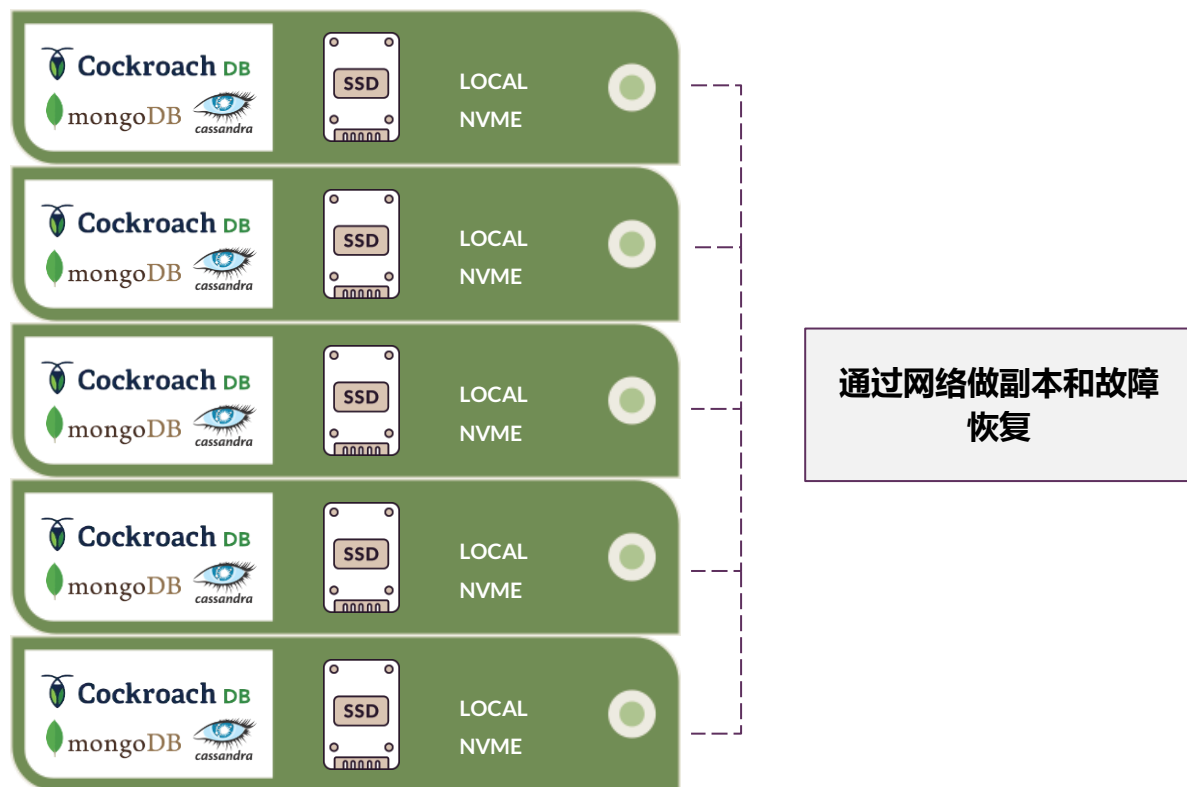




基于NVMe/TCP的超高性能存储解决方案

陈定宝 Lightbits 解决方案架构师

云原生应用: 新常态



NoSQL, 内存运行, 分布式

这些应用对存储的需求:

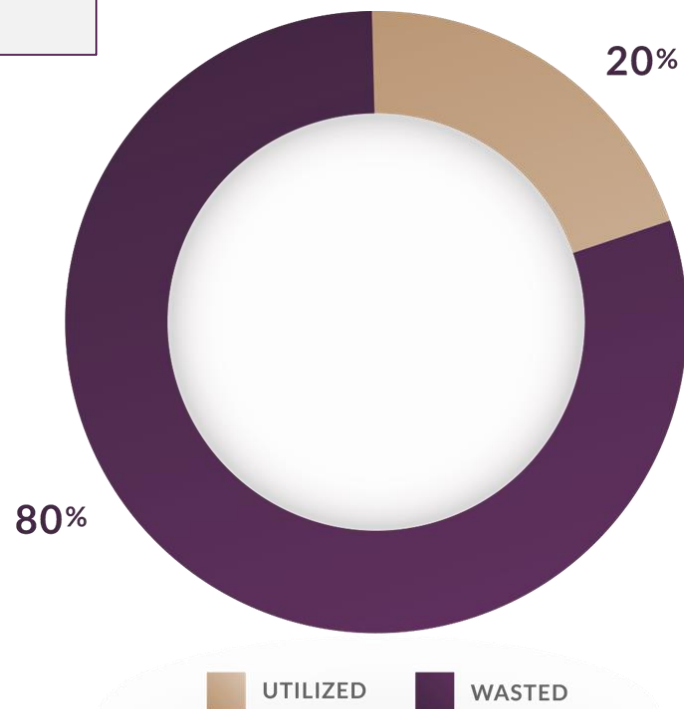
- 低延迟, 高带宽
- 稳定一致的操作响应时间
- 应用程序做数据保护
- 使用本地NVMe盘

面临的问题:

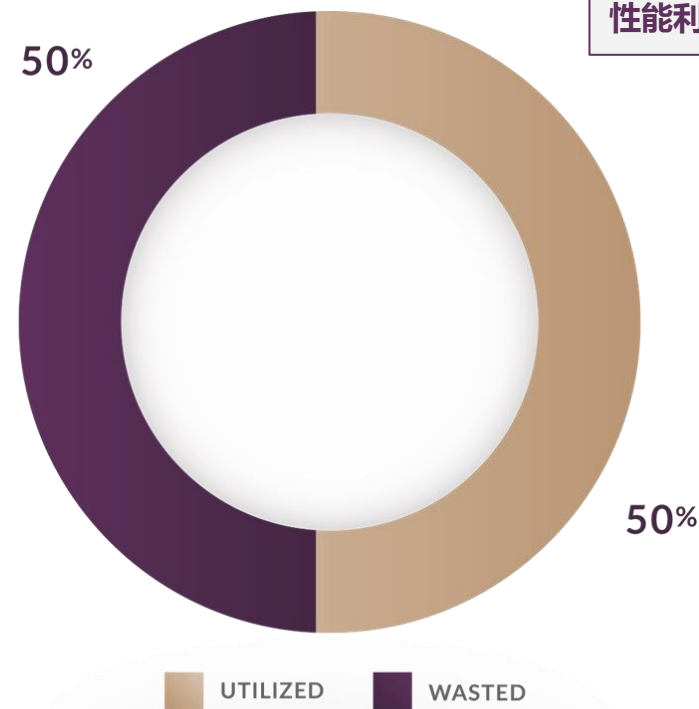
- 本地盘有效利用率低
- 故障恢复:
 - 需要很长时间, 服务能力下降
 - 严重影响网络性能
- 应用程序只能运行在某些存有数据的物理服务器上

