

美团云混合存储系统

李慧霸
2016.11.24

自我介绍

- 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

云计算管理平台

计算虚拟化

网络虚拟化

存储虚拟化

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

2015年
07月

开放为业界所熟知

整合的解决方案

2015年
10月

美团云大数据全景战略发布

信息等级安全保护三级

2016年
1月

共同信赖美团云

目录

- 介绍
- 云上的硬盘
- 成本优化方案
- 性能优化技术
- 性能评估
- 结论

云上的硬盘

	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

低 ← 性能、价格 → 高

云上的硬盘

- 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

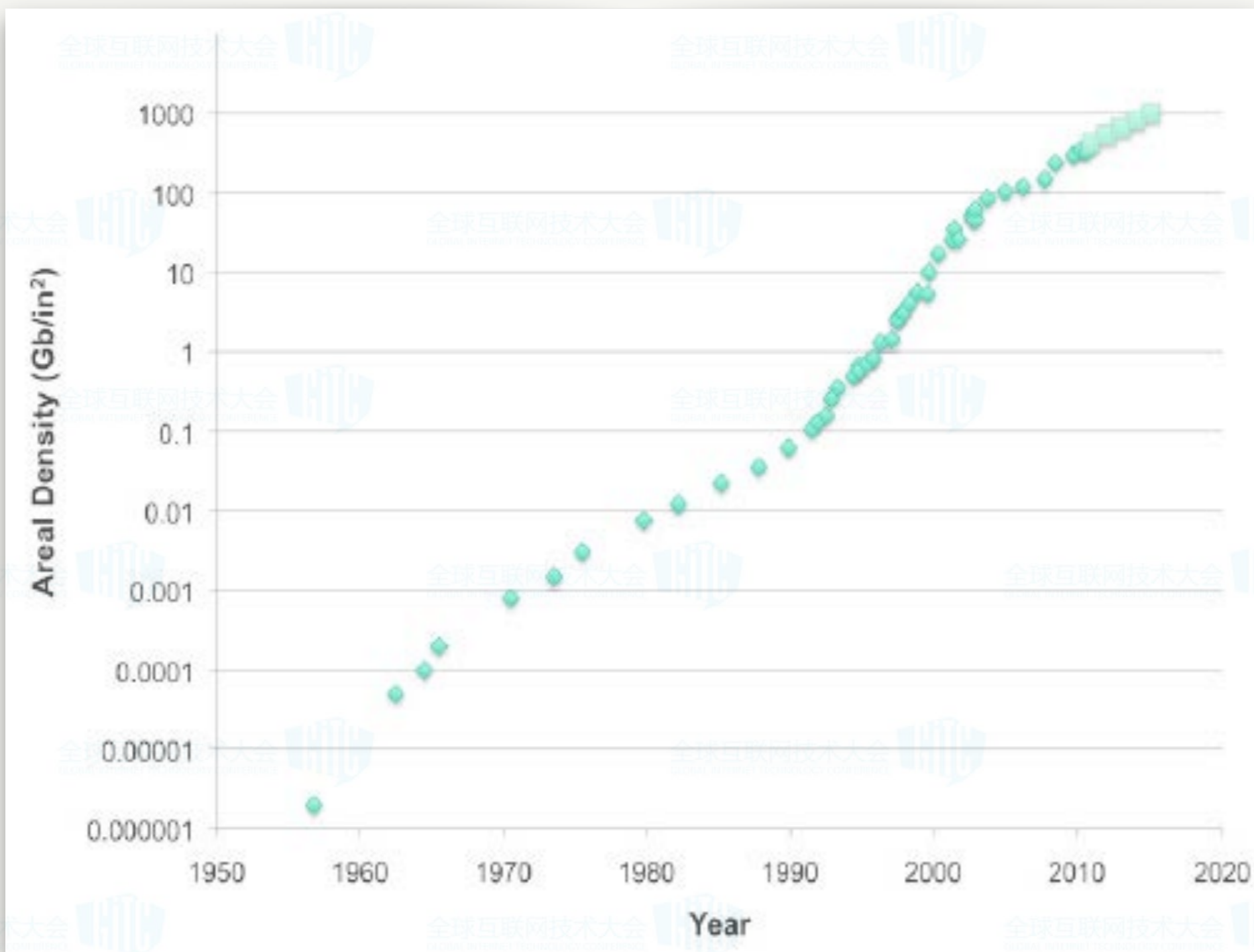


IBM交付5M硬盘，1956

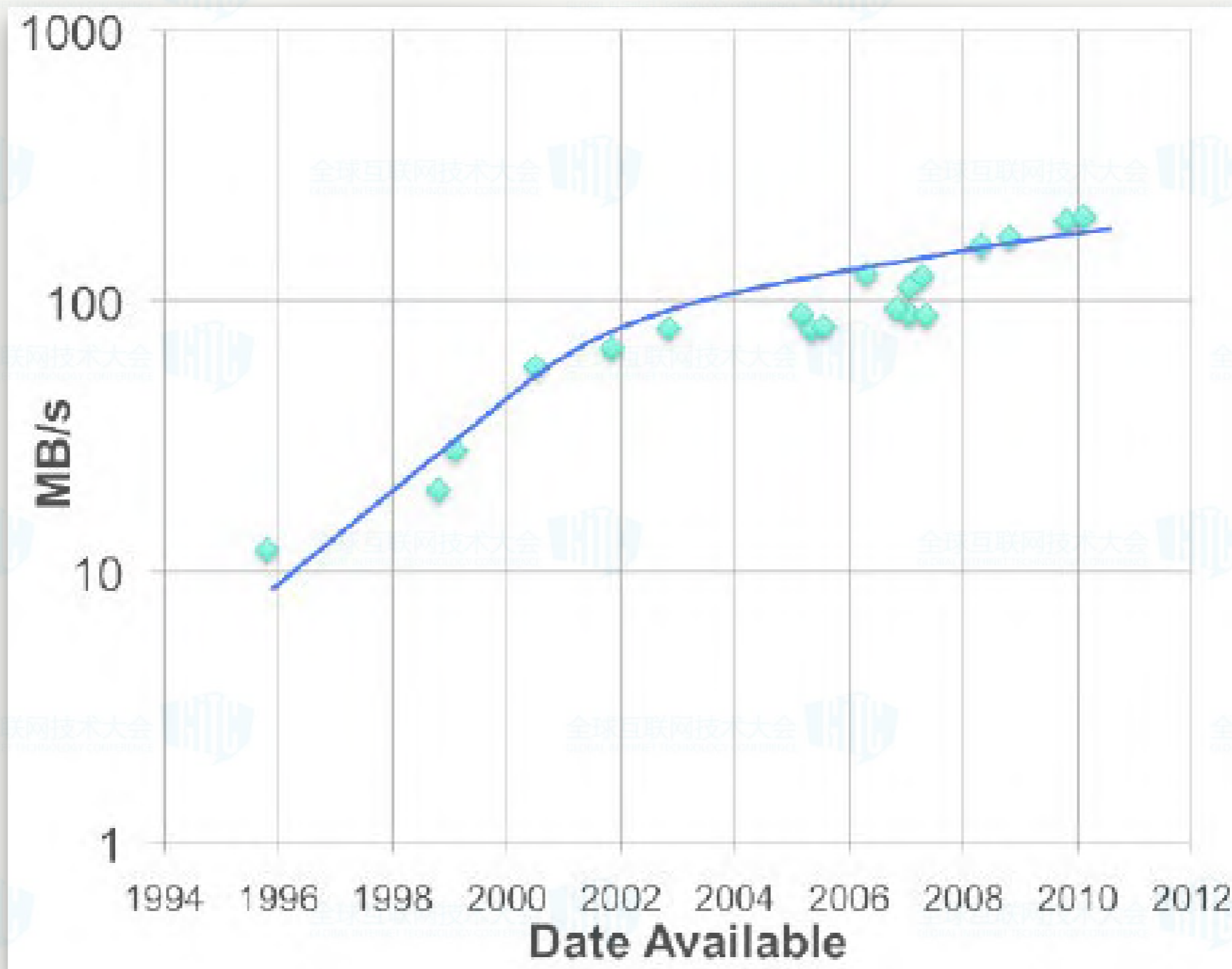


3.5寸10TB硬盘，2016

容量：60年增长9个数量级

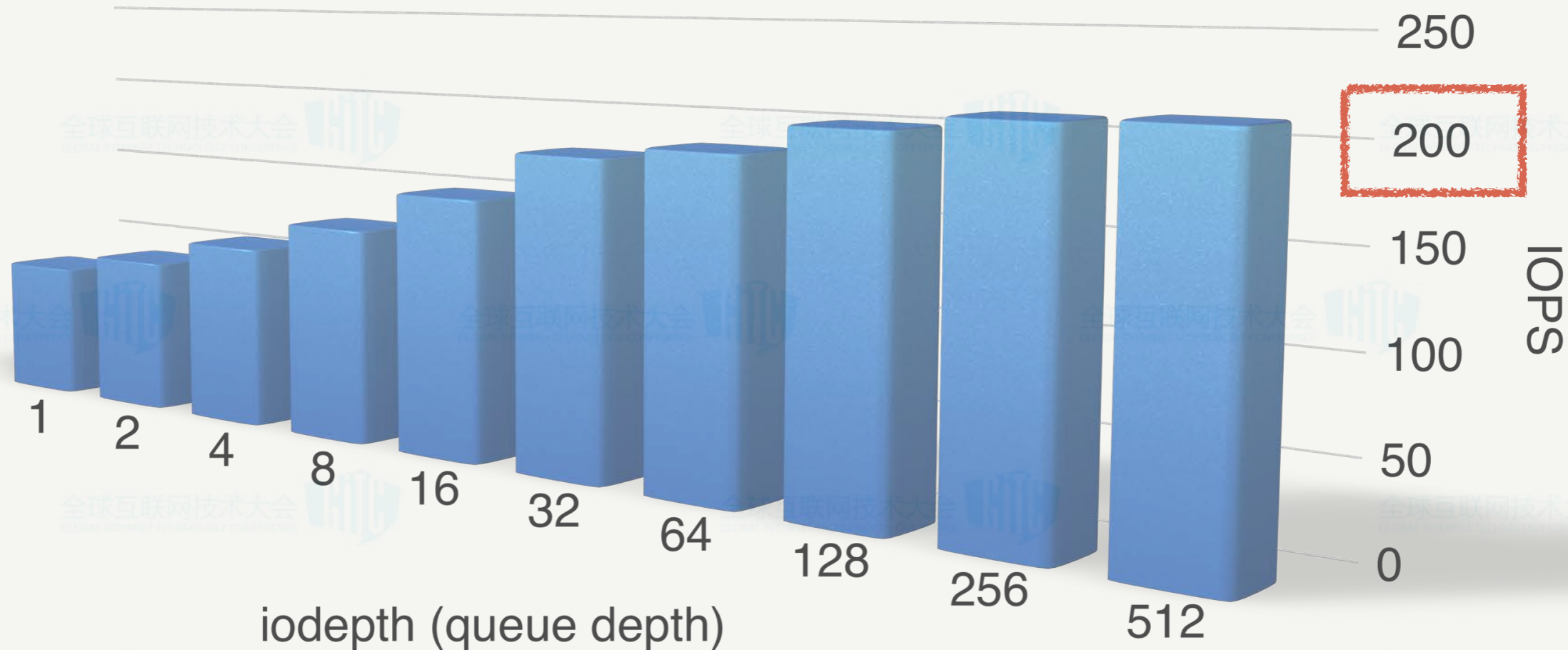


性能：呵呵呵.....



