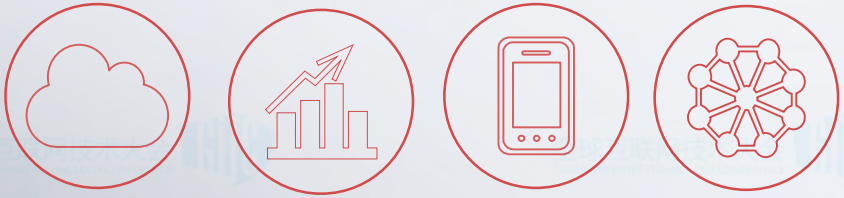
华为服务器在SSD领域的实践历程和展望

LEADING NEW ICT

华为技术有限公司 单彤

什么引发数据中心全面转向闪存化？

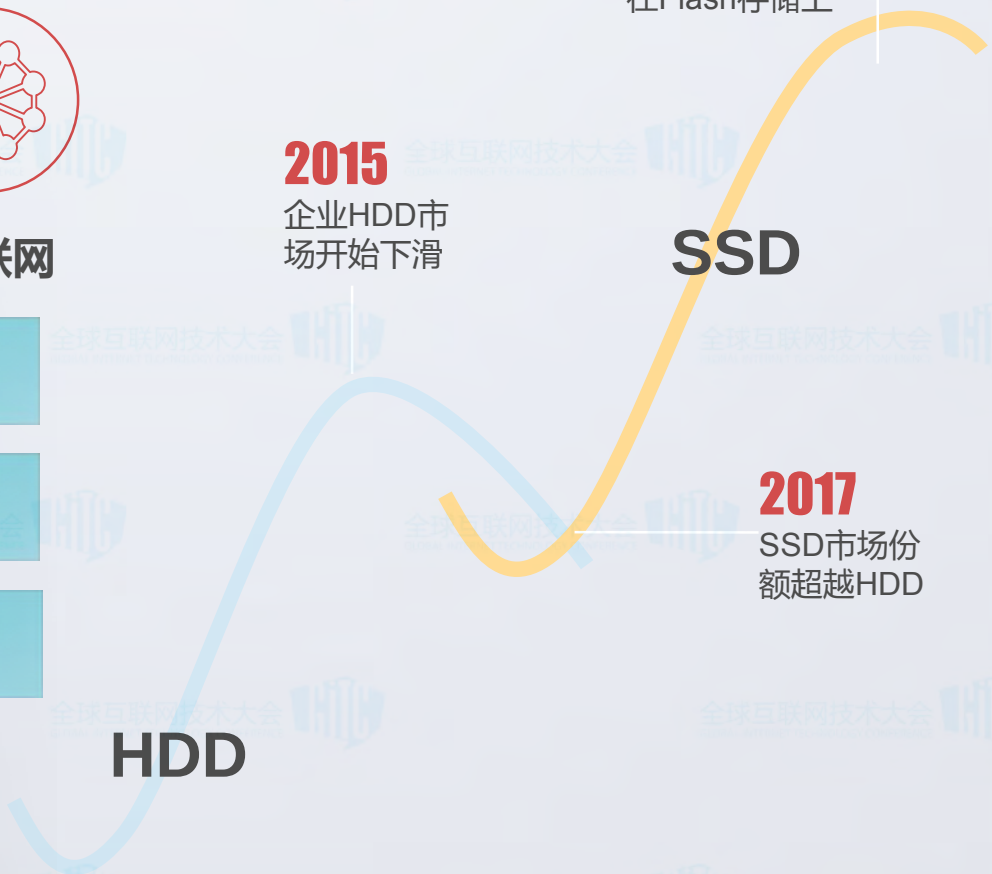


云　　大数据　　移动互联网　物联网

不可预期的业务高峰

日益增多的报表实时分析

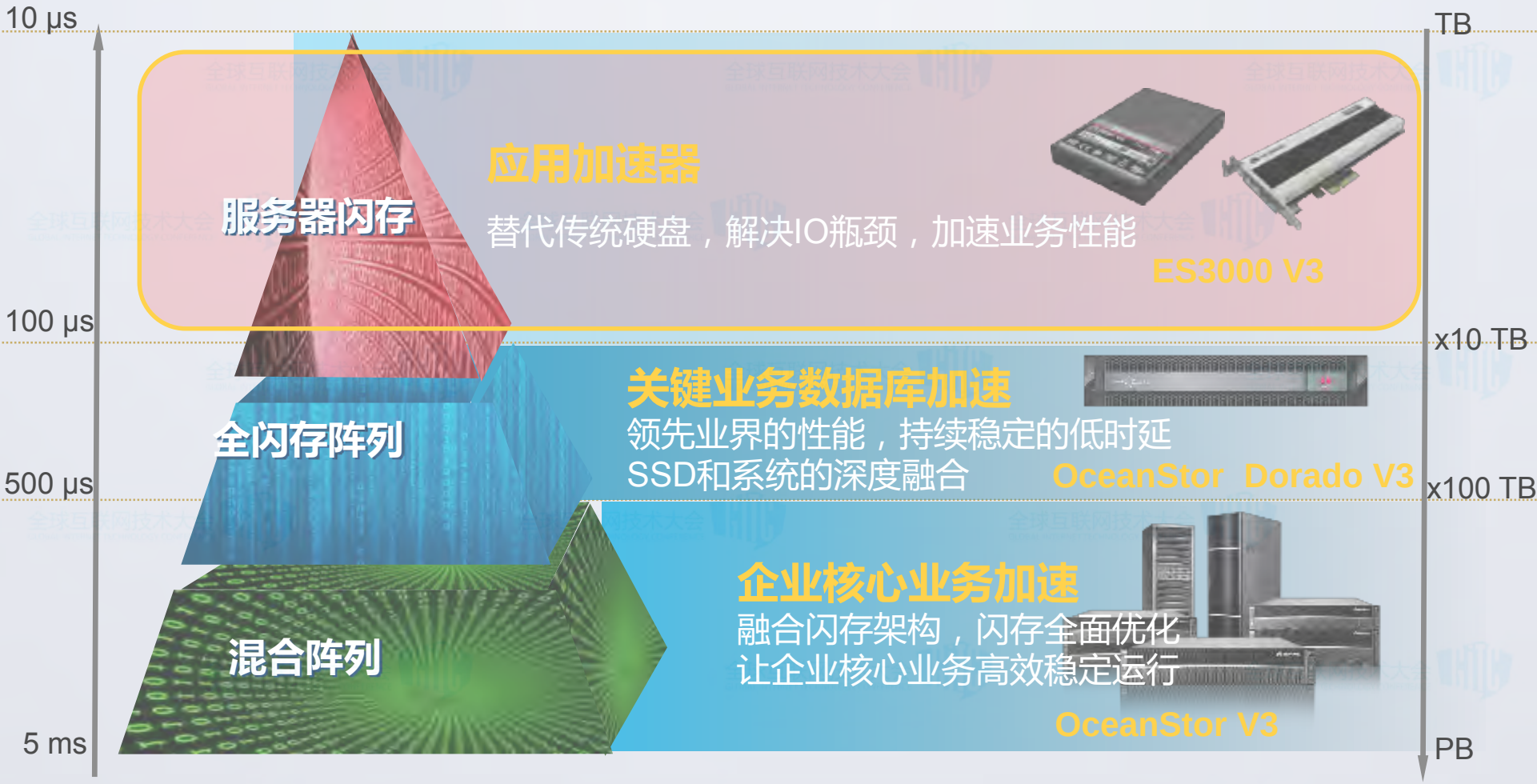
