SSD在新浪数据库平台应用实践

杨尚刚

微博 @zolker

自我简介

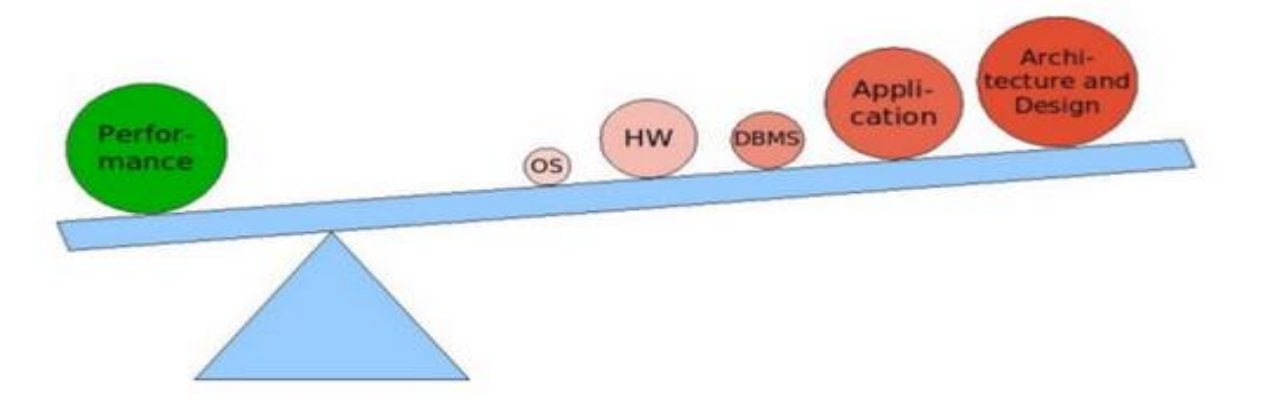
- 2011年加入新浪
- 负责新浪微博核心数据库架构设计和优化
- 负责新浪数据库平台底层软硬件平台优化
- 理念：设计简洁的架构

MySQL @ Sina

- MySQL版本官方社区版
- 数据量PB级
- 每日承担访问量百亿级
- 2011年开始使用SSD,数量在万片级别
- 集群上千台
- 承担新浪微博核心数据持久化存储