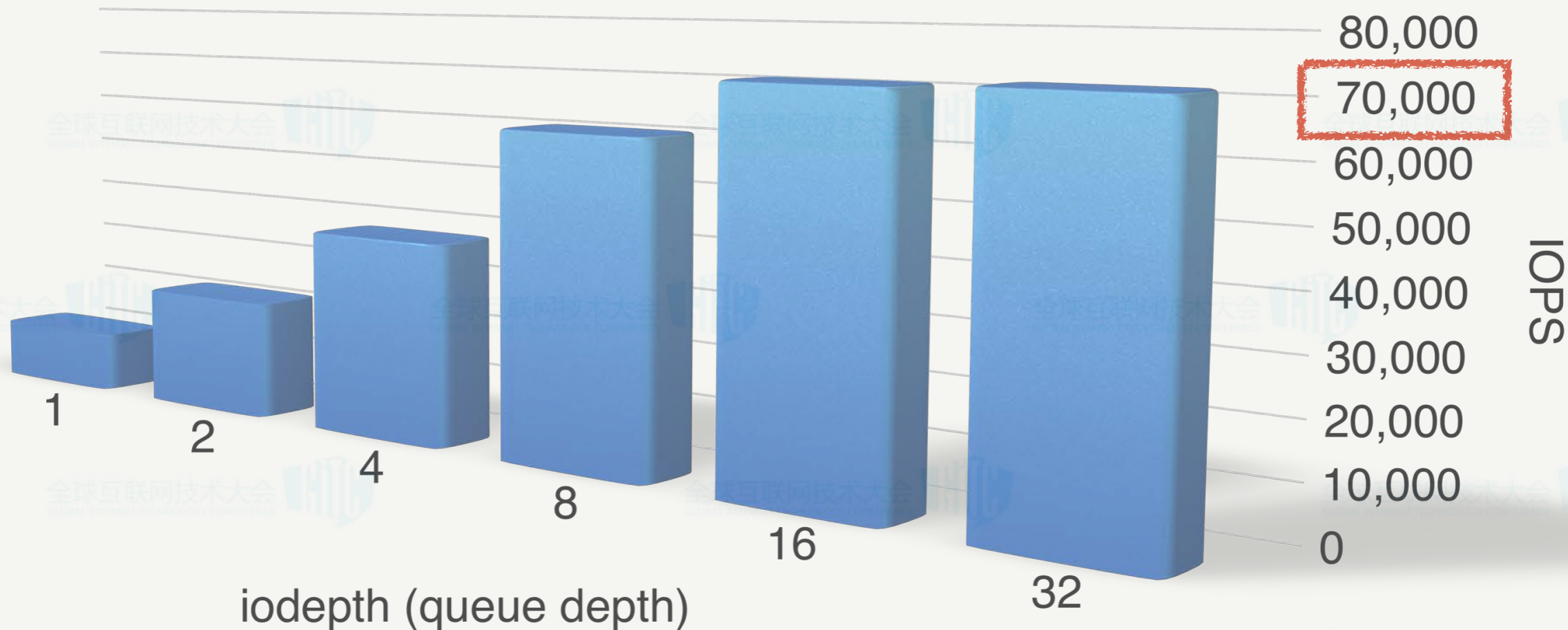
HDD性能与容量的失衡

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

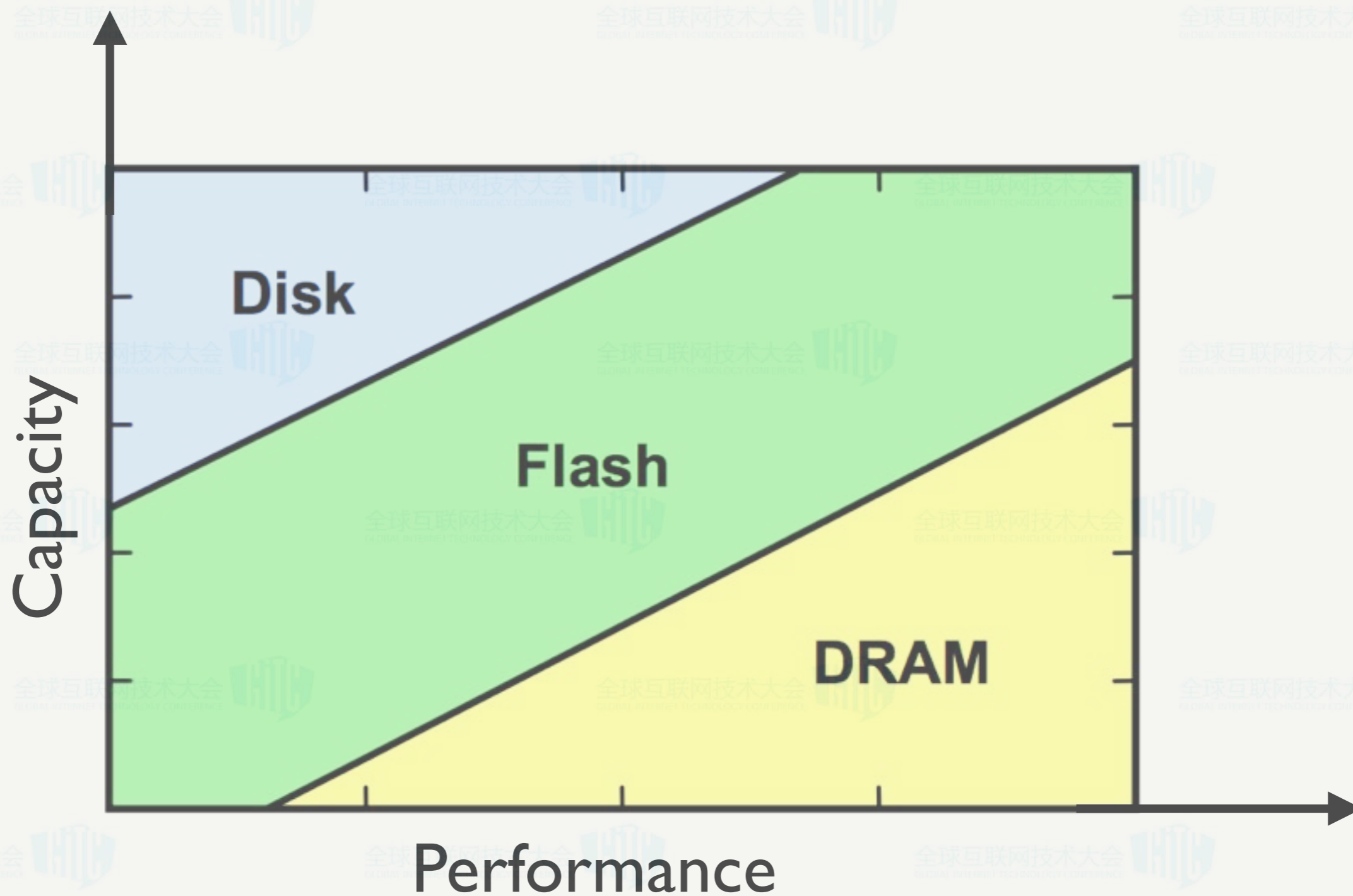


HDD性能与容量的失衡

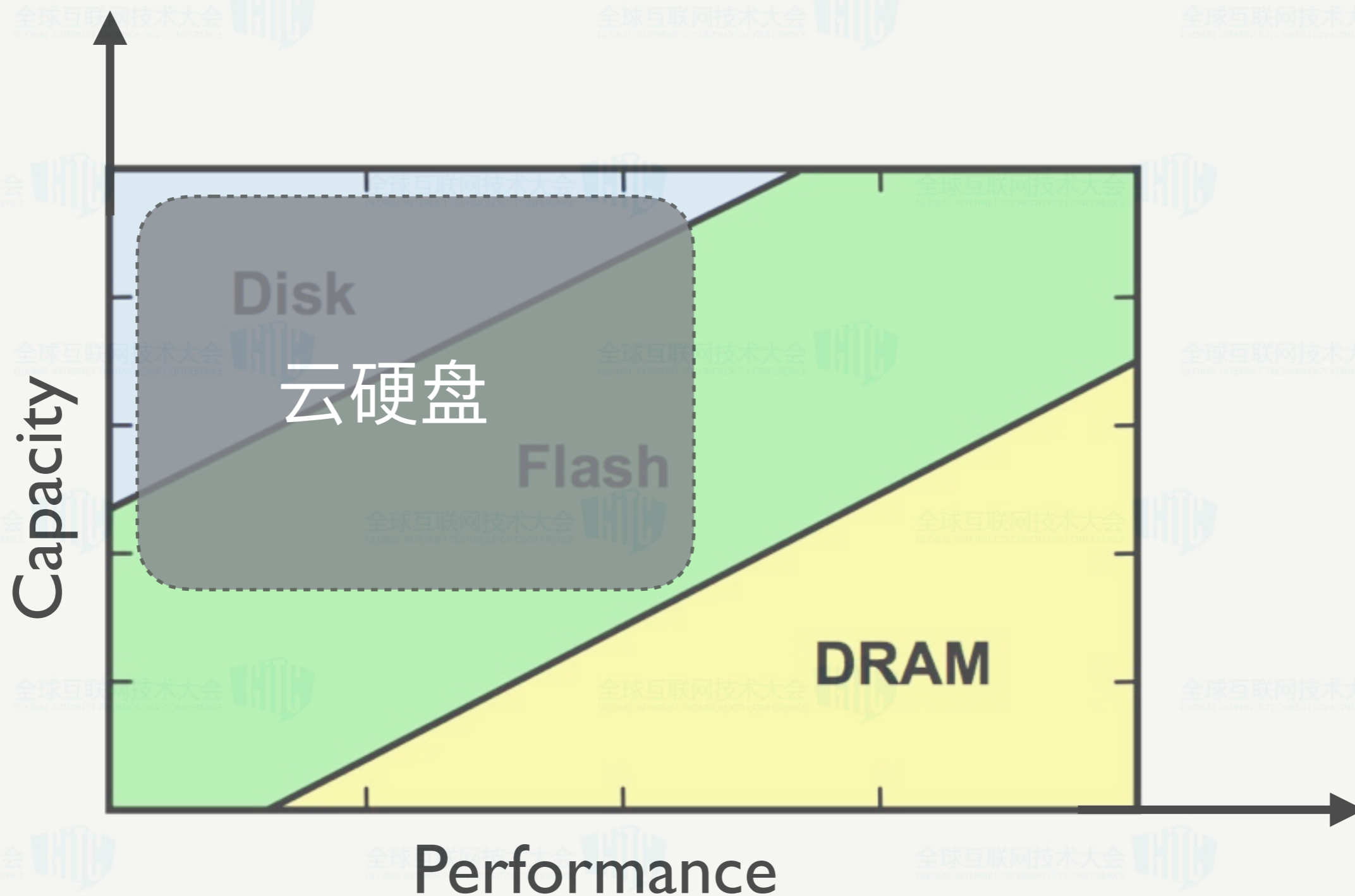
Random Read IOPS of a typical SSD (480GB, SATA)



3年TCO最低：容量与性能



3年TCO最低：容量与性能