需持续提升的客户体验



Source: IT Brand Pulse, 2015

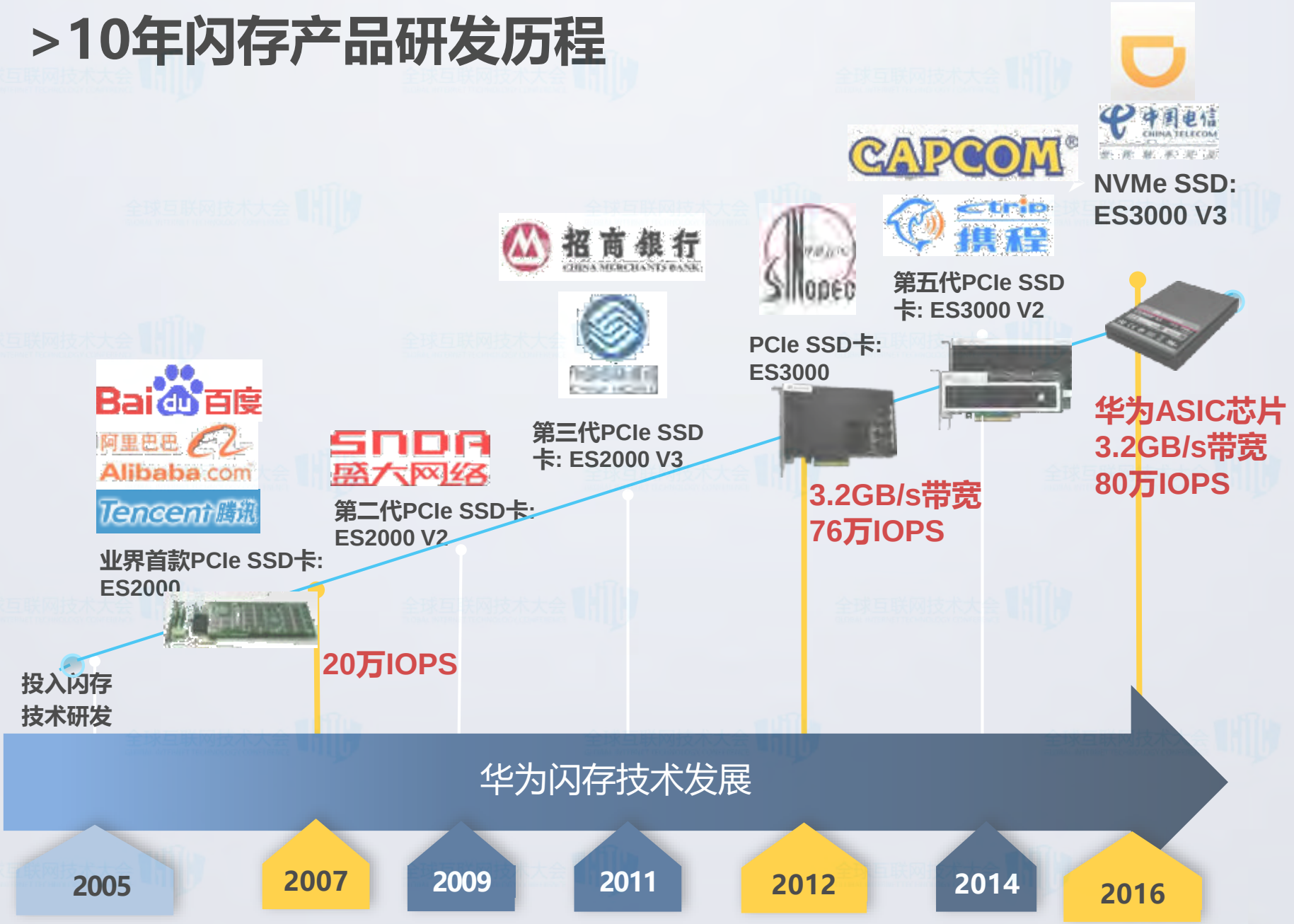
HUAWEI

华为端到端闪存解决方案



HUAWEI

>10年闪存产品研发历程



ES3000引领SSD走向NVMe时代

NVMe SSD盘

- **支持热插拔**
- **更高性能**

ES3000 V3

NVMe闪存风暴

- 华为SSD控制芯片和算法技术
- PCIe 3.0接口，NVMe 1.2标准
