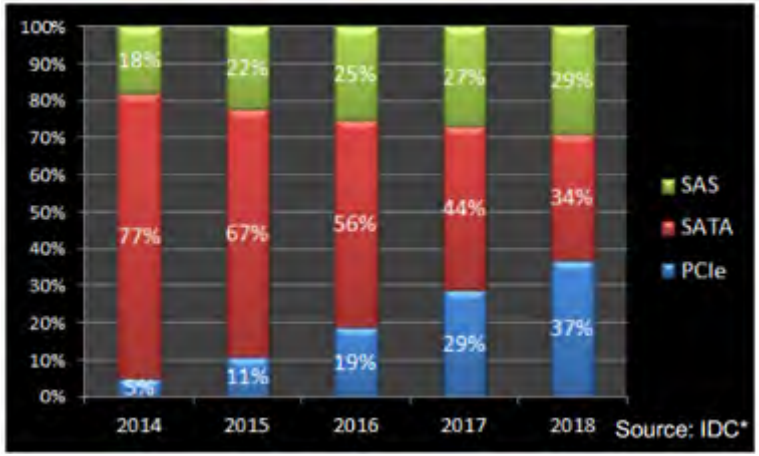
PCIe SSD 在数据中心应用的几点思考

张泰乐

软件定义的数据中心

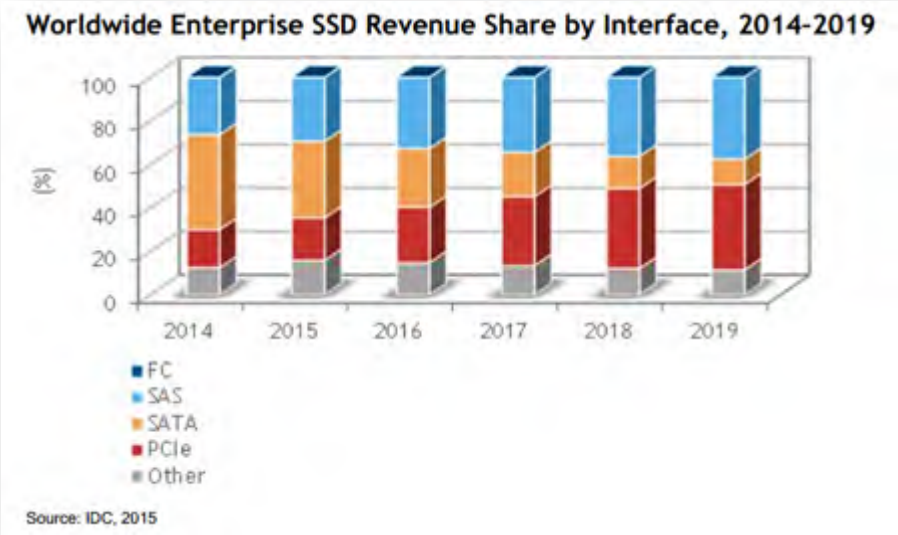
计算	存储	网络
▪ 多核 ▪ 虚拟化 ▪ 异架构计算	▪ 数据保护 ▪ 带宽 ▪ Flash	▪ RDMA ▪ Fabric网络

PCIE SSD 市场份额正在快速增加



To 2018, the proportion of PCIe SSD in datacenter will reach **37%** (from IDC:2014)

As the price point of PCIe SSDs decline, and as more SSD optimized software, servers, and storage systems are brought to market, it expects **PCIe interface SSDs will become a higher proportion** of enterprise SSD revenue over the 2014~2019 (from IDC:2015)



