GOPS2017
Shenzhen

GOPS

全球运维大会

2017

深圳站

指导单位：数据中心联盟 Data Center Alliance

主办单位：高效运维社区 GreatOPS Community　开放运维联盟 GOPSA Open OPS Alliance

SSD在云业务上的大规模应用

兰炜昌 宝存科技

SSD技术创新部分

定制化的经验一： 写放大跟踪

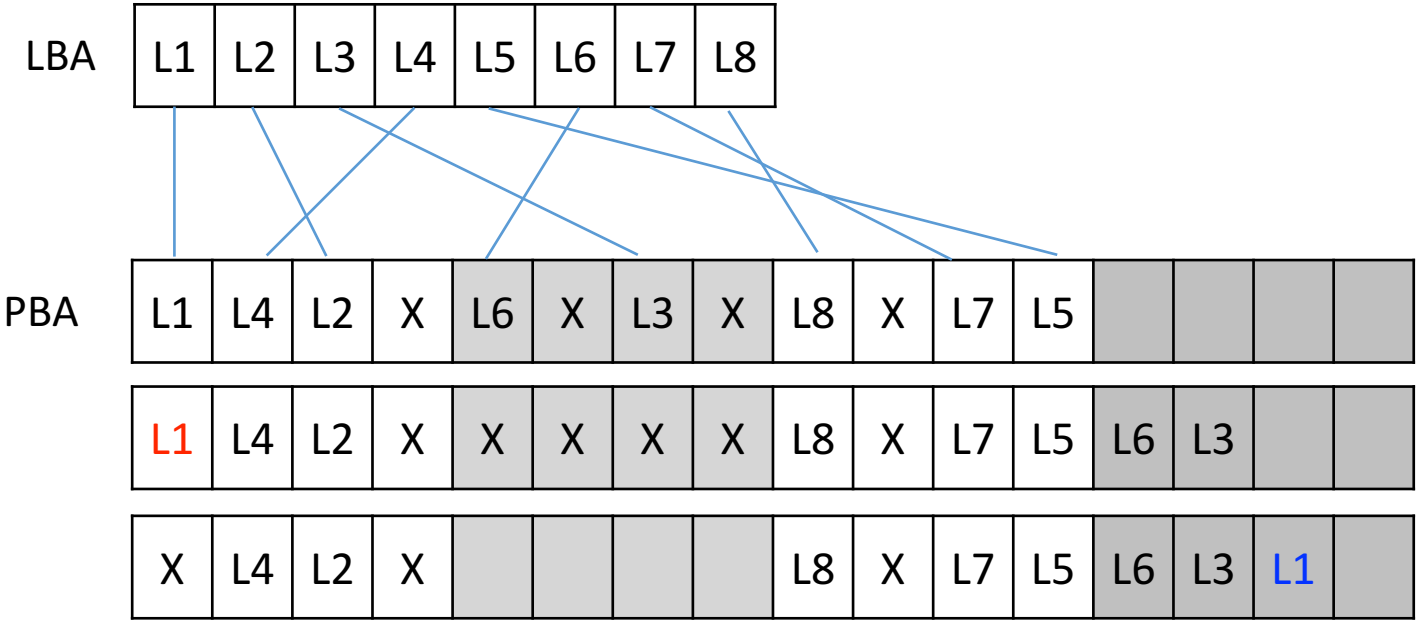
定制化的经验二： 原子写

定制化的经验三： 优先写

定制化的经验四： PCIe RAID技术

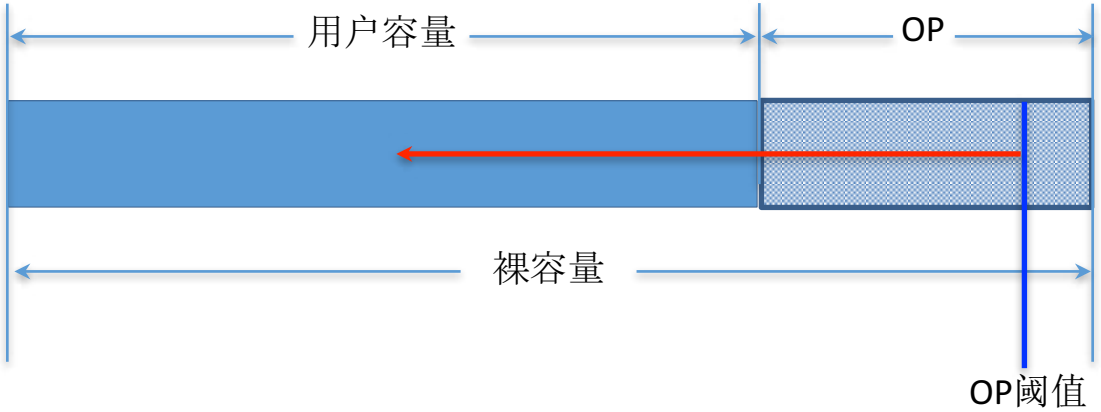
定制化的经验五： 运维关注

写放大跟踪


LBA
L1 L2 L3 L4 L5 L6 L7 L8
PBA
L1 L4 L2 X L6 X L3 X L8 X L7 L5
L1 L4 L2 X X X X X L8 X L7 L5 L6 L3
X L4 L2 X L8 X L7 L5 L6 L3 L1