- U.2盘(2.5寸)，800GB~3.2TB容量
- 高达**3.2GB/s带宽、80万IOPS**性能

SATA SSD盘

- **支持热插拔**
- 高性能

PCIe SSD卡

- **更高性能**
- 不支持热插拔

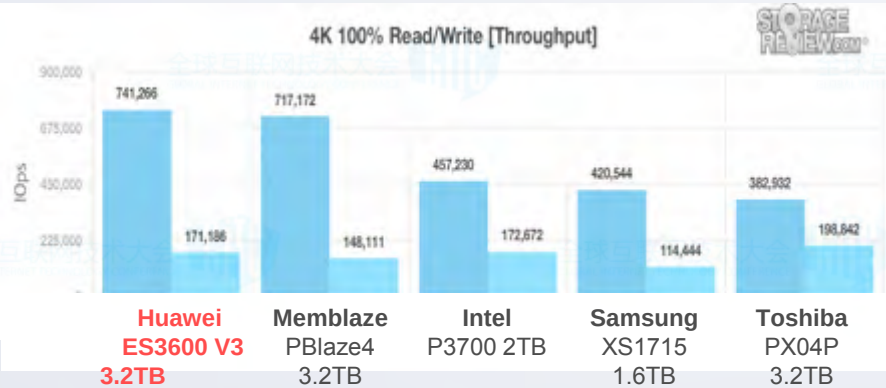
1块NVMe SSD	替代	4块SATA SSD
3,200 GB	总容量	3,200 GB
800,000	IOPS	336,000
22 W	最大功耗	45 W