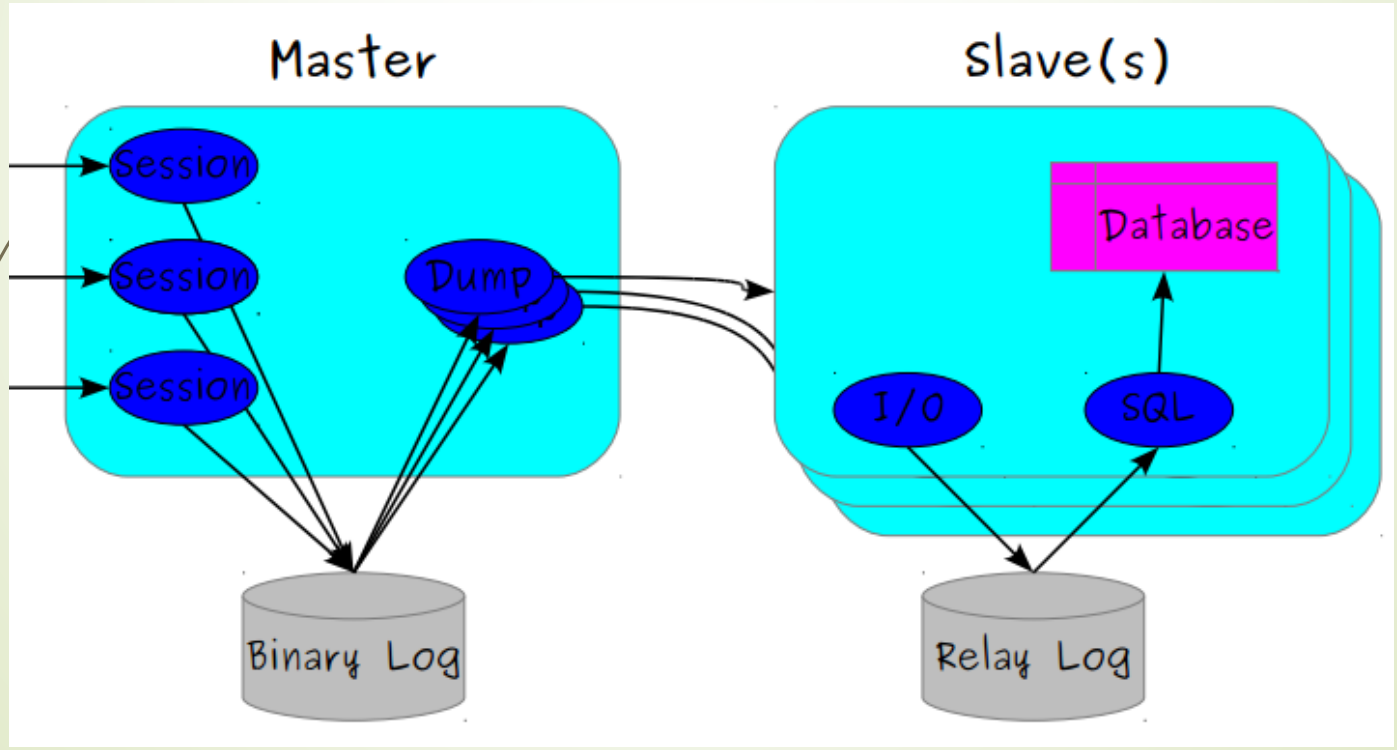
我们如何使用MySQL

- 读写分离
- Master-Slave 复制
- Sharding
- 多IDC容灾
- 完善的监控和容灾策略
- 良好的Schema设计



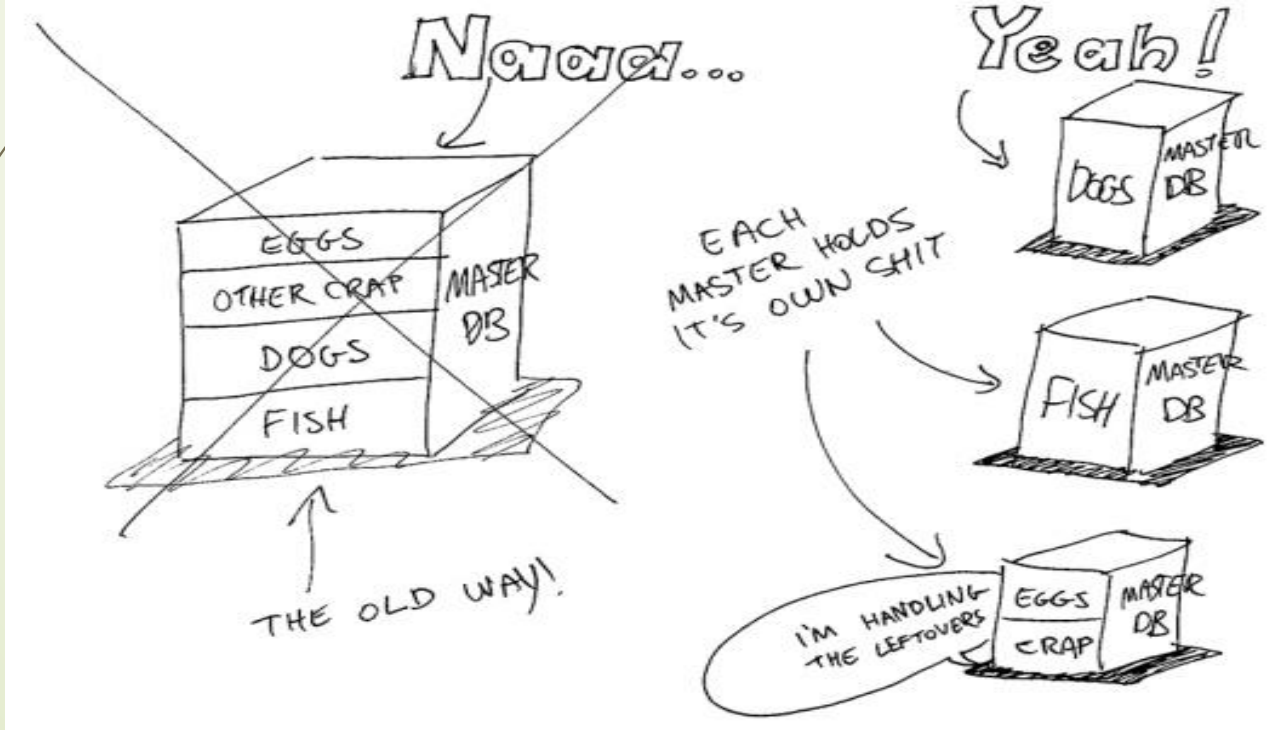
MySQL Replication

- 解决读性能问题
- 数据容灾
- 原生复制有明显的性能瓶颈



MySQL Sharding

- **Sharding is very complex, so it's best not to shard until it's obvious that you will actually need to!**
- 不要轻易做**Sharding**



