SACC

2020 中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2020

架构融合 云化共建

live 2020年10月22日 - 24日网络直播

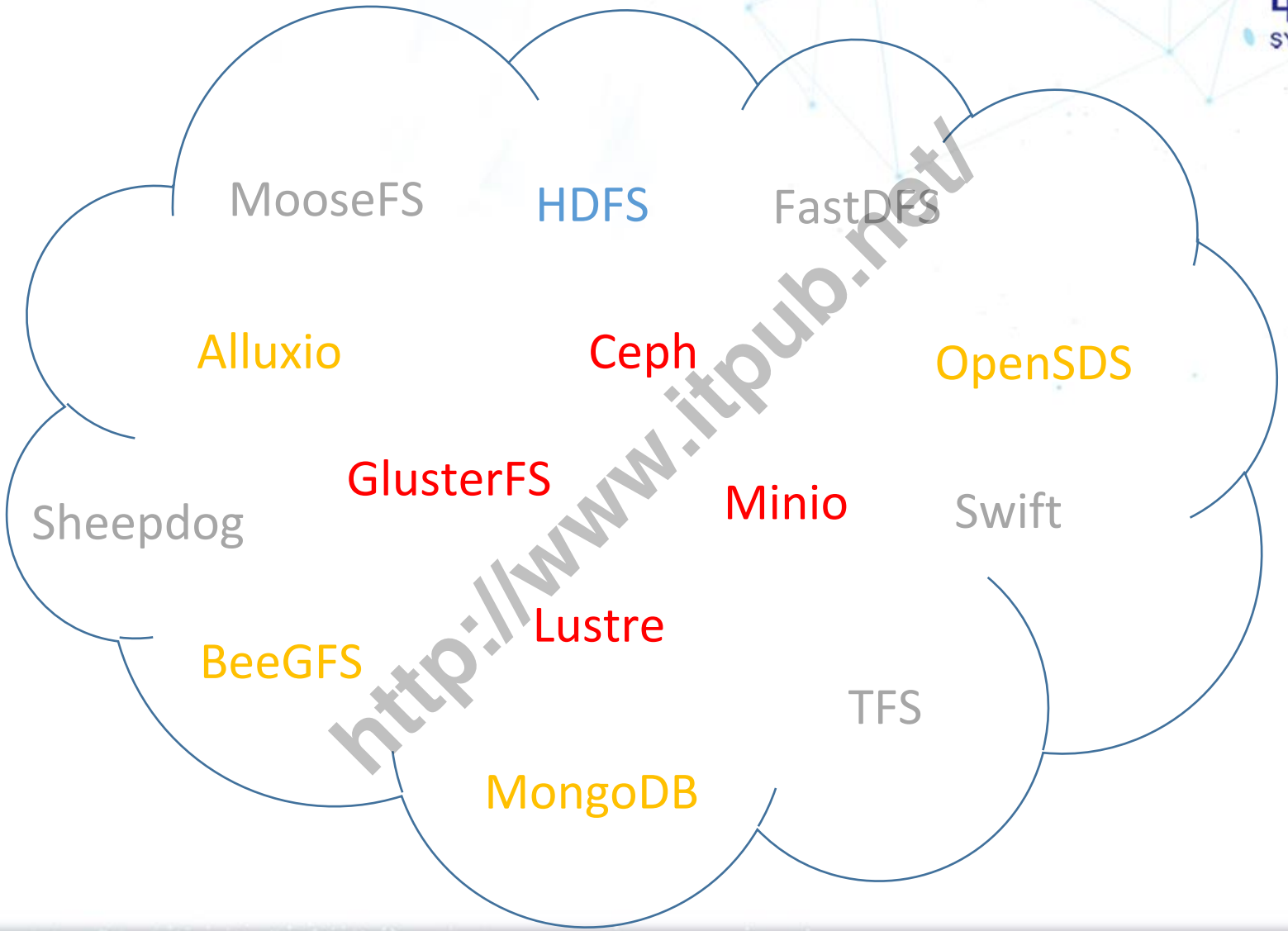
云上会议

架构融合
云化共建

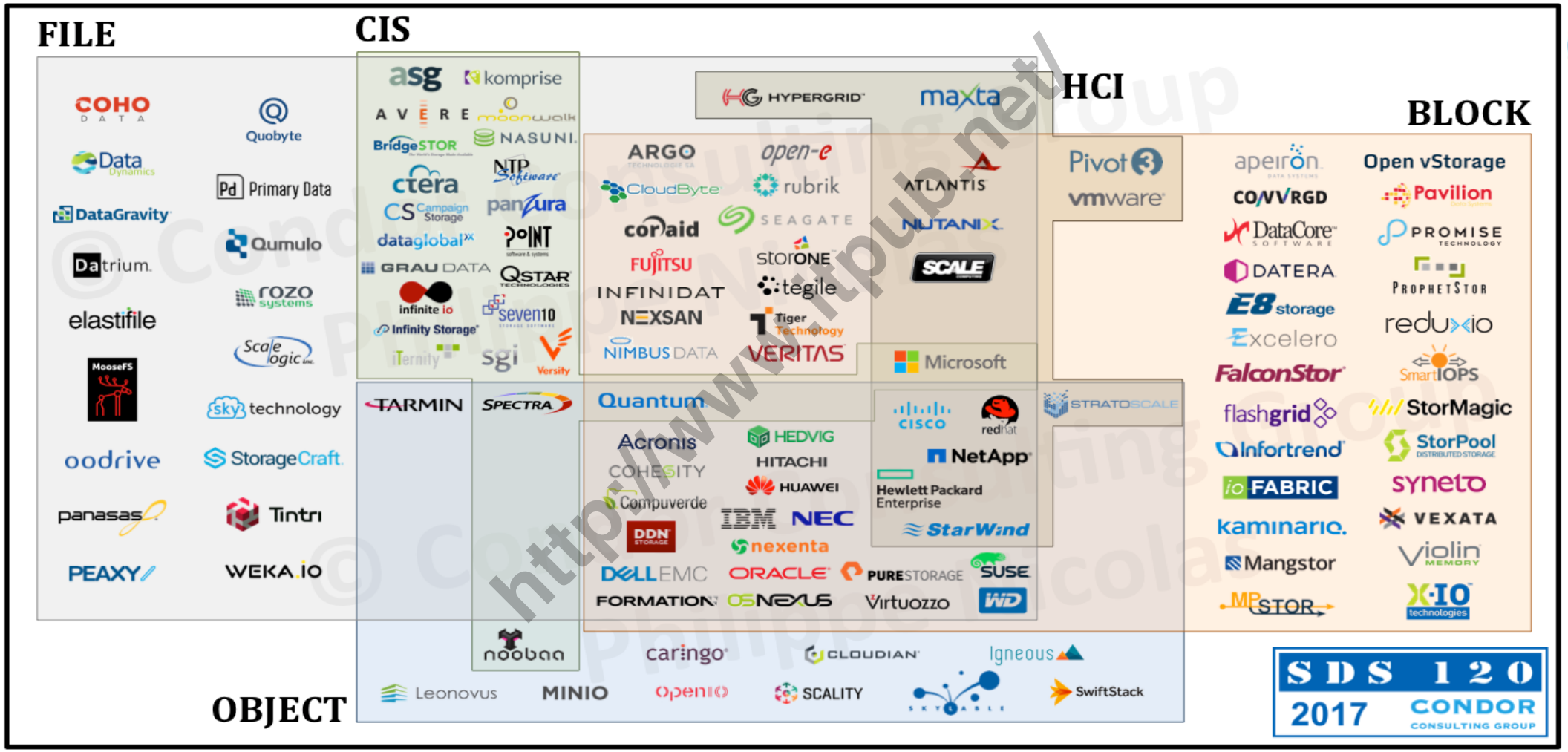
