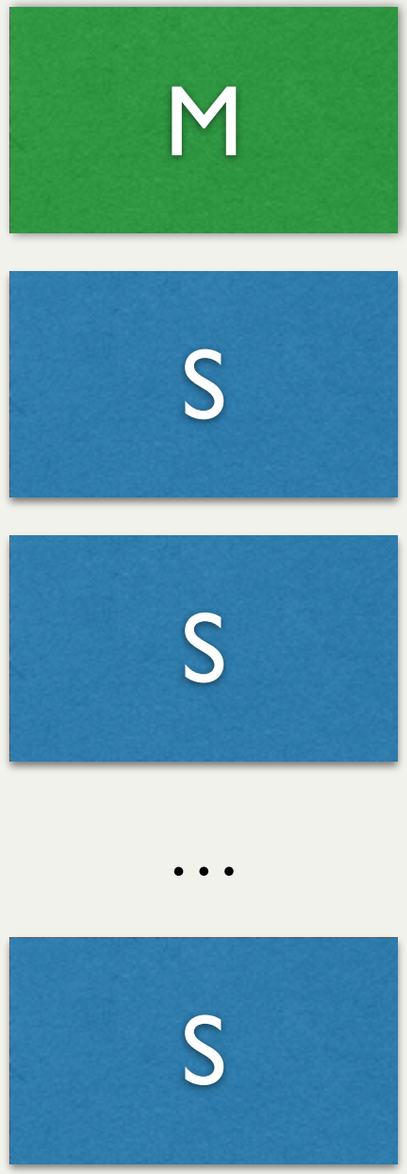
移形换位——热升级



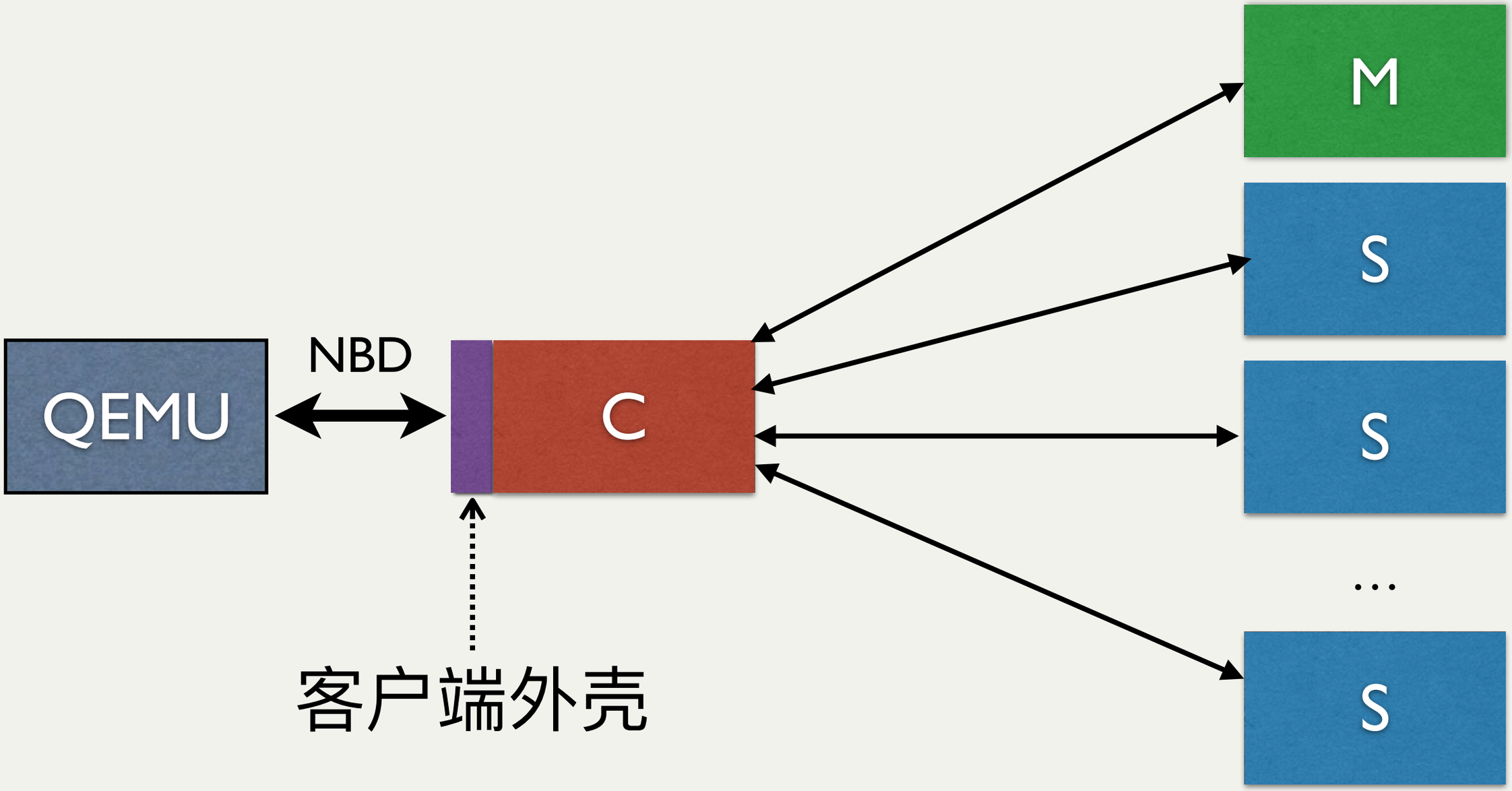
移形换位——热升级

美团云
Meituan Open Services





移形换位——热升级



