美团云
Meituan Open Services

美团云混合存储系统

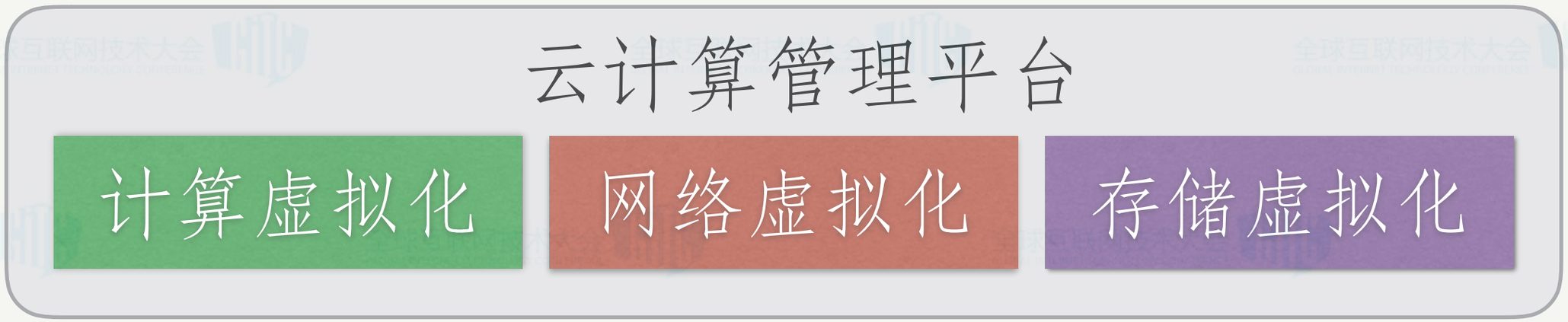
李慧霸

2016.11.24

自我介绍

美团云
Meituan Open Services

- 李慧霸 博士
- 美团云架构师
- 块存储系统负责人
- 曾经在国防科大计算机学院工作
  - 负责天河一号云平台建设
  -



新美大和美团云

美团云
Meituan Open Services

- 消费场景

到店

到店餐饮
到店综合
猫眼电影

到家

外卖配送

在途

酒店旅游

新美大和美团云

美团云
Meituan Open Services

美团开放云服务上线

2013年
5月

正式进入企业及IT服务领域
秉承美团技术输出核心价值
共建美团合作伙伴生态圈

全面承载美团业务

2013年
07月

全网IDC牌照

2015年
3月

自建机房上线
服务质量更上一层楼

新美大和美团云

美团云
Meituan Open Services

可信云认证

整合的解决方案

信息等级安全保护三级

2015年07月

2015年10月

2016年1月

开放为业界所熟知

美团云大数据全景战略发布

共同信赖美团云

目录

美团云
Meituan Open Services

云上的硬盘

美团云
Meituan Open Services

	容量型	效率型	性能型
IOPS	数百	小几千	两万
吞吐率(MB/s)	数十	一百多	两百多
延迟(ms)	10	3	3
单价(元/GB/月)	低	中	高

云上的硬盘

美团云
Meituan Open Services

- 如果要给100个云硬盘提供4,000并发IOPS能力

	单盘性能(IOPS)	需求量(三副本)	功耗(千瓦)	3年TCO(百万元)
HDD	200	6,000	48	7+3=10
SSD	30,000	40	0.1	0.2