新一代全闪SDS存储系统技术架构

刘爱贵@TaoCloud

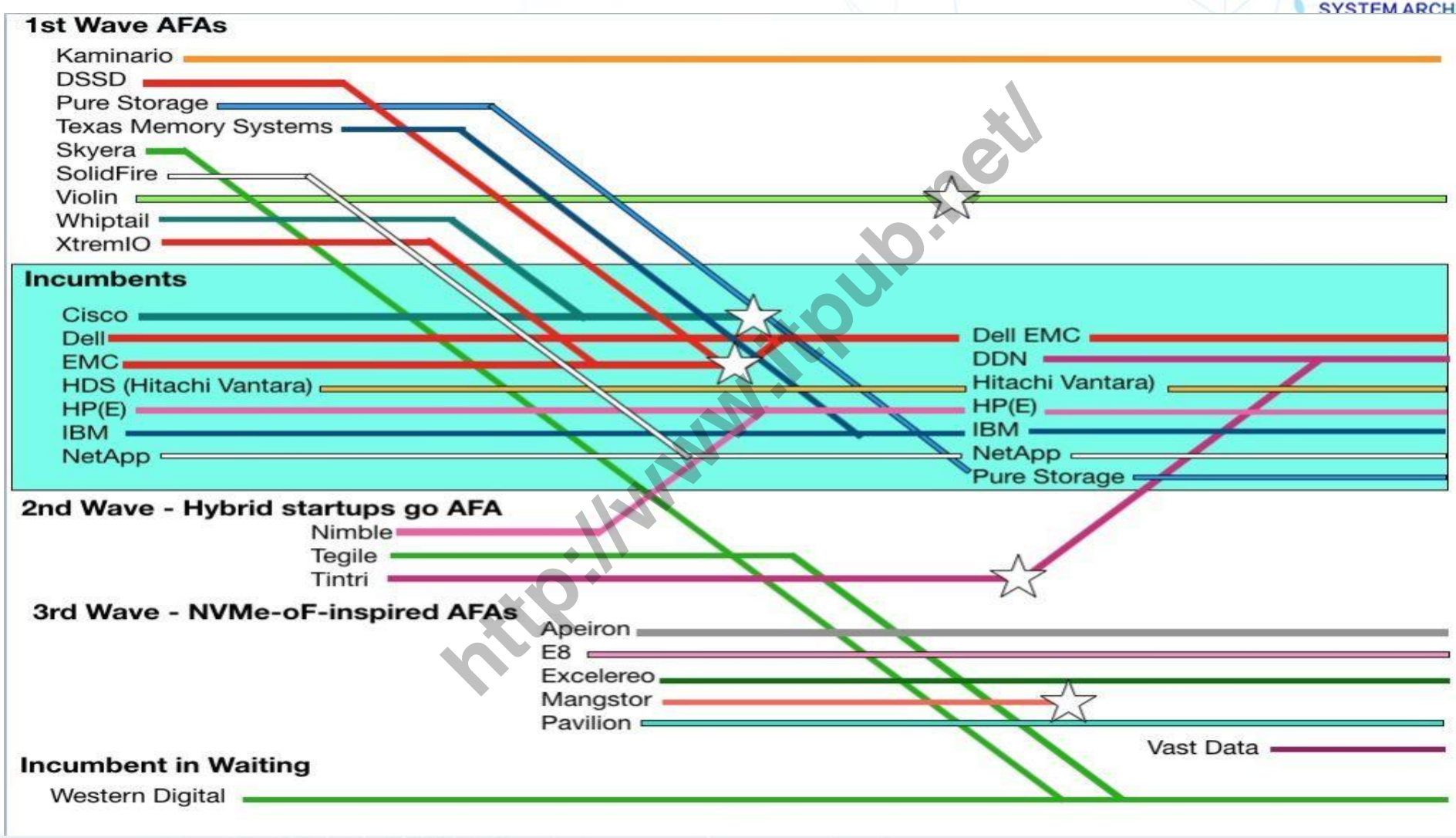
开源SDS系统



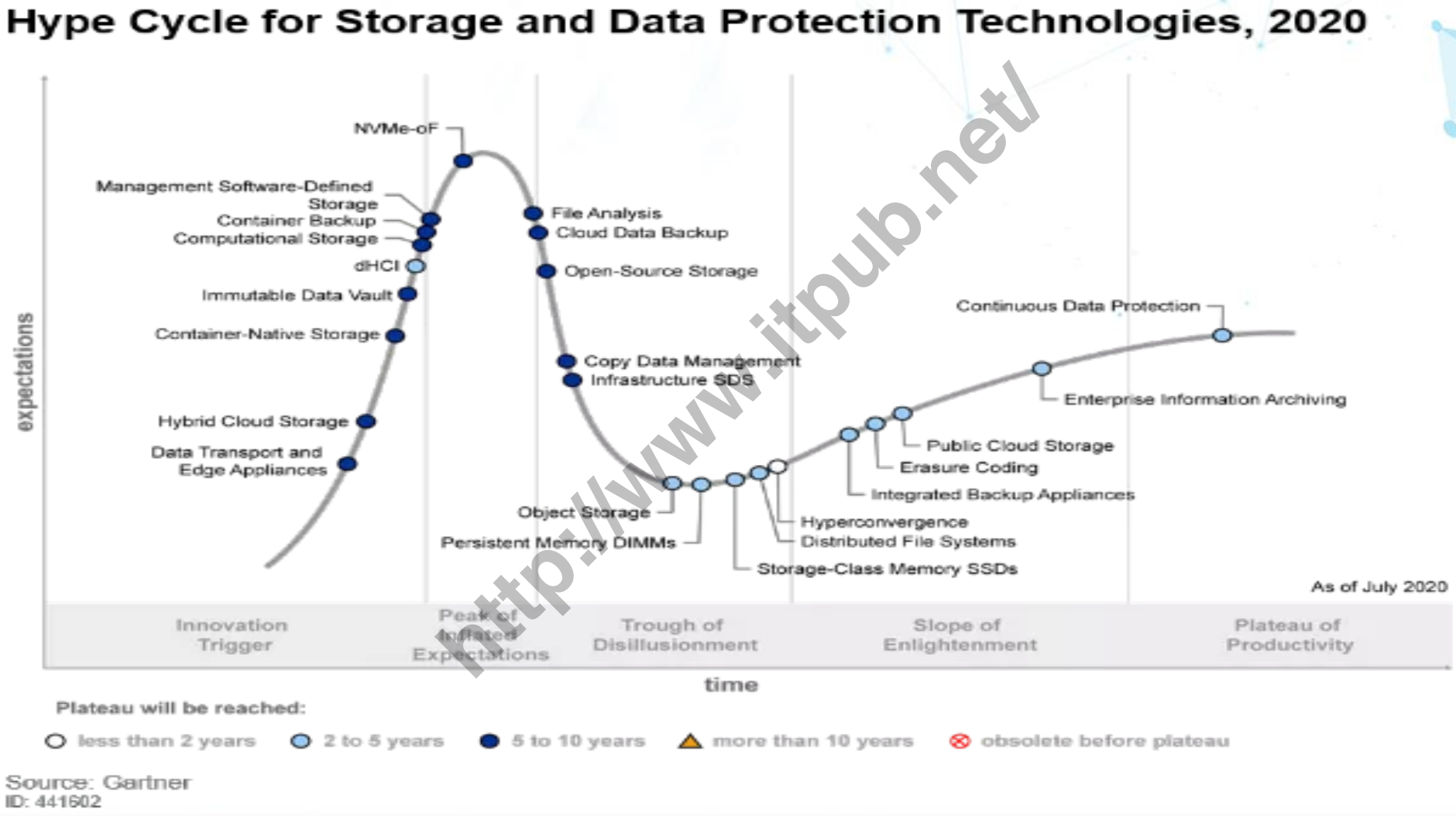
SDS大格局



AFA全闪阵列浪潮



Gartner：2020存储技术成熟度曲线



Source: Gartner
ID: 441602

为什么是全闪SDS?