Flash闪存利用率非常低

容量利用率 15-25%

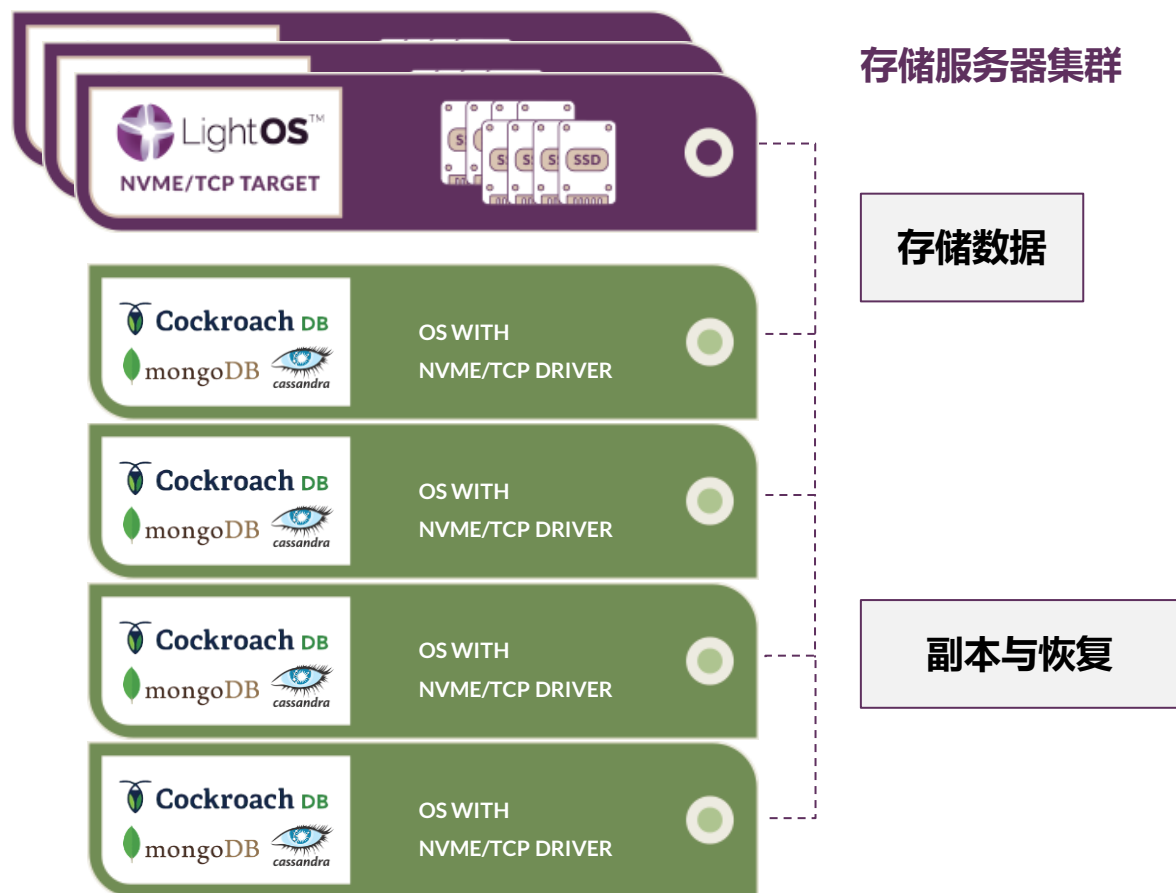


性能利用率 50%



50-85% 的Flash被浪费!

云原生应用+ Lightbits的 NVMe/TCP块存储



更高的服务灵活性，Flash使用率
， 更快的故障恢复 = 低TCO

优势:

- 类似于本地盘性能
- 任意大小的块存储，并实施精简配置
- SSD故障不会影响应用层
- 服务器故障:
 - 应用程序可以在任意一台服务器运行
 - 故障恢复时间以秒来计算
 - 更高的应用可用性，更高的服务连续性



Who and How?



Lightbits Labs: When, Where and How

- 成立于 2016年早期
- Offices: 硅谷, 纽约, 以色列, 阿姆斯特丹, 上海
- NVMe/TCP协议的先驱, 并且标准化了协议
- 2019年3月推出世界上第一个NVMe/TCP 产品
- 强大的投资人背景



超过10年的技术创新

Lightbits 团队在NVMe领域的工作和贡献

NVMe

Direct-attached
高性能的 PCIe SSDs

2009 - 2013

- **第一代NVMe SSD 控制器**
- Adopted by top Hyperscalers and all-flash-arrays
- First Linux & VMware drivers