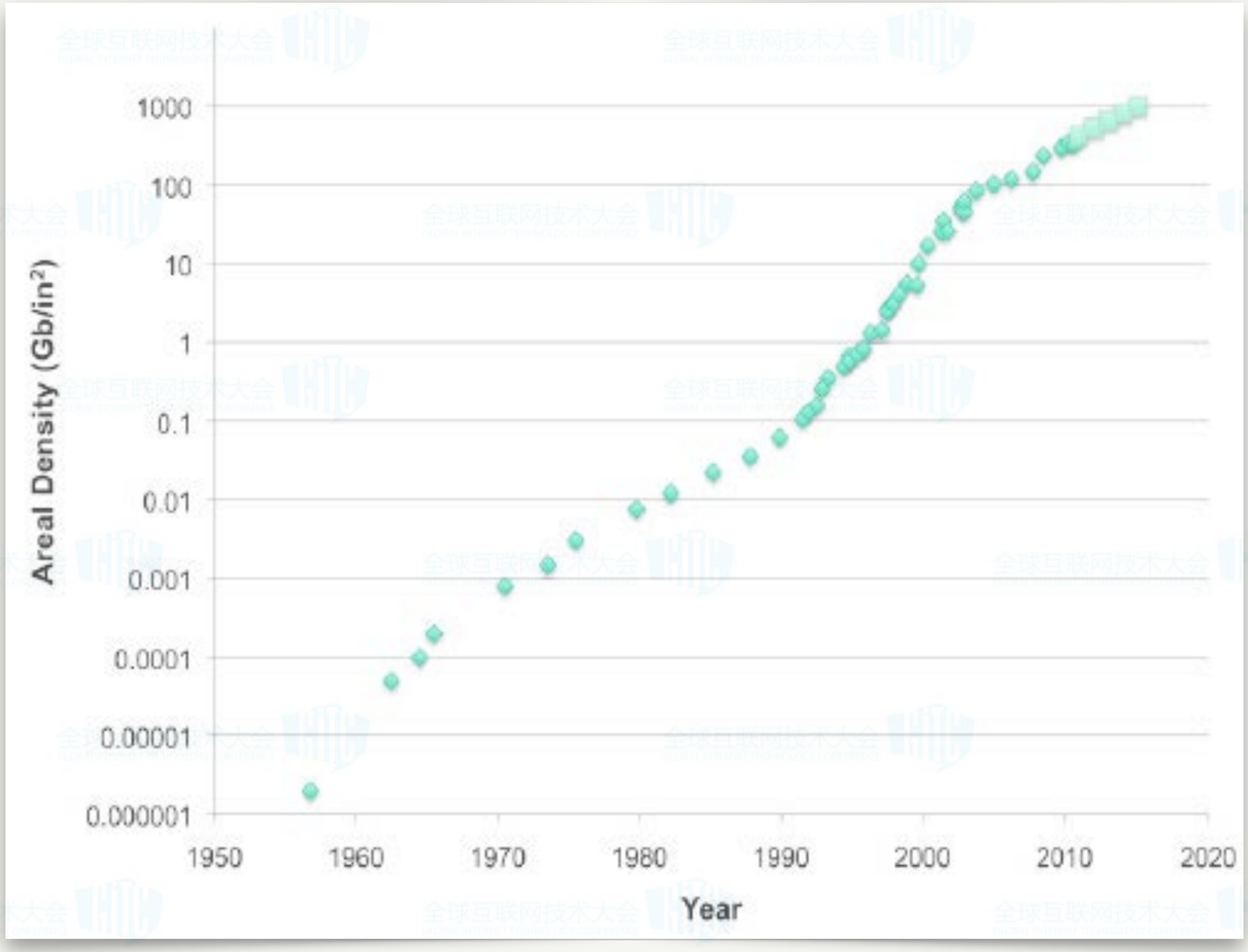
容量：60年增长9个数量级

美团云
Meituan Open Services

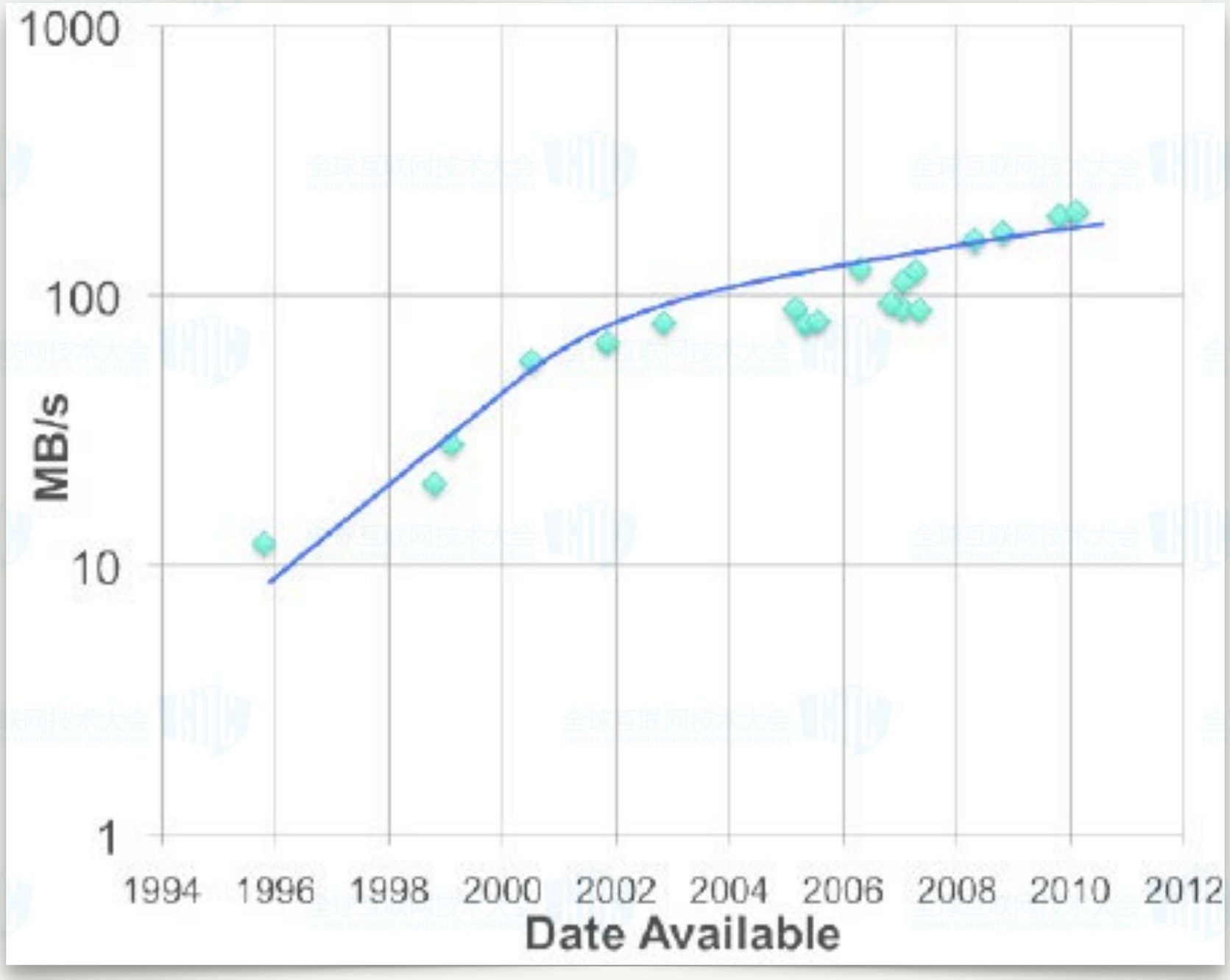
IBM交付5M硬盘，1956

3.5寸10TB硬盘，2016

容量：60年增长9个数量级



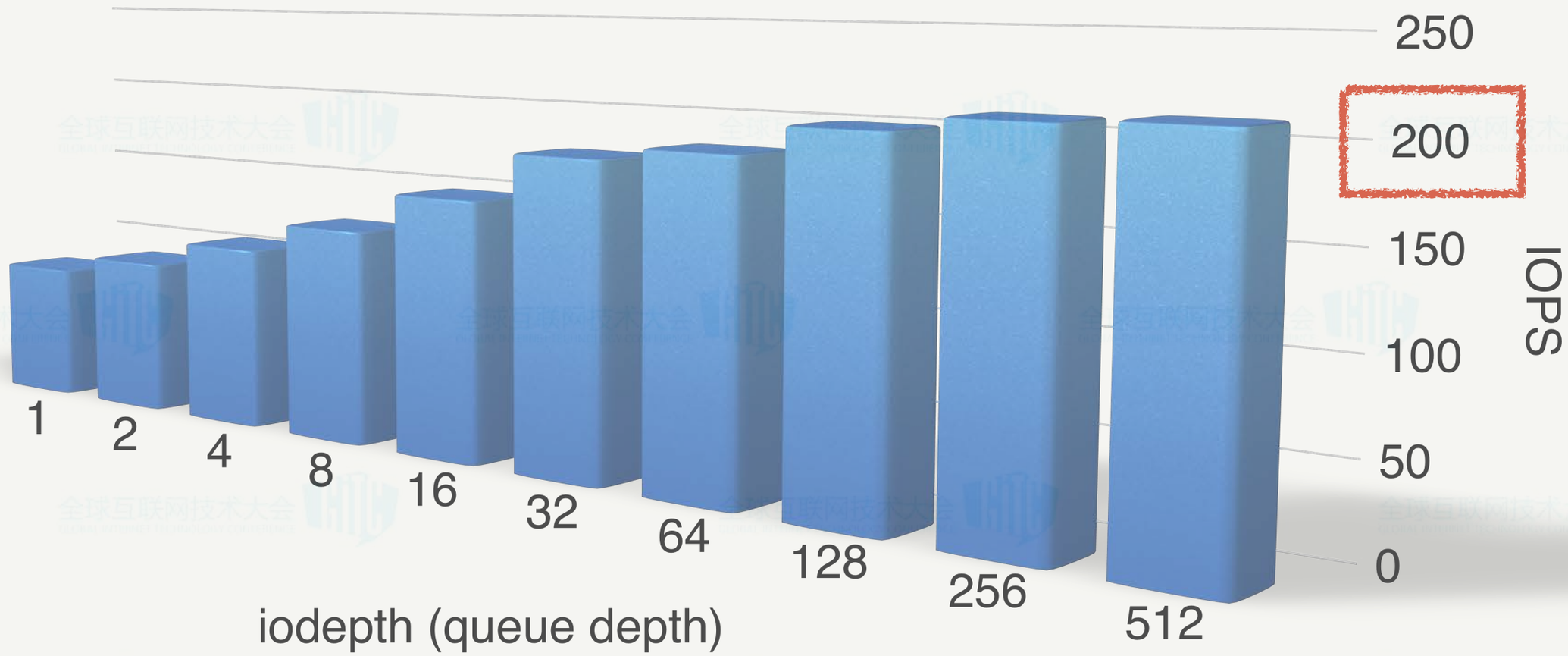
性能：呵呵呵……


1000
100
10
1
MB/s
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
Date Available

HDD性能与容量的失衡

美团云
Meituan Open Services

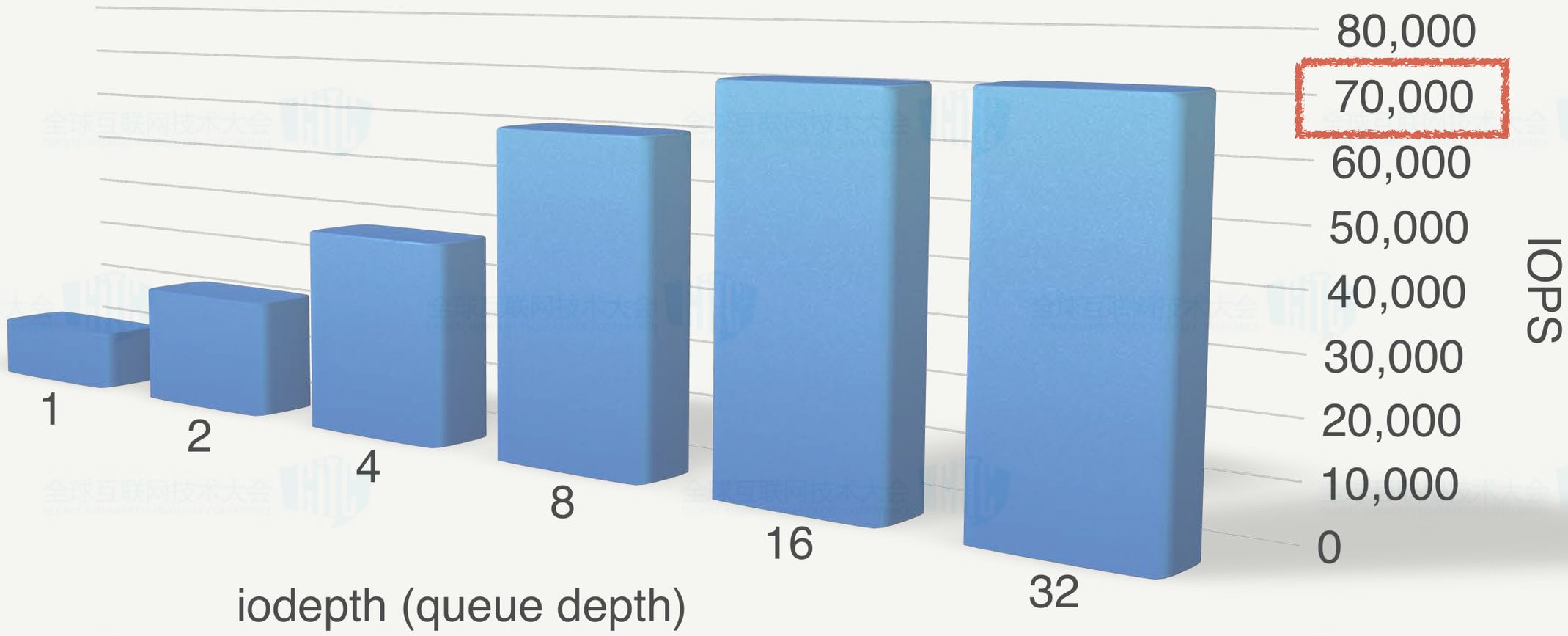
Random IOPS of a typical HDD (3TB, 7200 RPM)