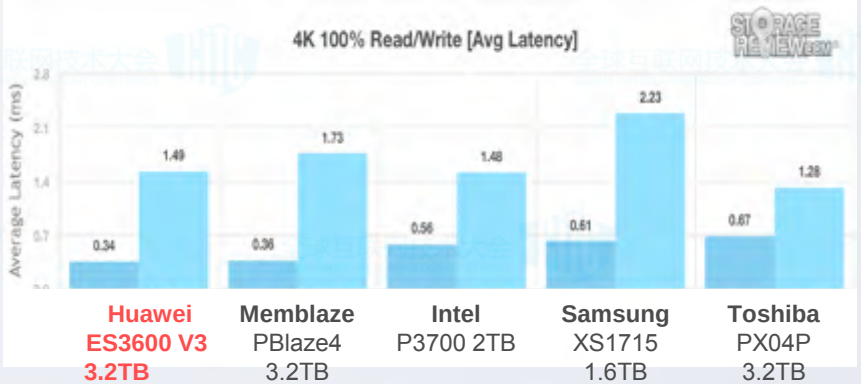
同等容量

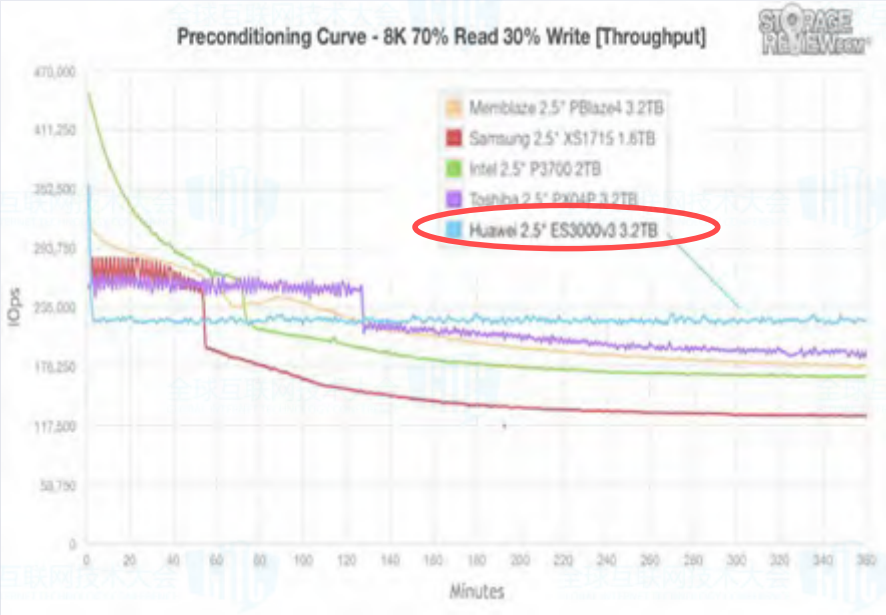
性能翻倍

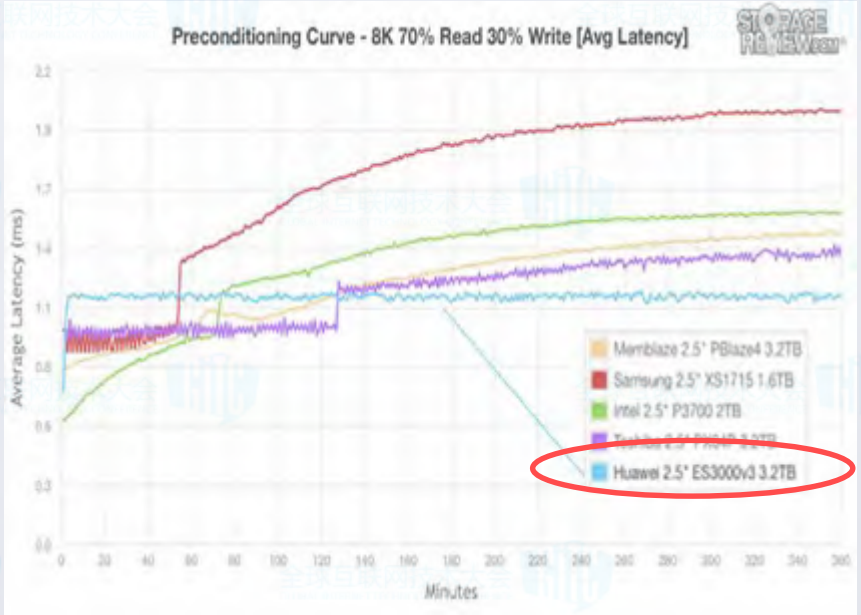
节能40%

Storagereview.com评测结果





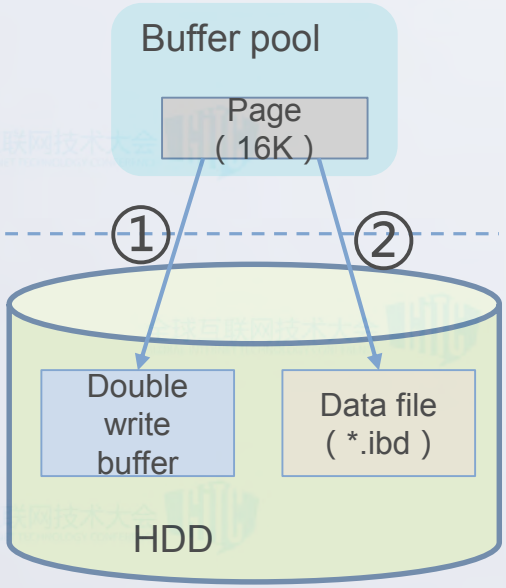




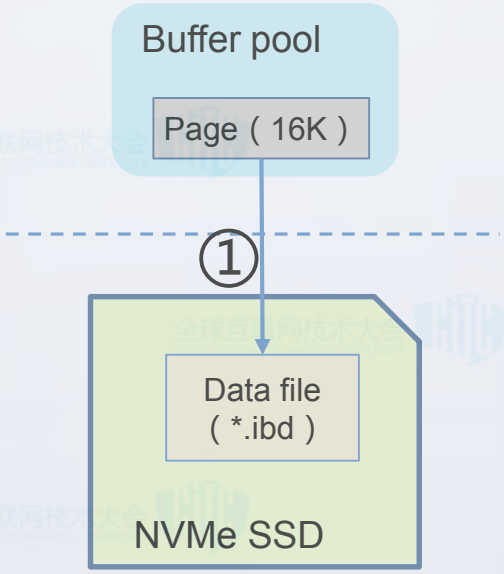
评测报告: http://www.storagereview.com/huawei_es3600_v3_nvme_ssd_review_25

HUAWEI

Atomic Write，优化MySQL与SSD的配合应用



为了防止Partial page write问题：
① Page先写到DWB
② 然后写到数据文件



外存储有atomic write特性：
① Page直接写到数据文件

- Atomic write特性可以减少SSD的写损耗、减少每个事务的写操作时间
- 使用Atomic write特性，在不同条件下MySQL的TPMC提升**7%**以上，SSD使用寿命延长**40%**以上

Multi-NameSpace，挖掘NVMe SSD的潜力

NVMe SSD

	容量	性能（业界均值）
2016	主流产品最大容量为3.2TB	600K read IOPS，100K write IOPS
2017	大部分厂家的路标产品最大容量6.4~8TB	750K read IOPS，150K write IOPS
2018	大部分厂家的路标产品最大容量16~32TB	NA