SSD成为存储介质主流

技术成熟，成本不断下降

SDS已经是存储市场主流

X86架构，横向扩展

全闪SDS是低延迟高性能存储

云计算、金融科技、新兴市场

自主可控存储替代

硬件、OS、存储软件的全国产化

SDS面临的挑战

低延迟：ms级延迟，时延型应用无法适用

高性能：IOPS/带宽低，性能型应用无法满足

存储效率：裸金属性能和容量50-85%被浪费

系统鲁棒：故障修复或系统扩容影响正常业务

当我们说全闪SDS，我们在说什么

软件定义
X86架构

NVMe
端到端

500万
高IOPS

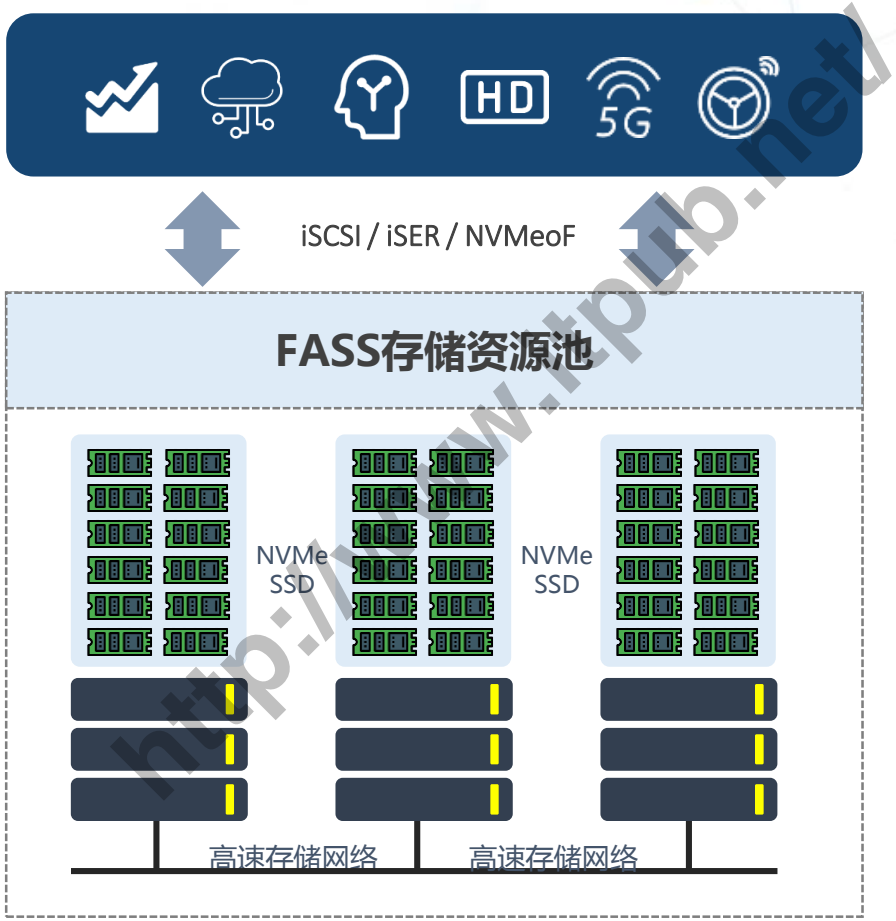
横向扩展
分布式，非AFA

RDMA
IB/ROCE网络

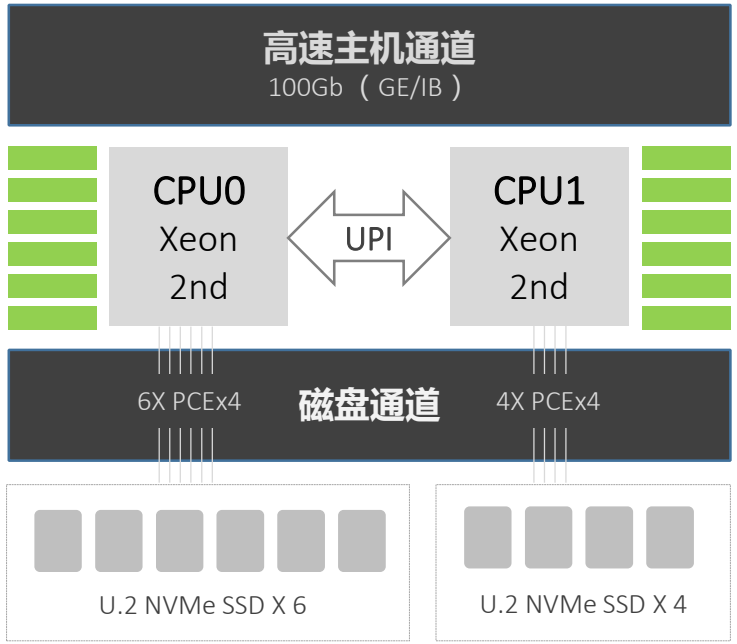
200us
低延迟

全闪SDS技术架构原则

一个实例：FASS全闪块存储



全NVMe高速硬件架构



全闪存硬件架构

- 高速硬盘通道
- 全NVMe硬盘
- 100Gb主机通道

48线程
处理能力

1.5TB
高速缓存

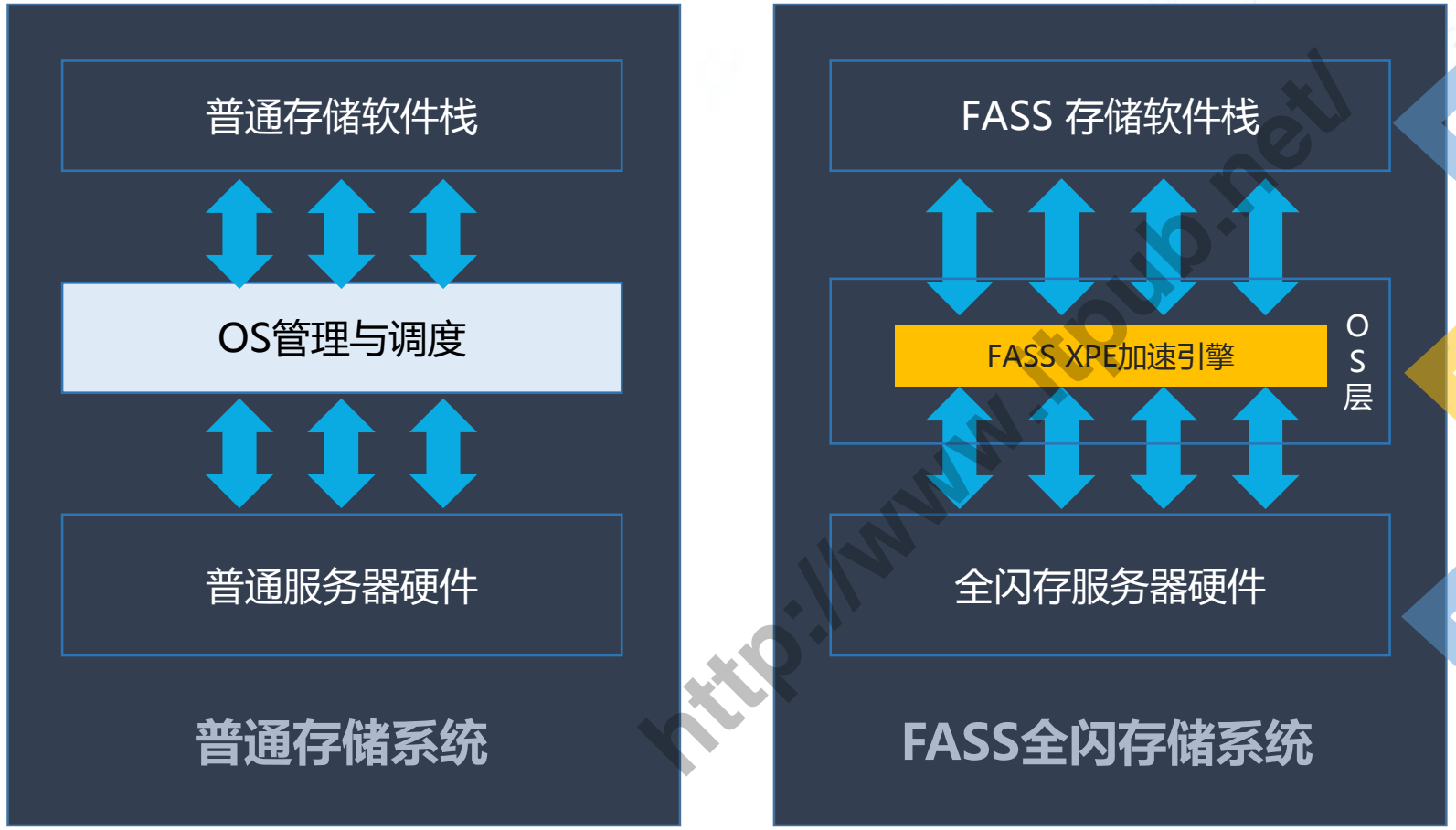
80GB/s
背板带宽

200 Gb/S
主机带宽

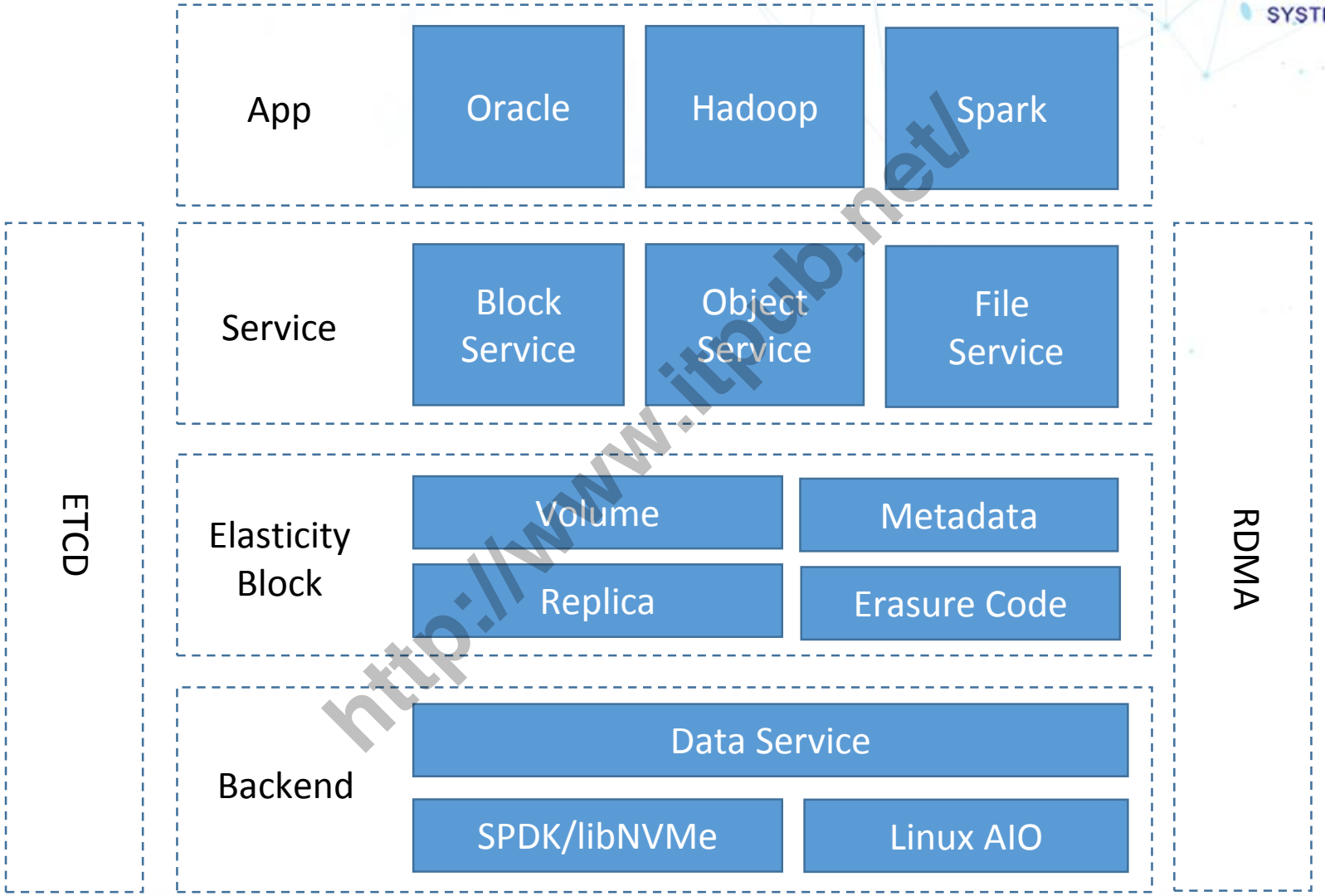
OS成了性能杀手

- CPU调度（core scheduling）
- 内存分配（Memory allocation）
- 锁竞争（Lock contention）
- 数据拷贝（Data Copies）
- 上下文切换（Context Switches）