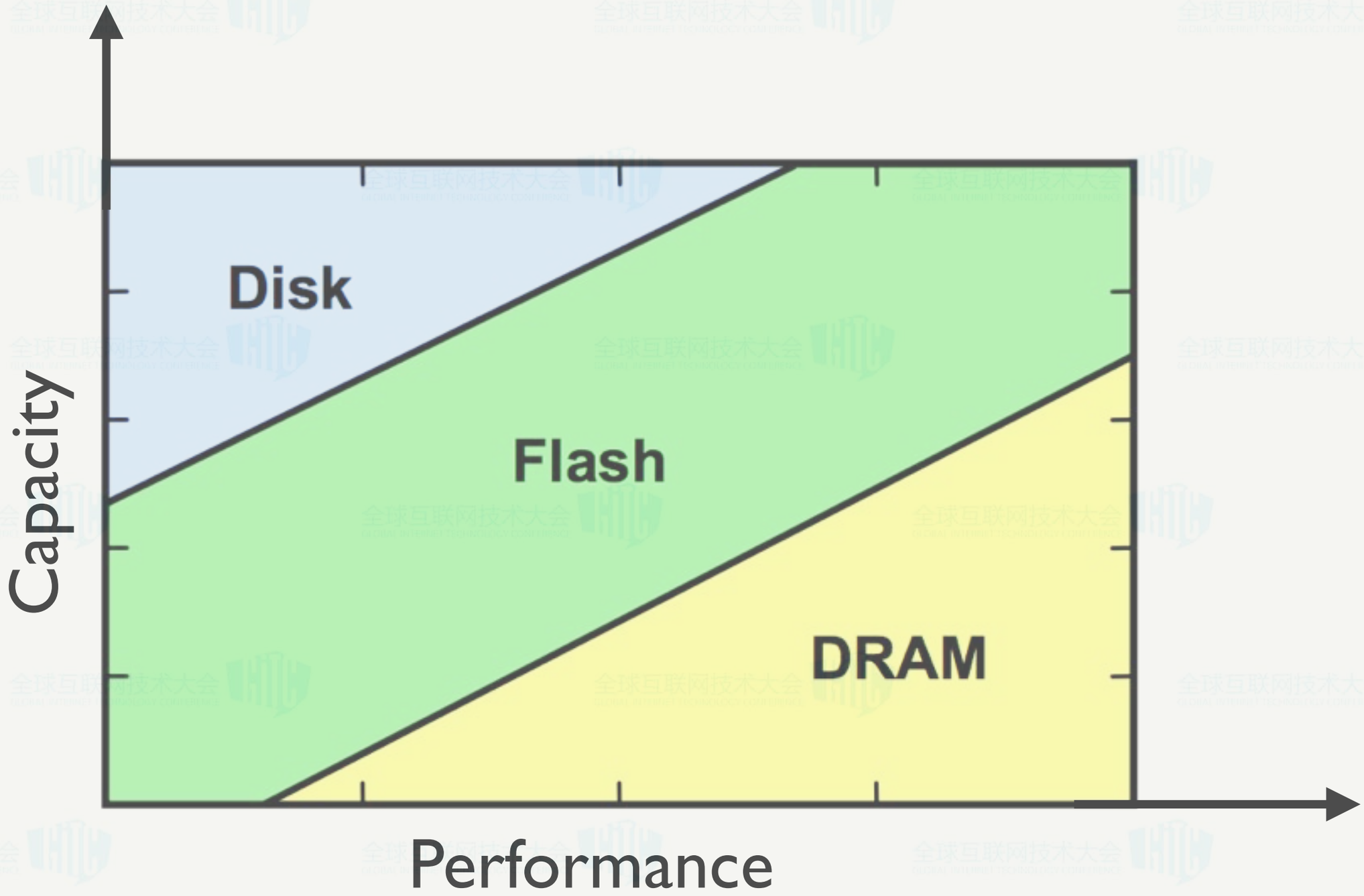
HDD性能与容量的失衡

美团云
Meituan Open Services

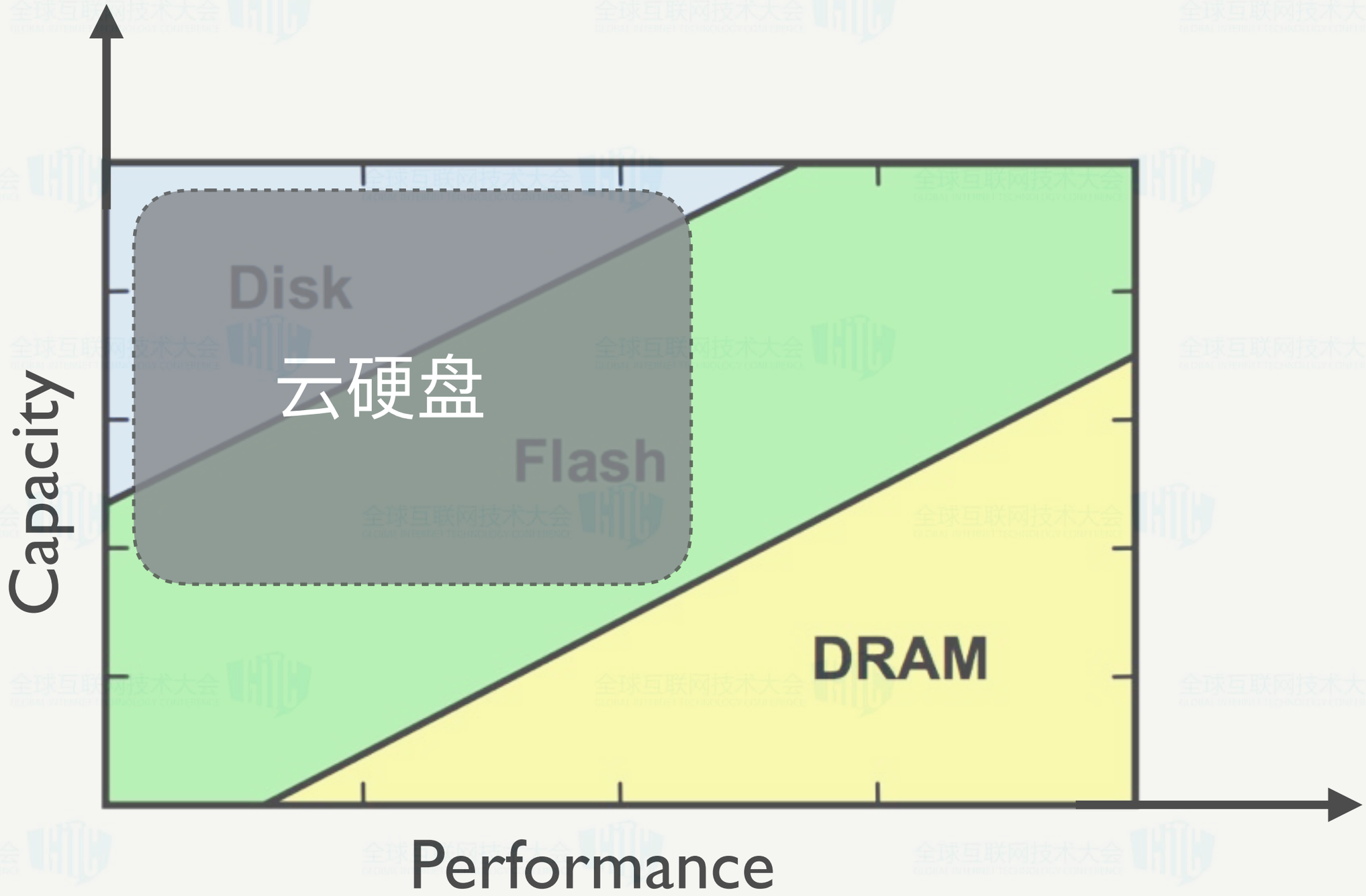
Random Read IOPS of a typical SSD (480GB, SATA)



3年TCO最低：容量与性能



3年TCO最低：容量与性能



目录

美团云
Meituan Open Services

选择：性能与成本

- 价差就是优化空间

¥5319.00

英特尔（Intel）DC S3520 数据中心系列 SSD固态硬盘MLC颗粒SATA3 1.6TB

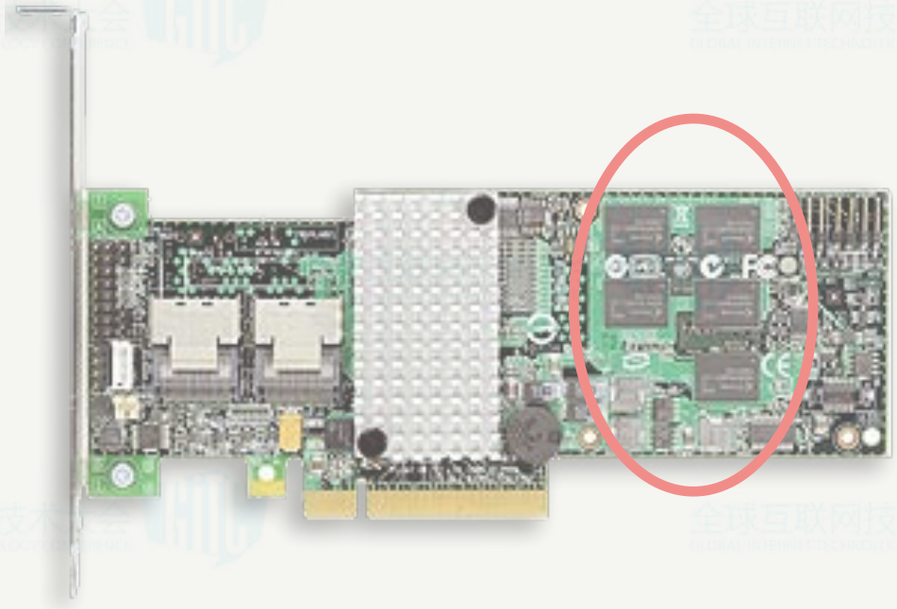
¥ 3324/TB

¥1299.00

希捷(SEAGATE)ES.3系列 4TB 7200转12

¥ 325/TB

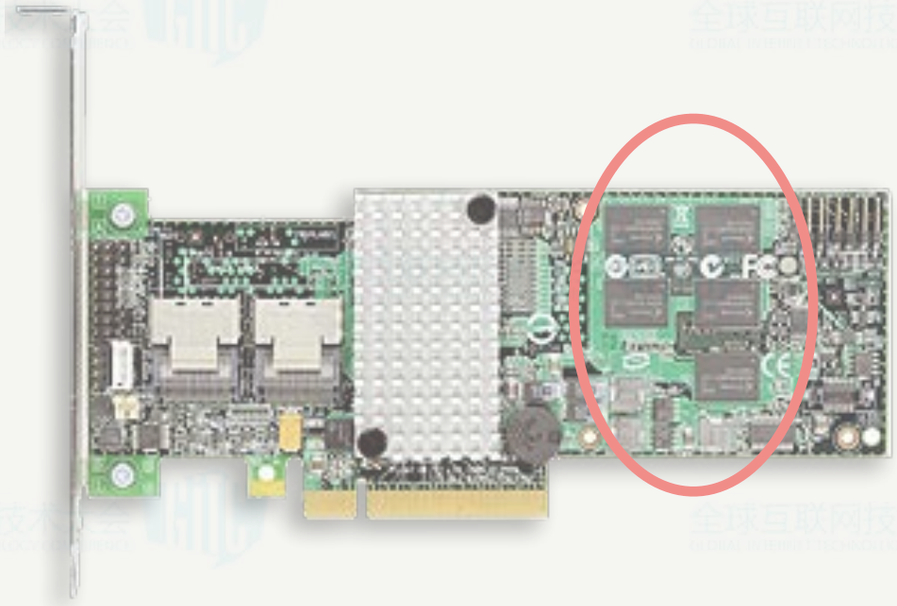
缓存优化

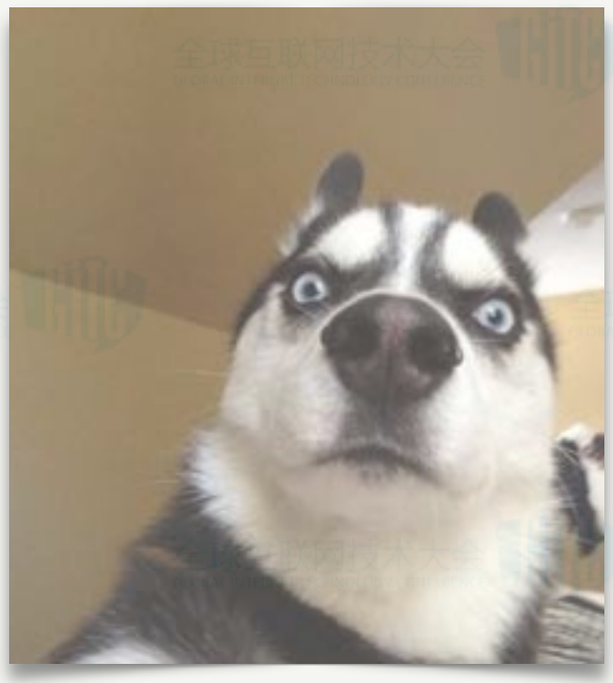


- RAID卡缓存
- 内核缓存模块
  - flashcache
  - lvmcache (dm-cache)
  - bcache
- 性价比高
- 效果受限于命中率
  - 对随机读无效
  - 无法实现产品承诺