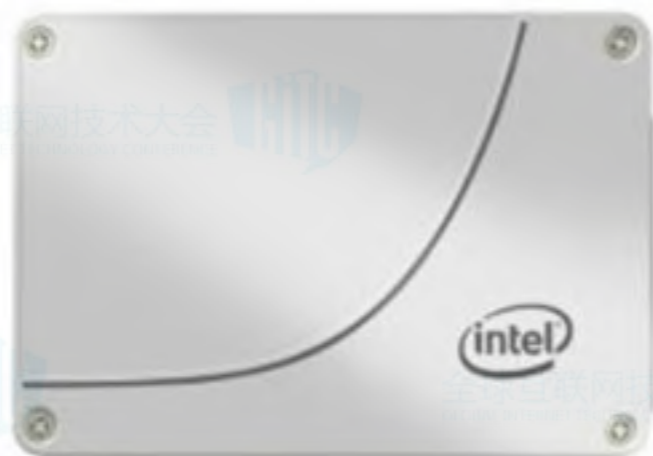


目录

- 介绍
- 云上的硬盘
- 成本优化方案
- 性能优化技术
- 性能评估
- 结论

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



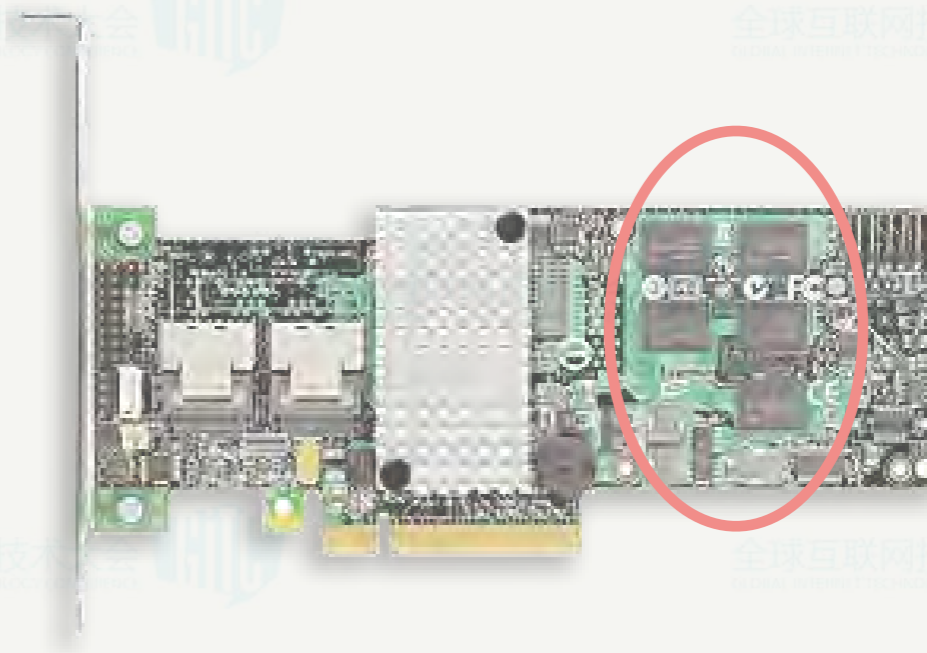
¥1299.00

希捷(SEAGATE)ES.3系列 4TB 7200转12

¥ 325/TB

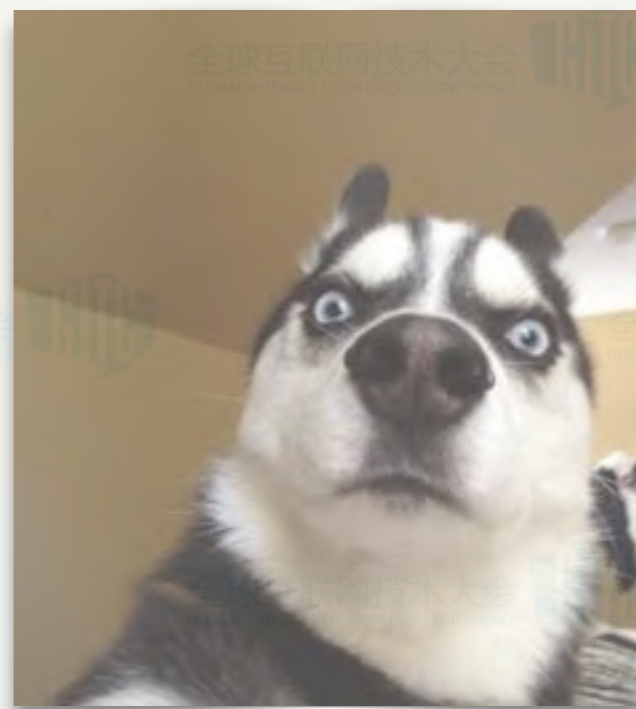
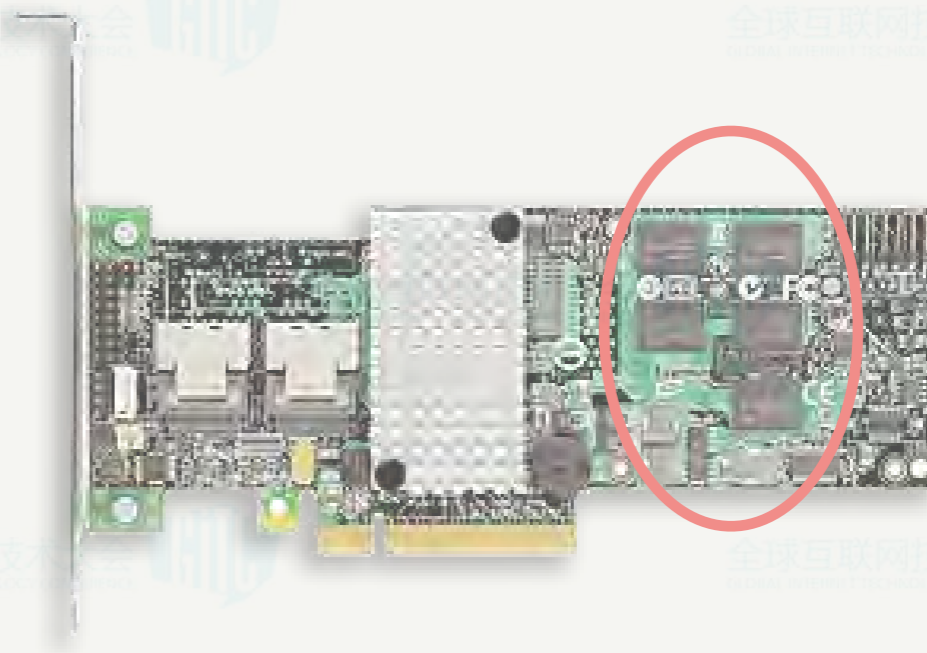
缓存优化