先锋盾——抵挡普通伤害

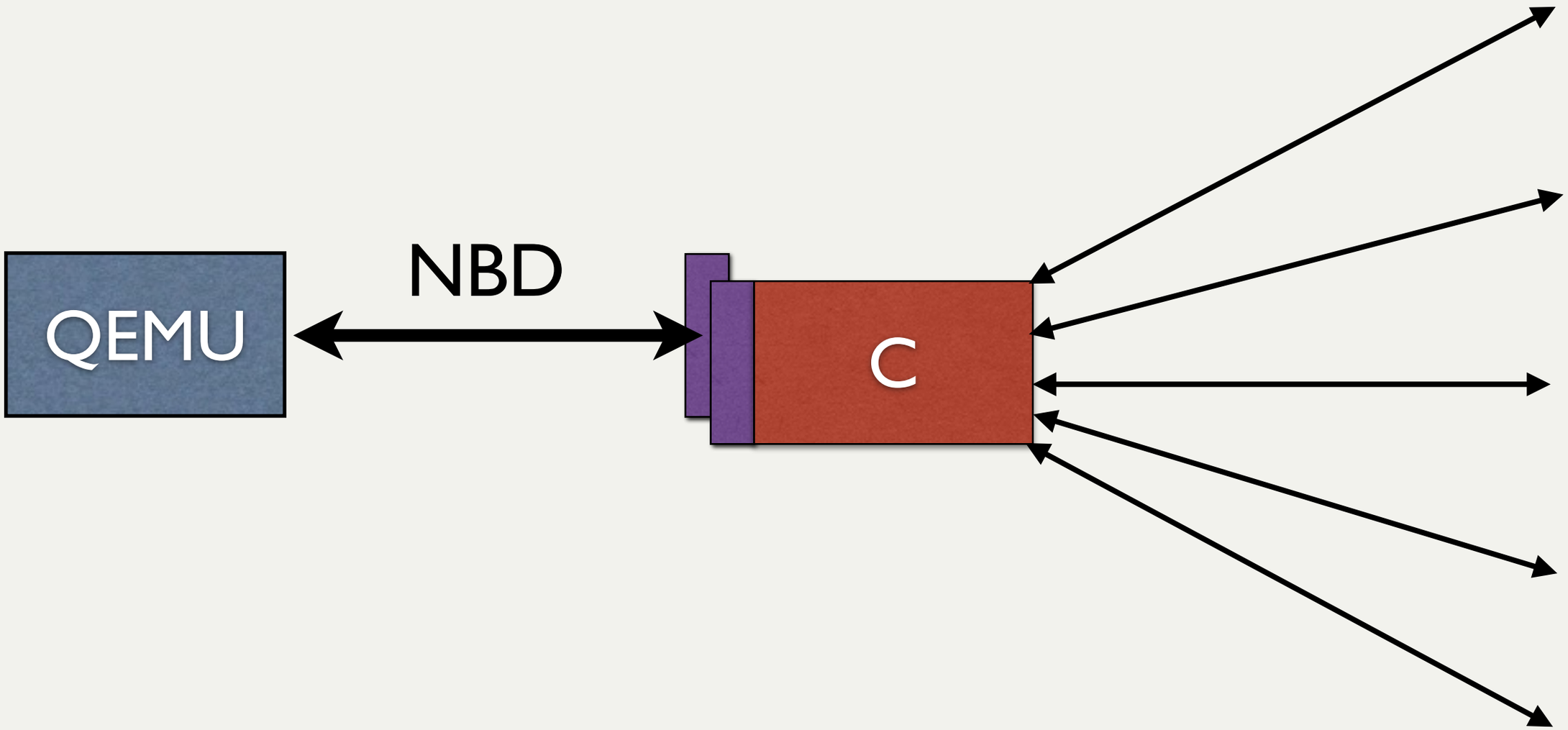
美团云
Meituan Open Services

- 捕获一切异常信号
  - ✤ 除以0
  - ✤ 内存非法访问
  - ✤ …
- 返回到正确场景（壳）
  - ✤ setjmp() / longjmp()
  - ✤ try { throw } catch(…) { }

不朽之盾——起死回生

美团云
Meituan Open Services

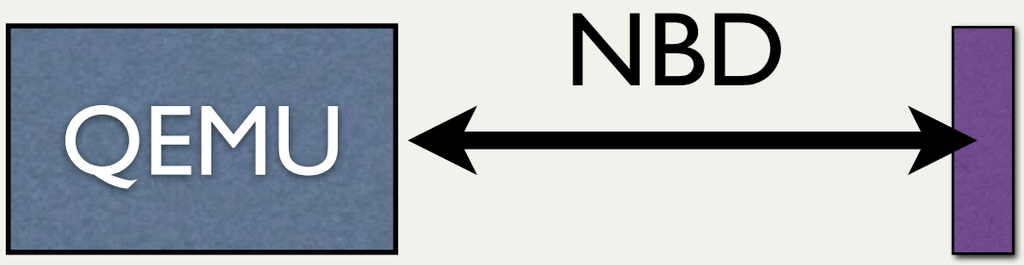
- 任何时候，恢复到绝对正确的运行状态!!!