缓存优化

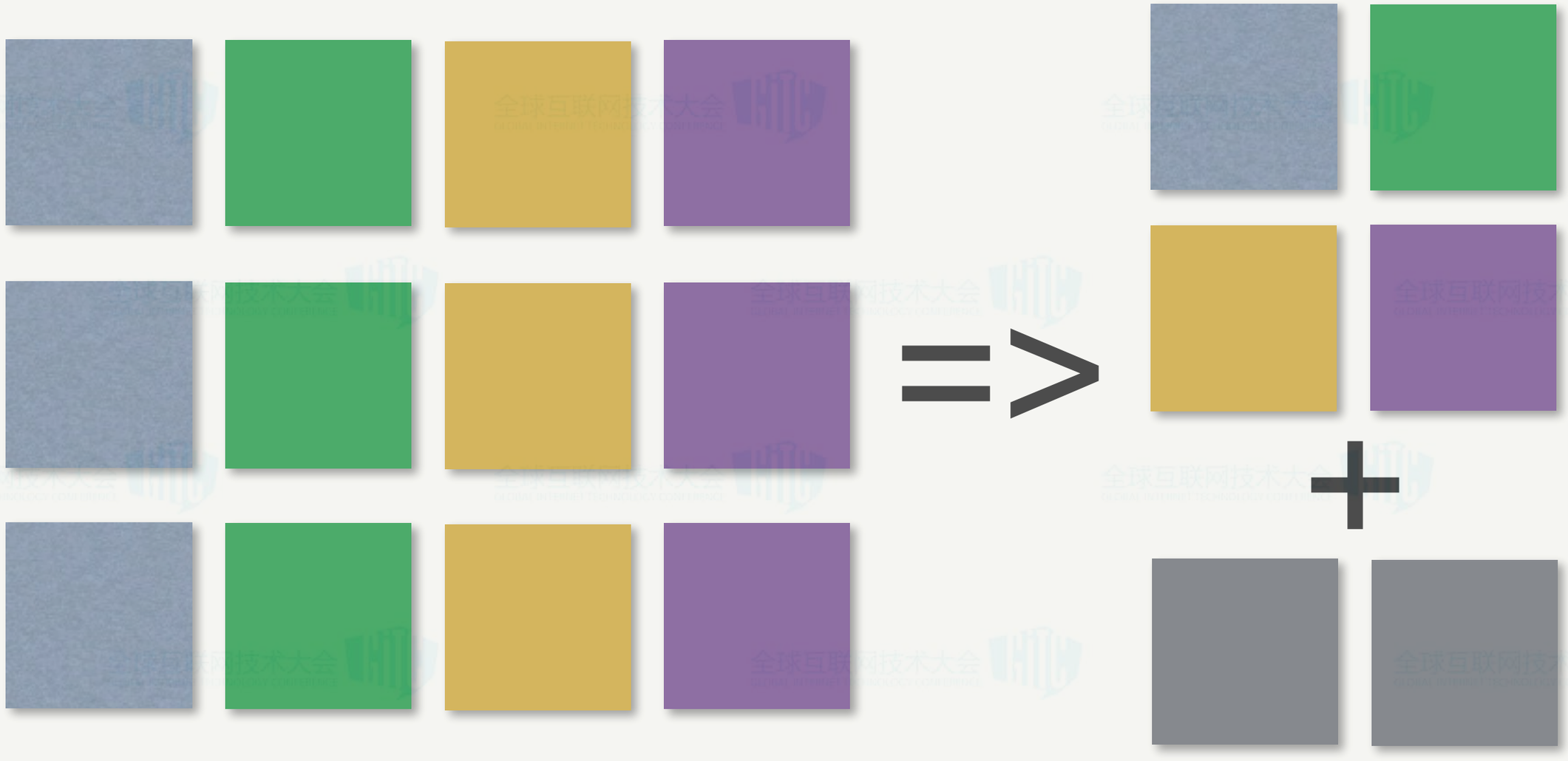
- RAID卡缓存
- 内核缓存模块
  - flashcache
  - lvmcache (dm-cache)
  - bcache
- 性价比高
- 效果受限于命中率
  - 对随机读无效
  - 无法实现产品承诺





纠删码(Erasure Coding)

美团云
Meituan Open Services



4*3=12

4+2=6

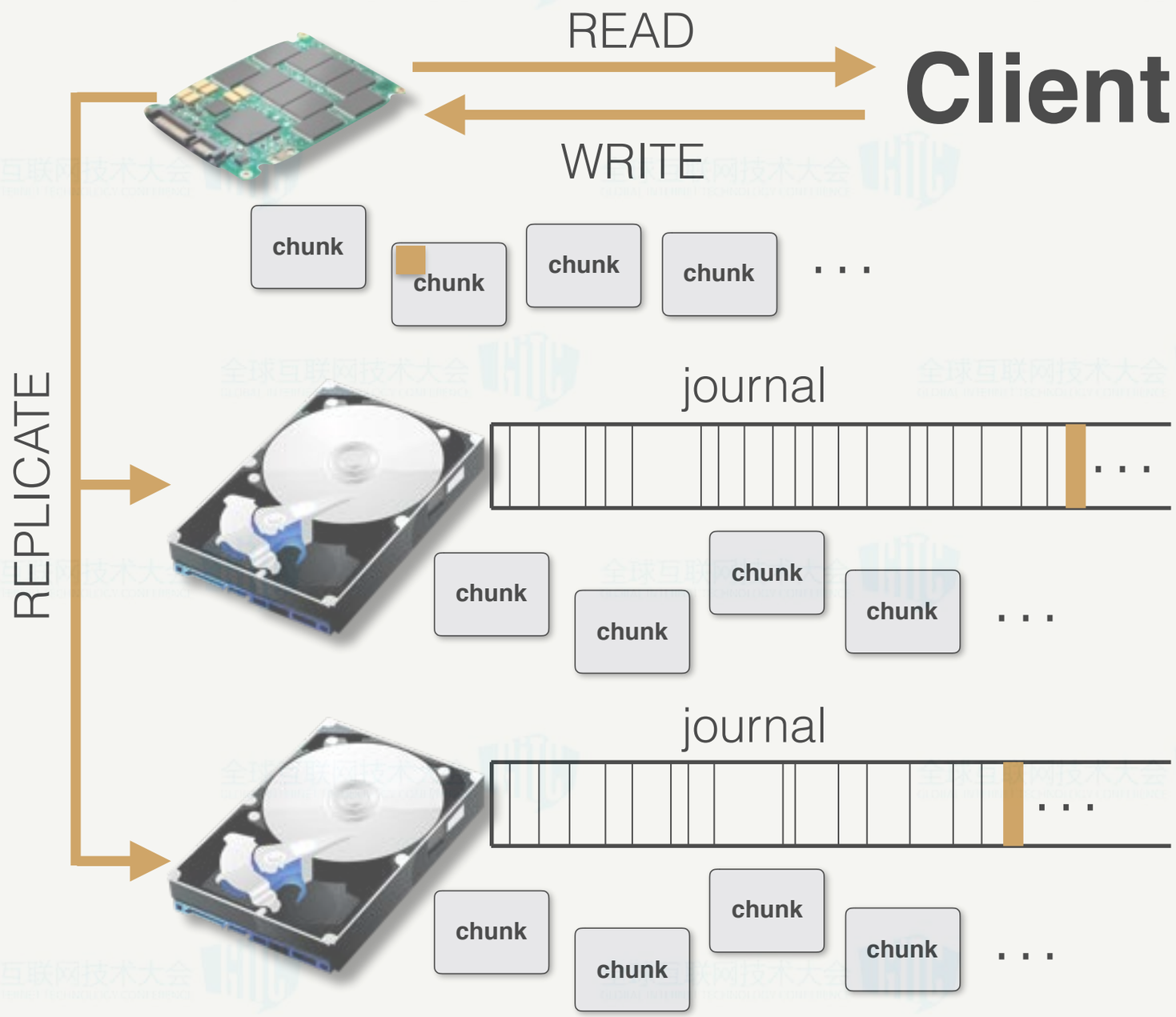
纠删码(Erasure Coding)

美团云
Meituan Open Services

- n个数据块，编码生成k个校验块
  - 可容忍任意k个数据块损坏
- SSD + EC
  - 技术复杂，写性能低
  - 需要多批次、品牌的盘片
  - 成本略低