后面会发生什么？

现实往往是很残酷的

预估的性能

实际的性能

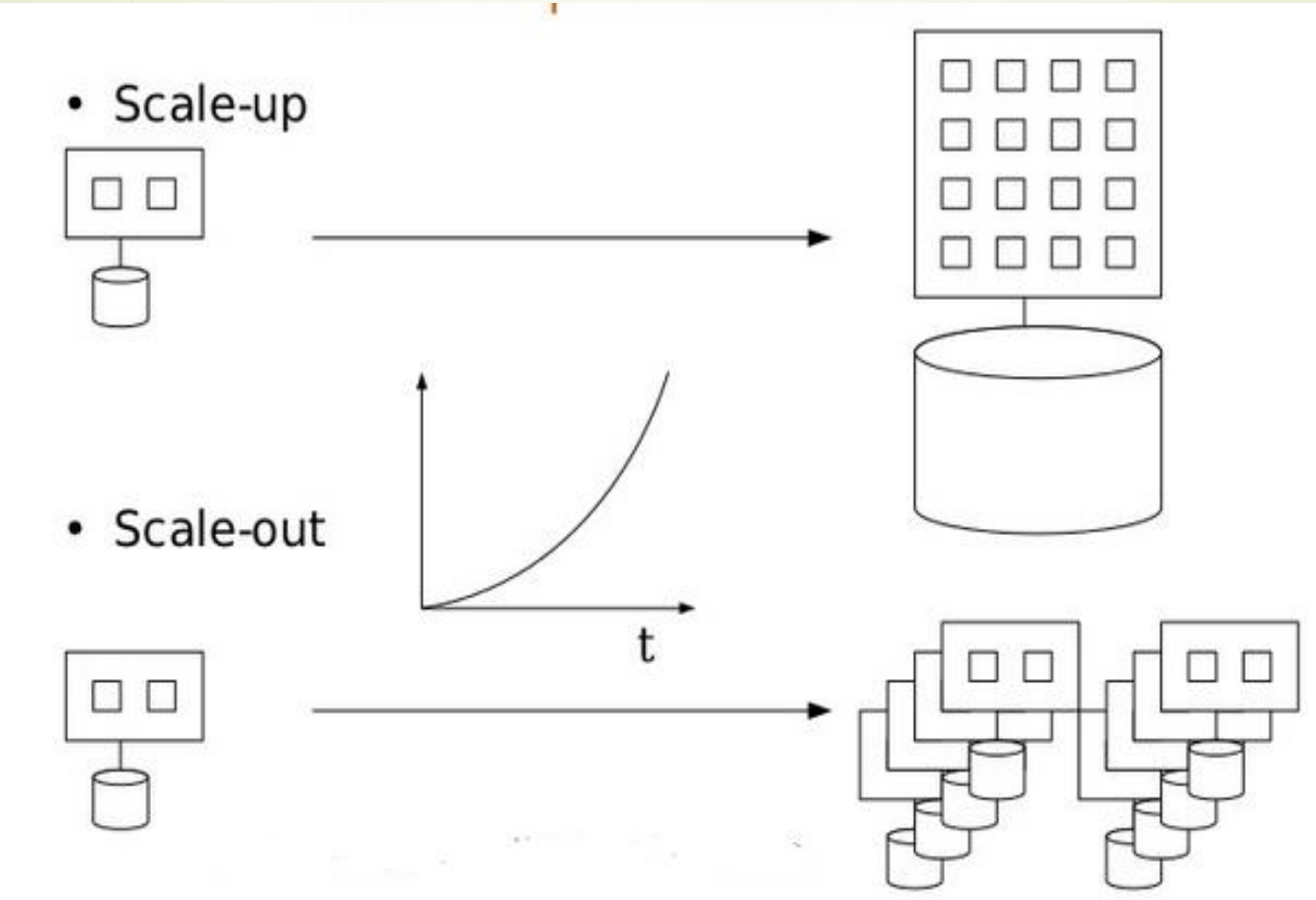
神马原因

- 服务峰值需求
- 服务SLA需求
- TCO的需求

我们该如何解决呢

- Scale UP VS Scale Out
- 继续Sharding

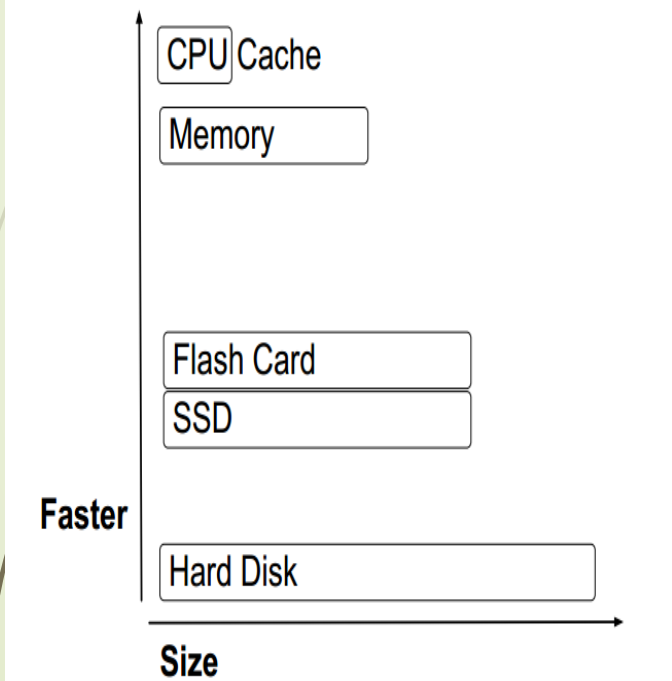
Scale UP VS Scale Out



当时的现状

真得无解了吗？
答案当然是No
因为SSD

必须知道的数字

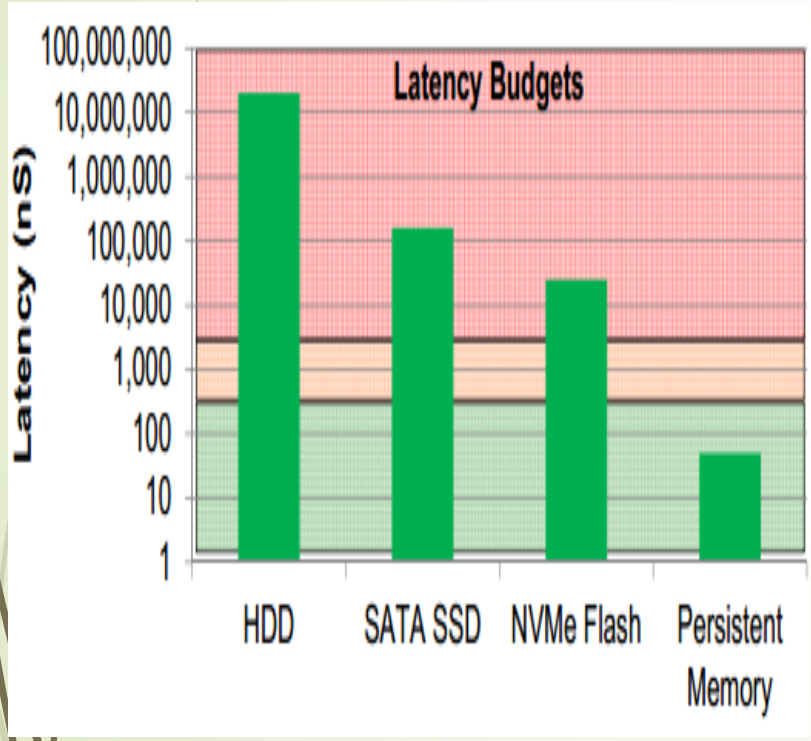


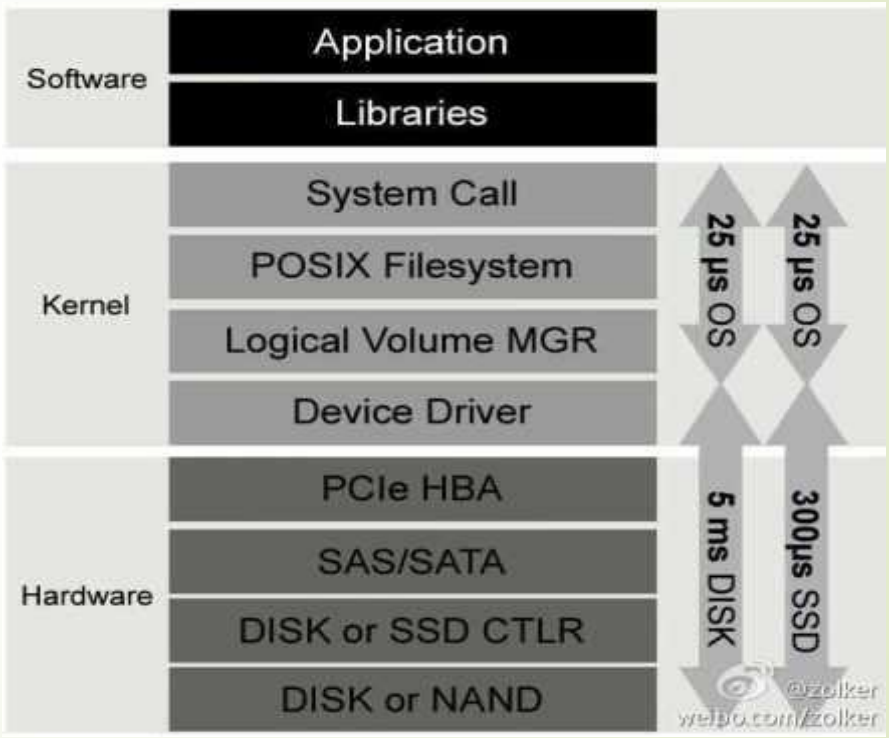
L1 Processor Cache Reference	0.5
Branch mis-predict	5
L2 cache reference	7
Mutex lock/unlock	25
Main memory reference	100
Compress 1K bytes with Zippy	3,000
Send 2K bytes over 1 Gbps network	20,000
MLC SSD Access Latency	50,000
Typical application IO response time from an SSD	150,000-450,000
Read 1 MB sequentially from memory	250,000
Round trip within same datacenter	500,000
Hard Drive Disk (HDD) seek	10,000,000
Read 1 MB sequentially from disk	20,000,000
Send packet California->Netherlands->California	150,000,000