NVMe-oF

Rack-scale
通过RDMA 或FC远程访问
NVMe SSDs

2014 - 2016

- **定义NVMeoF**
- First NVMe-rack-scale storage solution

NVMe/TCP

Cloud & Hyper-scale
通过TCP/IP数据中心级别访问NVMe

2017 - 2020

- **NVMe/TCP协议的领导者**
- Contributed initiator code to upstream kernel
- First NVMe/TCP product
- First Clustered/Failover NVMe SDS

Lightbits 核心价值

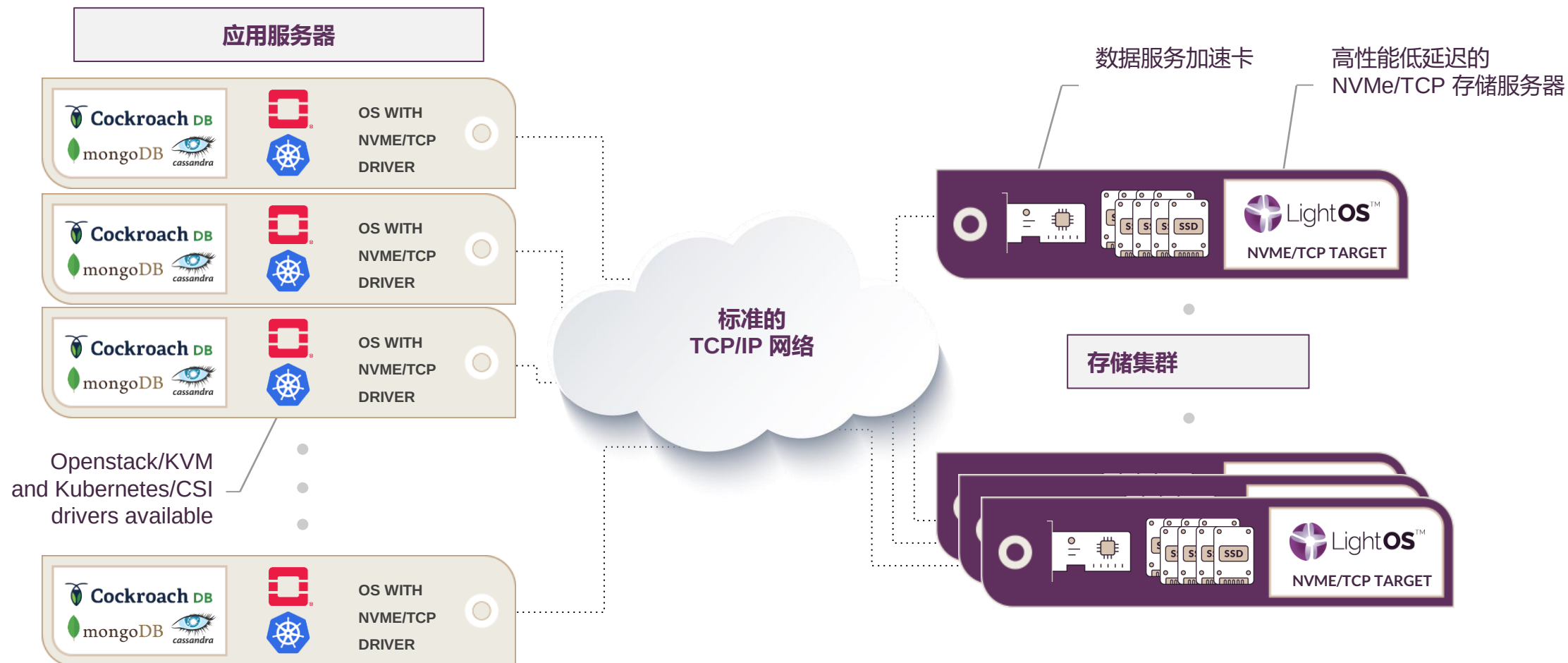


LightOS: Hyperscale Storage For All

- ➔ ■ **灵活性和运营效率**
 - 软件定义块存储，标准的硬件基础设施
 - 独立动态扩展存储和计算需求
 - 丰富的数据服务
- **高性能**
 - 低延迟，高IOPS，高带宽
- **降低成本**
 - 最大化使用率和投资回报比
 - 提高闪存的持久性

