Shannon Systems
宝 存 科 技

固态硬盘的前世今生

钟勇

售前技术顾问
上海宝存信息科技有限公司

PostgreSQL

本人简介（#^_^#）

- 80后企业IT基础架构从业人员，从售后工程师进阶到售前顾问
- 5年惠普企业产品售后（原厂工业标准服务器、存储、方案实施）
- 3年惠普企业产品售前（惠普刀片系统、3PAR存储中国总代理）
- 2年戴尔企业产品售前（服务器、存储及解决方案铂金伙伴）
- 以上10年时间都在同一家公司，担任不同的角色，逐渐成长

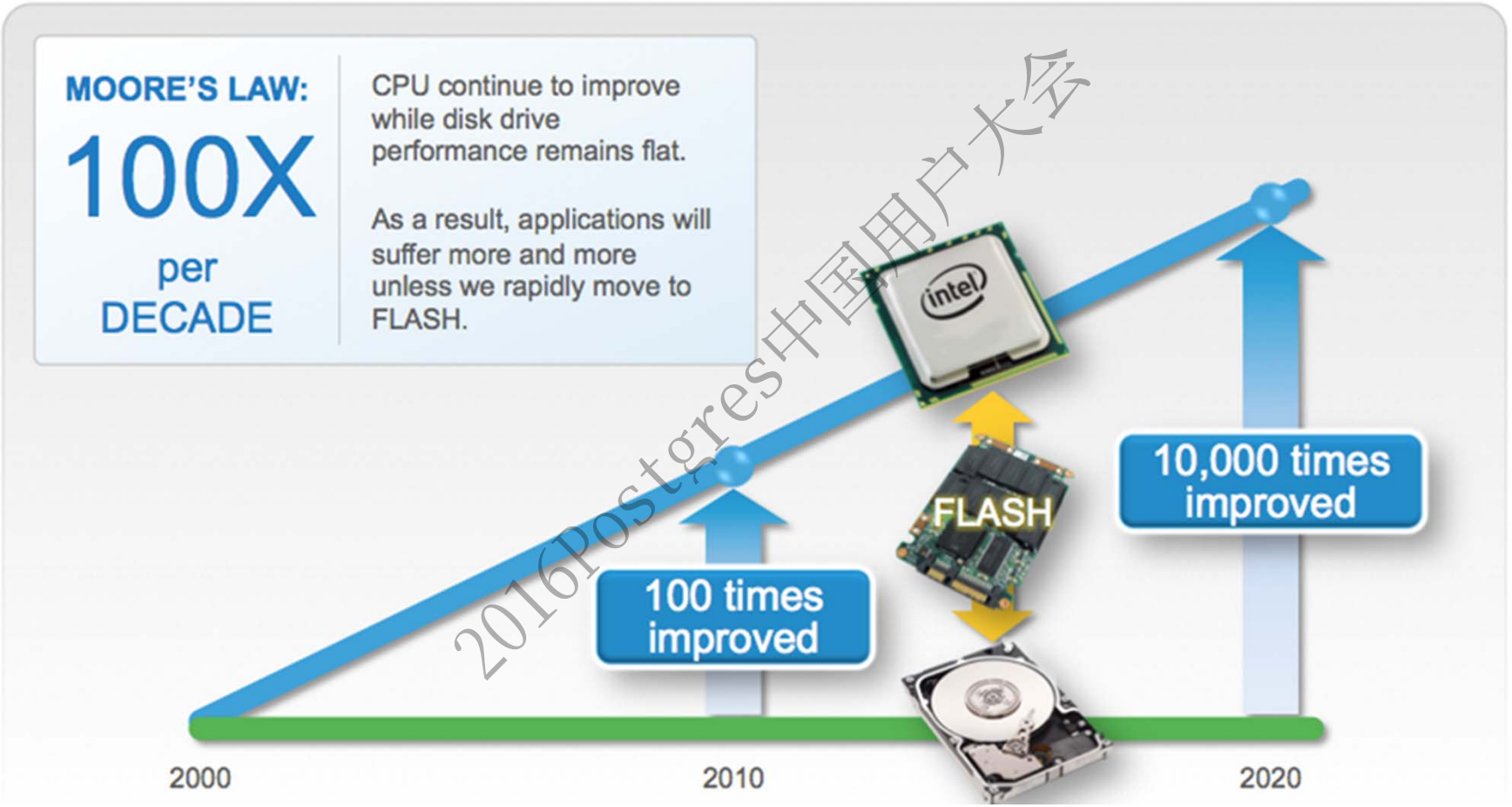
----------命运的分割线----------

- 2015年加入宝存科技，司职华东区售前技术顾问，Base上海
- 致力于帮助用户更好的利用闪存技术，在数据库、虚拟化、分布式存储等场景下最大化IT基础架构的价值。

硬盘----数据存储之地

- 1956年，IBM发明硬盘。
- 1973年，IBM推出了Winchester密封结构硬盘。
- 1996年，希捷推出首款10K转速的硬盘，3.5英寸。
- 1998年，日立推出首款12K转速的硬盘，2.5英寸。
- 2000年，希捷推出15K转速的SCSI硬盘，3.5英寸。
- 2007年，希捷推出其15K转速的SAS硬盘，2.5英寸。
- 直到今年，主流服务器本地存储仍然是10K和15K的SAS硬盘。

为什么需要固态硬盘



性价比到底差多少

固态硬盘的类型

物理接口	外观形态	传输协议
SATA	2.5/3.5英寸盘/M.2	AHCI协议