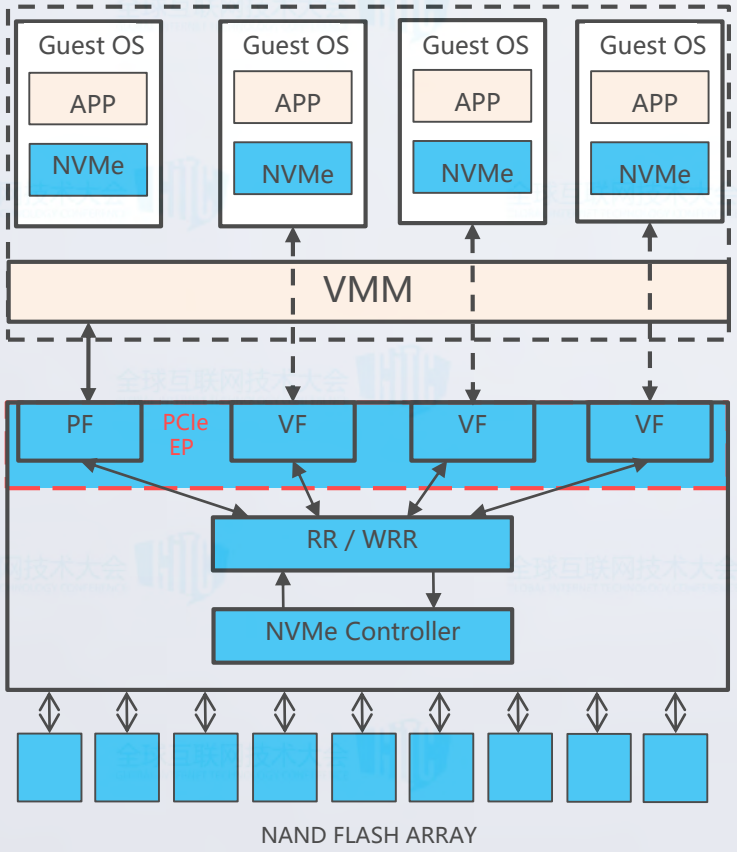
•容量越来越大
•性能（IOPS、QoS）逐步提升

应用调研显示：

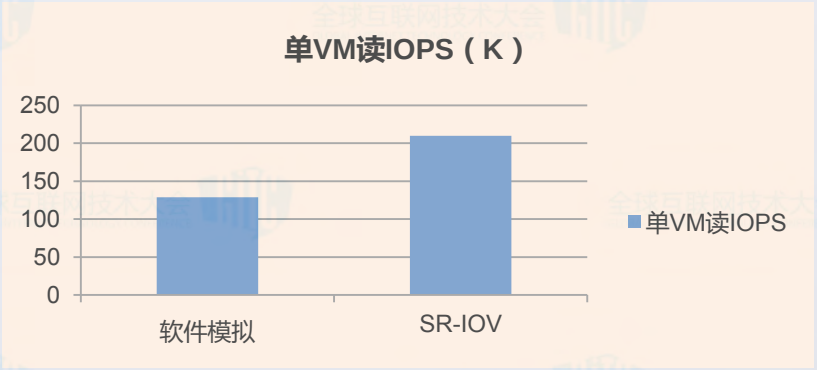
- 大部分App通常会给SSD下发<16QD的负载，远未发挥出SSD的并发能力
- 中小客户在单盘部署的数据量大部分<2.4T

- NVMe针对存储系统整合、多路径等特性提出NameSpace
- NameSpace可以配置不同容量、不同的数据格式、不同的访问模式，更进一步，可以跟App适配，配置不同的读写性能
- ES3000 V3产品是业界第一个支持多个NameSpace能力的NVMe SSD产品

SR-IOV，优化SSD在虚拟化中的应用



- 允许多个虚拟机高效共享PCIe设备
- 与CPU的VT-x、VT-d等虚拟化技术配合，虚拟机可以获得能够与物理机性能媲美的I/O 性能
- ES3000 V3产品支持1个PF和15个VF
- SR-IOV特性可以对单VM内的IOPS提升约**40%**的性能：