FASS 高性能设计

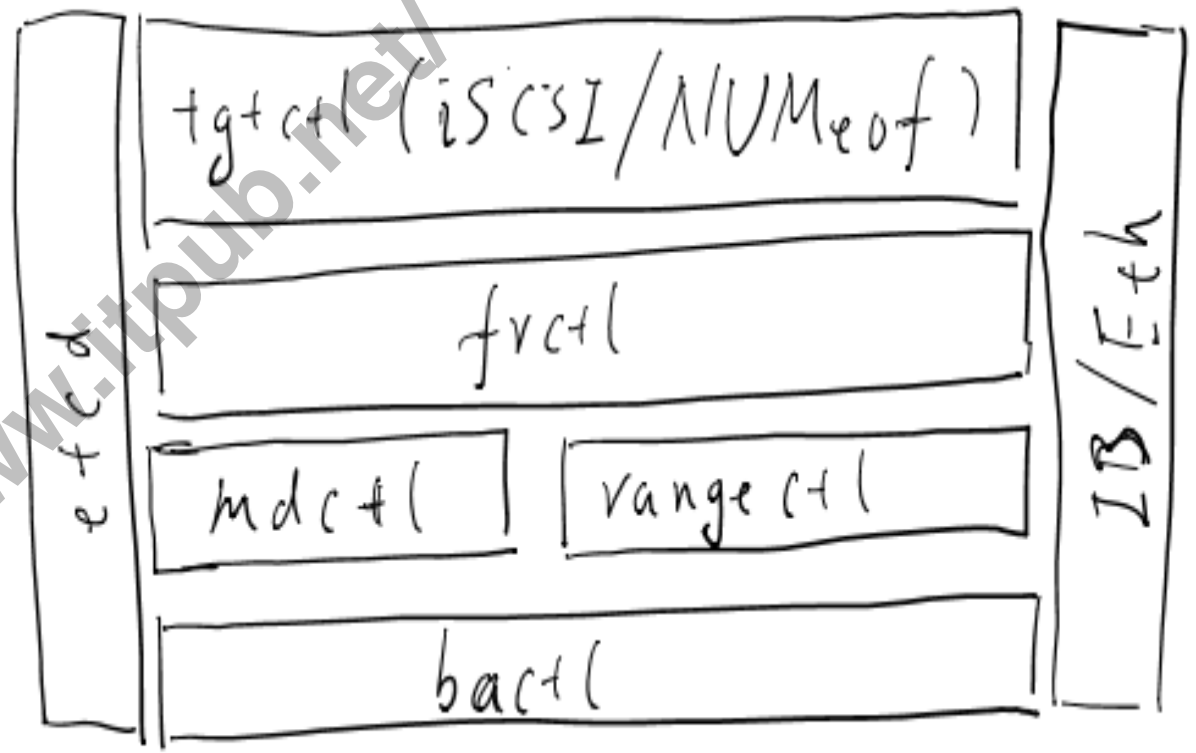


全闪SDS统一架构

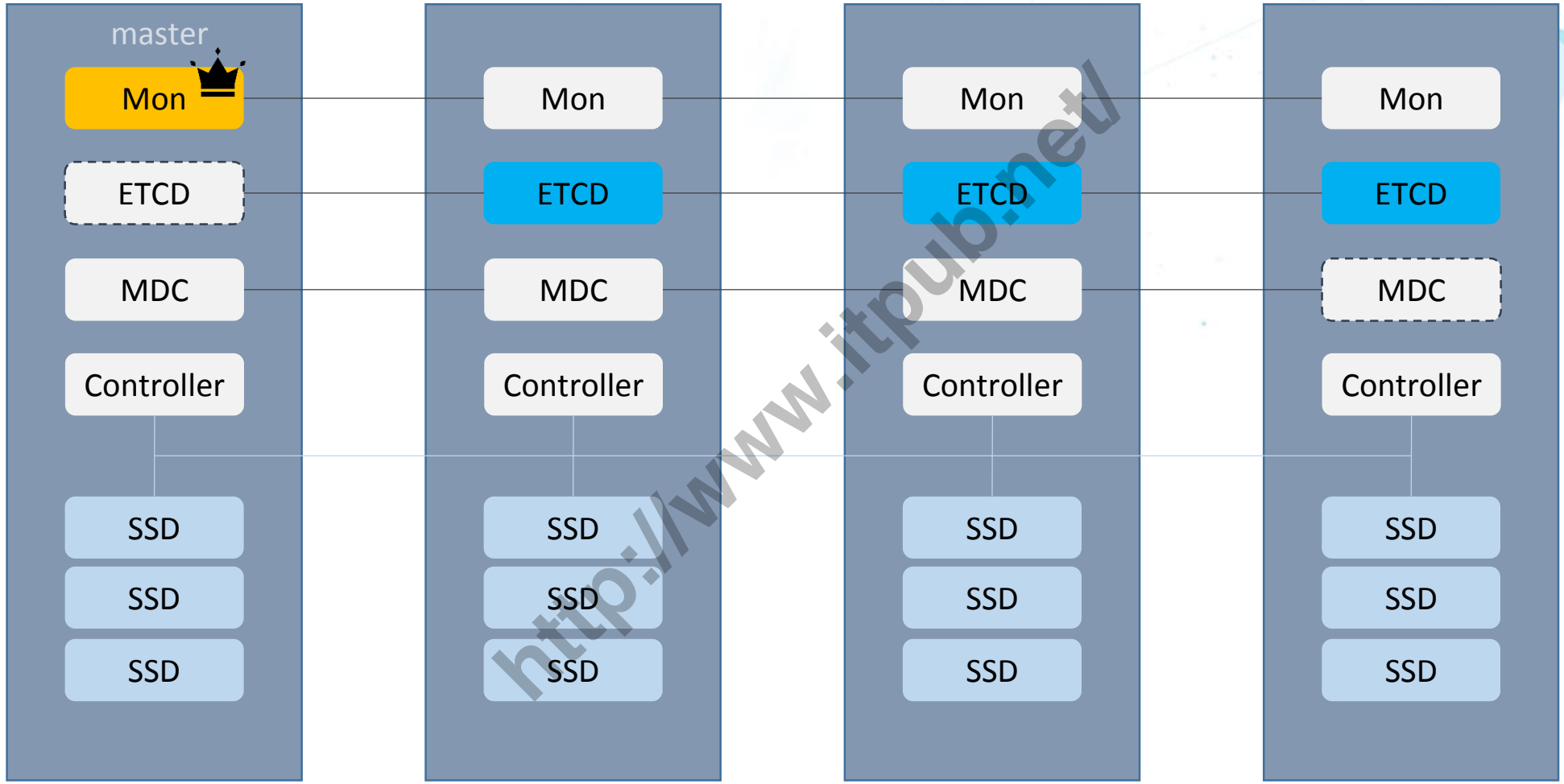


关键核心技术

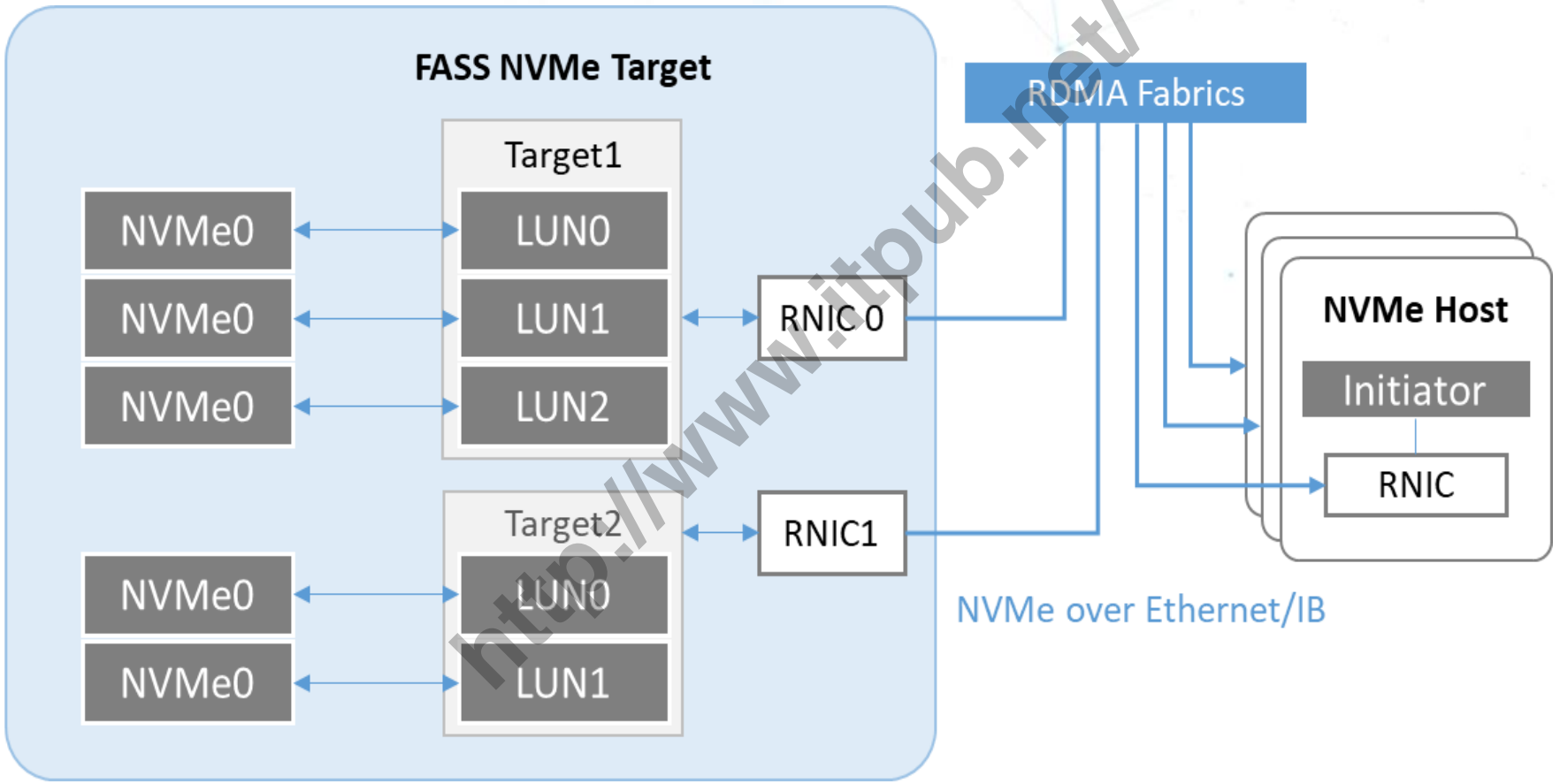
- 元数据（非完全DHT）
- 全用户态（User Space）
- Core绑定（PMD：polling mode driven）
- Hugepage-based内存管理
- 裸盘管理（libnvme/libaio+KV）
- 异步通信（RDMA/TCP）
- 访问协议（NVMf/iSER/iSCSI）
- 编程模型：协程