SAS	2.5/3.5英寸盘	SCSI协议
PCIe	PCIe卡/2.5英寸盘	私有协议/NVMe

PS：目前仍然有部分特殊行业（军事、工业控制）使用其他非主流SSD

SATA HDD和SSD比较-随机读写

	HDD	SSD	For SSD…
读写速度	Up to 150/150 MB/s	Up to 550/500 MB/s	Up to **3.X** 倍
4KB 随机读写延迟	~1.5ms/1.7ms	0.02ms/0.02ms	Up to **85** 倍
4KB随机读写IOPS	~80/200	80K/70K	Up to **1000** 倍
耗电(使用中)	~15W	~3W	**~1/5** 倍
(UBER)不可修复的错误比特率	< 1 in 10^{14}	< 1 in 10^{17}	Up to **1000** 倍
$/GB	< $0.1/GB	$0.3~0.4/GB	**成本已经接近!**

15K SAS HDD和PCIe SSD比较-随机读写

	15K SAS HDD	3.2TB Direct-IO
4KB随机读/写延迟	1000~2500微秒	<100微秒
稳态4KB IOPS	读400/写400	读590,000/写480,000
每个IOPS成本	~6RMB	~0.2RMB
可写入数据量/天 (4K IOPS)	140GB~560GB (4K IOPS ~16K IOPS)	16TB @ 5DWPD
3年总写入数据量	1.5PB~6PB （速度限制）	17.5PB （寿命限制）
故障模式	随机，不可预测	可预测