全NVMe闪存服务器，大幅提升业务性能

全NVMe服务器

华为RH2288H V3

12 / 24 盘NVMe SSD

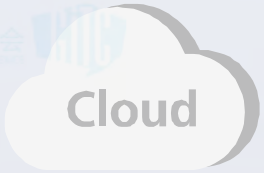
业务性能加速

OLTP 数据库

每秒处理事务TPS **10倍**

分布式存储

数据读写IOPS性能 **5倍**



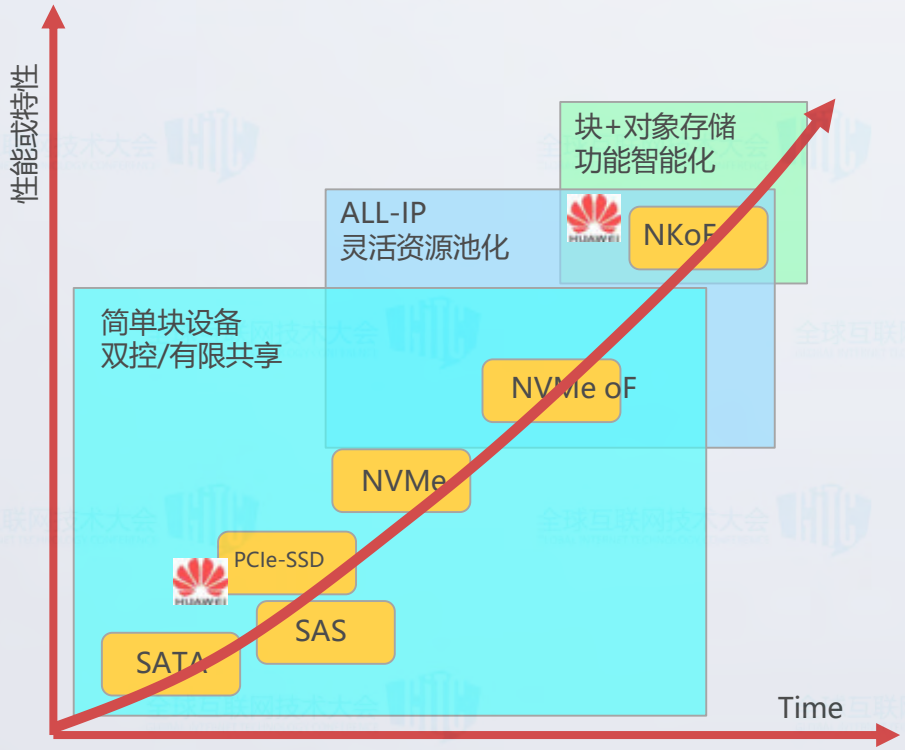
Cloud / VDI

虚拟机部署密度 **2倍**

数据搜索 / 大数据分析

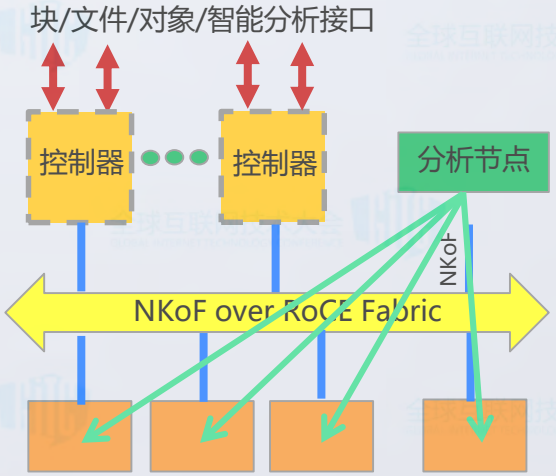
每秒数据查询性能 **6倍**

NVMe 未来展望（接口IP化，功能智能化）



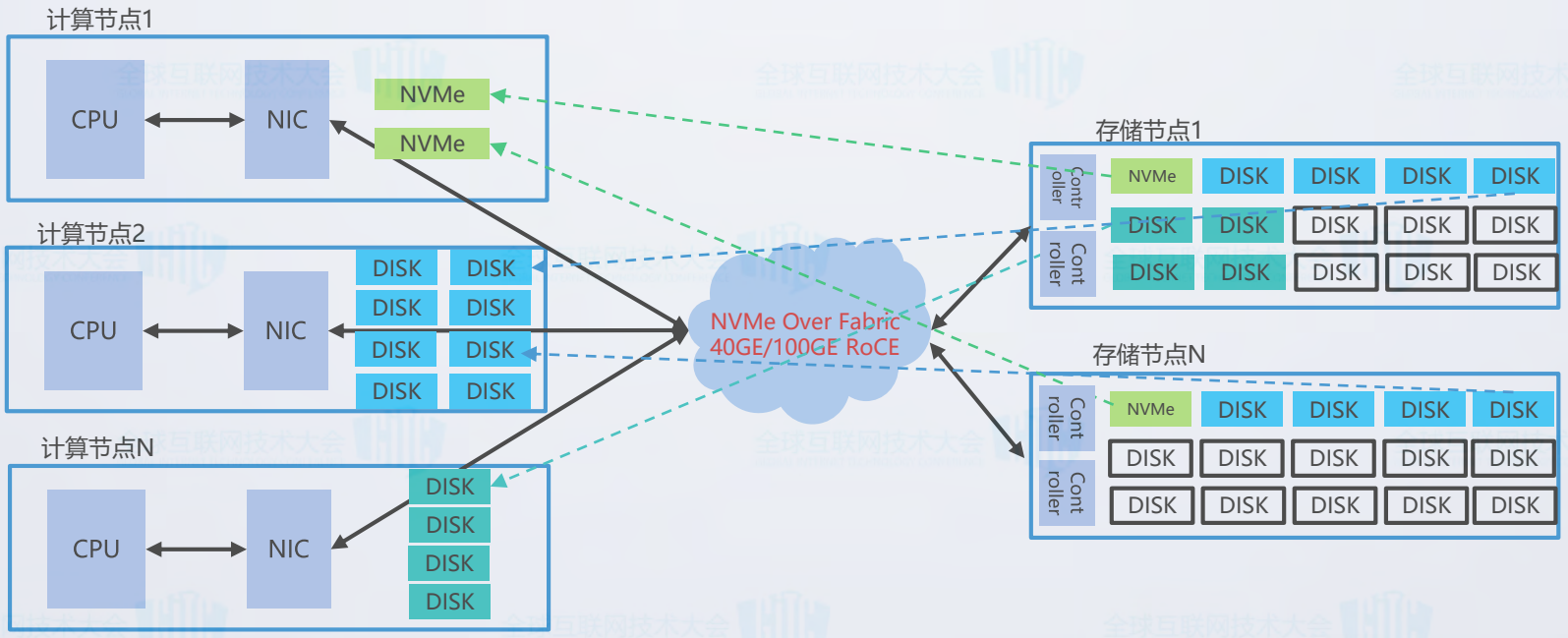
*NKoF: NVMe & Key-Value Store over Fabric with Analytics extension

- NVMe SSD共享资源池和可编排服务器
- 全共享后端高效存储系统
- 性能与容量解耦、按需扩展存储系统
- 计算靠近数据，分布式并行搜索和分析，大幅度减少数据搬迁