SSD特点

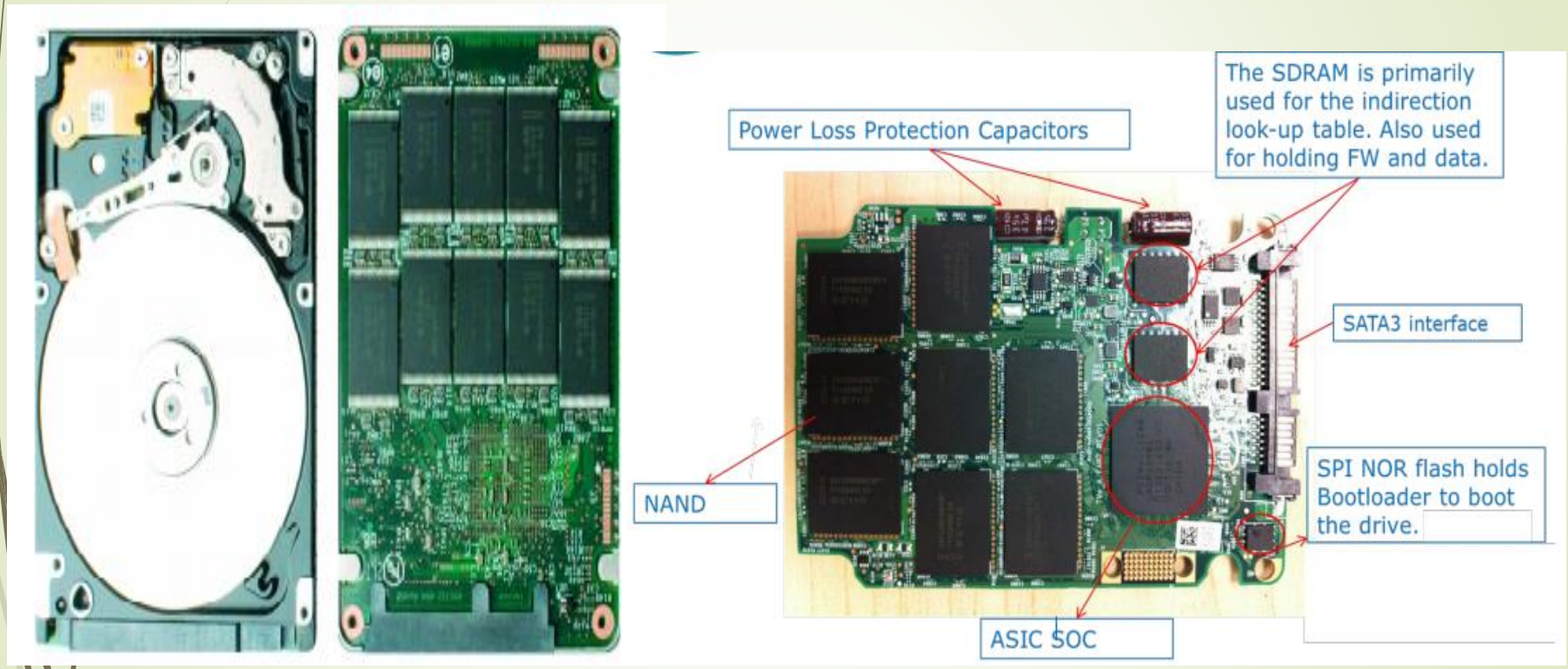
- 使用电子元件实现数据的永久存储，没有机械组件
- 主要有基于DRAM和Flash两种
- 兼容目前HDD的接口
- 随机读写性能非常好
- 低延时，延时在微秒级

SSD有多快


Latency Budgets
Latency (nS)
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
100
10
1
HDD
SATA SSD
NVMe Flash
Persistent Memory


Software
Application
Libraries
Kernel
System Call
POSIX Filesystem
Logical Volume MGR
Device Driver
Hardware
PCIe HBA
SAS/SATA
DISK or SSD CTLR
DISK or NAND
25 µs OS
25 µs OS
5 ms DISK
300µs SSD
@zolker
weibo.com/zolker

HDD VS SSD

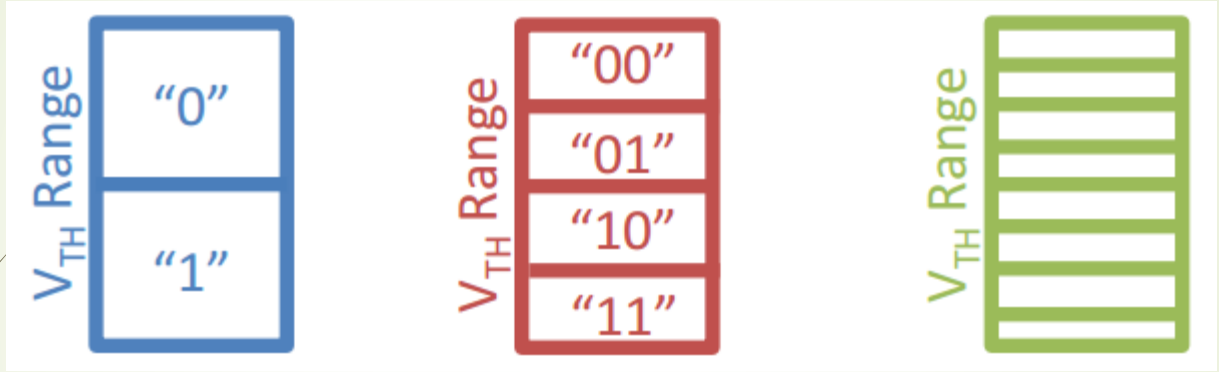

The SDRAM is primarily used for the indirection look-up table. Also used for holding FW and data.
Power Loss Protection Capacitors
SATA3 interface
NAND
SPI NOR flash holds Bootloader to boot the drive.
ASIC SOC

SSD关键技术

- 1.Read &Write
- 2. Flash translation layer
- 3. WL(Wear leveling)
- 4. Garbage collection & Trim
- 5. Over –provisioning &Write amplification
- 6. Bad block management & ECC