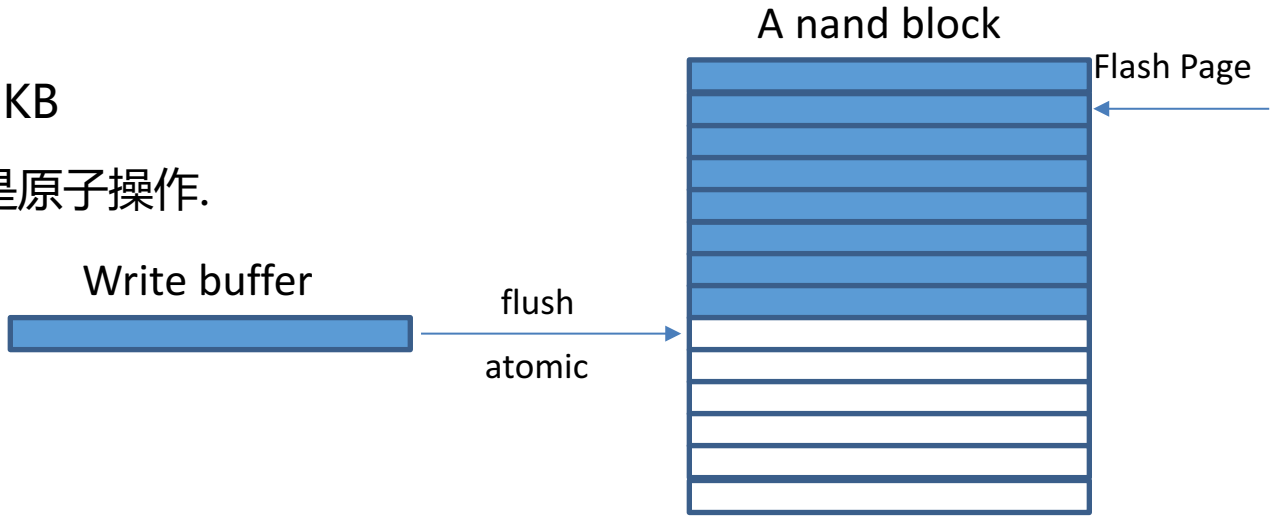
写放大跟踪

- 增加可用容量，动态压缩OP
- 收益：寿命和性能范围内,提高单机承载能力



原子写

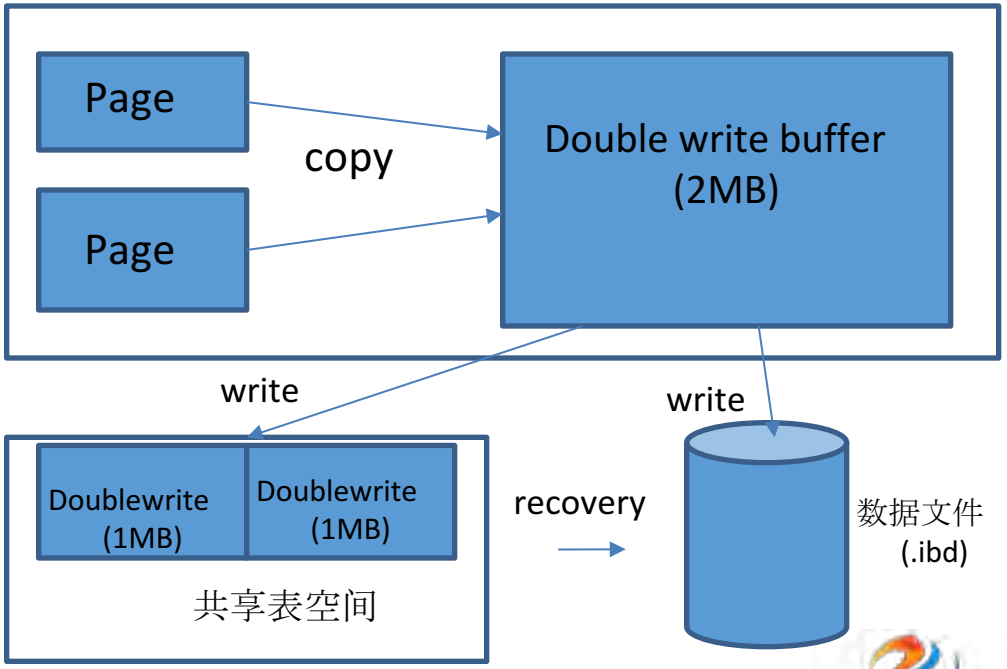
- Flash Page Size: 32KB
- NAND flash write 是原子操作.



Double Write