Memblaze PCIe SSD 产品历史

2010

PBlaze1

100GB SLC

Device-based

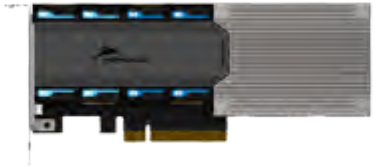
2012

PBlaze2

800GB MLC

Device Based

2013



PBlaze3

4.8TB MLC

Device Based

PBLAZE IV

MemSpeed2.0

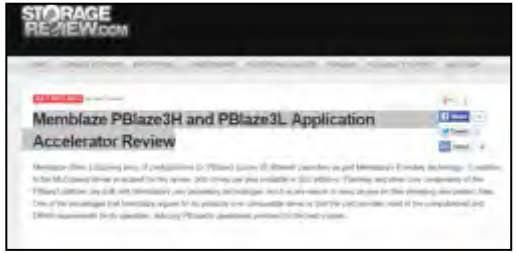
Memsolid2.0

容量	800GB、1.2TB、1.6TB、2.4TB、3.2TB
顺序读（128KB）	UP to 2800MB/s
顺序写（128KB）	Up to 2200MB/s
随机读（4KB）IOPS	Up to 740K
随机写（4KB）IOPS	Up to 200K
延时	20 μs
总线接口	PCIe 3.0 × 4
功耗	⩽ 25W

国际领先的技术和产品

全球最值得关注10大存储初创公司
第一位

产品被美国顶级存储评测机构评为“必须拥有”与最佳功能奖



产品被美国顶级存储评测机构评测为最佳性能奖励

* Memblaze reserves the modification rights without notice

PCIe SSD在数据中心中大规模应用的四个挑战

- 运维成本高
  - PCIe SSD使用私有驱动仍然存在
  - PCIe 闪存卡缺乏热插拔支持
- PCIe SSD的强悍性能对系统提出新要求
  - CPU与网络成为新的瓶颈
  - IO协议栈需要重新优化
- 缺乏高效的数据保护方案
- 容量增加的速度超过了单机应用的需求

Can we fix them?

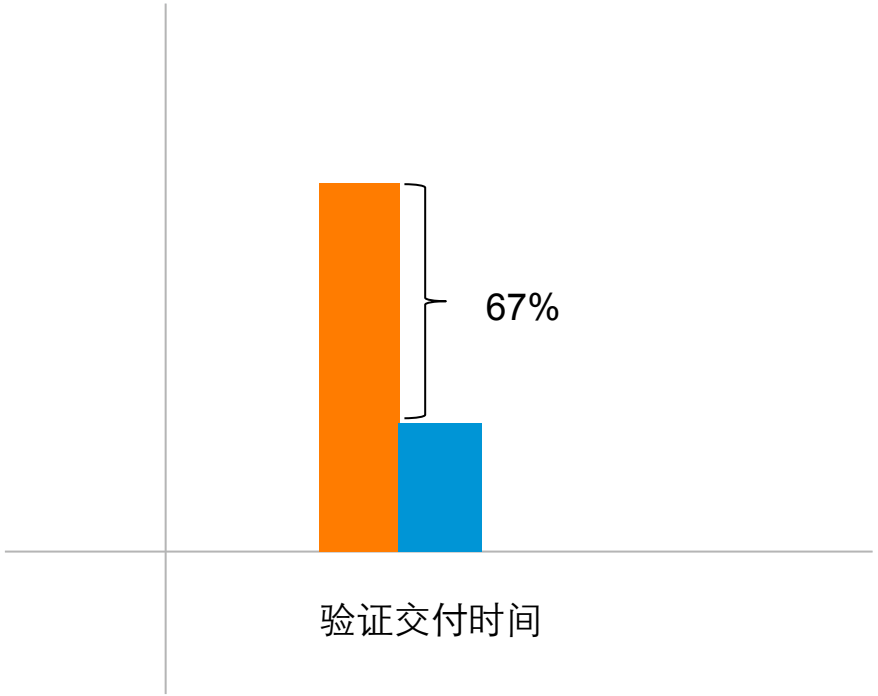
降低PCIe SSD的运营成本的方法

- 采用NVMe标准
  - 降低超大规模部署验证交付时间
  - 支持原生驱动，易维护，降低维护成本
    - Linux RHEL 6.5 and Above
    - Ubuntu 13.04 and Above
    - Windows 2012 R2
  - 多核多队列性能提升
  - 越来越多的应用基于 NVMe进行优化，性能提升
- 2.5寸PCIe SSD可以实现热插拔
  - 降低维护成本
  - 单台服务器部署更多SSD