- 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

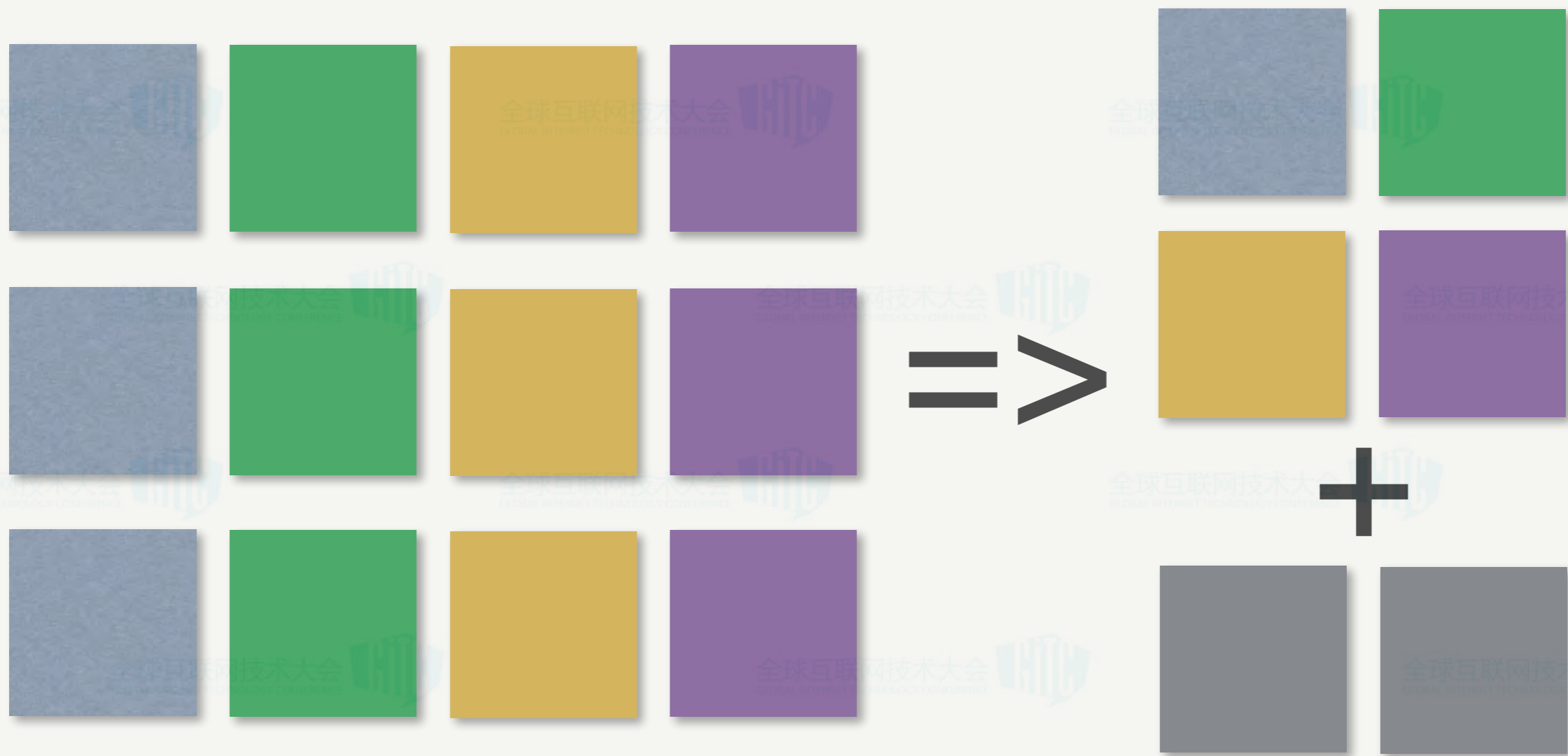


缓存优化

- 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)



$$4 * 3 = 12$$

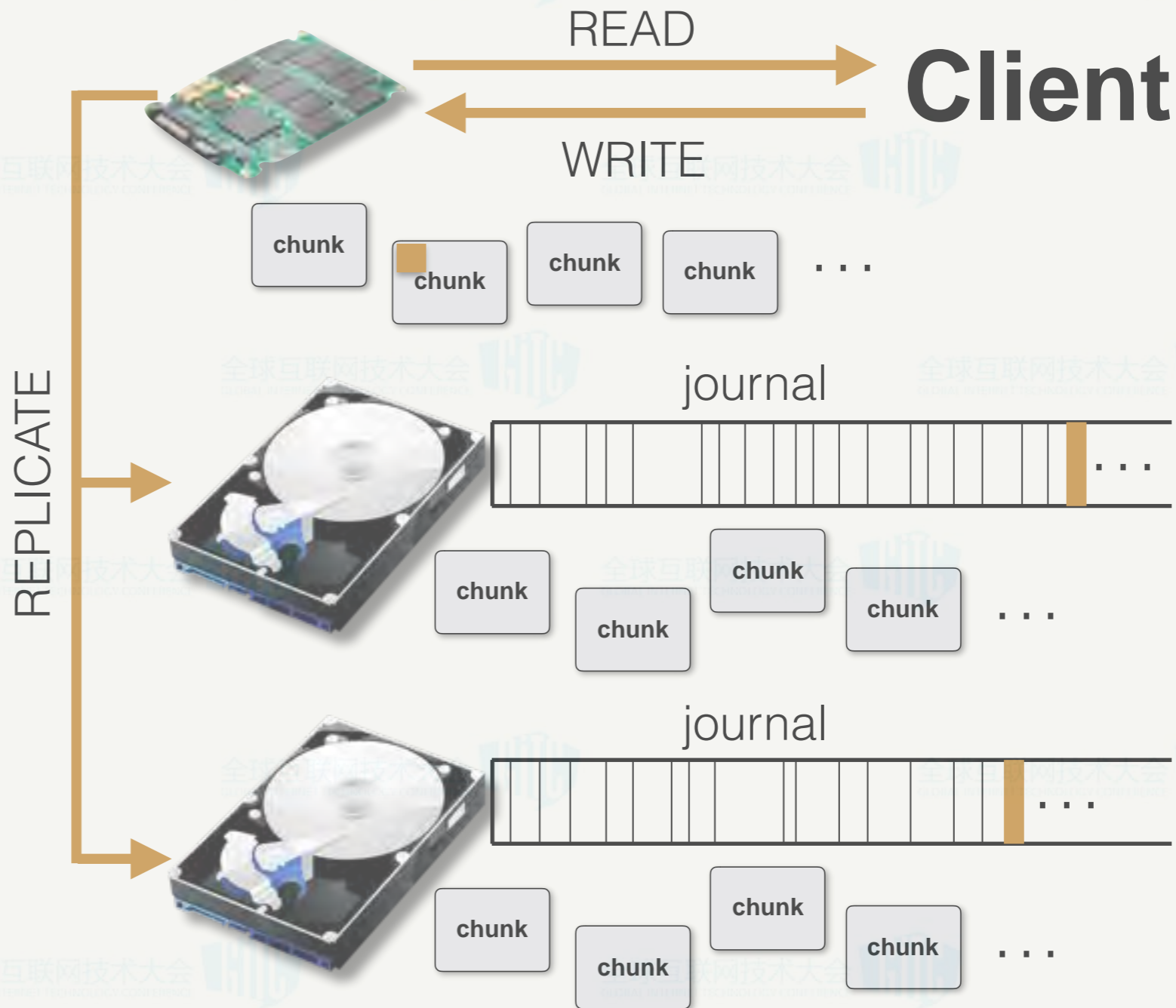
$$4 + 2 = 6$$

纠删码 (Erasure Coding)