- 传统硬件和操作系统不能保证InnoDB Page 写入这一操作的原子性
- InnoDB使用Double write这个特性来避免InnoDB Page写入不完整的问题.

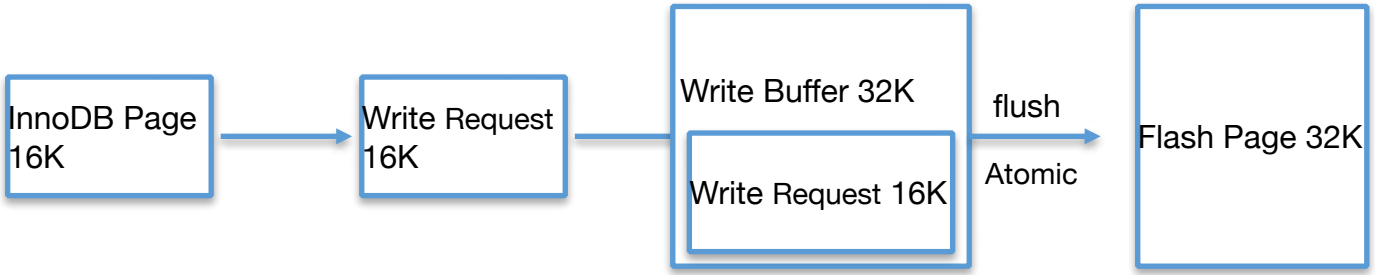
- Double Write 缺点:
- 数据写入量加倍(Bad For Flash)
- 增加了写入负载(不是2倍)