OS Native Support

SFF8639 Hot Pluggable

奇虎360应用 PBlaze4大幅降低运营成本



PBlaze3

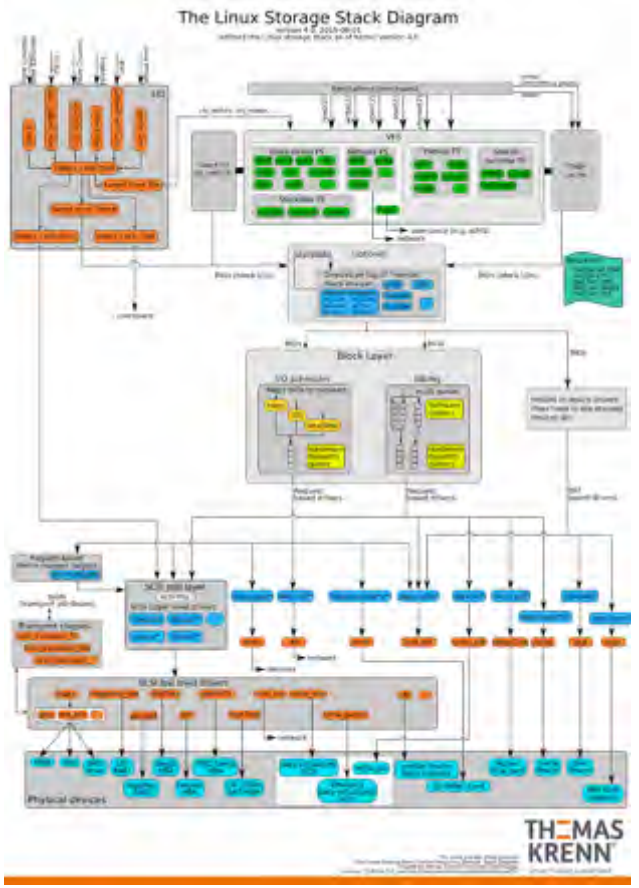
PBlaze4

* 此数据以1000 片计算

优化操作系统，发挥PCIe SSD优势

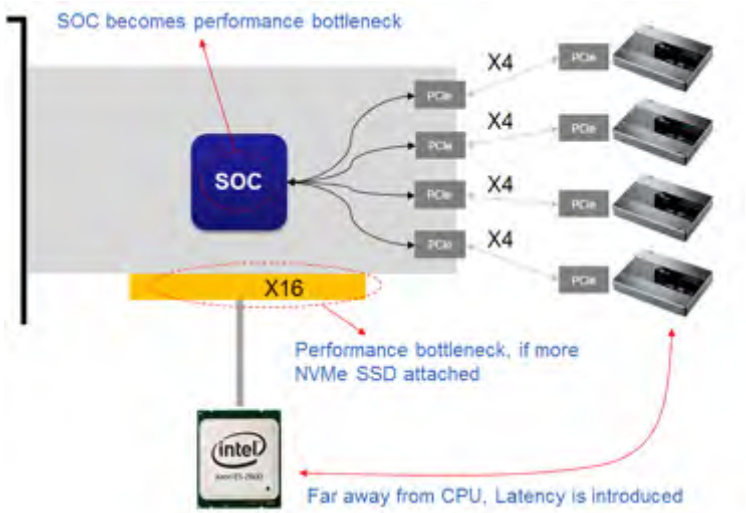
- 多核心优化与中断绑定
- PCIe and NUMA
- Multi-queue for the Block Layer
- Polling Mode Drive