NAND Type

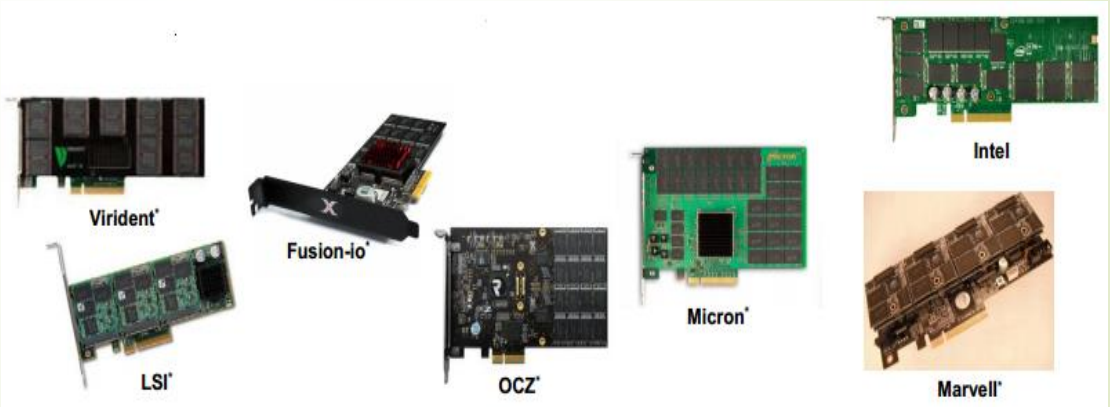
SLC **MLC** **TLC**



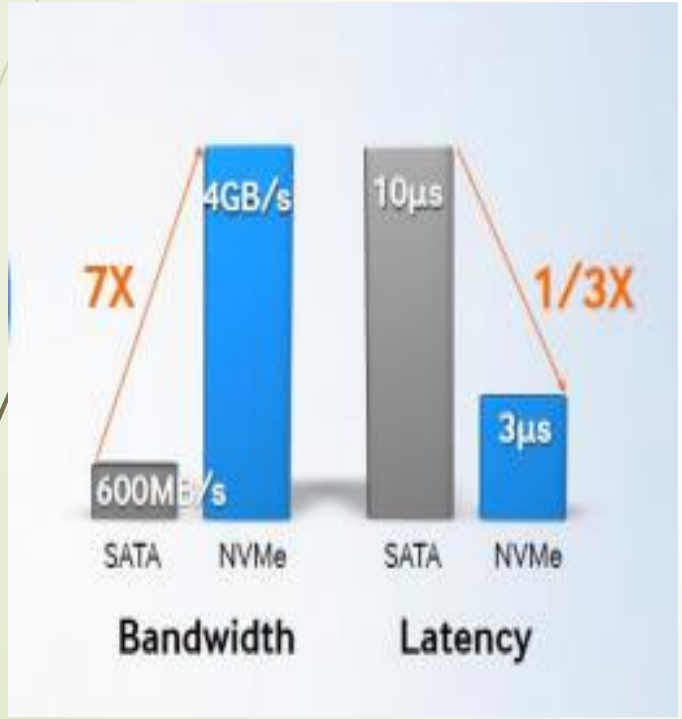
类型	性能	耐久度	价格	适合对象
SLC				稳定性和性能 要求苛刻的应用环境
MLC				对价格敏感又要 性能和稳定兼顾的环境
TLC				价格是首位， 性能和稳定性有基本保证即可

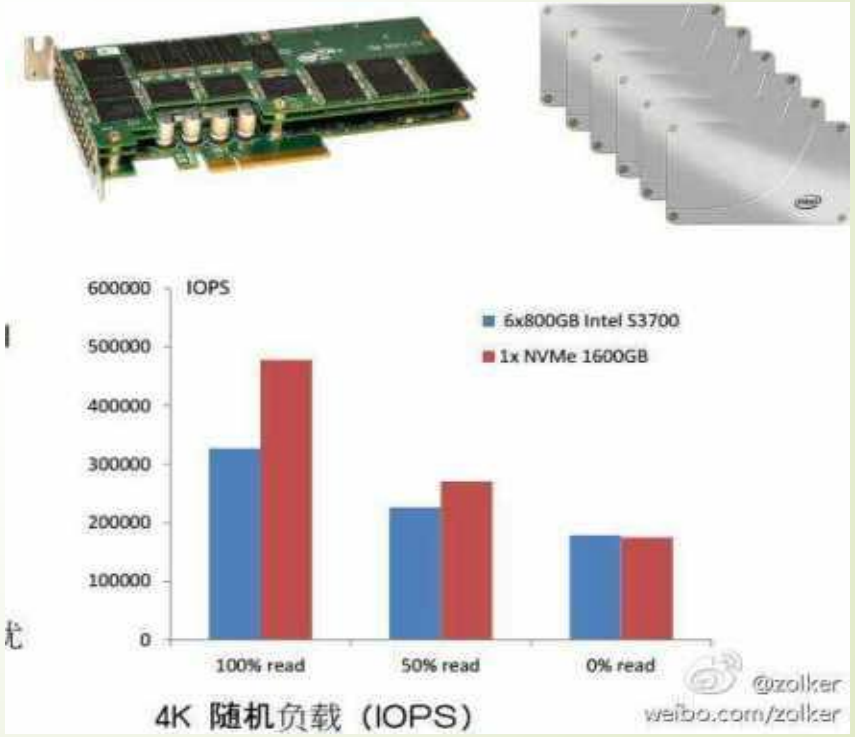
PCIe VS SATA SSD

	容量	4K随机读/写	每GB	带宽	延时
PCIe	1TB+	70w+/20w+	2-3$	8GB/s	10-40 μs
SATA	<1TB	7w/2w	1-2$	6Gb/s	50-100μs