PCIe SSD拥有超高IOPS和带宽且功耗极低，有效降低IT系统整体拥有成本和复杂度

半高半长PCIe SSD

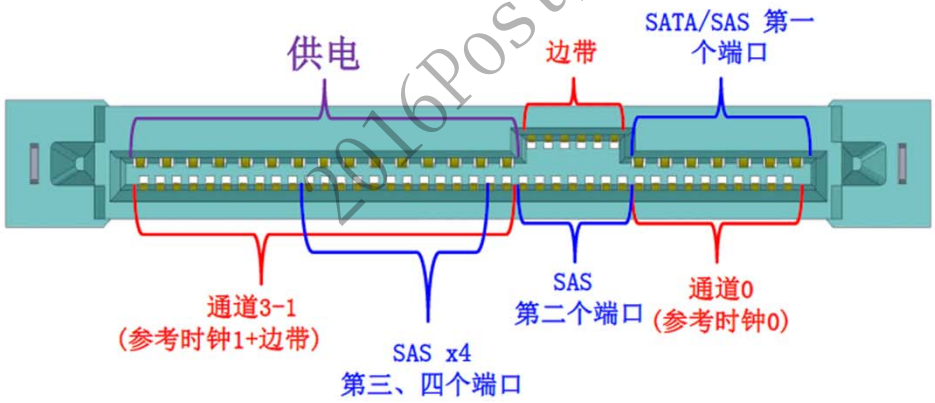
全高半长PCIe SSD

2.5英寸SATA SSD

2.5英寸PCIe SSD

固态硬盘接口趋势：U.2（SFF-8639）

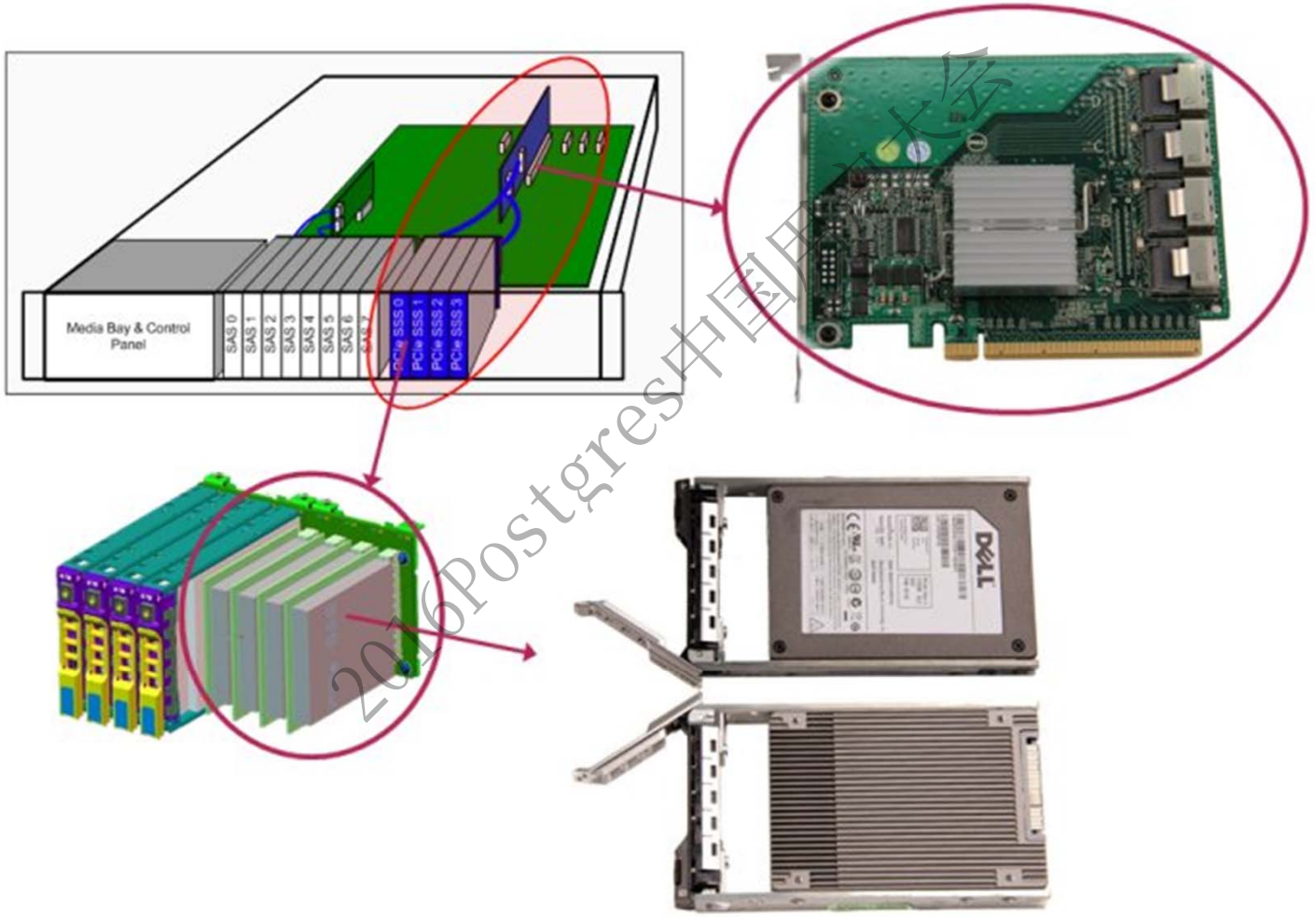
- U.2一种用于2.5英寸固态硬盘的企业级背板接口，涵盖了PCIe、SATA和SAS
- U.2拥有6个通道
- 4个通道 (如下红色标示) 专供PCIe使用，以直连CPU
- 2个通道 (如下蓝色标示) 专供SAS和SATA来连接HBA/RAID控制器或者芯片组