SSD-HDD混合存储方案

- SSD+HDD
  - SSD副本*1
  - HDD副本*2
    - 写journal

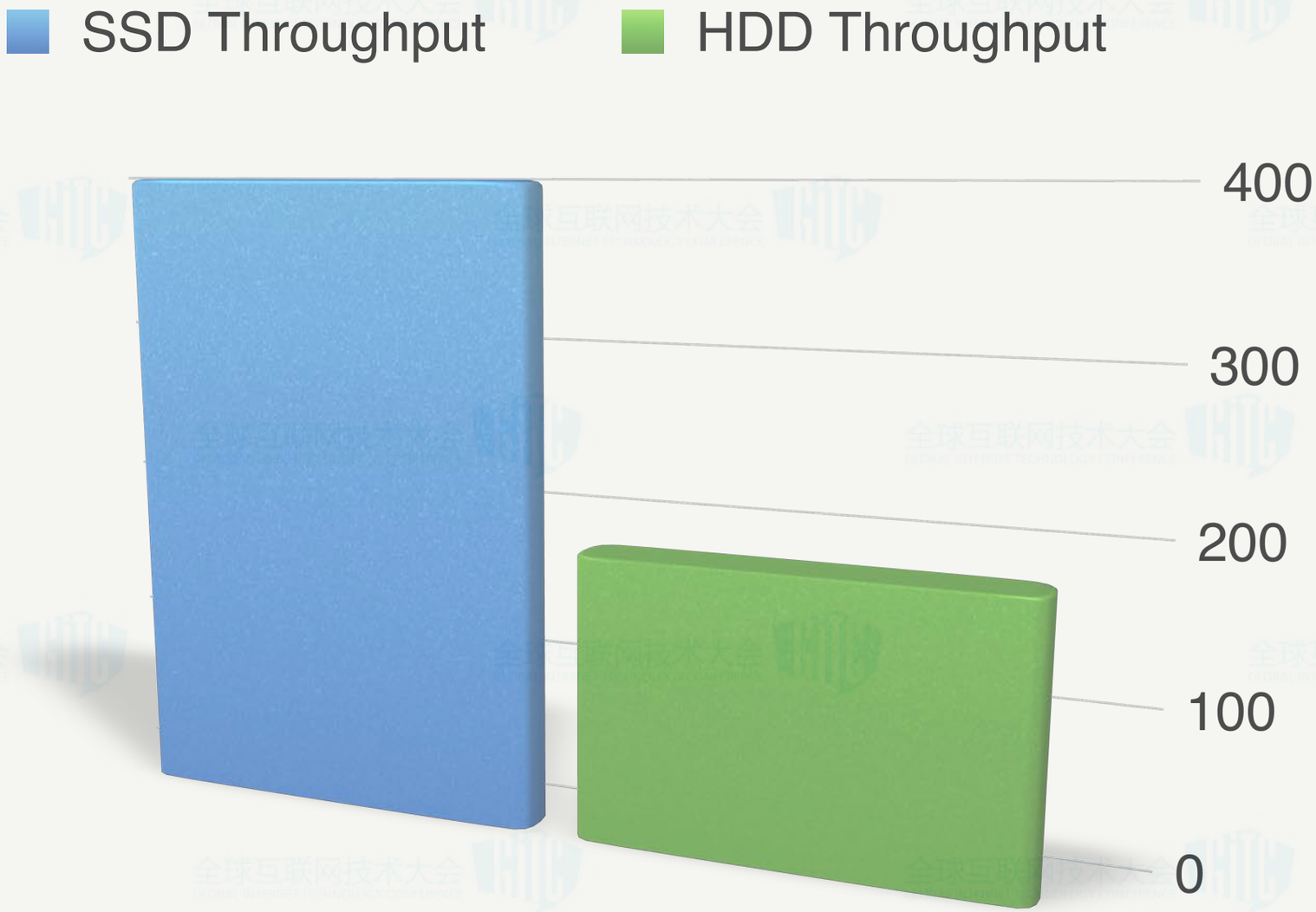


写入性能：IOPS

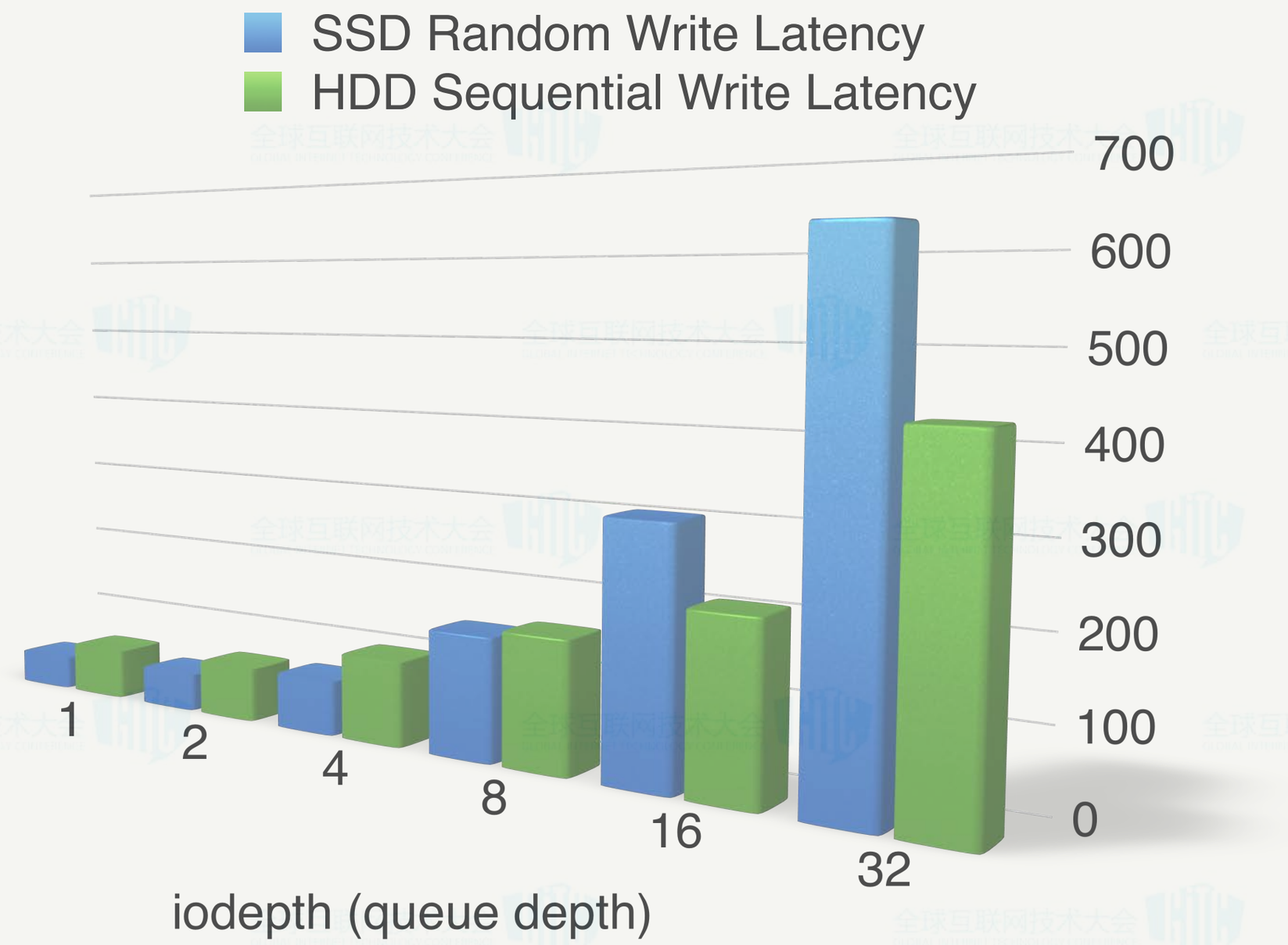

SSD Random Write IOPS
HDD Sequential Write IOPS
60,000
50,000
40,000
30,000
20,000
10,000
0
1
2
4
8
16
32
64
iodepth (queue depth)

写入性能：吞吐率

美团云
Meituan Open Services



写入性能：延迟


SSD Random Write Latency
HDD Sequential Write Latency
700
600
500
400
300
200
100
0
1
2
4
8
16
32
iodepth (queue depth)

写入性能

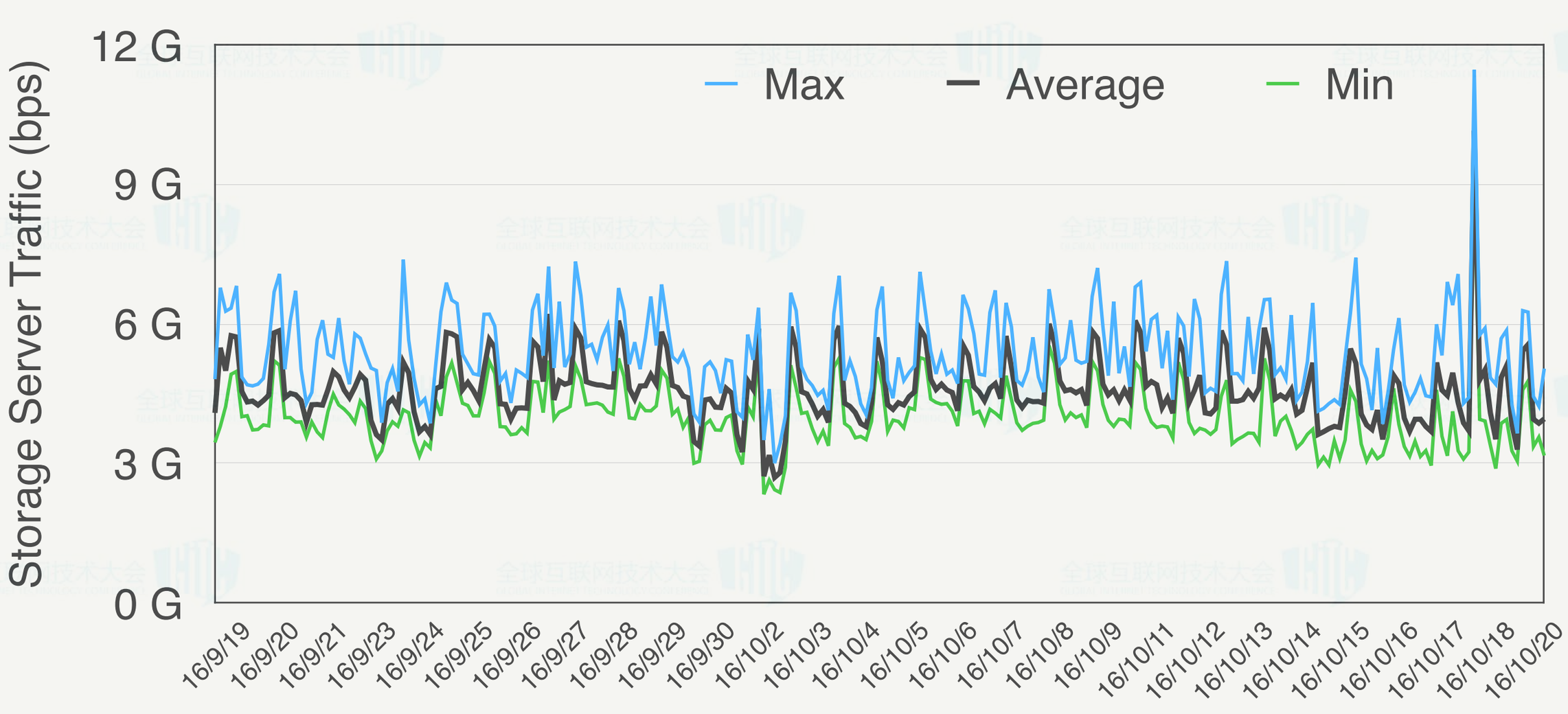
	SSD(随机)	HDD(顺序)
IOPS	HDD略低	
吞吐率	HDD低一半	
延迟	HDD略好	