NVMe is coming

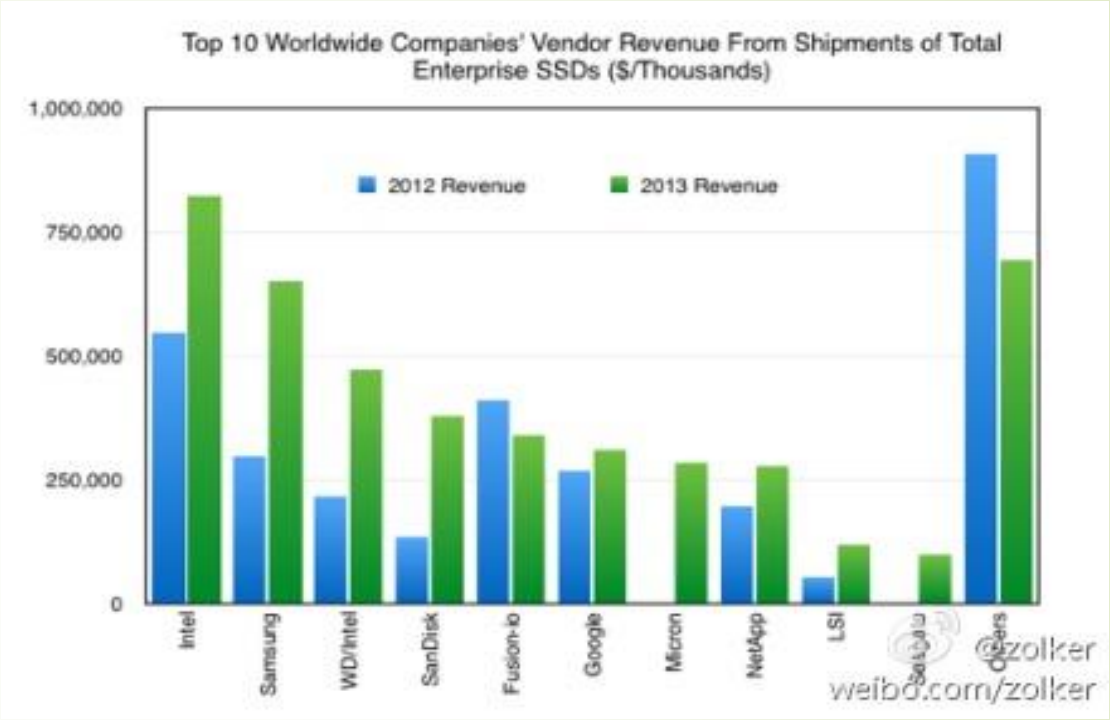




选择SSD

- 价格
- 容量
- 性能
- 可靠性

  MTBF，UBER，TBW
- 稳定性
- IO 延时
- 可维护性
- 功耗

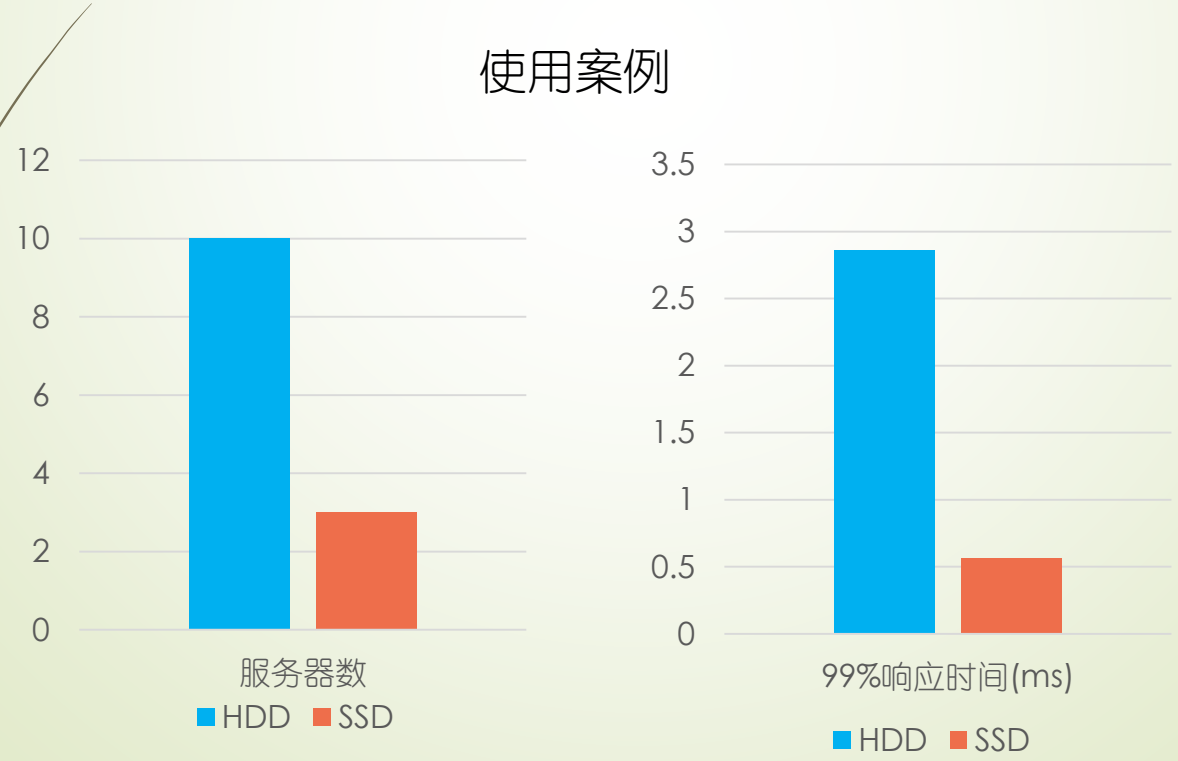


我们如何使用SSD

- 作为持久化存储，替换原有15K HDD
- 混合存储
- 在缓存场景作为扩展内存使用

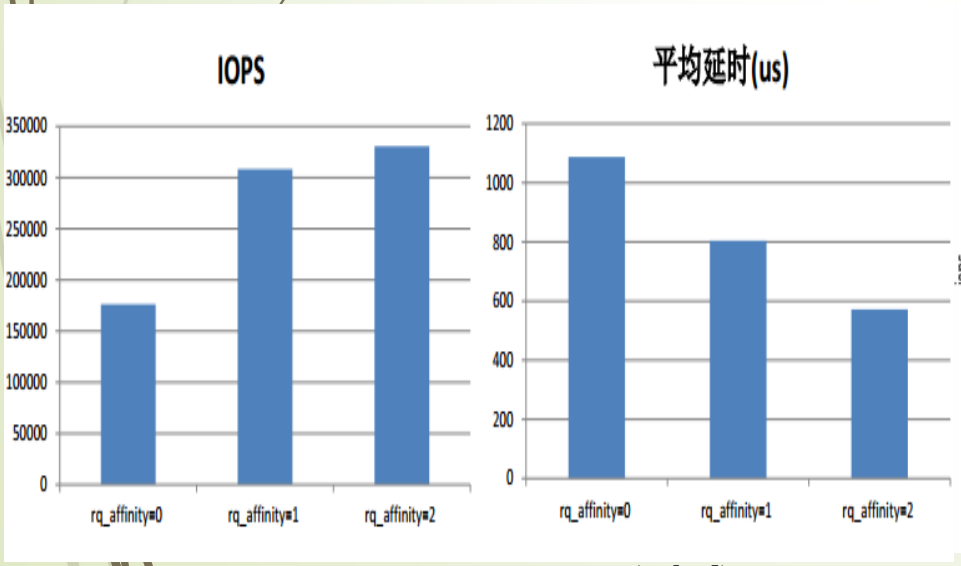
SSD存储

- 主要是单机10块SATA SSD
- 硬Raid+Raid5
- 从库优先使用SSD，峰值写入高主库也会采用SSD
- 采用PCIe SSD

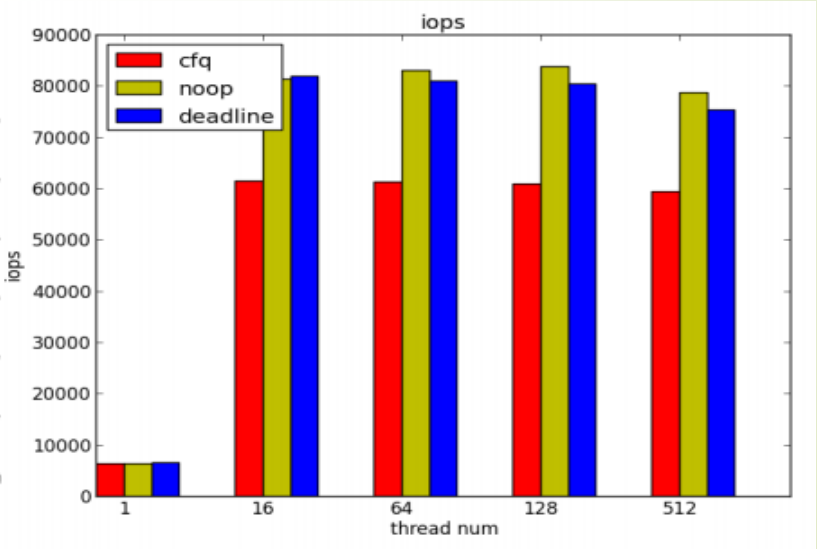


针对SSD的IO 优化

- echo noop/deadline >/sys/block/[device]/queue/scheduler