不朽之盾——起死回生

美团云
Meituan Open Services

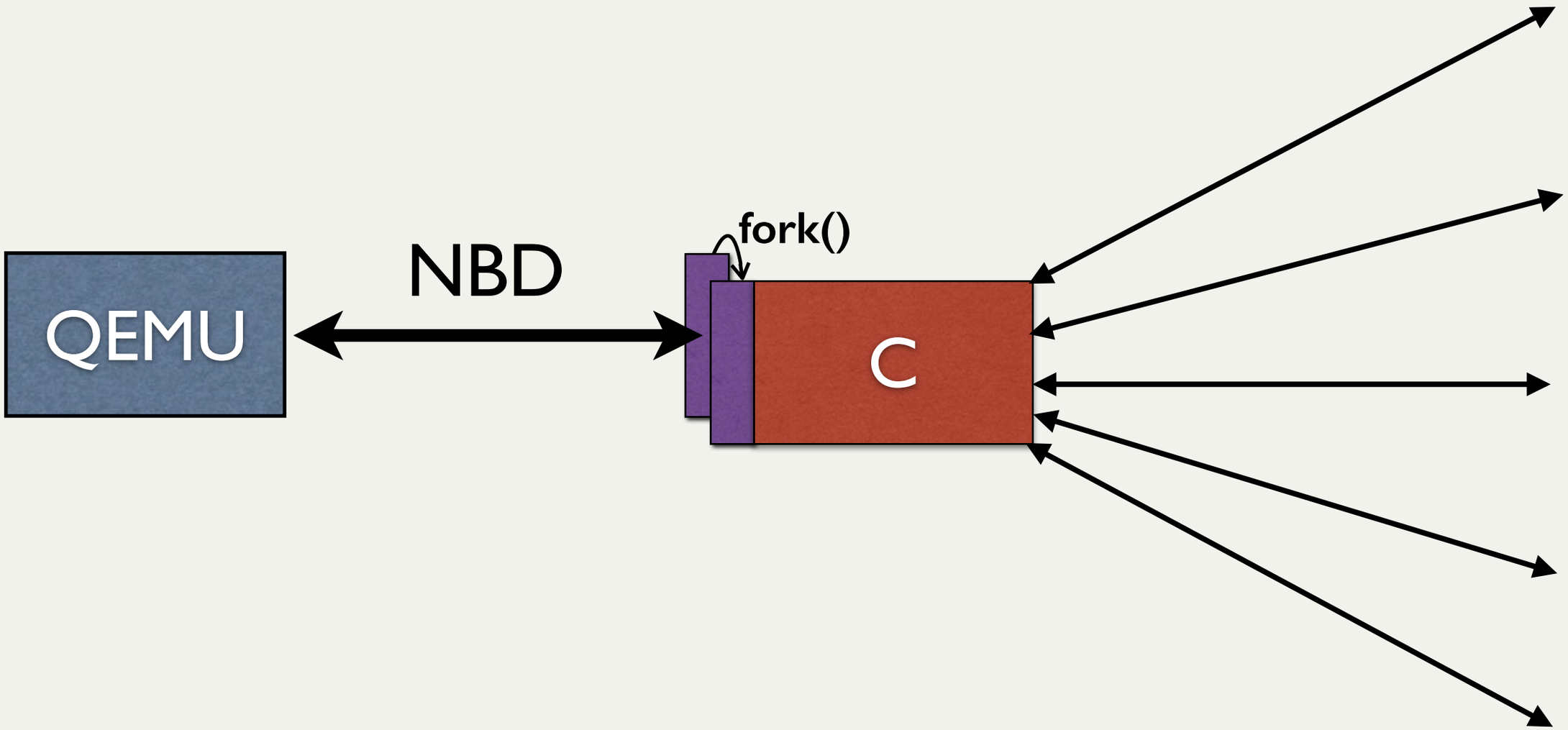
- 任何时候，恢复到绝对正确的运行状态!!!



不朽之盾——起死回生

美团云
Meituan Open Services

- 任何时候，恢复到绝对正确的运行状态!!!