结论：HDD的顺序写入性能与SSD匹配

回写

美团云
Meituan Open Services

- Journal/buffer回写
  - “出来混迟早要还的”
  - 时机很重要：线上workload有明显波峰、波谷



Ursa：美团云块存储系统

美团云
Meituan Open Services

- 完全自主研发
- 稳定可靠高性能
- 功能强大
- 实现混合存储，内部上线
- http://tech.meituan.com/block-store.html

目录

美团云
Meituan Open Services

性能优化

美团云
Meituan Open Services

- 发挥出SSD应有水平，降低成本
  - ✓优化代码效率
  - ✓发觉利用并行性

性能优化：代码效率

美团云
Meituan Open Services

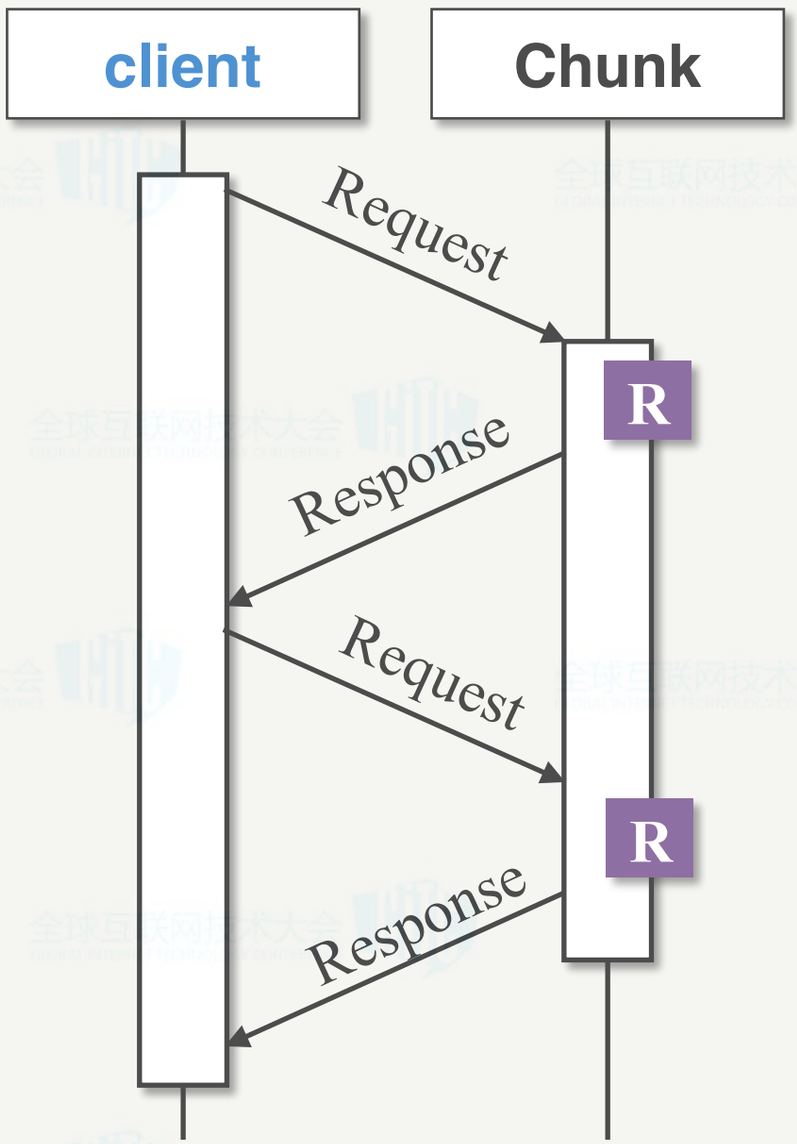
- ~~iostream~~
- stack vs new/delete
- resource pooling & caching
- logging
- CRC、EC（SSE / AVX）
- ~~零拷贝~~
- 保持可维护性

性能优化：并行

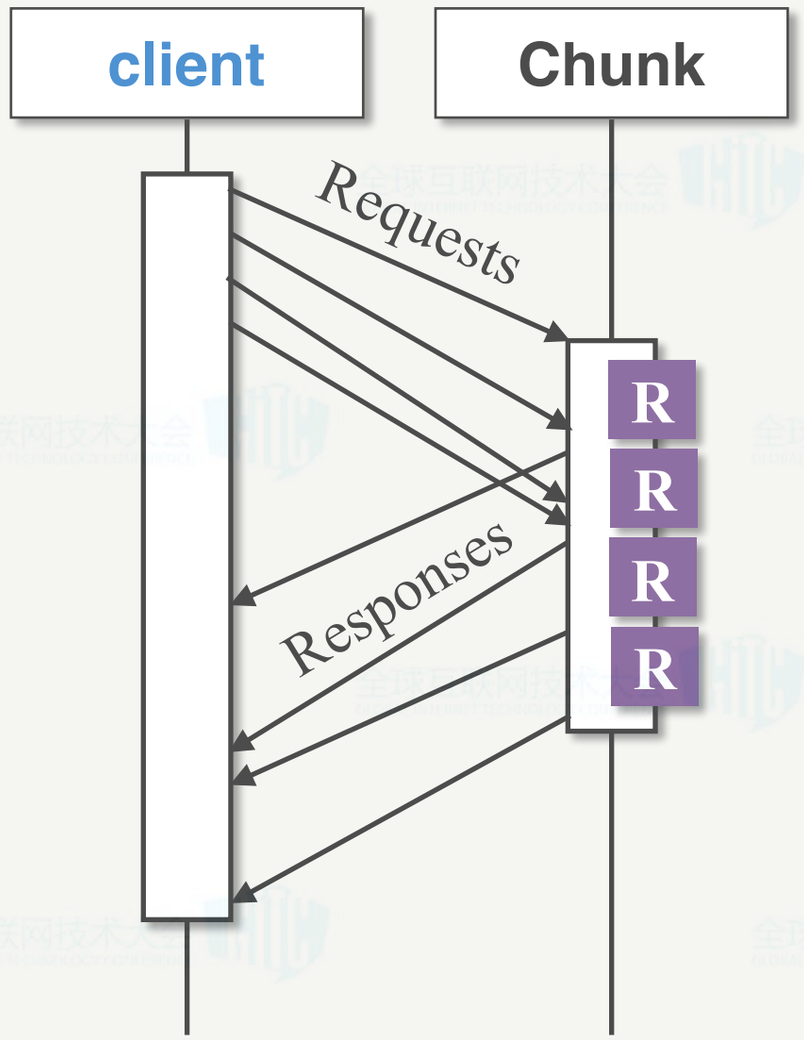
美团云
Meituan Open Services

- 无关任务独立并行执行
  - 服务端：每disk一个或多个服务进程
  - 客户端：每virtual disk一个服务进程
- 盘内并行：使用异步API（libaio）
- 任务流出：流水化处理
  - 磁盘和网络并行运行
- 任务完成：乱序完成
  - 慢请求不阻塞其他请求

流水化处理



(a) non-pipelined



(b) pipelined
(with network jitter)

性能优化：SMP亲和性

美团云
Meituan Open Services

- 核间负载均衡
  - 启用MSI-X，并将不同中断设置到不同核心
- 减少task的核间迁移
  - 网卡IRQ、网卡SoftIRQ、硬盘IRQ、硬盘SoftIRQ、存储服务进程、CPU核心
- 减少资源争抢
  - 多个物理CPU + 1个SAS卡 ==> 资源争抢