- 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

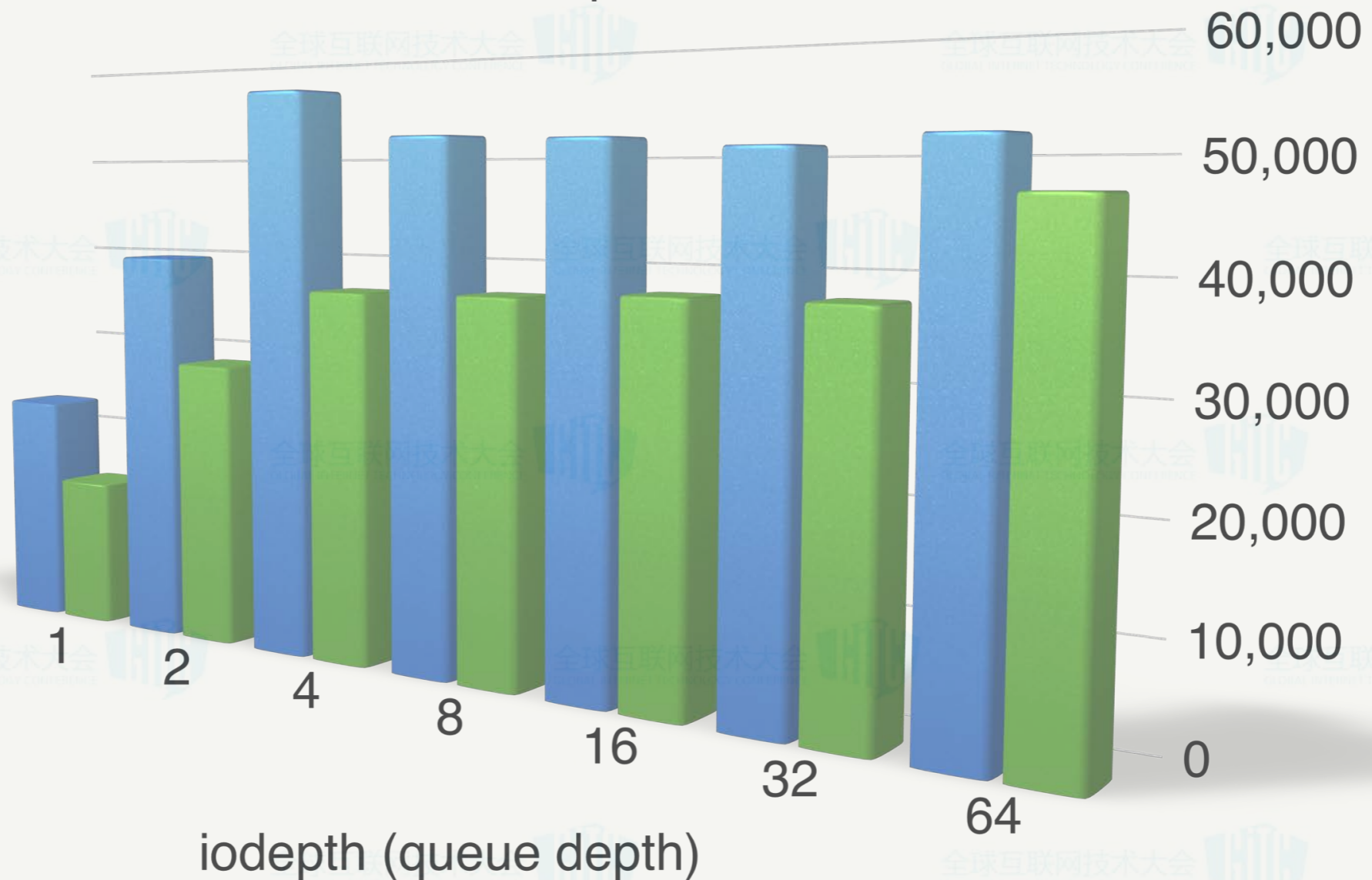
SSD-HDD混合存储方案

- 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



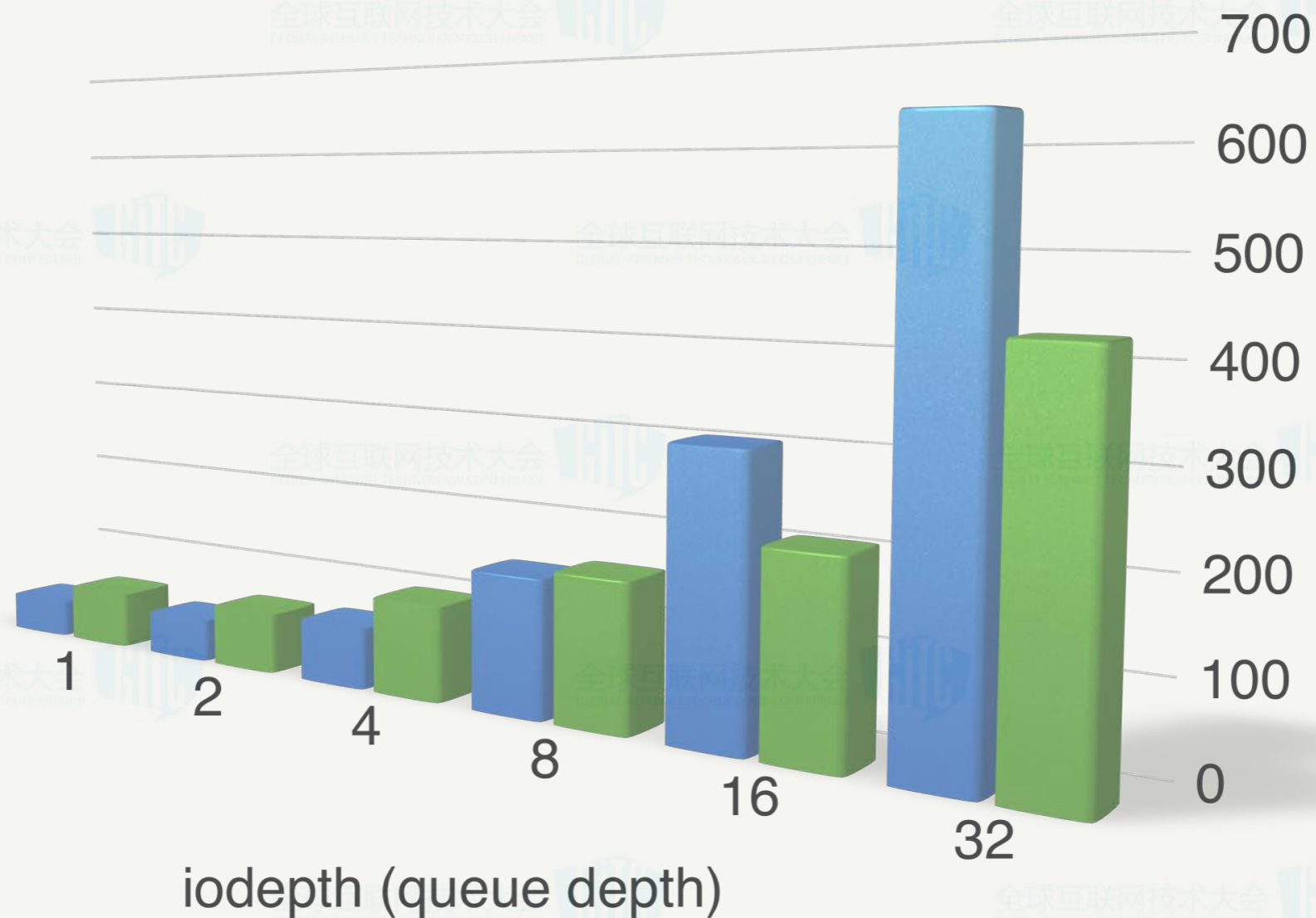
写入性能：吞吐率

■ SSD Throughput ■ HDD Throughput



写入性能：延迟

- SSD Random Write Latency
- HDD Sequential Write Latency



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	

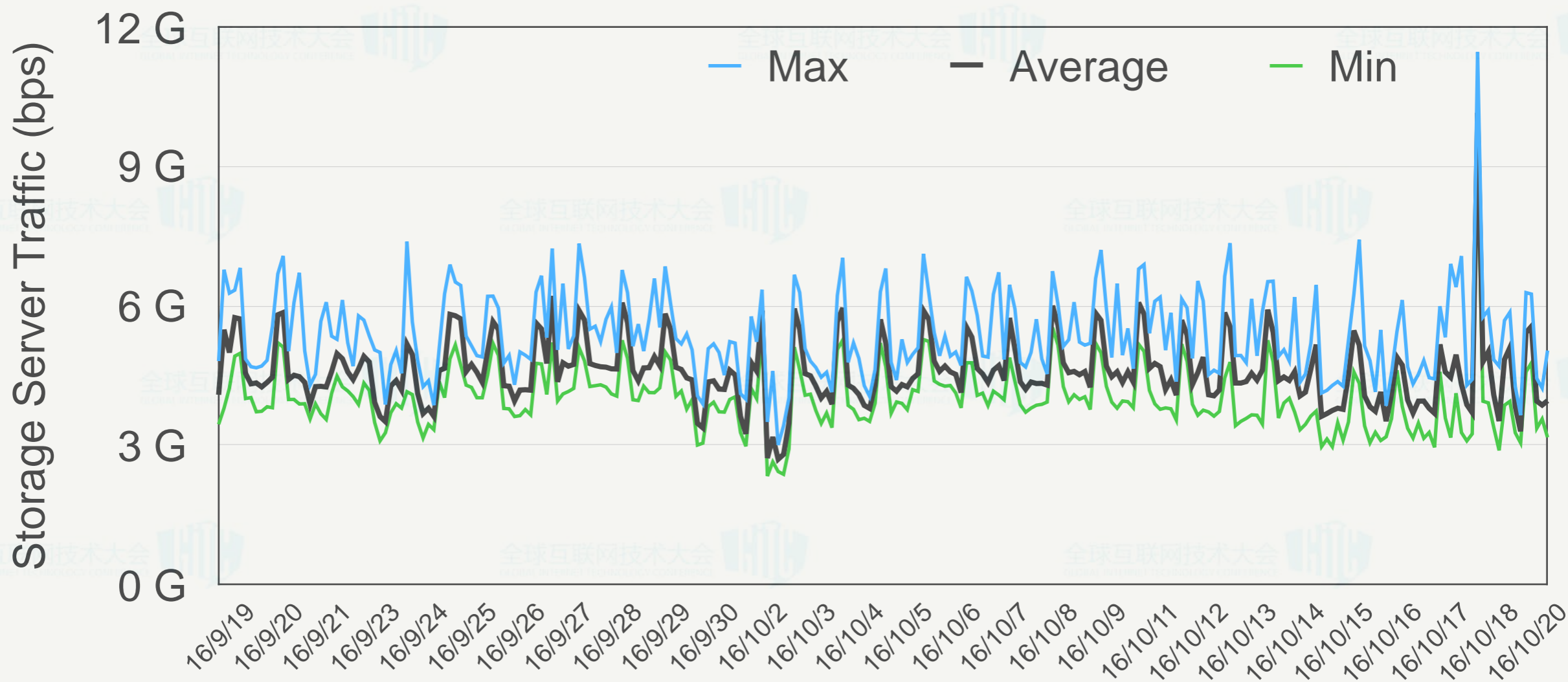
结论：HDD的顺序写入性能与SSD**匹配**

回写

- 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

- 介绍
- 云上的硬盘
- 成本优化方案
- 性能优化技术
- 性能评估
- 结论

性能优化