IP化&智能化基本存储单元
搜索/分析功能下移, 并行操作

NVMe Over Fabric存储资源池的逻辑架构图

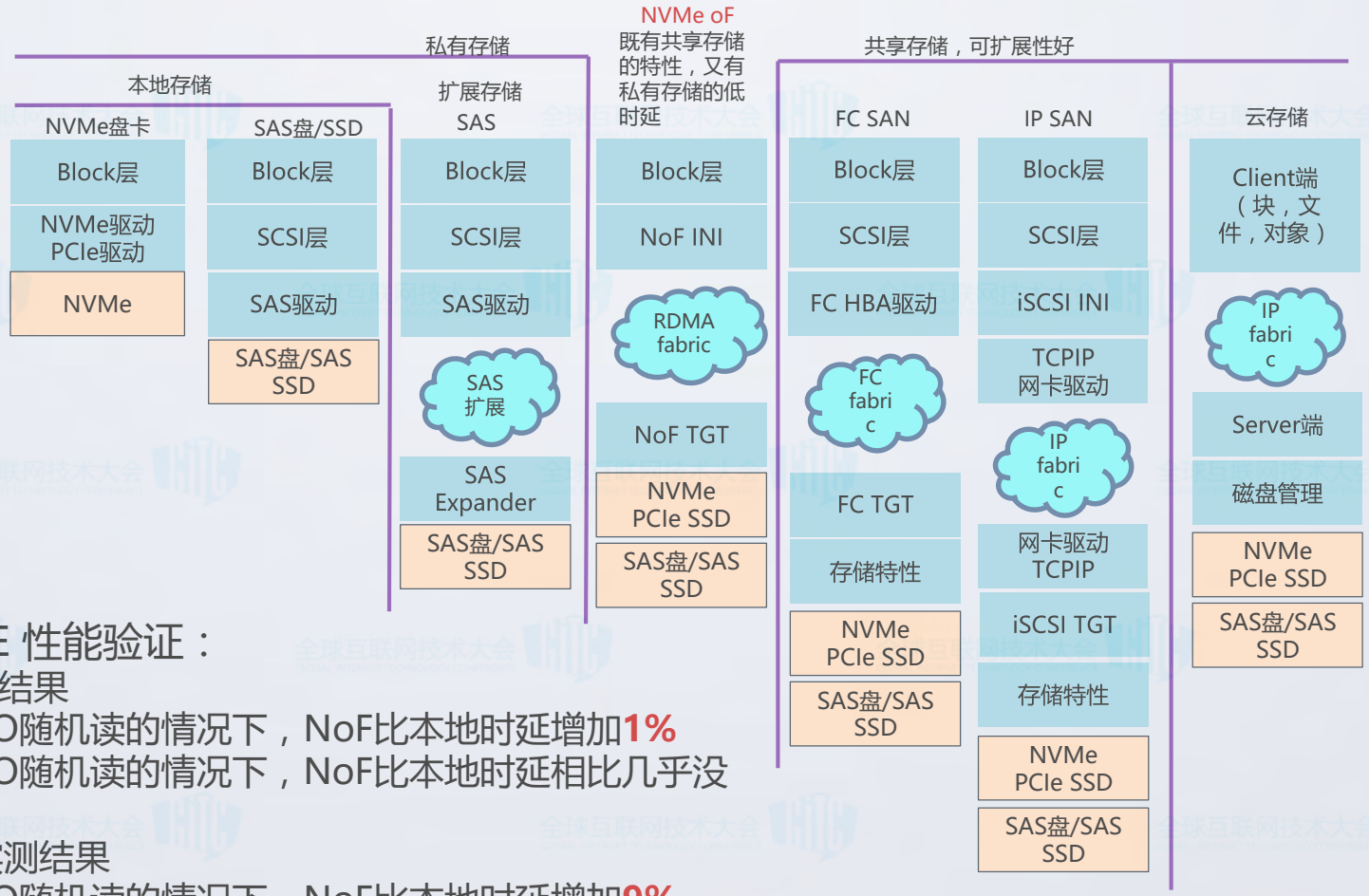


- 可以像本地存储一样使用
- 通过NVMe提供低时延
- Fabric带来的时延相比磁盘来讲可以忽略
- 可以实现灵活的计算、存储资源配比

云数据中心痛点：
业务的需求复杂性和多变性使固定配比的计算和存储的套餐产生资源浪费，或性能不足

直接部署NVMe oF场景：
- 提高业务性能的Cache或buffer
- 快速分析类业务的主存访问

各种块设备存储的协议栈对比



40GE RoCE 性能验证：

- SAS盘实测结果
  - 4k 单IO随机读的情况下，NoF比本地时延增加**1%**
  - 4k 多IO随机读的情况下，NoF比本地时延相比几乎没有差别
- NVMe盘实测结果
  - 4k 单IO随机读的情况下，NoF比本地时延增加**9%**，绝对时延增加**10us**左右
  - 4k 多IO随机读的情况下，NoF比本地时延相比几乎没有差别

Thank You