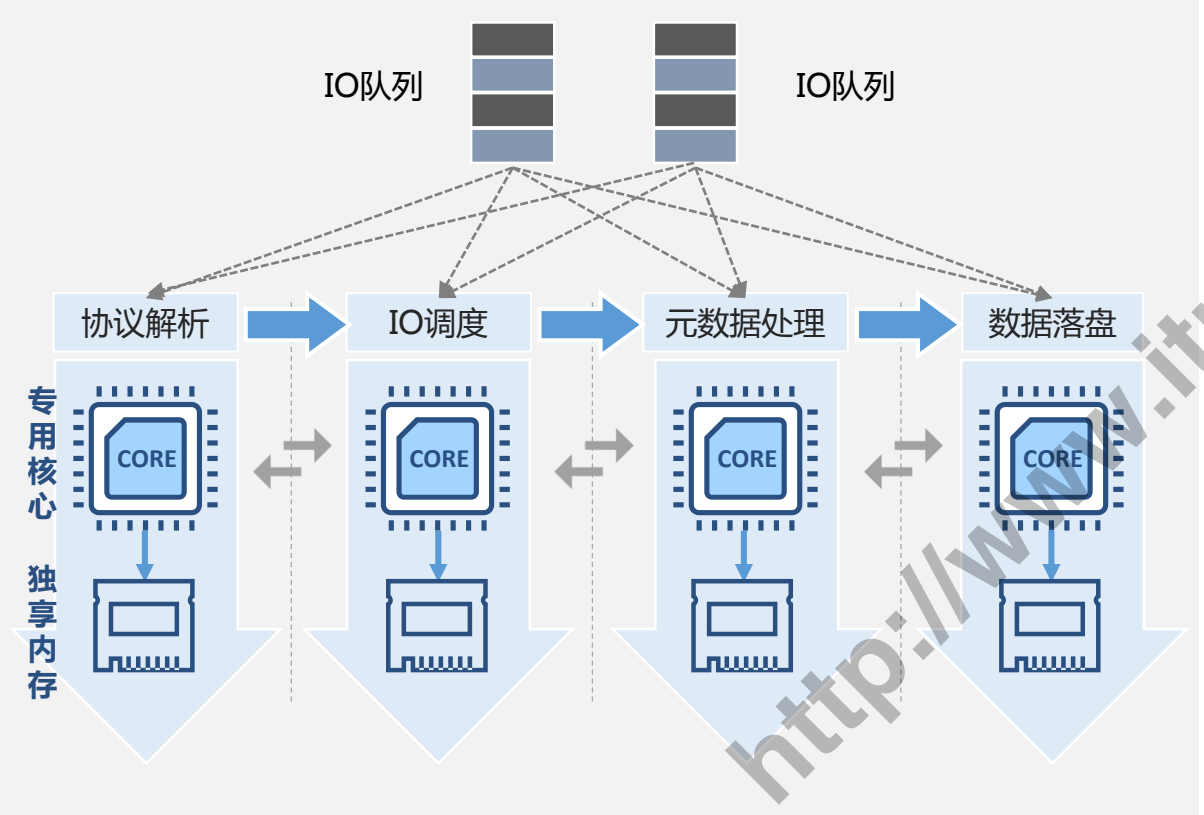
分布式集群设计



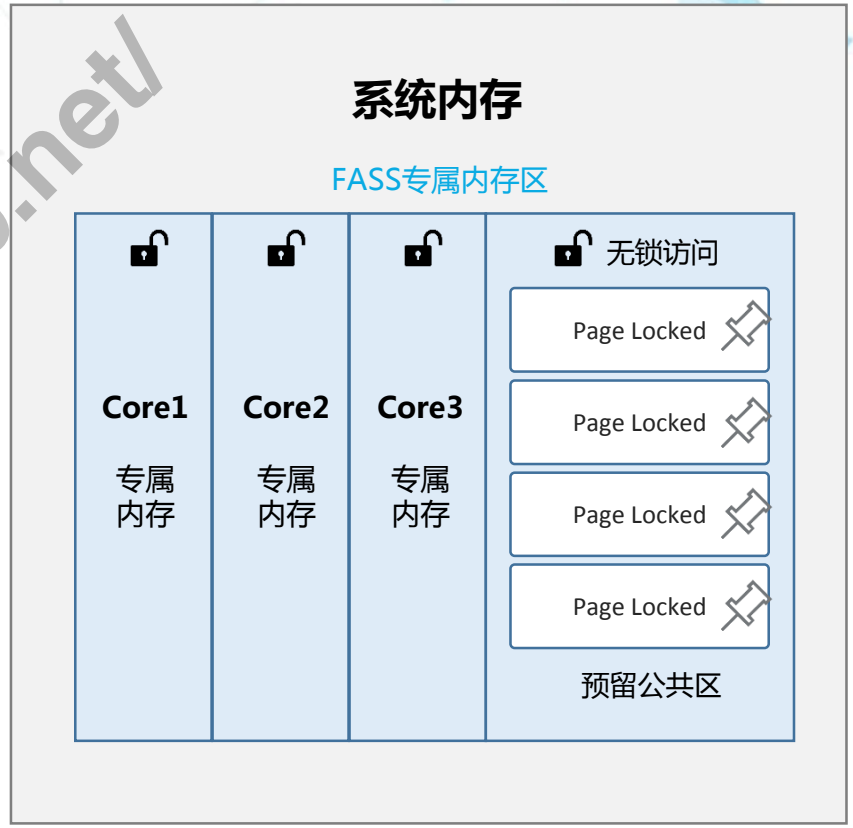
端到端NVMe



XPE加速引擎

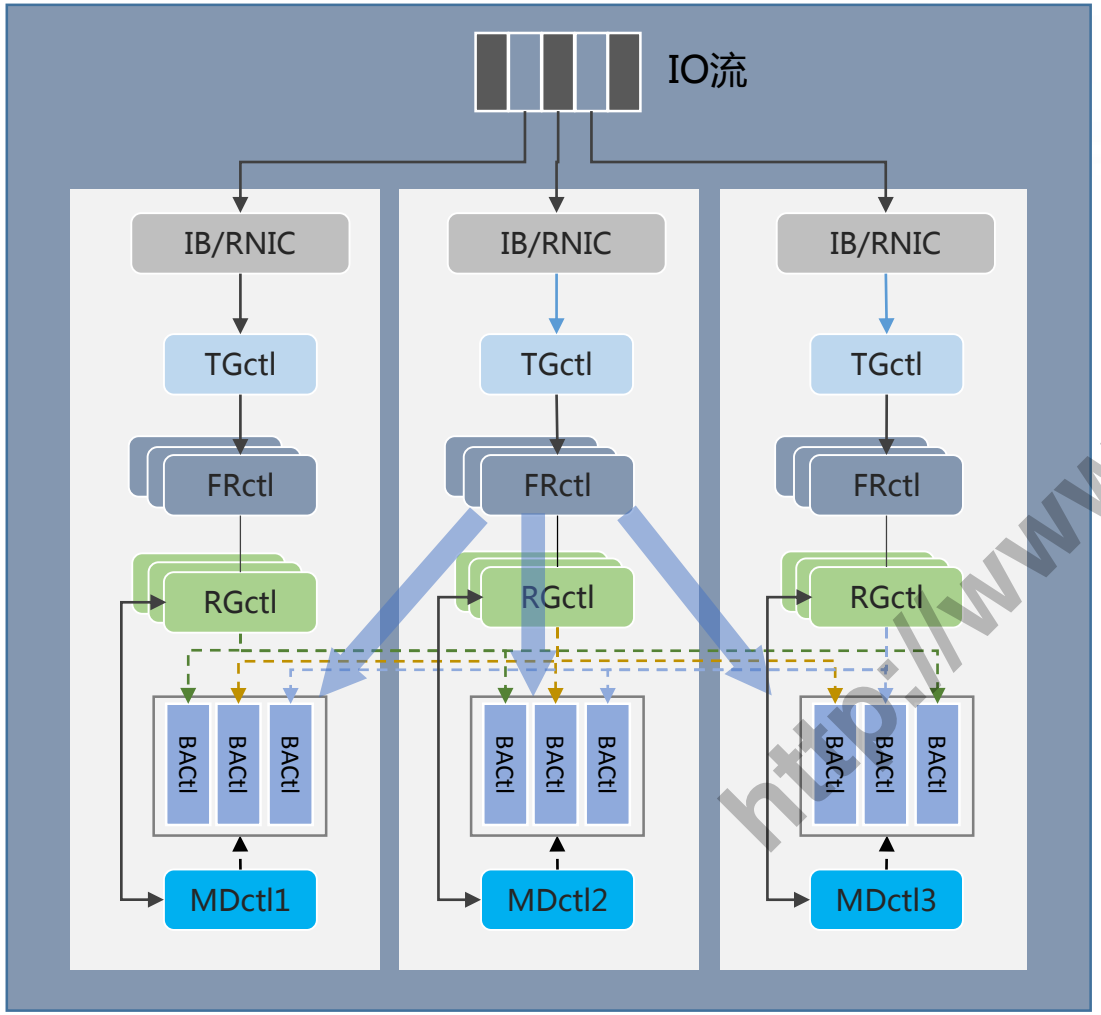


FASS CPU调试器



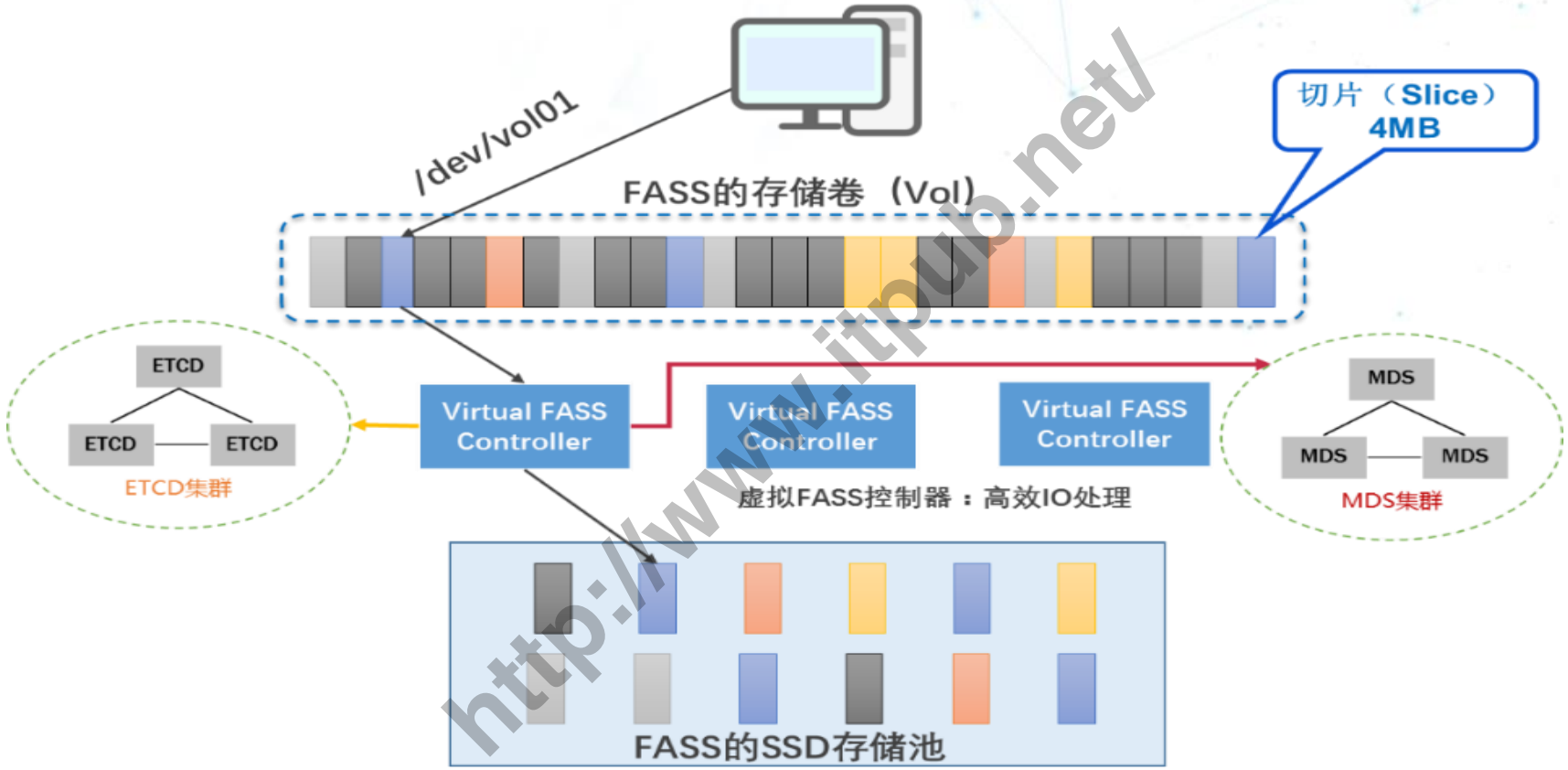