- 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

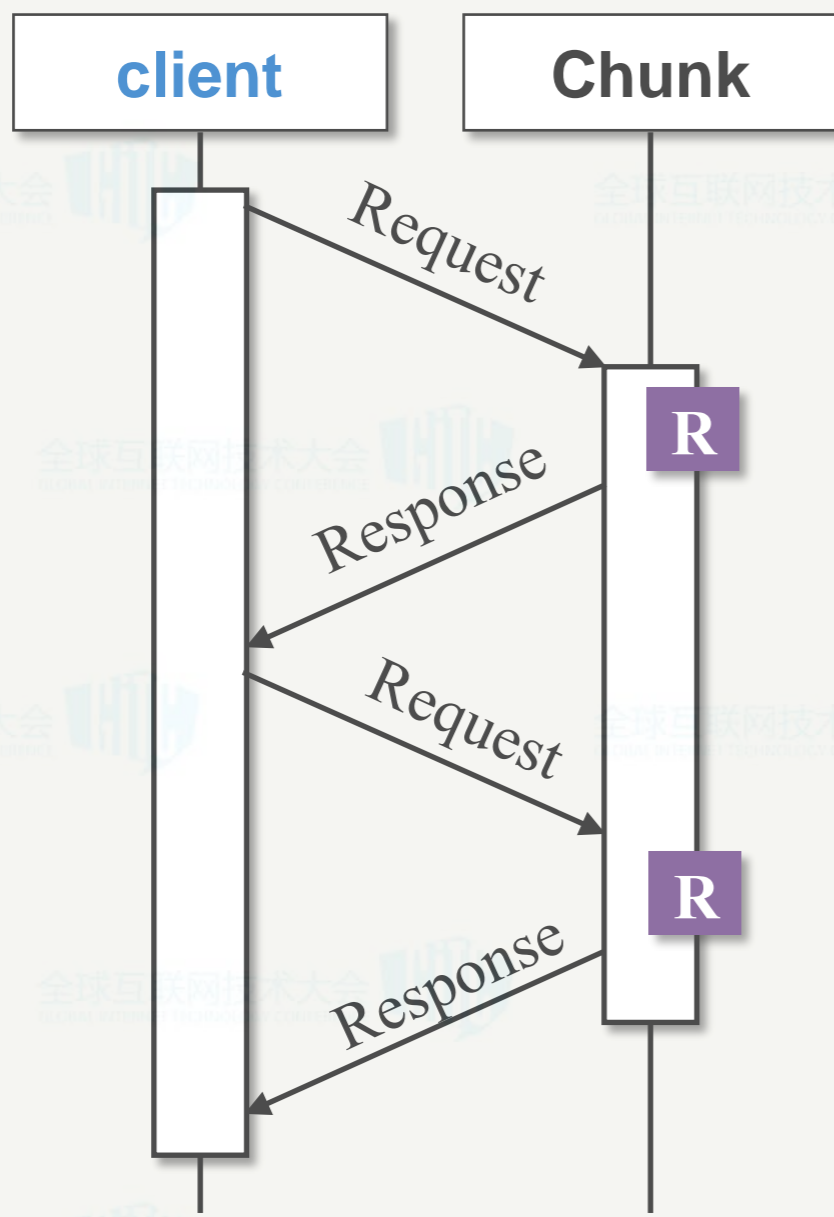
性能优化：代码效率

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

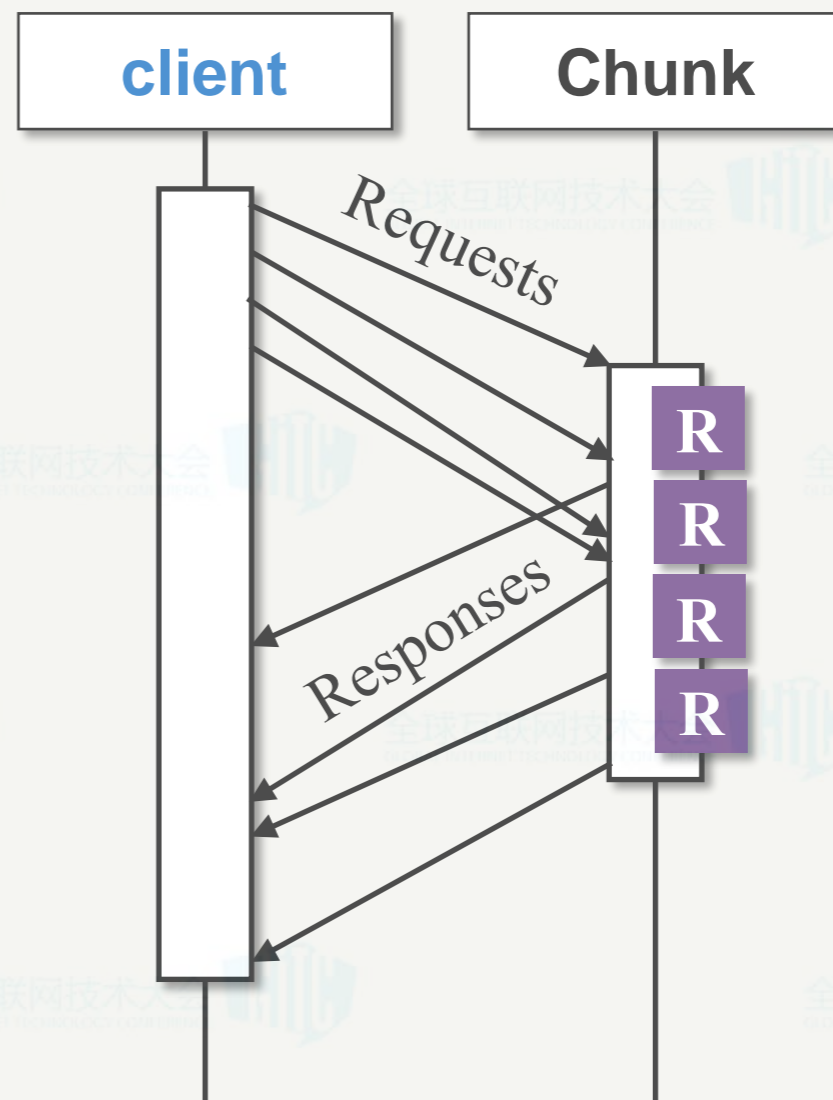
性能优化：并行

- 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

- 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

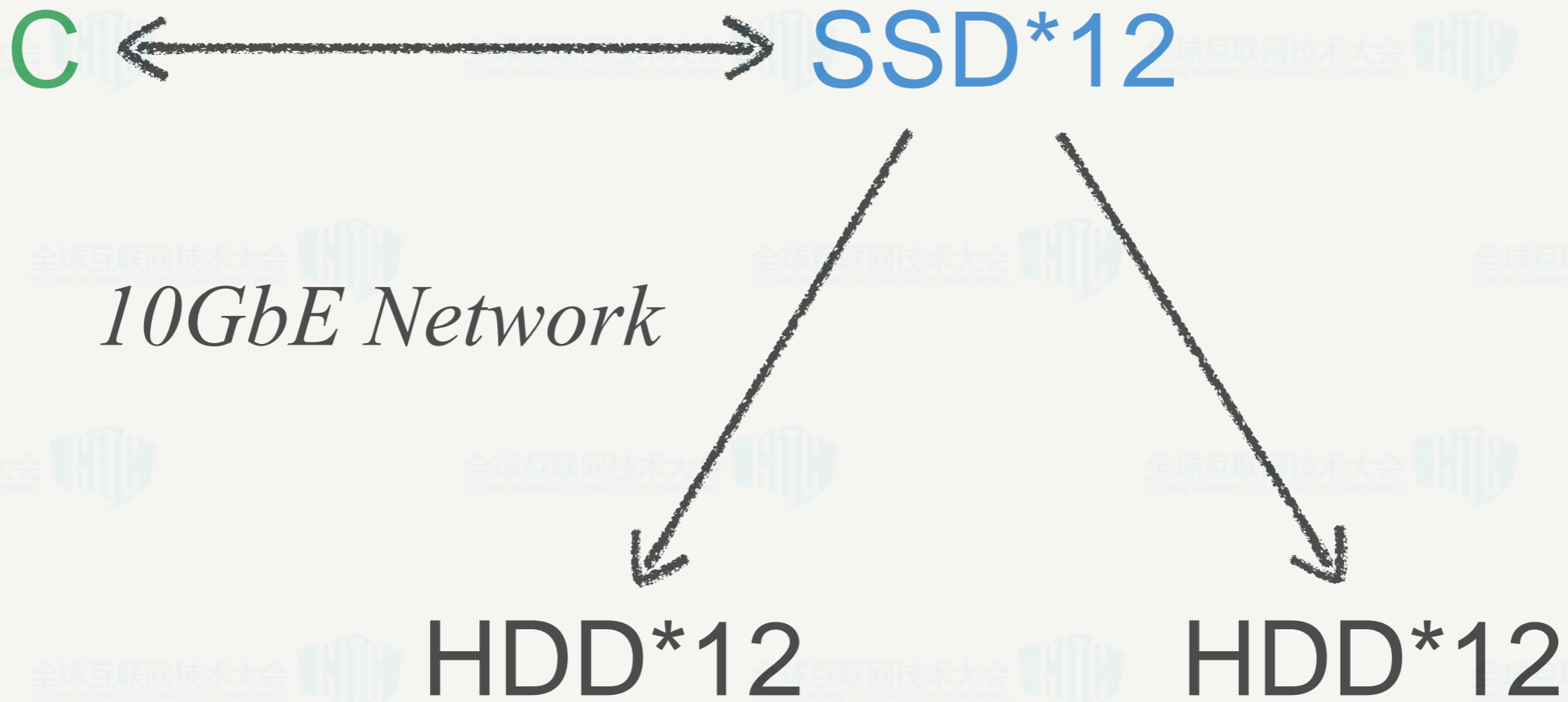
内核功能不够用了

- 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

- 介绍
- 云上的硬盘
- 成本优化方案
- 性能优化技术
- 性能评估
- 结论

Ursa-on-SSD性能初步测试



初步性能测试

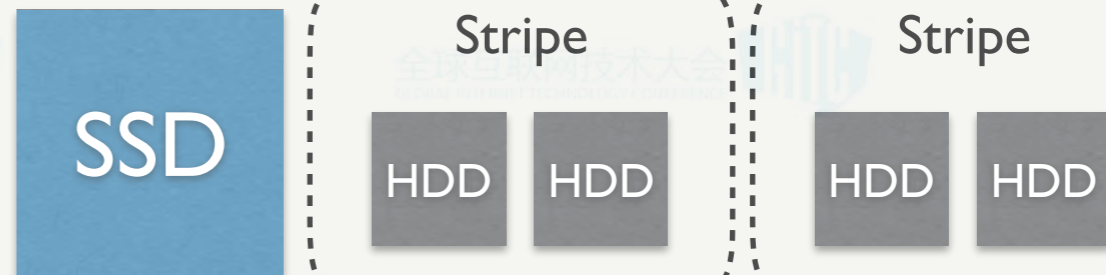
- 补齐HDD吞吐率短板 (SubChunk)