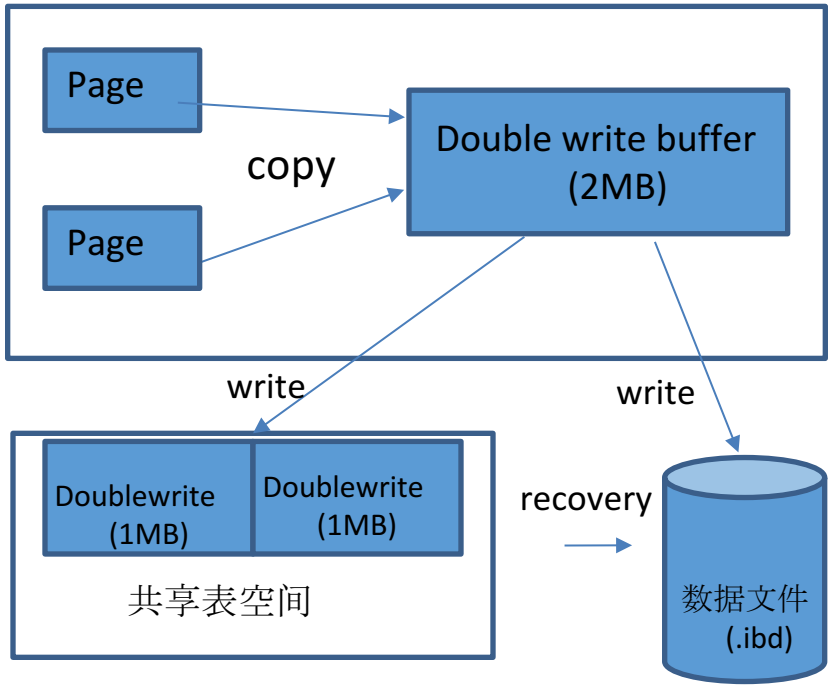
原子写

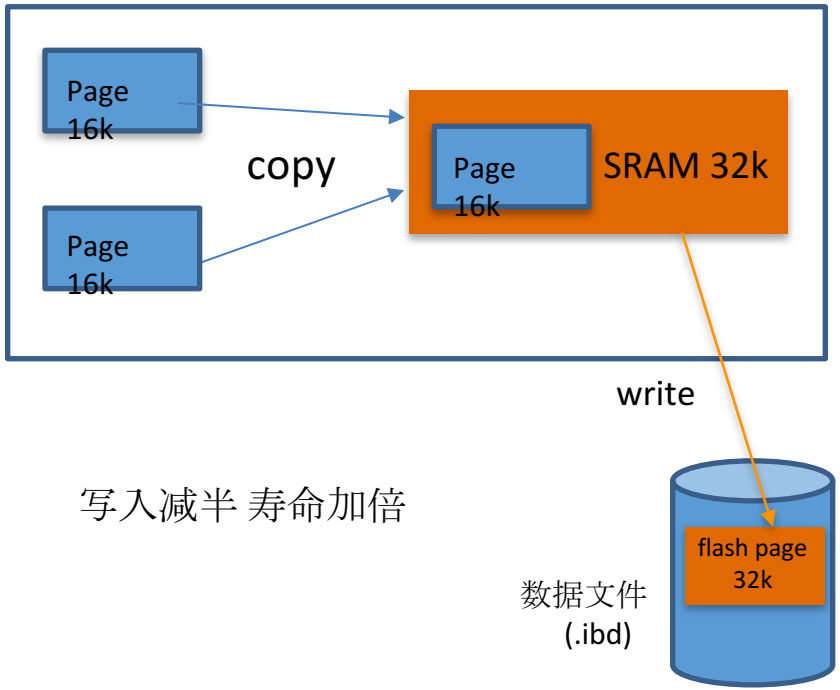
- InnoDB Page Size: 16KB(default)
- Flash Page Size: 32KB