FASS 虚拟内存管理

分布式微控制器并行作业

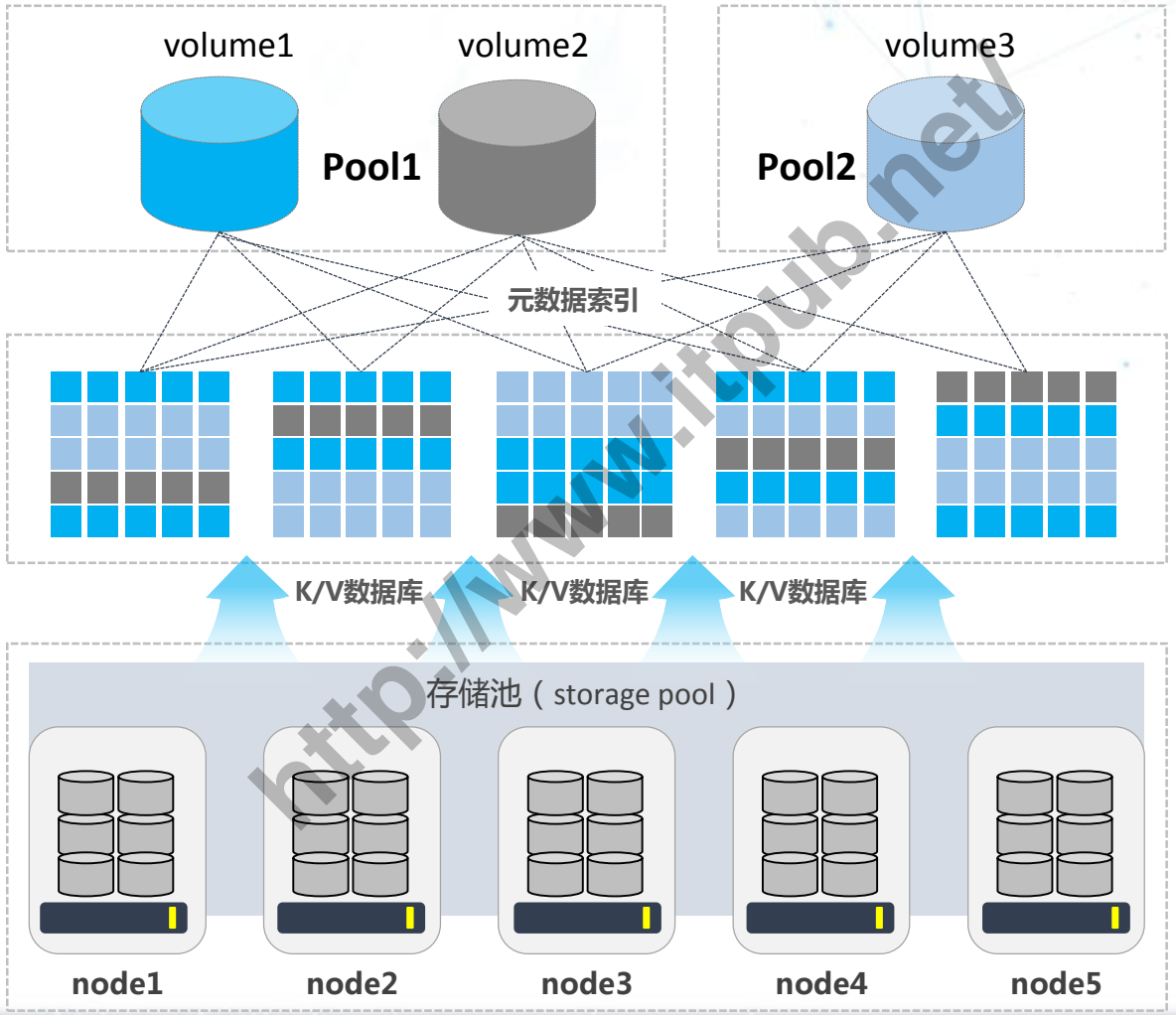


- 多种微控制器角色，高并发，流水线作业
- Target控制器把多个IO请求分发到多个FRCtl
- FRCtl下发请求到子卷控制器，获取子卷信息
- FRCtl将数据写入多个后端控制器（数据节点）
- 每控制器分配专属CPU核心