IO协议栈优化，
欢迎联系memblaze技术支持部门
support@memblaze.com

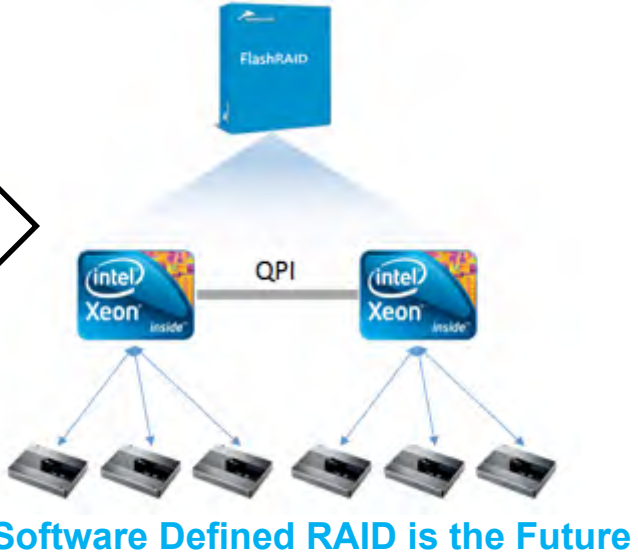


PCIe SSD 单机数据保护方案

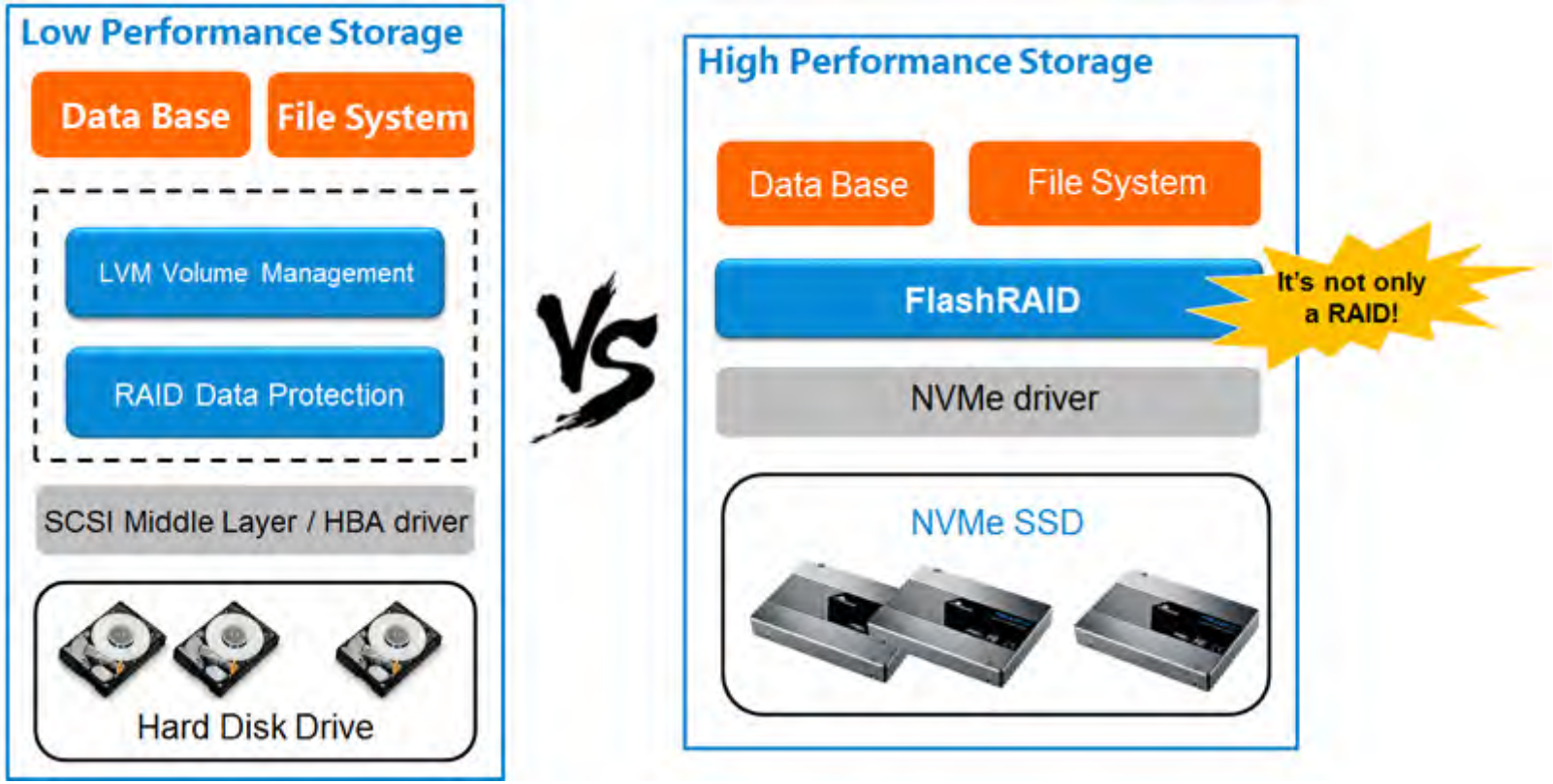
- 硬件RAID不再适合PCIe SSD环境
- 操作系统所含的数据保护技术， 没有为高性能设备优化