SFF-8639

蓝色 = HBA (SAS, SATA)
红色 = PCIe

U.2硬盘物理链路示例

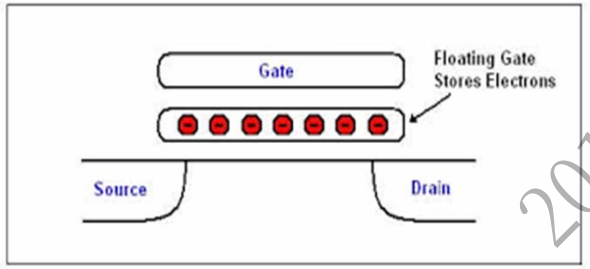


固态硬盘的核心----主控芯片

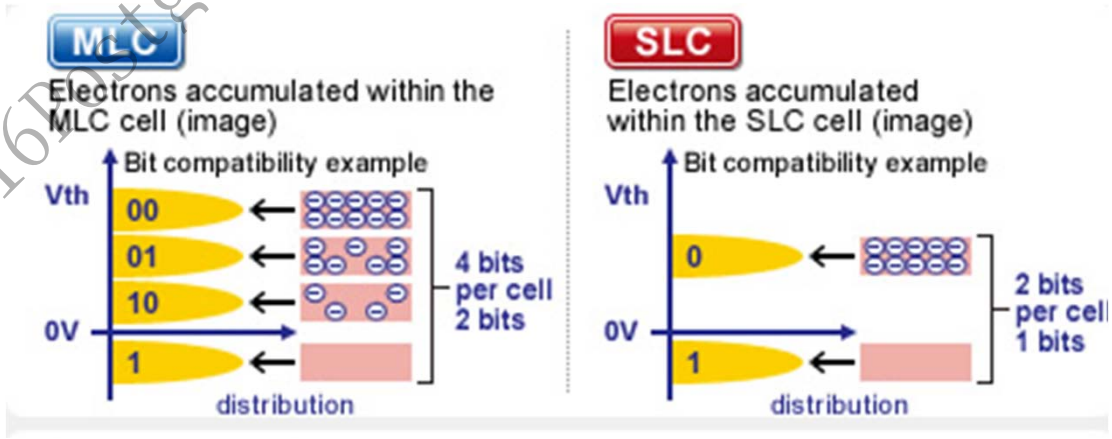
- 固态硬盘的大脑
- 国际大厂长期垄断
- 国内厂商已经迎头赶上
- 大致分为ASIC和FPGA两种形态
- 主控和固件是固态硬盘的基石
- 不同类型的固态硬盘，主控芯片承担的工作内容并不完全相同
- 国内拥有SSD主控完全自主研发能力的仅有宝存、华为等少数几家厂商

固态硬盘的细胞--闪存颗粒（NAND）

- 在固态硬盘中扮演着关键重要角色，是数据存储的最终场所
- 半导体制程工艺越来越先进，在同一物理Cell单元中存储更多的bit



Flash Cell结构



NAND闪存颗粒的五大特性

- 存在坏块，出厂时会有，使用过程中会动态产生
- 读写必须以page为单位
- 不可覆盖写，必须擦出后才可写，擦除以block为单位
- 擦除次数有限制，主流MLC颗粒大约3000至10000次