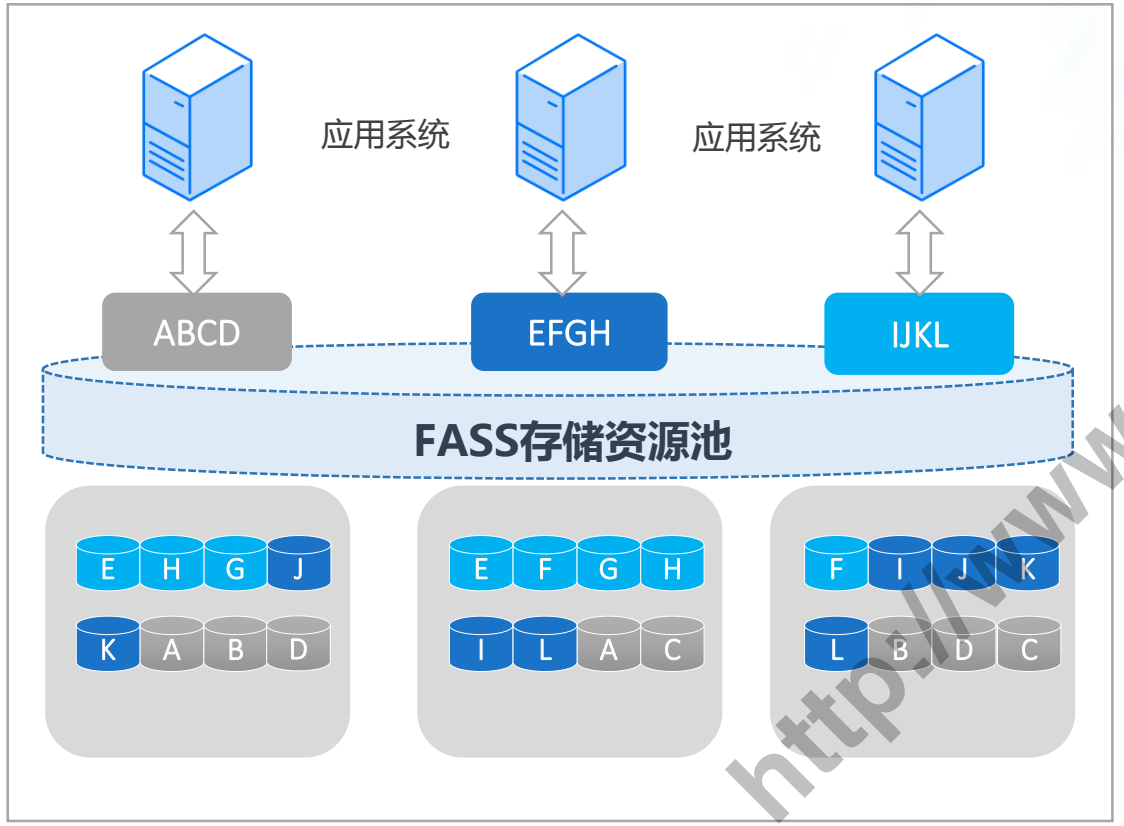
数据布局

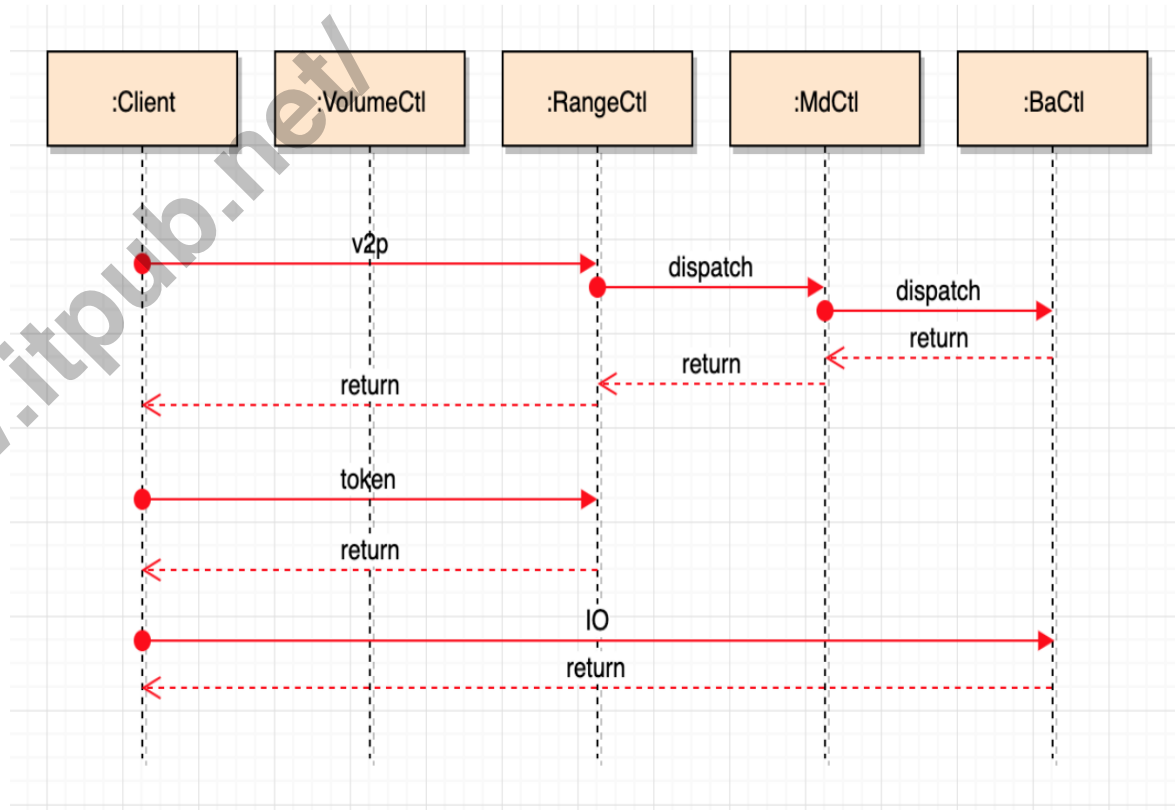


元数据管理

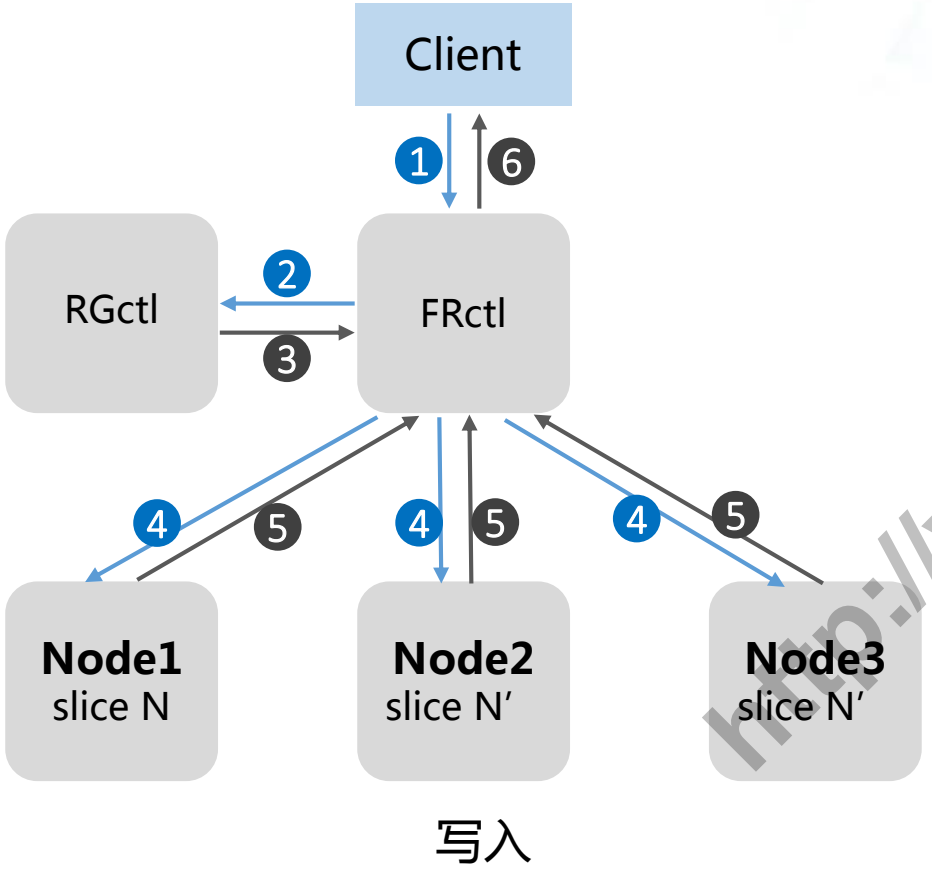


副本数据一致性（Raft like）





数据写入流程

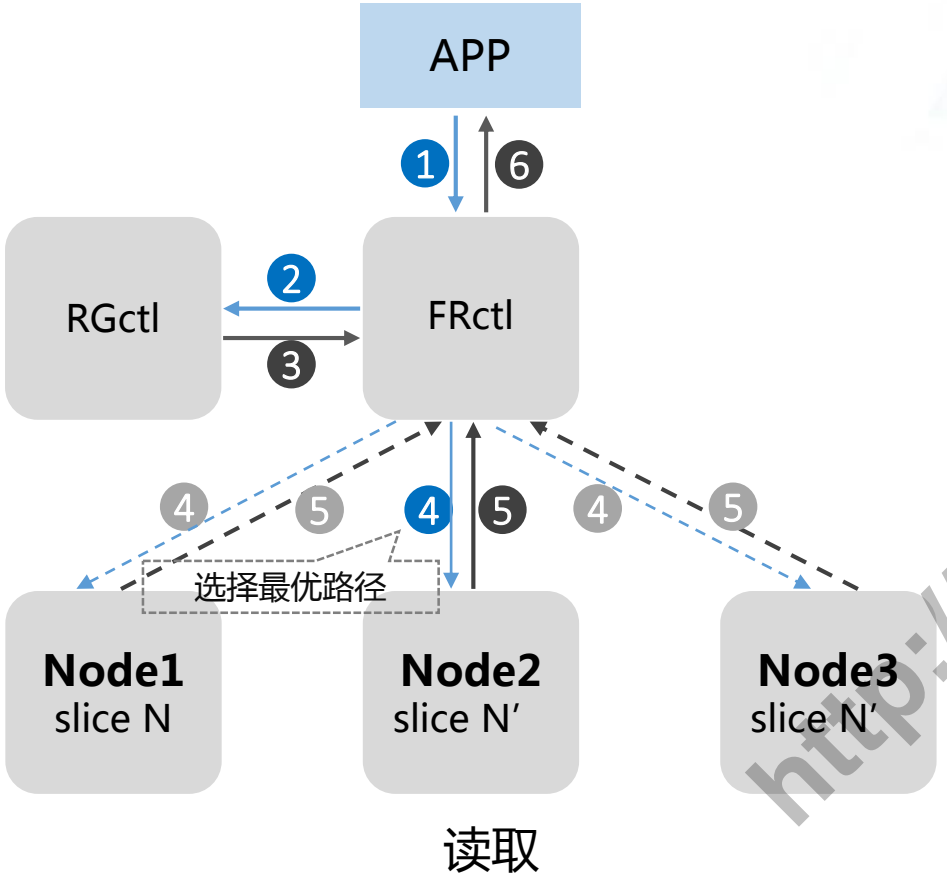


写入

多副本写

1. 应用主机发起写入操作，请求转发至前端控制器（FRctl）
2. 前端控制器访问子卷控制器（RGctl）
3. FRctl从RGctl获取目标卷的元数据，得到数据卷的位置信息
4. FRctl同时向多个节点（按副本策略）发起写入操作
5. 各副本节点数据写入完成后，返回确认消息
6. FRctl向前端返回写入完成信息，并更新相关Slice位置信息

数据读取流程



读取

数据读取

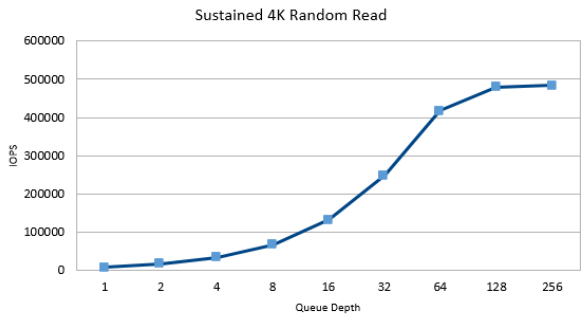
1. 应用主机发起读取操作，读取请求被转发至前端控制器（FRctl）
2. 前端控制器访问子卷控制器（RGctl）
3. FRctl从RGctl获取目标卷的元数据信息，得到数据卷的位置信息
4. FRctl选择时延最优的副本节点发起读取请求
5. FRctl从目标节点上获取数据
6. FRctl将数据传回TGctl，由TGctl发送给应用主机

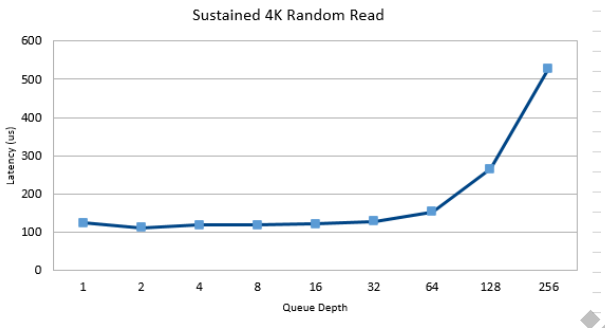
实测性能（100G+NVMe SSD）

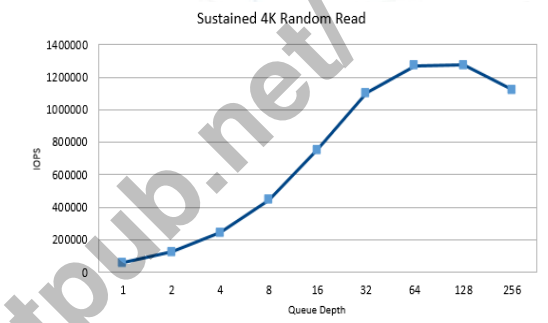
	ceph	FASS	
	优化的Ceph集群（5 OSDs）	FASS集群（4服务器）	*FASS优势*
4K随机读	2,270,000	10,800,000	~5×倍IOPs
4K 70/30随机读写IOPs	691,100	6,080,000	~9×倍IOPs
4K随机写IOPs	463,800	3,370,000	~7×倍IOPs
4K随机读延迟	3000μs(6ms)	166μs	18×倍低延迟
4K 70/30随机读写延迟	6000μs(6ms)	165μs	36×倍低延迟
4K随机写延迟	11000μs(11ms)	150μs	73×倍低延迟
方案估算成本	$185,320(HW only)	$80,000(HW+SW)	43%成本

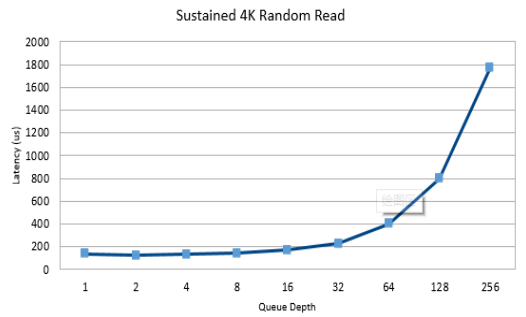
存储节点配置（四台）	
CPU	Intel(R) Xeon(R) Gold 6240 CPU @ 2.60GHz X2