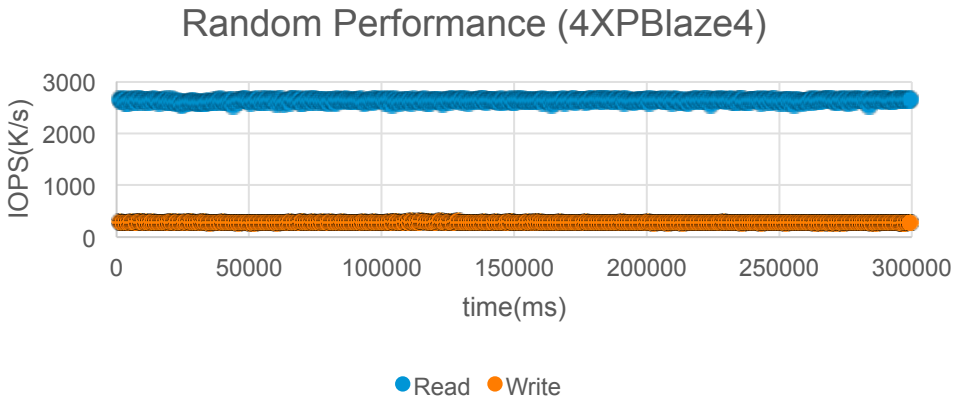
Trends

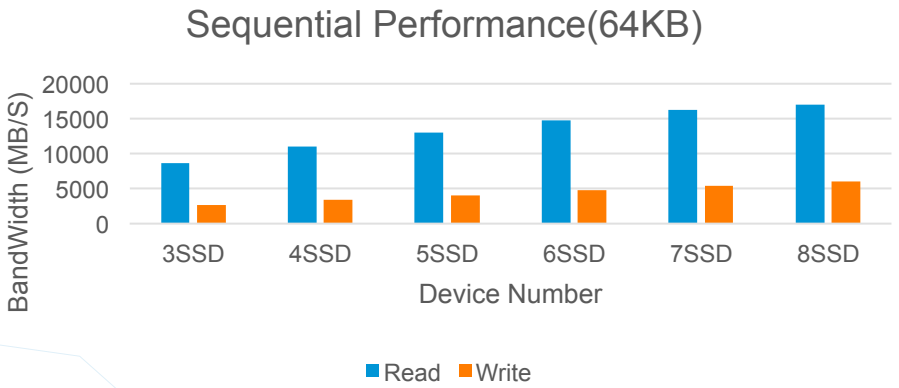


Memblaze的FlashRAID方案



FlashRAID Performance





Test Configuration

- Supermicro Platform
  - 2 Sockets, 32 Cores(Intel(R) Xeon(R) CPU E5-2695 v3 @ 2.30GHz)
  - Pblaze IV 1.6T(4disks)
  - RAID5
- CentOS7.0
- Test Tool: fio
- Disk performance

	IOPS	BandWidth(GB/s)
RandRead(4KB)	742.783K	-
RandWrite(4KB)	150.843K	-
SeqRead(64KB)	-	2.72
SeqWrite(64KB)	-	1.14

某海外客户实测 FlashRAID 性能

	IOPS	BandWidth	Latency(us)
4k_rand_write(16_depth)	416.765K		1226.40
4k_rand_write(8_depth)	325.256K		785.30
8k_rand_write(16_depth)	332.420K		1538.22
8k_rand_write(8_depth)	323.708K		789.10
4k_rand_read(16_depth)	1913.1K		266.00
4k_rand_read(8_depth)	1295.4K		196.09
8k_rand_read(16_depth)	1338.5K		380.82
8k_rand_read(8_depth)	983.595K		258.55
4k_rand_write(1_depth)			241.39
4k_rand_read(1_depth)			197.06
64k_seq_write(64_depth)		3.4GB/s	
64k_seq_read(64_depth)		11GB/s	