- echo 0 > /sys/block/[device]/queue/add_random
- echo2 > /sys/block/[device]/queue/rq_affinity(CentOS 6.4以上)
- echo 0 > /sys/block/[device]/queue/rotational
- 文件系统关闭barrier
- FastPath技术



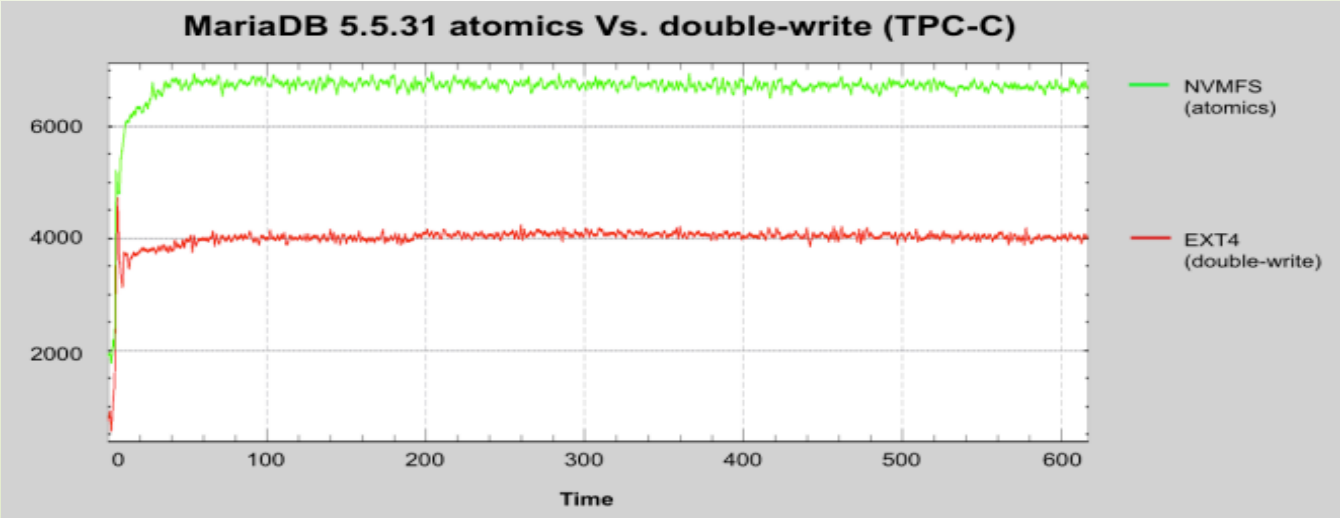
CPU rq_affinity测试



IO调度算法测试

针对SSD的MySQL优化

- 在5.5以上，提高innodb_write_io_threads和innodb_read_io_threads
- innodb_io_capacity需要调大
- 日志文件和redo放到机械硬盘，undo放到SSD
- atomic write,不需要Double Write Buffer
- InnoDB压缩
- 单机多实例+cgroup