内核功能不够用了

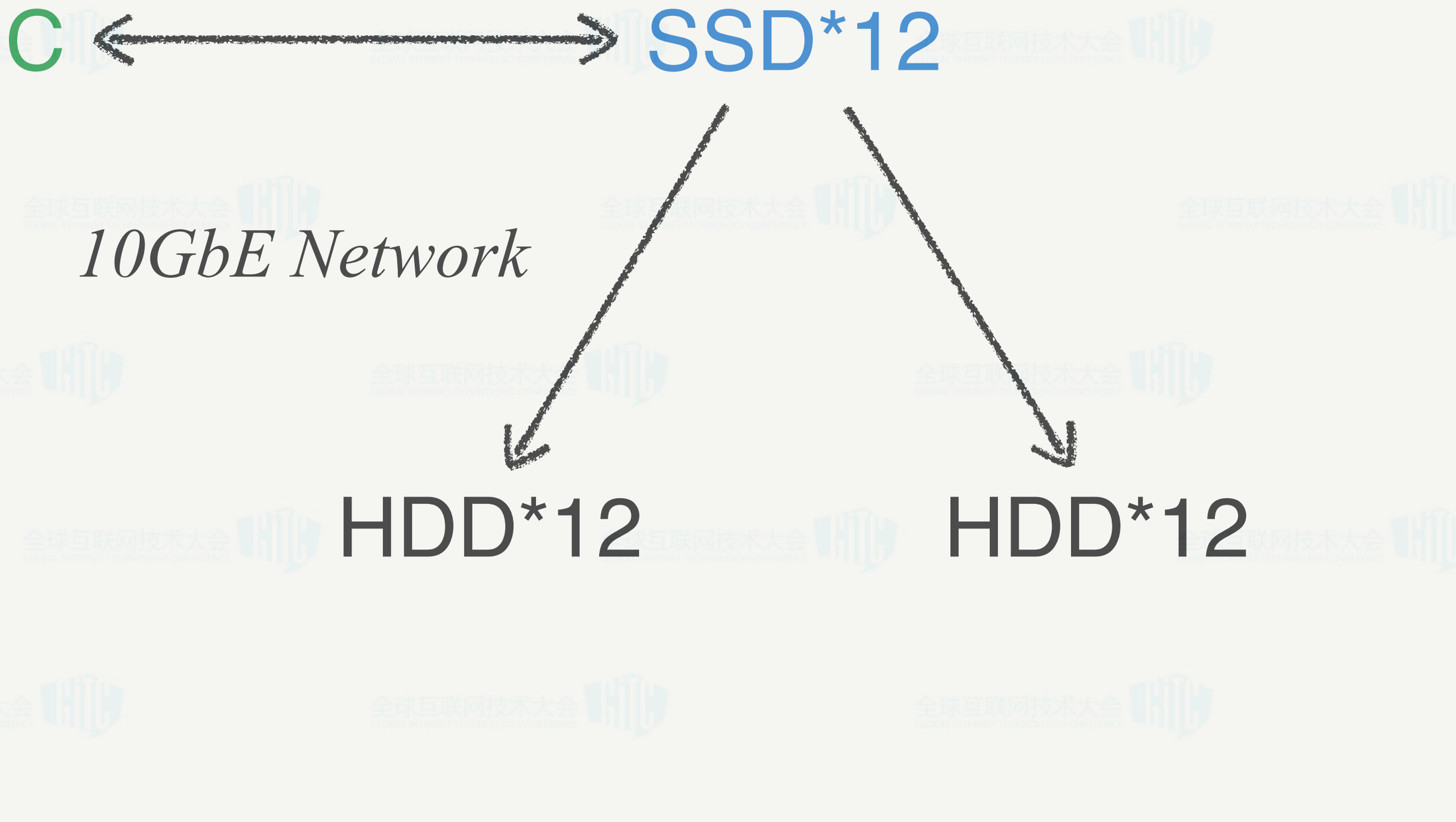
美团云
Meituan Open Services

- 零拷贝（sendfile, splice, ...）
  - 功能弱，不能异步
- 缓存/缓冲机制
  - 需要：不阻塞、优化write-through、两态共享、两阶段提交、惰性补齐、...
- 异步I/O（libaio）
  - 不能零拷贝，不能缓冲
- 磁盘QoS
  - 不能缓冲，需要CFQ（不适合SSD）