高性能高可靠的NVMe/TCP的块存储

看起来像SAN的逻辑卷，用起来像本地NVMe盘



与Intel战略合作提供优化的存储平台

Intel® Ethernet 800 系列网卡带Application Device Queues (ADQ) 功能

- 优化NVMe/TCP的延迟

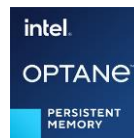


数据服务加速卡

- 压缩
- 纠删码保护
- 闪存优化

Intel® QLC 3D NAND SSDs

- 延长QLC寿命, 降低TCO
- 优化 \$/GB的同时保持高性能和低延迟



Intel® Optane™ Persistent Memory

- 提供高速非易失写缓存及metadata
- 没有电池, 电容等要求
- 更大的memory, 更低的TCO



Intel® Xeon® Scalable Processors

- VMD: 企业级SSD 热插拔和LED 管理
- 高性能
- 为存储软件优化



LightOS: Hyperscale Storage For All



■ 灵活性和运营效率

- 软件定义块存储，标准的硬件基础设施
- 独立动态扩展存储和计算需求
- 丰富的数据服务



■ 高性能

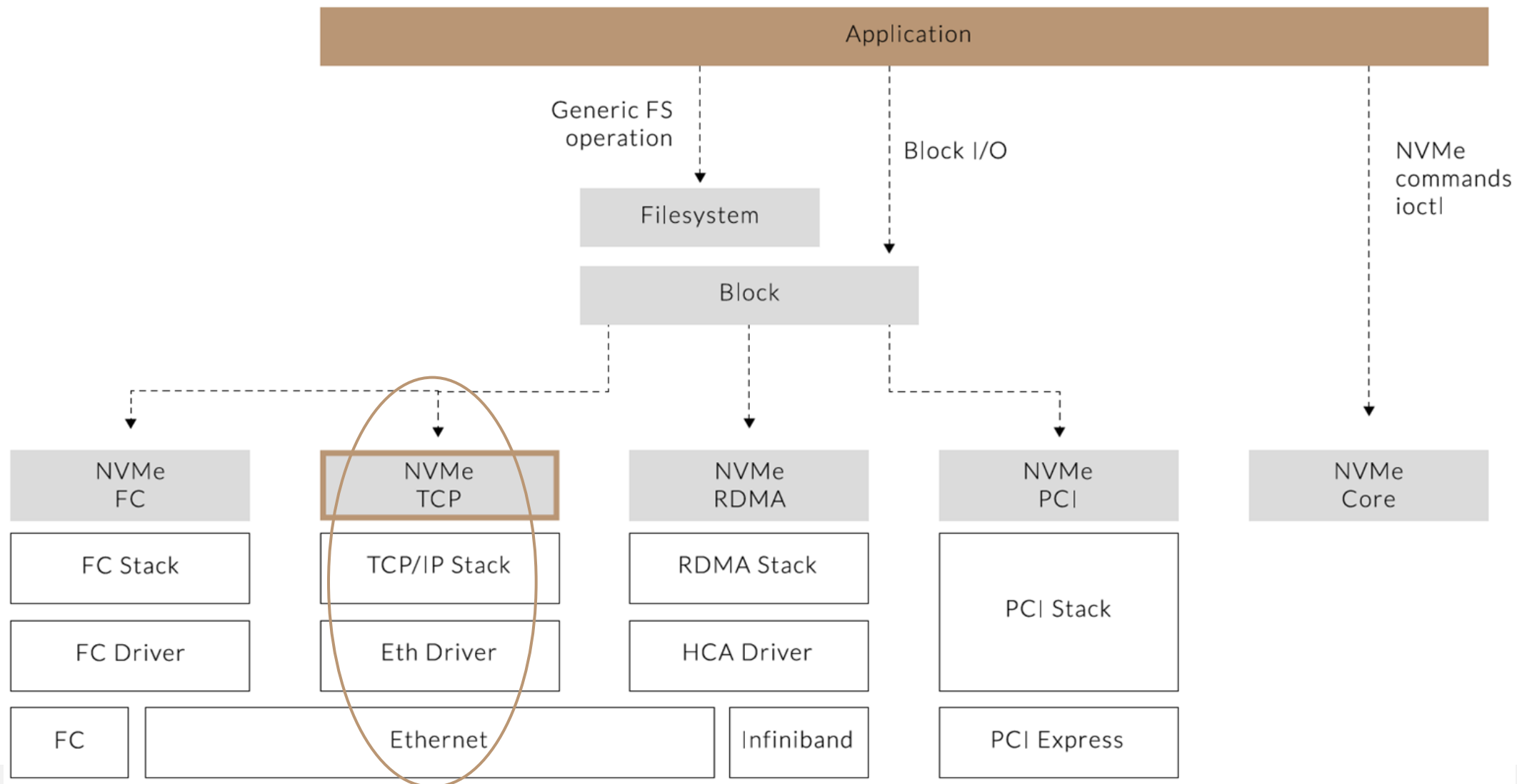
- 低延迟，高IOPS，高带宽

■ 降低成本

- 最大化使用率和投资回报比
- 提高闪存的持久性

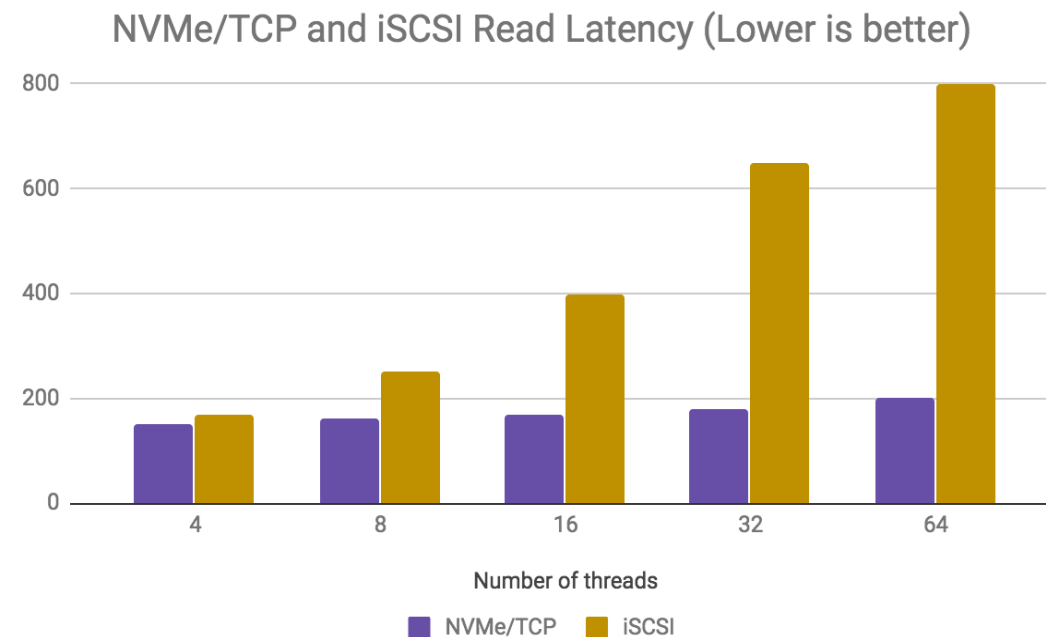
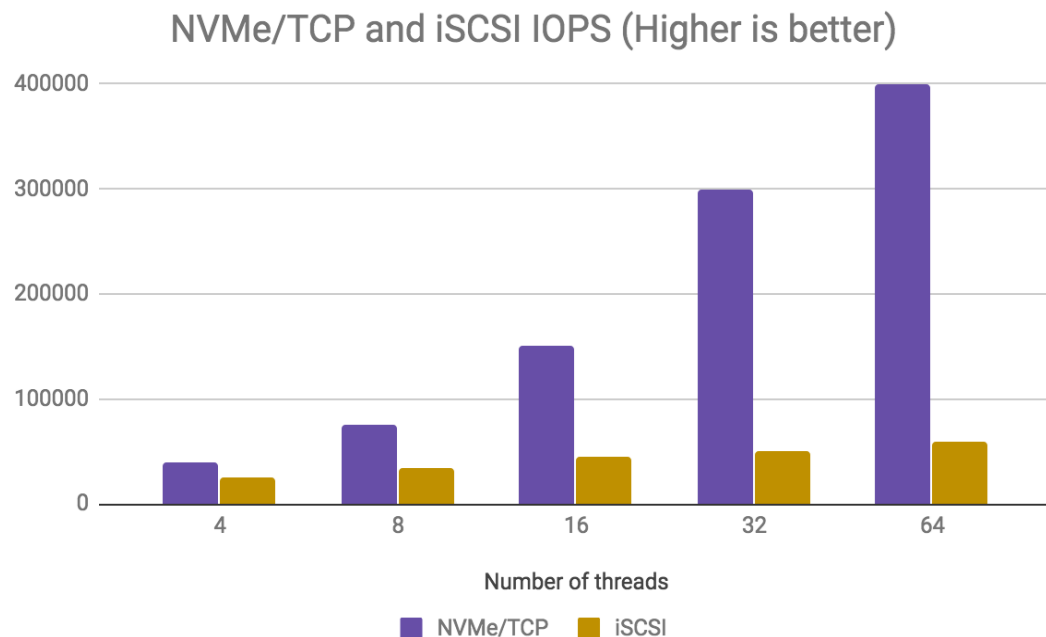
解耦合方案：标准的NVMe/TCP