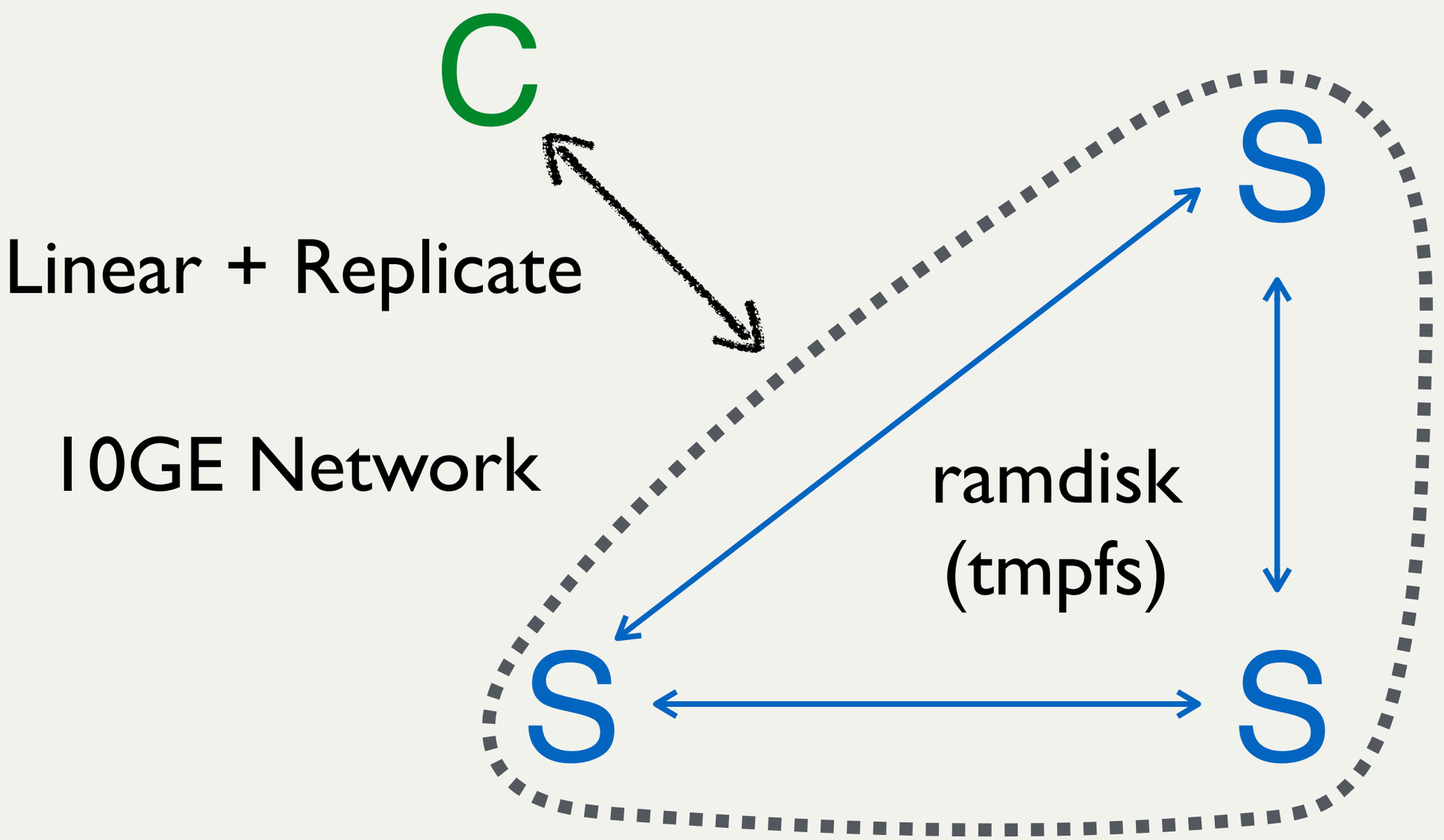


内容提要

美团云
Meituan Open Services

- 个人简介
- 云计算@美团
- 云计算对块存储的需求
- 设计与实现
  - ✤ 设计理念、存储结构、并发模型、
  - ✤ 系统架构、写入策略、错误防护
- 性能测试

测试环境



fio测试结果

美团云
Meituan Open Services

	吞吐率(KB/s)	客户端CPU	服务端CPU（x3）
读	1,168,282	98%	15%
写	976,428	90%	120%

	IOPS	客户端CPU	服务端CPU（x3）
读	61,340	96%	43%
写	40,438	95%	177%

与Ceph对比效率

美团云
Meituan Open Services

服务端：

21282	root	20	0	809m	80m	11m	S	37.4	0.1	6:38.88 ceph-osd
17771	root	20	0	802m	80m	10m	S	27.8	0.1	4:54.46 ceph-osd
19448	root	20	0	839m	79m	10m	S	27.1	0.1	6:42.14 ceph-osd
18580	root	20	0	824m	82m	10m	S	26.5	0.1	5:52.29 ceph-osd
20392	root	20	0	798m	77m	10m	S	13.9	0.1	4:02.15 ceph-osd

123.7% @ 4,101读IOPS

对比：43% @20,446读IOPS

提升**14**倍效率

客户端：5xx% @16,407读IOPS

对比：96% @61,340读IOPS

提升**21**倍效率

测试结果

	延迟(ms)	
	均值	标准差
读	0.166	0.05911
写	0.279	0.04703

对比：

```
[root@ /]# ping
PING  56(84) bytes of data.
64 bytes from : icmp_seq=1 ttl=61 time=0.141 ms
64 bytes from : icmp_seq=2 ttl=61 time=0.152 ms
64 bytes from : icmp_seq=3 ttl=61 time=0.169 ms
64 bytes from : icmp_seq=4 ttl=61 time=0.091 ms
64 bytes from : icmp_seq=5 ttl=61 time=0.117 ms
64 bytes from : icmp_seq=6 ttl=61 time=0.169 ms
^C
```

Q&A

QCon

全球软件开发大会

美团云

Meituan Open Services

谢谢！