目录

美团云
Meituan Open Services

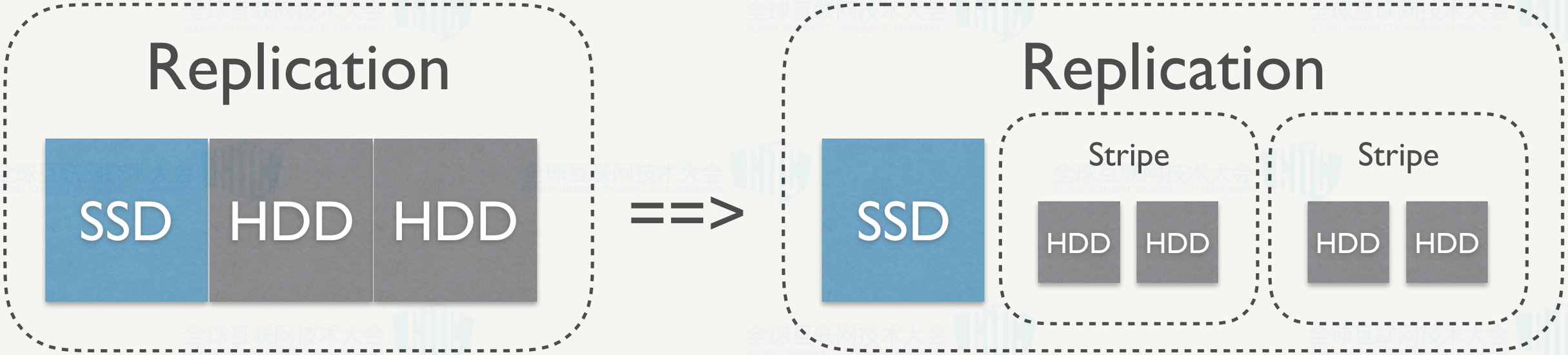
Ursa-on-SSD性能初步测试



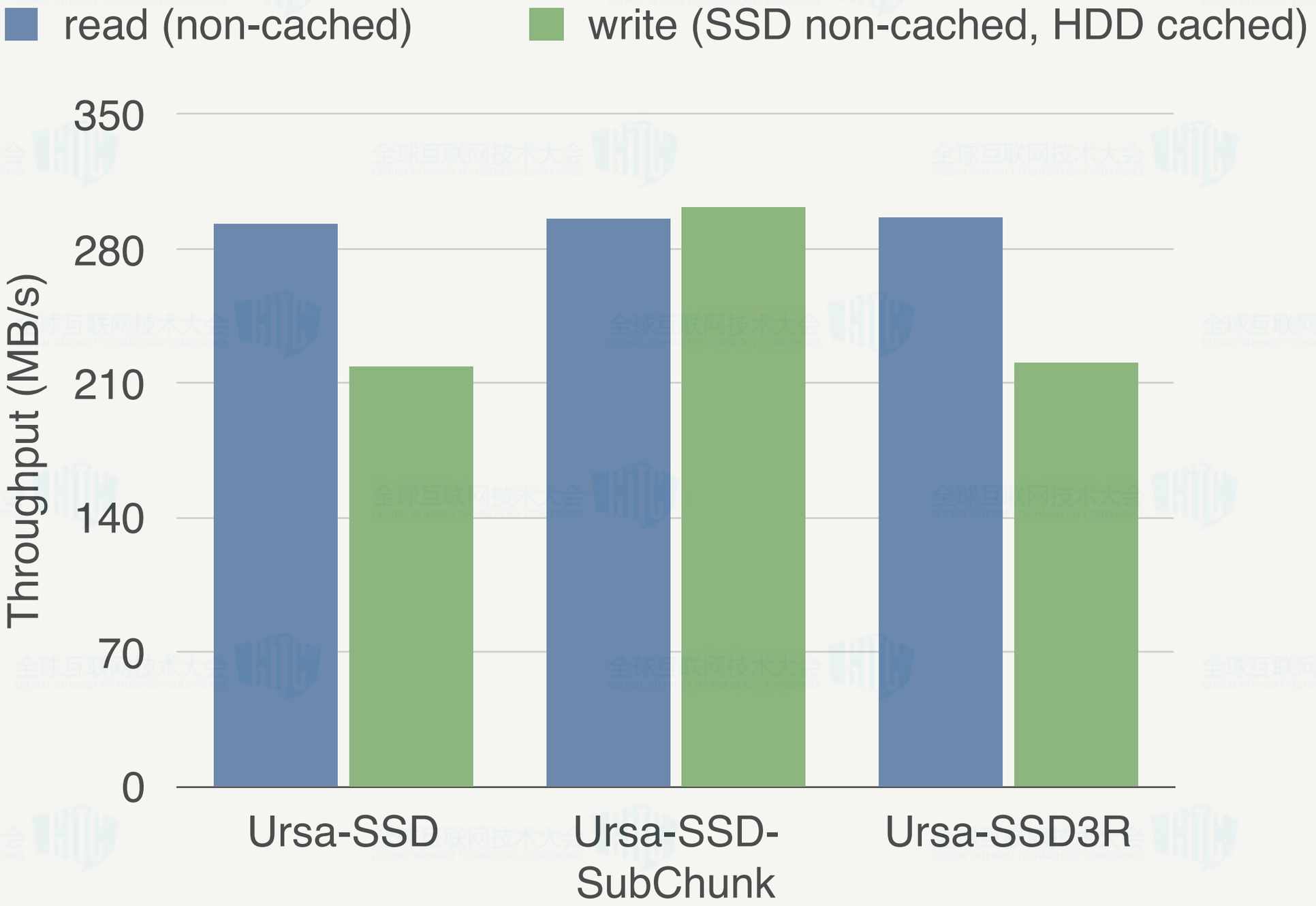
初步性能测试

美团云
Meituan Open Services

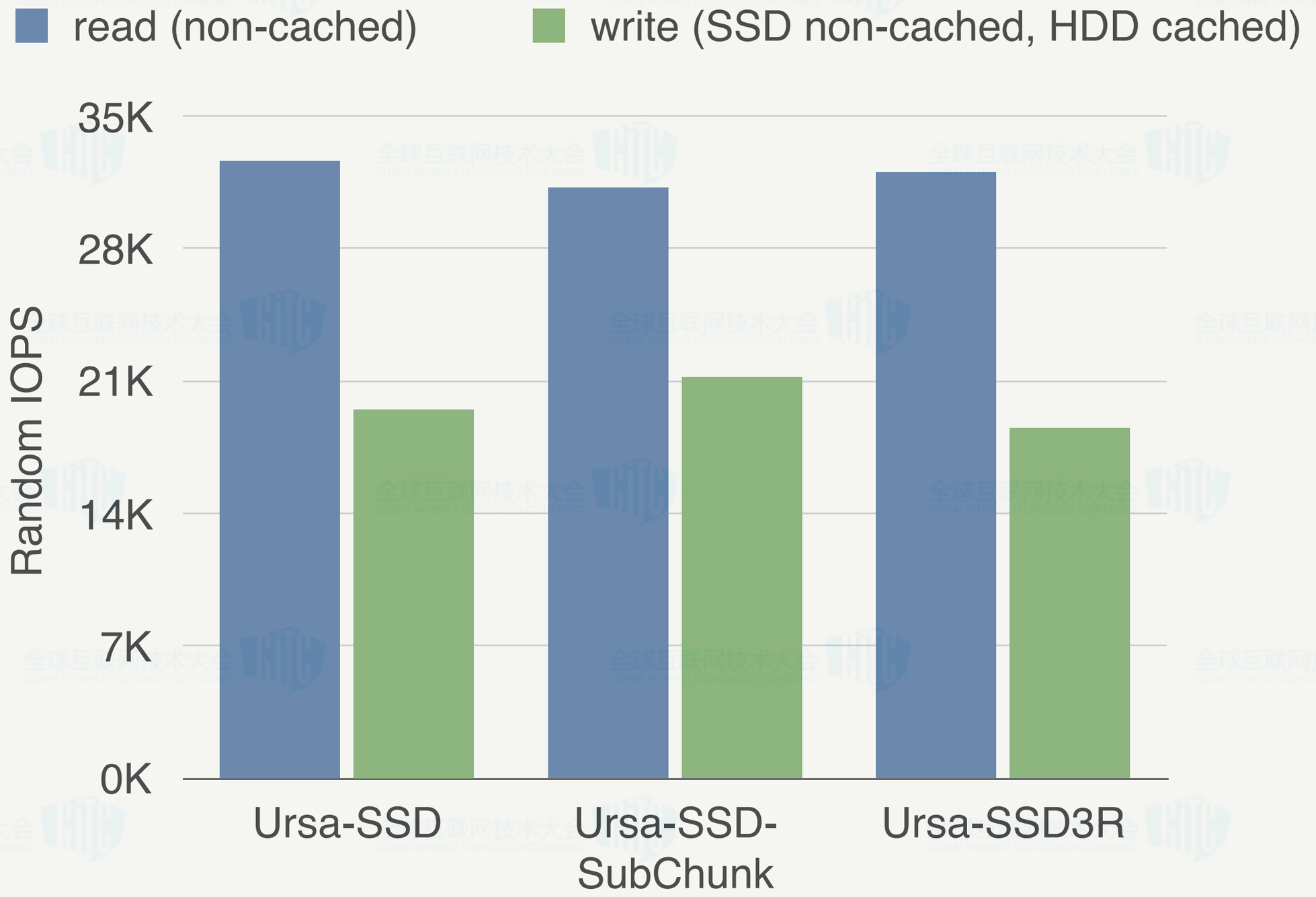
- 补齐HDD吞吐率短板（SubChunk）



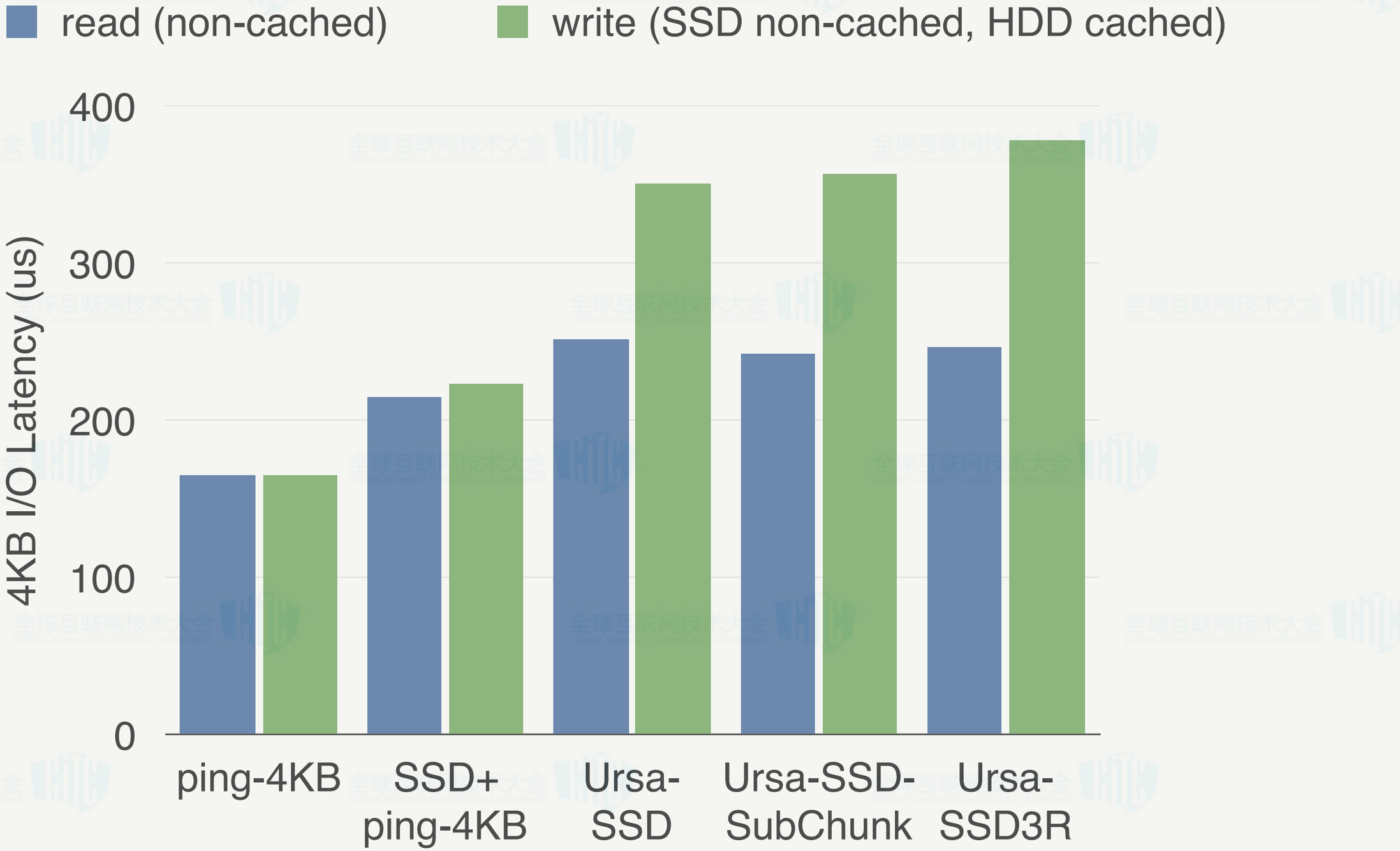
初步性能测试


read (non-cached)
write (SSD non-cached, HDD cached)
Throughput (MB/s)
350
280
210
140
70
0
Ursa-SSD
Ursa-SSD-SubChunk
Ursa-SSD3R

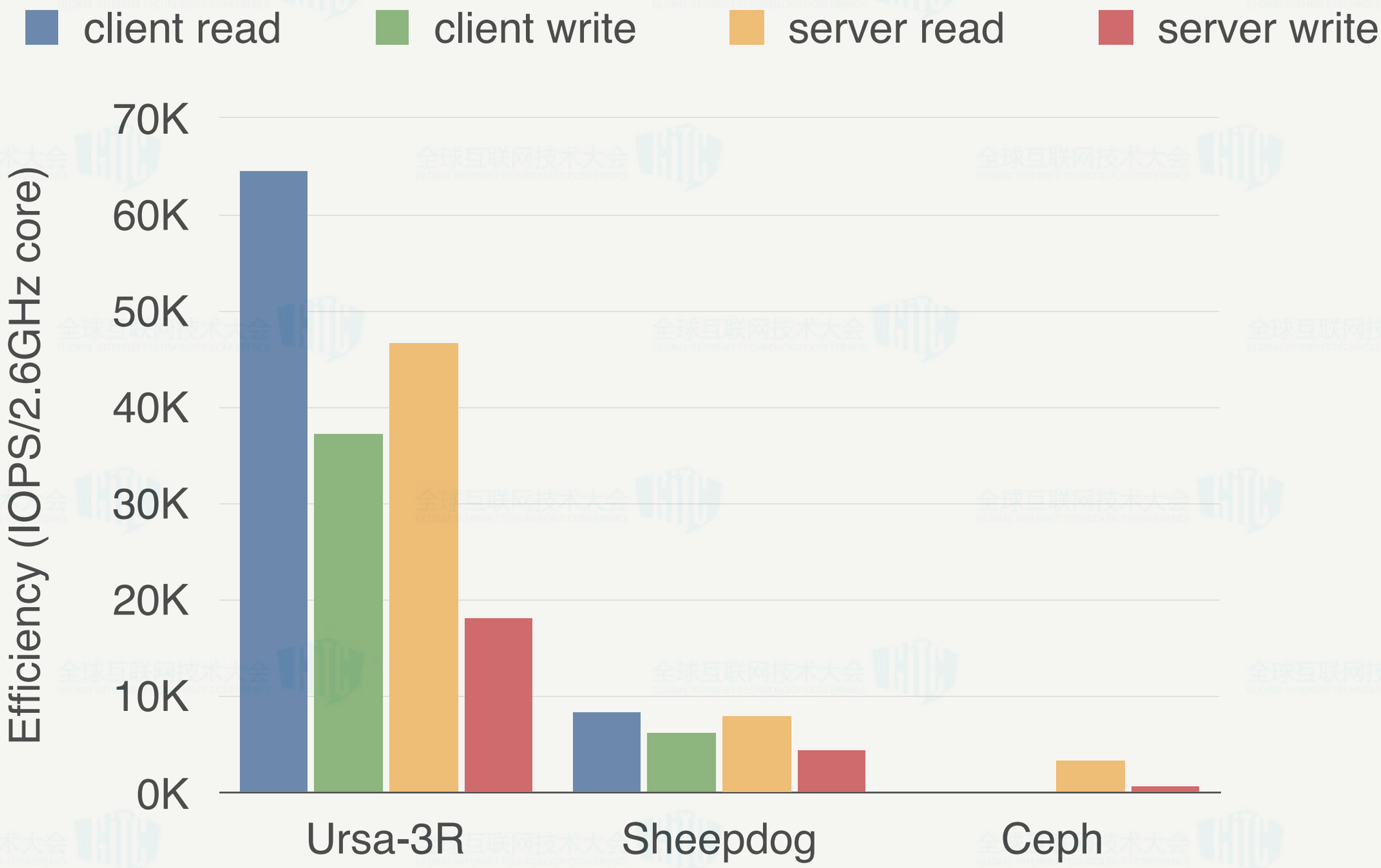
初步性能测试


read (non-cached)
write (SSD non-cached, HDD cached)
Random IOPS
35K
28K
21K
14K
7K
0K
Ursa-SSD
Ursa-SSD-SubChunk
Ursa-SSD3R

初步性能测试


read (non-cached)
write (SSD non-cached, HDD cached)
4KB I/O Latency (us)
400
300
200
100
0
ping-4KB
SSD+ ping-4KB
Ursa-SSD
Ursa-SSD-SubChunk
Ursa-SSD3R

系统效率


client read
client write
server read
server write
Efficiency (IOPS/2.6GHz core)
70K
60K
50K
40K
30K
20K
10K
0K
Ursa-3R
Sheepdog
Ceph

结论

美团云
Meituan Open Services

- 混合存储方案性能符合产品需求
- 性能仍有潜力可以挖掘
- 测试结果仅供参考
  - 环境与生产环境不相等
  - 优化持续进行中

美团云
Meituan Open Services

谢谢！

Q&A

https://mos.meituan.com