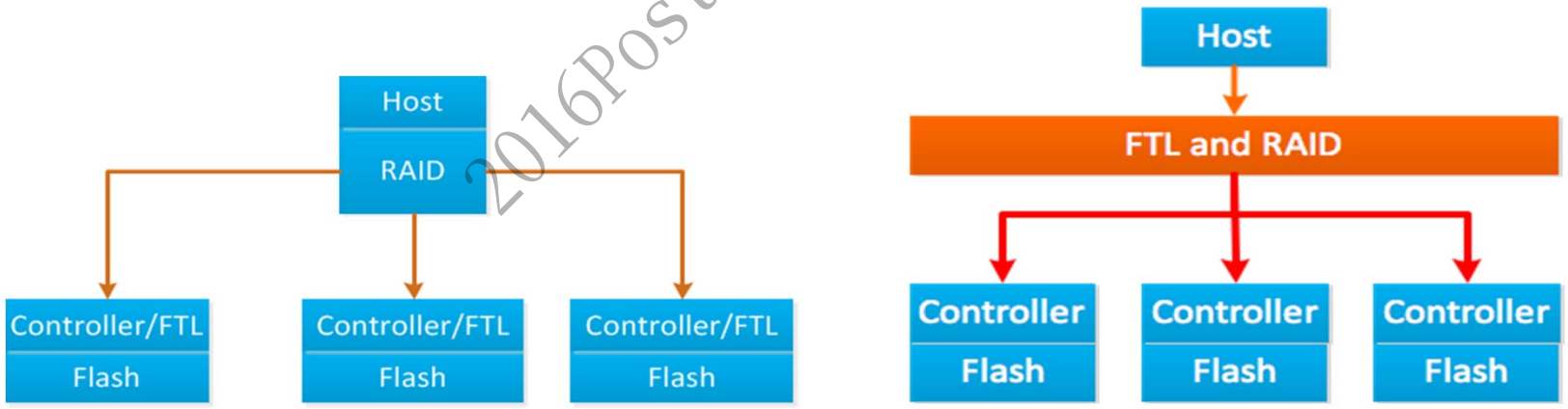
FTL (Flash Translation Layer)
屏蔽Flash特性，模拟成计算机通用存储设备

- 写入数据存在cell上存在bit翻转现象，即数据出错 → ECC/RAID

FTL算法是SSD的核心价值之一

FTL (Flash Translation Layer)的几个概念

- 垃圾回收，Garbage Collection (GC)
- 磨损均衡，Wear Leveling (WC)
- 写放大系数， Write Amplification (WA)
- Over-Provision (OP)



Device-Based与Host-Based SSD

- FTL软件算法在主机上实现（主要用于PCIe SSD）

特点：利用主机CPU、内存资源
优点： 延迟低，容量大，效率高，灵活性好
缺点：需要消耗少量主机资源
代表： 宝存、Fusion-IO的PCIe SSD

- FTL软件算法在设备上实现（主要用于SATA、SAS SSD）

特点：基于嵌入式系统，设备自带cpu、内存等
优点：消耗主机资源较少
缺点：容量小，效率低，功耗高，灵活性差
代表：所有SATA、SAS接口SSD，所有遵循NVMe标准的PCIe SSD

固态硬盘也可以软件定义

- 宝存采用软件定义闪存架构
- 提供完全可深度定制化接口
- Atomic write
- Redo-log 优化
- PCIe RAID
- etc…

数据库+固态硬盘

关于宝存科技

- 宝存科技成立于2011年9月，在北京、深圳、广州、厦门、重庆设有销售、技术支持中心，总部及研发中心位于上海；
- 自主研发并拥有**全部知识产权**的Direct-IO™ PCIe Flash系列产品，**全球第一块**单卡6.4TB PCIe Flash产品；
- **中国第一块**U.2接口Flash存储盘，全球领先的**2.5英寸PCIe接口**技术；
- 全球第一个基于**全局FTL**的通用**PCIe RAID**系统；
- 已提交和正在申请的PCT**专利数量近20件**；
- 互联网、金融、政企、教育行业**客户超200家**；
- **2015年7月**加入Silicon Motion (NasdaqGS: SIMO)大家庭；
- **2015年全年营业额达1亿人民币**；
- 2016年第一季度订单金额超过1.5亿人民币。

Thanks!

Q & A