- 原理：将InnoDB Page包在一个NAND Page里



原子写

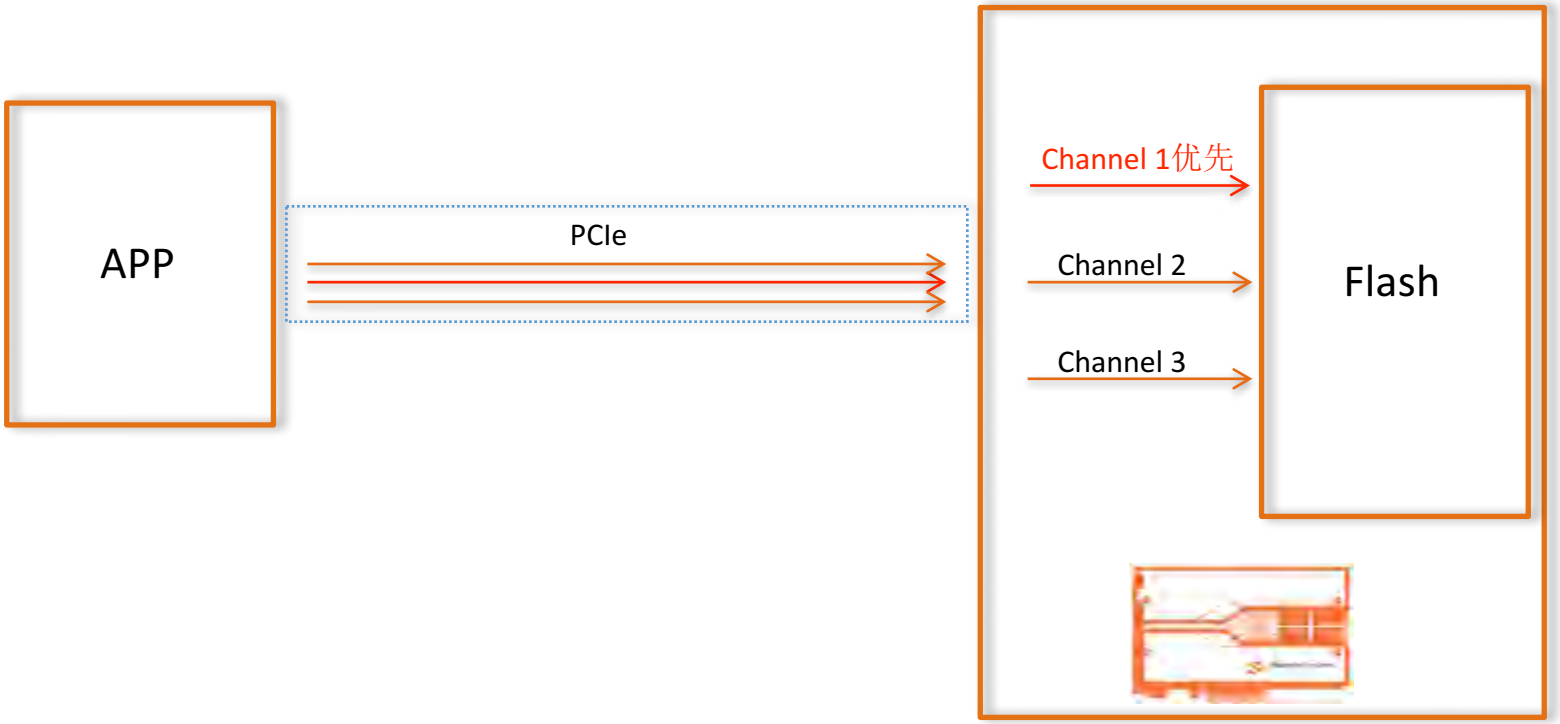

Page
Page
copy
Double write buffer
(2MB)
write
write
Doublewrite
(1MB)
Doublewrite
(1MB)
共享表空间
recovery
数据文件
(.ibd)


Page
16k
Page
16k
copy
Page
16k
SRAM 32k
write
写入减半 寿命加倍
flash page
32k
数据文件
(.ibd)

性能优化

- **原子写**
  - MySQL关闭Double Write；
  - Direct-IO打开原子写支持；
  - 30%左右的写性能改善
  - 写延迟降低50%；
  - 闪存寿命加倍。

redo log优化（优先写）



Redo Log 优化(Redo log optimization)

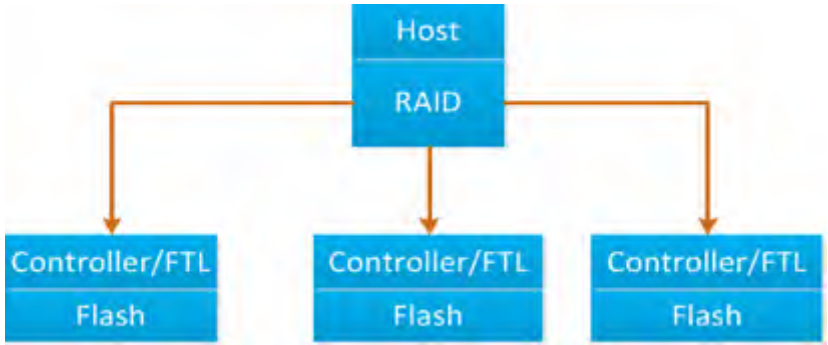
- 判断MySQL/InnoDB 传下来的flag
- if (EXT4_I(inode) ->i_flags & EXT4_PRIO_FL)
- 设置Flag 传给Direct-IO Driver
- bio->bi_flags = bio->bi_flags | (1 << BIO_RW_PRIO);