内存	128G DDR4
硬盘	紫光得瑞 D5457 NVMe SSD 3.2T X 5（存储盘） 512GB SSD X2（系统盘）
网卡	Mellanox MCX516A-CCAT 100GB IB 卡（双口） X 2 （前端网、后端网） 万兆网卡 X2（管理网）

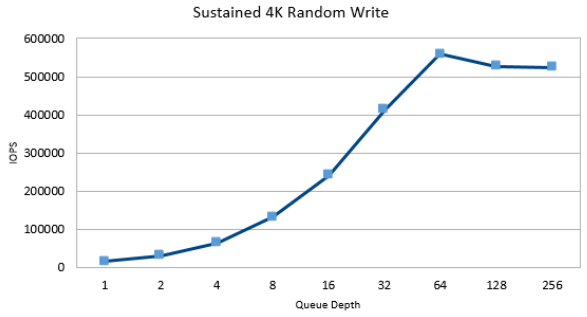
实测性能（25G+NVMe SSD）

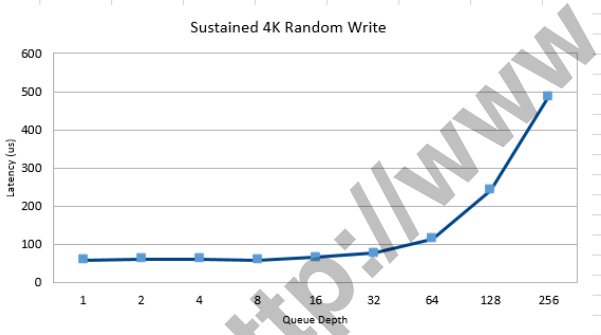


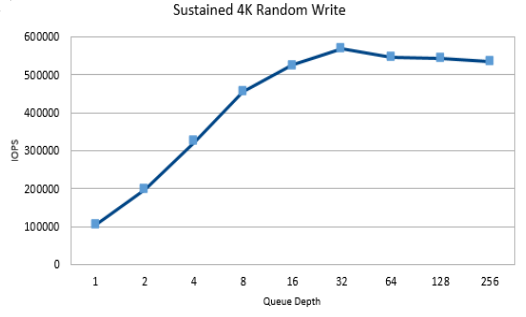


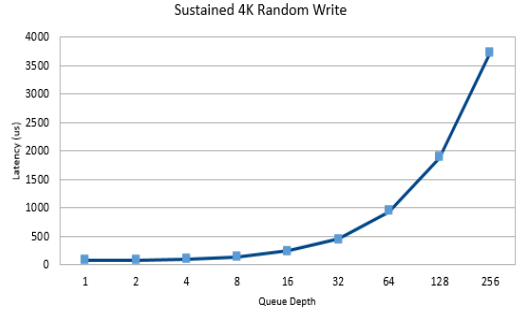












单卷：48万(RR)/55万(RW) IOPS　　单客户端：120万(RR)/55万(RW) IOPS　多客户端多卷：360万(RR)/150万(RW) IOPS

实测性能（10G+SATA SSD）

协议	测试项		IOPS	BW	平均延时（μs）	numjobs	iodepth	服务端节点	客户端节点
iSCSI	4k	随机写	29万	\	\	4	64	4server*5盘	4client*4volume（3副本）
	4k	随机读	55万	\	\				
	1M	顺序写	\	811MiB/s	\				
	1M	顺序读	\	1.9GiB/s	\				
	4k latency	随机写	\	\	712	1	1		
	4k latency	随机读	\	\	542				
iSER	4k	随机写	27万	\	\	4	16	4server*5盘	4client*4volume（3副本）
	4k	随机读	131万	\	\				
	1M	顺序写	\	1.95GiB/s	\	1	2		
	1M	顺序读	\	4.6GiB/s	\				
	4k latency	随机写	\	\	205	1	1		
	4k latency	随机读	\	\	167				

存储节点 硬件配置	节点数量：4 个 节点处理器：Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz *2 节点内存：128G 前/后端网口：10Gbps，2 个 / 25 Gbps，2 个 数据盘类型：SATA SSD、INTEL SSDSC2KB96 数据盘容量：894G 数据盘数量：5 块

典型应用场景

 云计算	 金融科技	 新基建	 其他
云基础设施加速 高速云存储	数据库加速 高性能容器存储	5G、人工智能 工业互联网、物联网	4K/8K非编 HPC、海量小文件

Thanks