Replication



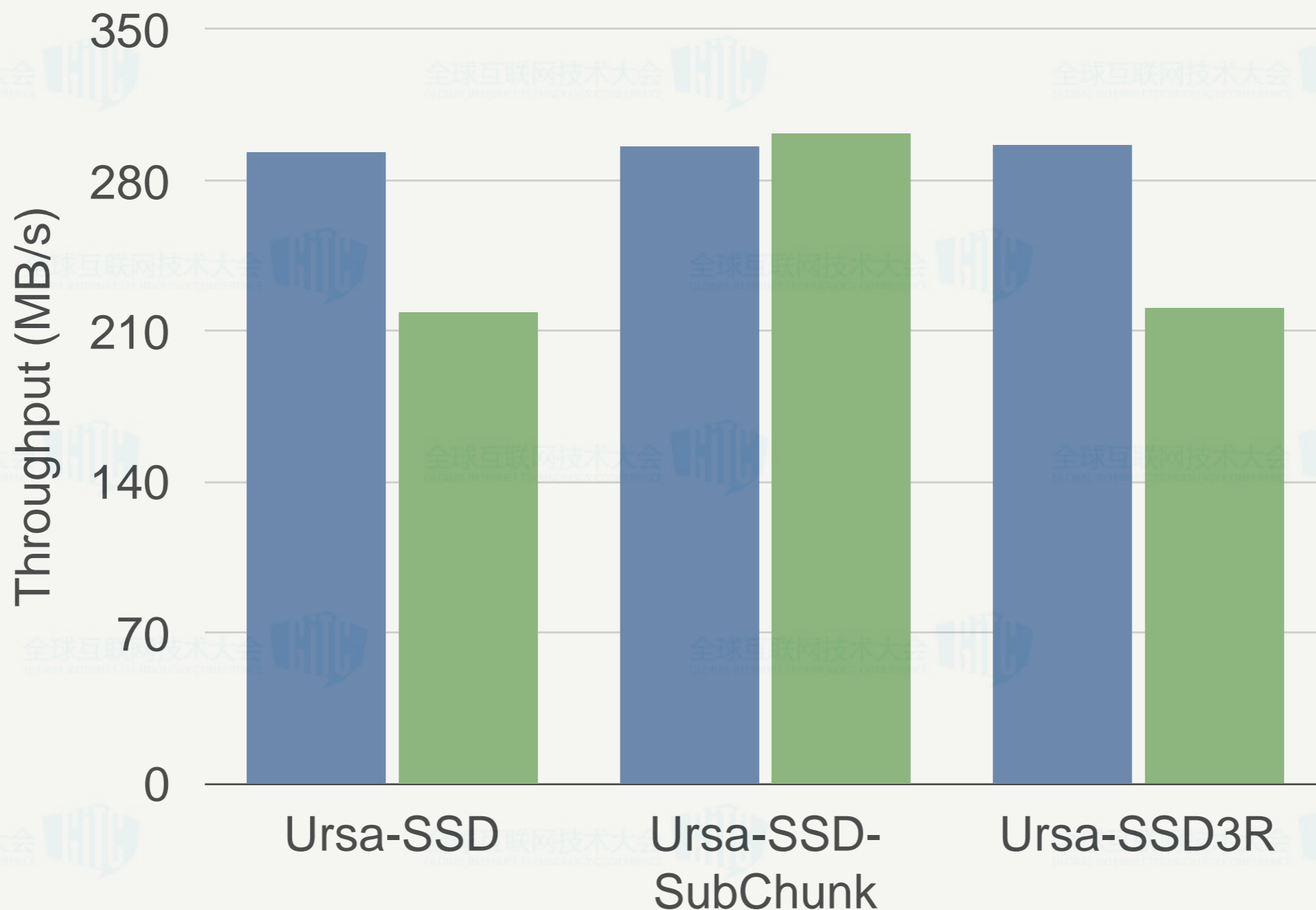
==>

Replication



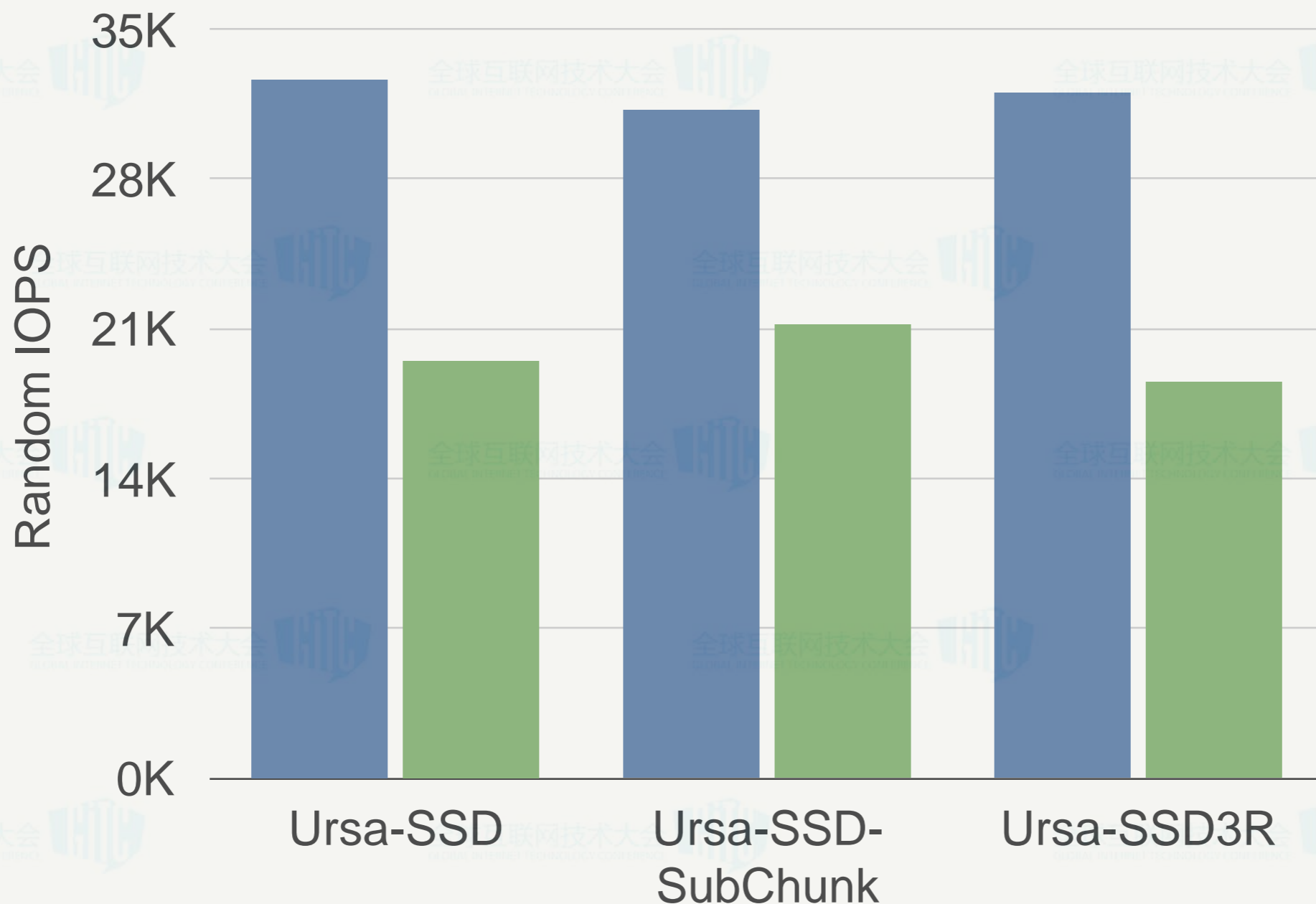
初步性能测试

■ read (non-cached) ■ write (SSD non-cached, HDD cached)



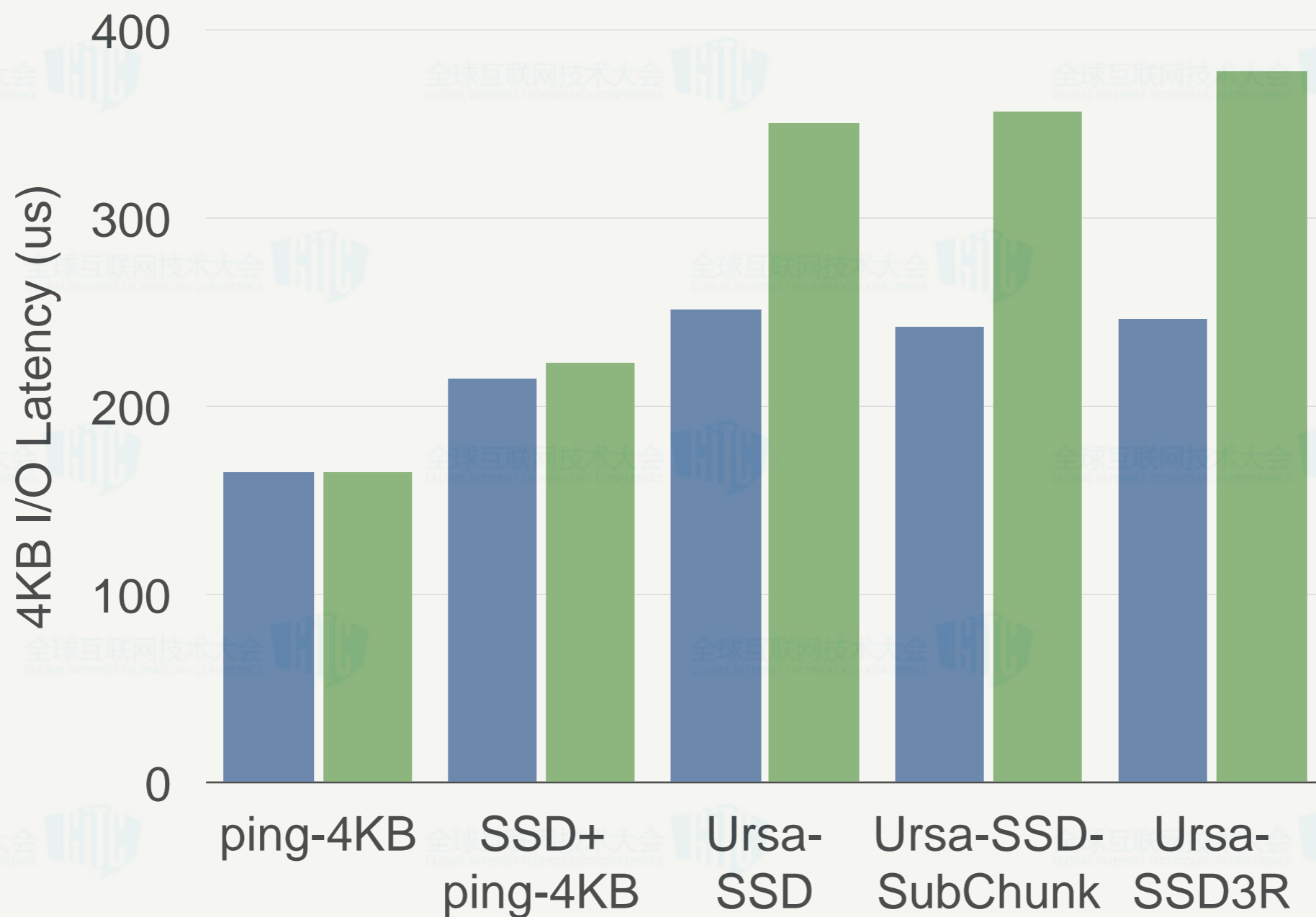
初步性能测试

■ read (non-cached) ■ write (SSD non-cached, HDD cached)

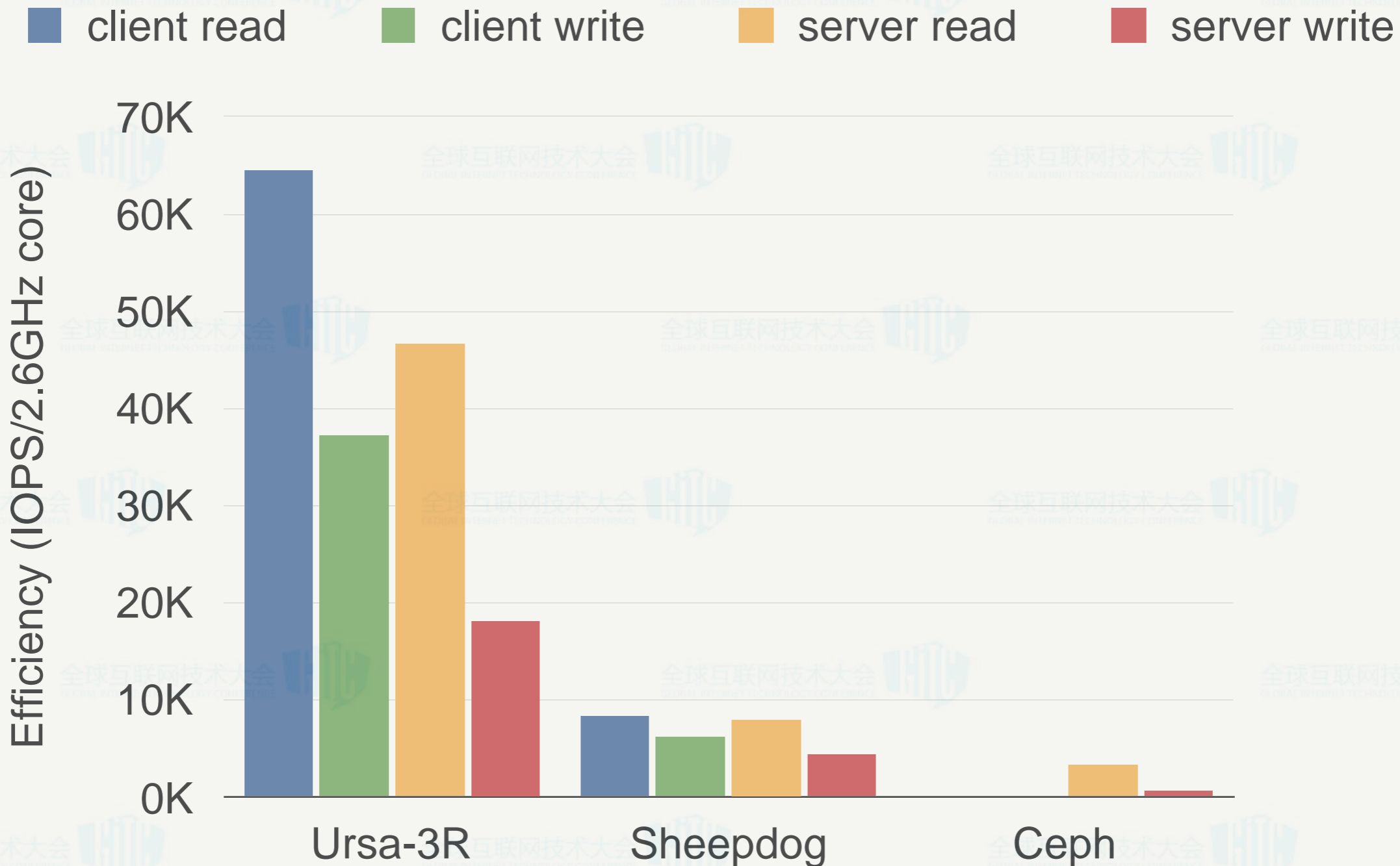


初步性能测试

■ read (non-cached) ■ write (SSD non-cached, HDD cached)



系统效率



结论

- 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

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

美团云
Meituan Open Services

谢谢!

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE



Q&A

<https://mos.meituan.com>