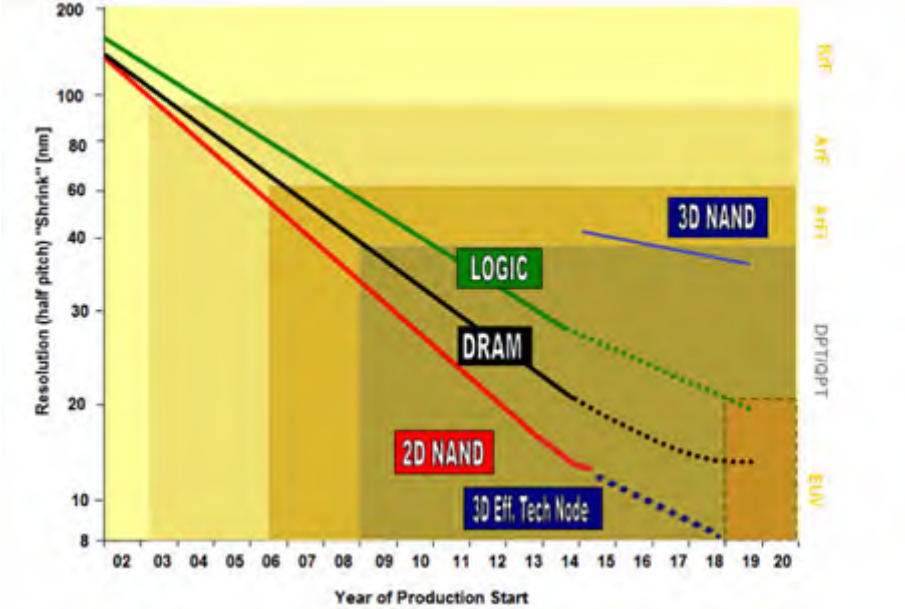
Test Configuration

- Platform
  - 2 Sockets, 32 Cores(Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz)
  - Pblaze IV 1.6T(4disks)
  - RAID5
- CentOS7.1
- Test Tool: fio
- Disk performance

	IOPS	BandWidth(GB/s)
RandRead(4KB)	742.783K	-
RandWrite(4KB)	240.843K	-
SeqRead(64KB)	-	2.72
SeqWrite(64KB)	-	1.34
RandMixRW (r:w=1:2)(4KB)	105.239K(r)/ 204.337K(w)	

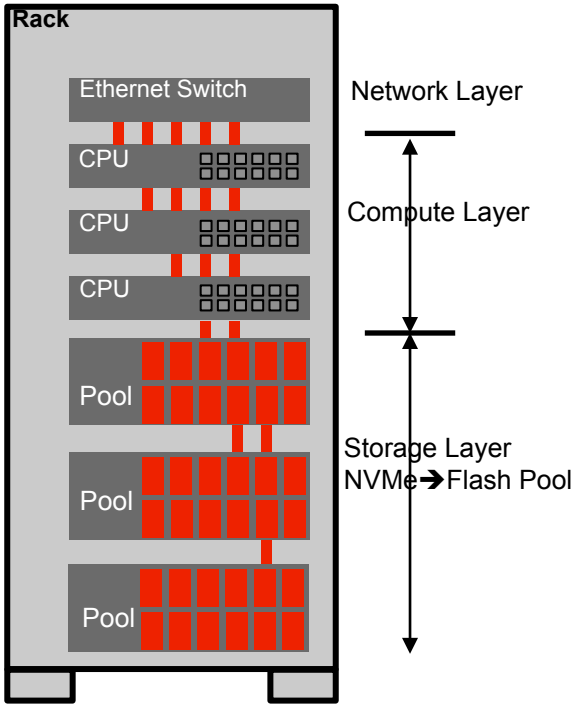
SSD的容量增长速度会超越单机的需求

- SSD的晶体密度的增加速度超过了DRAM与CPU
- 单机需求会受限于木桶的最短板--CPU的增长速度
- 预计在2016-2017年，10-20T容量的SSD成为主流，超过了大部分单机的需求。



资源池化是发挥SSD密度优势的必由之路

机架级资源池化



- NVMe 1.2 多NameSpace技术
  - NameSpace QoS 管理
  - NameSpace 共享技术
- NVMe Over Fabric导出技术
- 存储技术在资源池化中将发挥重要作用
  - 数据去重
  - 分布式数据保护

总结

- PCIe SSD 成为数据中心的主流存储部件
- 巨大的创新与优化空间

关注 Memblaze 官方微信公众平台，获取更多资讯！

Thank you！