性能优化

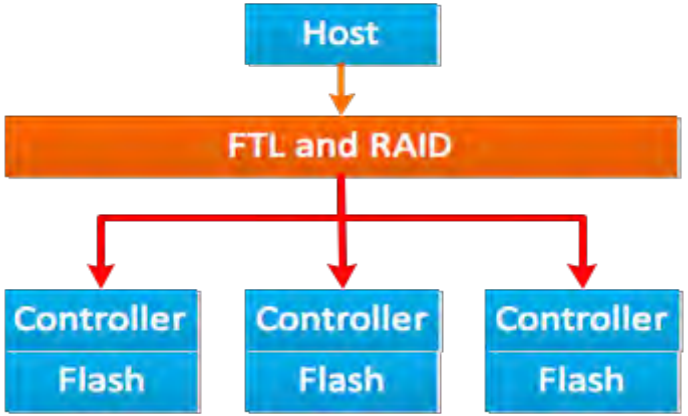
优先写

- 标记流量独占优先写硬件通道；
- redo log流量标记；
- tps提升一倍以上。

Shannon Direct-IO™ PCIe RAID



传统SSD RAID5架构



宝存科技创新PCIe RAID架构

Shannon Direct-IO™ PCIe RAID

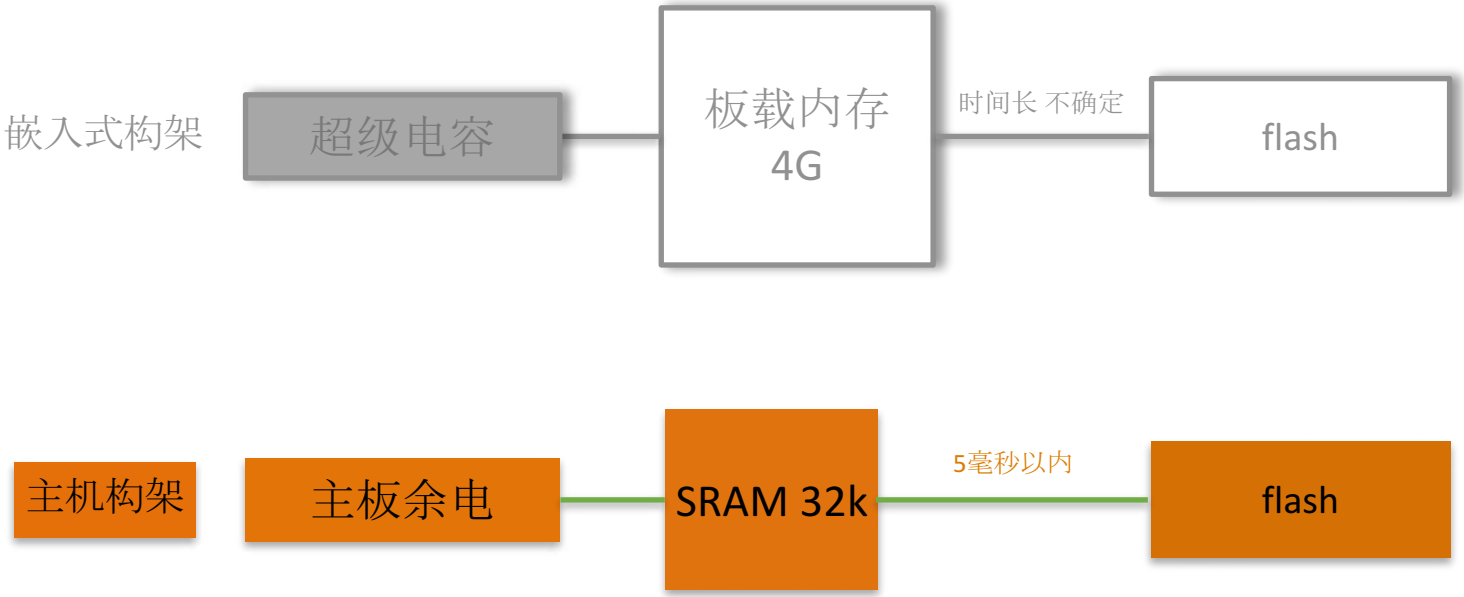
提供一个集大容量,高性能和高可靠性于一体的逻辑块设备.