用最新标准的 NVMe over TCP/IP



iSCSI vs. NVMe/TCP

同样的硬件，大不一样的结果



NVMe/TCP 的IOPS随线程数线性增加(在同样线程下超过SCSI 6倍) 同时保持低延迟, 最多比iSCSI低4倍 - 基于同样的硬件环境

更好的性能, 更低的成本

	 ceph	 LightOS™	
	优化的 Ceph 集群 (5 OSDs)	LightOS 集群(2 服务器)	LightOS 优势
4K 随机读	2,270,000	6,000,000	~3X 倍 IOPs
4K 70/30 随机读写 IOPs	691,100	4,000,000	~6X 倍 IOPs
4K 随机写 IOPs	463,800	1,600,000	~3X 倍 IOPs
4K 随机读延迟	3000μs (3ms)	170μs	17X 低延迟
4K 70/30随机读写延迟	6000μs (6ms)	270μs	22X低延迟
4K随机写延迟	11,000μs (11ms)	450μs	24X低延迟
方案估算成本	\$185,320 (HW only)	\$83,272 (HW+SW)	45% 成本

在相同可用容量情况下 (~70TB): 基于LightOS的解决方案提供远超Ceph的性能,
并且成本只有Ceph的约1/2

LightOS: Hyperscale Storage For All



■ 灵活性和运营效率

- 软件定义块存储，标准的硬件基础设施
- 独立动态扩展存储和计算需求
- 丰富的数据服务



■ 高性能

- 低延迟，高IOPS，高带宽



■ 降低成本

- 最大化使用率和投资回报比
- 提高闪存的持久性



LightOS 使能 QLC 闪存:

- 随机=>顺序
- 100Gb线速的压缩
- 纠删码(Erasure Code)保护
- 最多提高5倍的持久性

降低TCO

QLC = MORE DENSITY PER NAND CELL

BETTER \$/PER GB



FEWER WRITES PER CELL

	Intel P4510 (8T TLC)	Intel P4320 (8T QLC)
随机写DWPD	0.9	0.2
顺序写DWPD	3.0	0.88

LightOS: Hyperscale Storage For All

- ✓ ■ **灵活性和运营效率**
 - 软件定义块存储，标准的硬件基础设施
 - 独立动态扩展存储和计算需求
 - 丰富的数据服务
- ✓ ■ **高性能**
 - 低延迟，高IOPS，高带宽
- ✓ ■ **降低成本**
 - 最大化使用率和投资回报比
 - 提高闪存的持久性

Lightbits 存储集群适用场景

- **数据库**
 - 单实例的数据库 (MySQL, Postgres, Maria DB)
 - 分布式/副本机制的数据库 (MongoDB, CockroachDB, Cassandra)
- **数据分析处理**
 - 传统行业, FinTech
 - 日志处理(ELK, Splunk)
 - Apache Kafka streams
 - 人工智能AI/ML
- **私有/边缘/混合云基础配置**
 - 虚拟机
 - 容器
- **公有云服务提供**
 - IaaS (块存储) -nvme as service

同一种技术, 两种提供方式

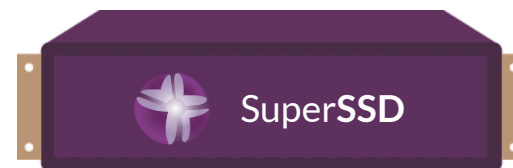
LightOS: 带硬件加速卡的软件定义的解耦合存储方案



加速数据服务

NVMe Over TCP/IP
软件定义全闪解决方案

SuperSSD: 完整的存储硬件及服务
Turnkey 方案



易于管理和部署的SDS一体机方案

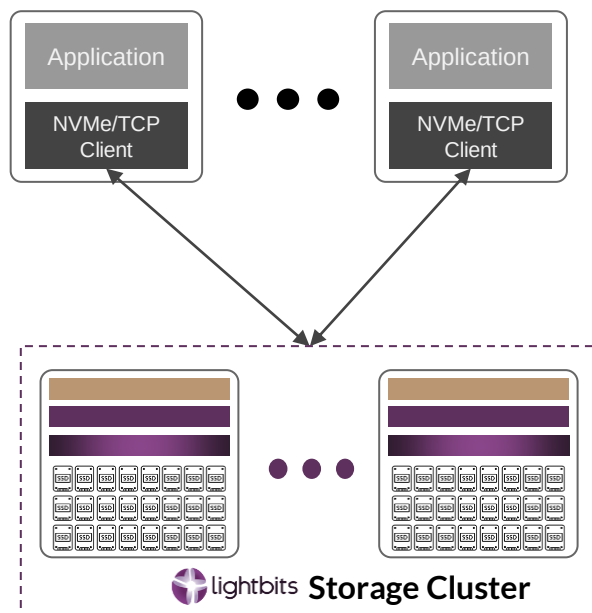
LightOS核心功能介绍



- **100Gbps 线速压缩**，不增加延迟
- Erasure coding 做数据保护
- 针对TLC, QLC的先进的Flash Management以提高使用寿命
- Thin provisioning精简配置
- 自动分配到不同故障域的卷 Replication
- 标准化的高可用（ANA），自动failover的集群方案



加速卡



高性能，低延迟

- 每个目标服务器的最大IOPS:
 - **4K随机读: 3M IOPS**
 - **4K随机写: 800K IOPS**
- 每个目标服务器的延迟（2副本时）：
 - 在210万IOPS时，4K随机读的平均延迟 $\leq$ **200 μ s**
 - 在60万IOPS时，4K随机写的平均延迟 $\leq$ **300 μ s**
- 每个目标服务器的最大带宽
 - 16GB/s的读带宽
 - 8GB/s的写带宽

Thank you!



敬请关注!

Next Steps

- 技术深入探讨
- 在线演示
- Proof of Value (PoV) 测试