混合存储

- Flashcache

  SSD作为读写的缓冲，数据最好有热点，对使用场景有要求

  Facebook的Mohan开发，并且线上一直在用，最新版本3.0　　对淘汰算法和读写效率做了优化

  微博核心数据库之前线上有使用

- LSI Cachecade & Nytro MegaRaid

混合存储优点非常明显

热点集中下性能表现优异

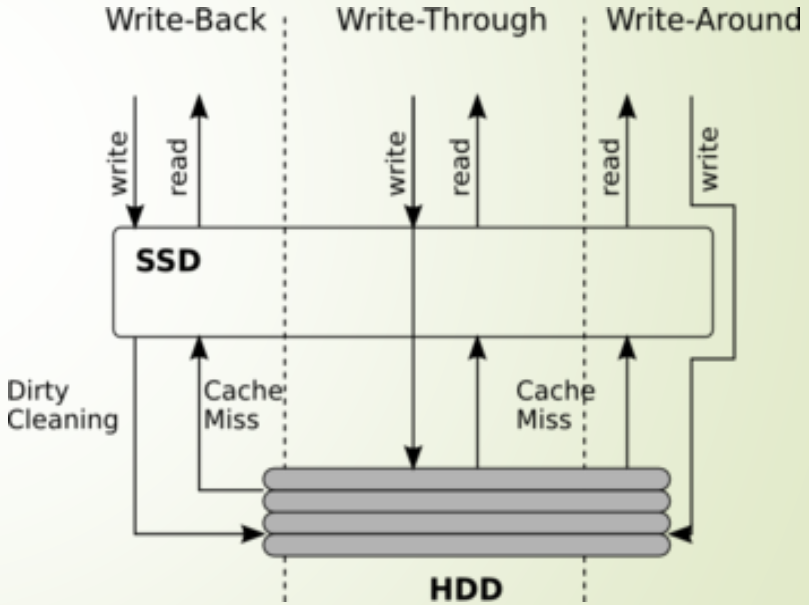
节省成本

不过缺点也非常明显

运维成本高

对SSD寿命损害大

性能不稳定



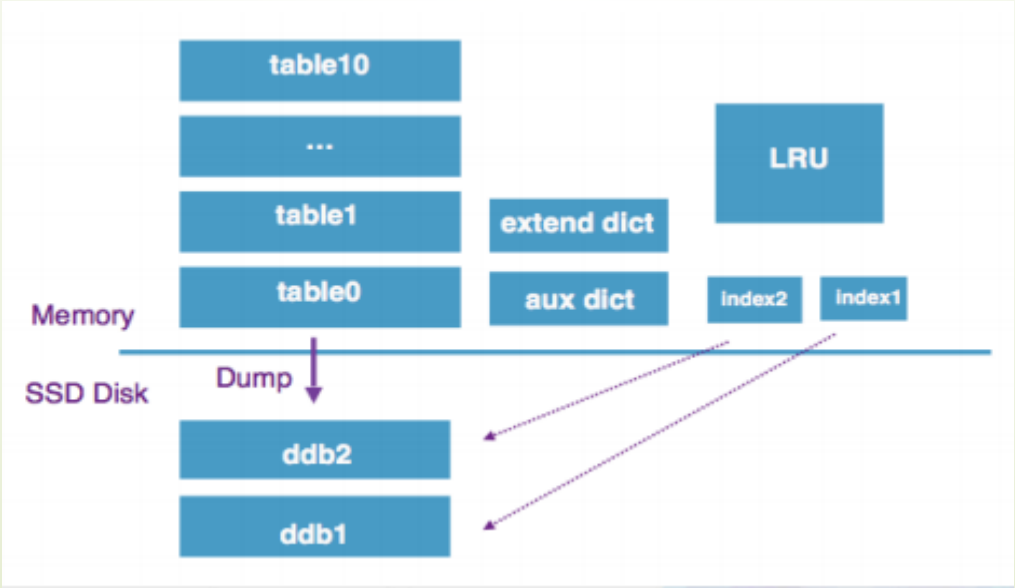
Flashcache原理

cache的特性决定了它并不适合高并发OLTP

缓存存储

- 把内存和SSD虚拟成一个设备来使用，实现缓存数据的持久化
- 缓存容量突破单机内存的限制，减少后端的压力
- 降低成本
- 提高服务可用性

Counter 落地SSD，容量提升20倍，8个月→10年



面临的问题

- SATA SSD做Raid后Raid卡性能瓶颈
- SSD寿命监控和生命周期管理
- SSD自动化测试，测试流程和指标的规范化
- 挖掘利用SSD性能

SSD In Future

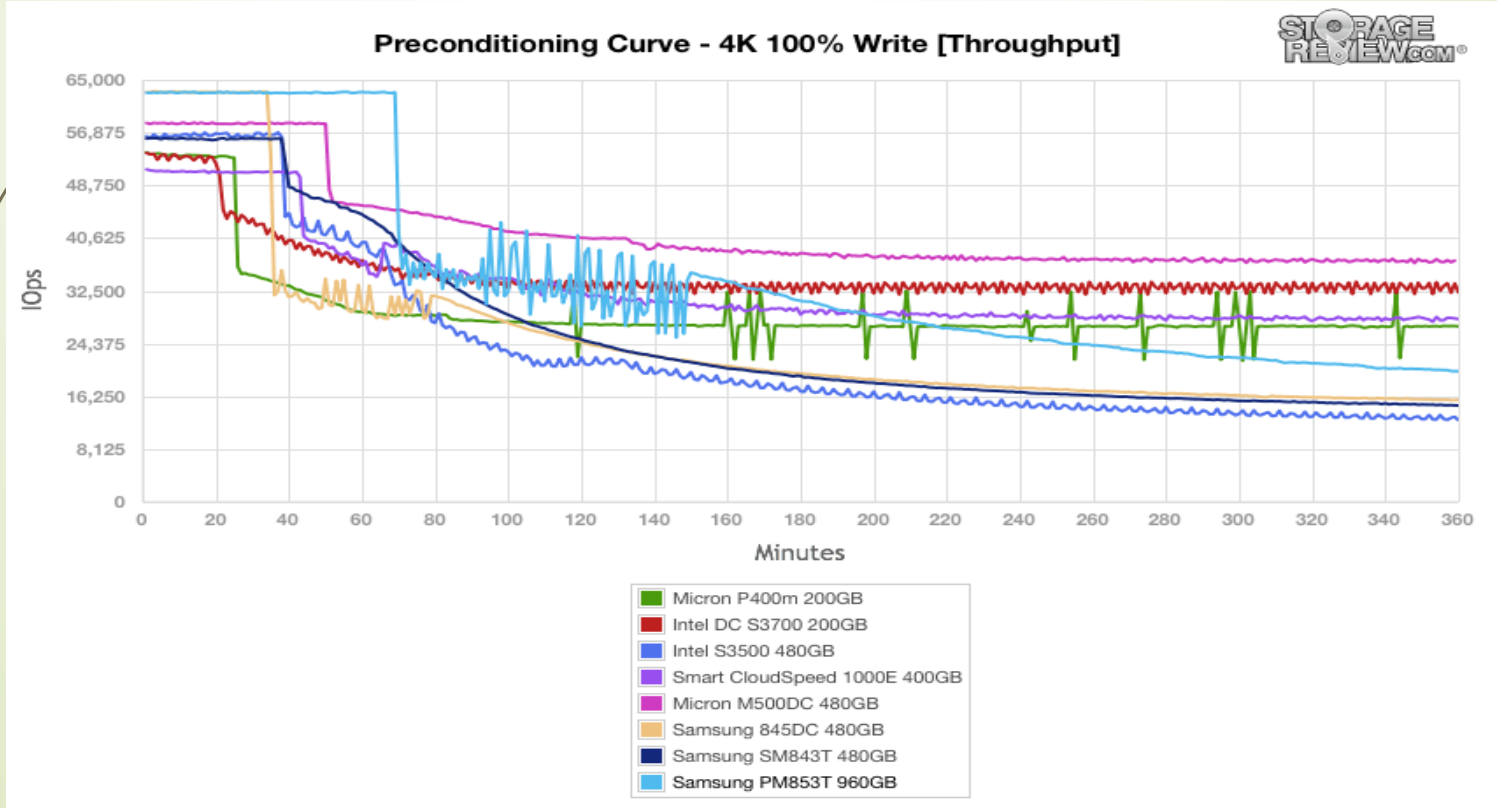
- 12G Raid卡
- No Raid，HBA卡直连
- NVMe
- 使用TLC & 3D V-NAND, 打造全flash数据库平台
- 定制SSD
- 更低的成本，更大的容量，更高的性能

No Raid

- 现有Raid卡很容易成为SSD性能瓶颈
- Raid有容量损失
- SSD可靠性已经得到验证，上层应用有多级容错策略
- Raid卡也可能出问题的

TLC在企业级应用

- 寿命如何
- 性能稳定性
- 故障率怎么样
- 能否取代高速机械硬盘，完成使命




It's FLASH
Stupid

SSD 性能极限

联系方式

微博或微信搜索 zolker