- 大容量需求, ~90T
- 避免单点故障
- 解决传统的RAID5阵列写性能极差的问题
- 解决传统的闪存盘阵列会有很大的写放大现象.
- 解决传统SSD硬盘性能稳定性和性能一致性不足的问题
- 支持原子写和redo log优化

闪存在运维中的关注点

- 掉电保护
- 异常处理
- 超低功耗

闪存在运维中的关注点：掉电保护



FIO 测试 PCIE- SSD 掉电

测试掉电步骤：

./plverify_with_fio.sh /dev/xxx prepare　　#　 prepare 阶段

./plverify_with_fio.sh /dev/xxx run　　# 在写阶段拔掉电源

./ plverify_with_fio.sh /dev/xxx check　　# 重新上电 check 阶段

闪存在运维中的关注点：监控接口

- **内部延迟观察**
  - 提供区间内最大延迟打印
- **后台流量监控和调节**
  - 提供GC/WL流量监控接口；
  - 提供GC/WL流量控制接口；
- **介质失效处理(坏块处理)**
  - 主动探测功能；
  - 主动标记屏蔽功能；
  - 内部修复速率调节

闪存在运维中的关注点：功耗（寿命）

- 标准温度下数据维持标准时间长度
- 很大程度取决于擦除次数
- 与工作温度强相关

商用级闪存的正常工作范围为0至70°C，

TOSHIBA TOSHIBA CONFIDENTIAL

2.4. Operating Temperature Condition

Table 6 Operating Temperature Condition

Symbol	Parameter	Part Number	Rating	Unit
TOPER	Operating Temperature Range for Commercial	TH58TFG8DFKBA4K	0~70	°C
	Operating Temperature Range for Commercial	TH58TFG9DFKBA8K	0~70	
	Operating Temperature Range for Commercial	TH58TFT0DFKBA8J	0~70	
TSOLDER	Soldering Temperature (10 s)		260	

NOTE:

1) Operating Temperature (T_{OPER}) is the case surface temperature on the center/top side of the NAND.

2) Operating Temperature Range specifies the temperatures where all NAND specifications will be supported. During operation, the NAND case temperature must be maintained between the range specified in the table under all operating conditions.

闪存芯片可靠性与闪存芯片的温度关系

Storage Temp (°C)	Acceleration Factor Relative to 55°C	Bake Time (Hrs) Equivalent to 1 Yr at 55°C
125	939	10
85	26	360
70	5	1712
55	1	9390 (~1 year)
25	0.01988	442,380 (~50 years)

宝存对Flash卡使用的建议

- 进一步加强监控
  - Flash卡作为硬件产品，无论概率多小，均存在失效可能
  - 需要在系统级提供有力监控机制，最小化失效时对业务的影响
  - 提供多种自动化监控插件和工具
- 系统冗余机制
  - 防止单点故障
  - 故障自动切换，减小对业务的影响。

高效运维社区
GreatOPS Community

会议

- 3月18日 DevOpsDays 北京
- 8月18日 DevOpsDays 上海
- 全年 DevOps China 巡回沙龙
- 4月21日 GOPS深圳
- 11月17日 DevOps金融上海

培训

- EXIN DevOps Master 认证培训
- DevOps 企业内训
- DevOps 公开课
- 互联网运维培训

咨询

- 企业DevOps 实践咨询
- 企业运维咨询

商务经理：刘静女士
电话 / 微信：13021082989
邮箱：liujing@greatops.com

Thanks

高效运维社区

开放运维联盟

荣誉出品

想第一时间看到

高效运维社区公众号

的好文章吗？

请打开高效运维社区公众号，点击右上角小人，如右侧所示设置就好