Alibaba Group 阿里巴巴集团 | 基础架构事业群 Alibaba Infrastructure Service

用户态本地存储引擎

— 百万IOPS背后的故事

刘攀
基础设施事业群存储技术专家
2018/3月

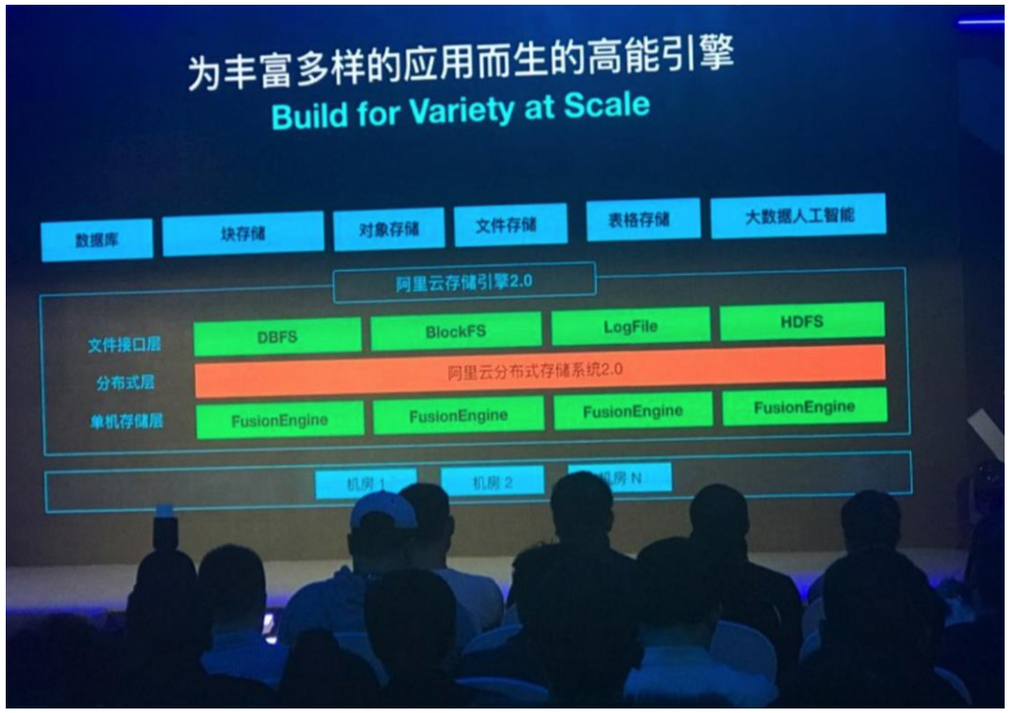
提纲

- FusionEngine项目背景
- USSOS介绍
- USSFS介绍
- 工具支持
- 后续计划

项目背景 — 问题的提出

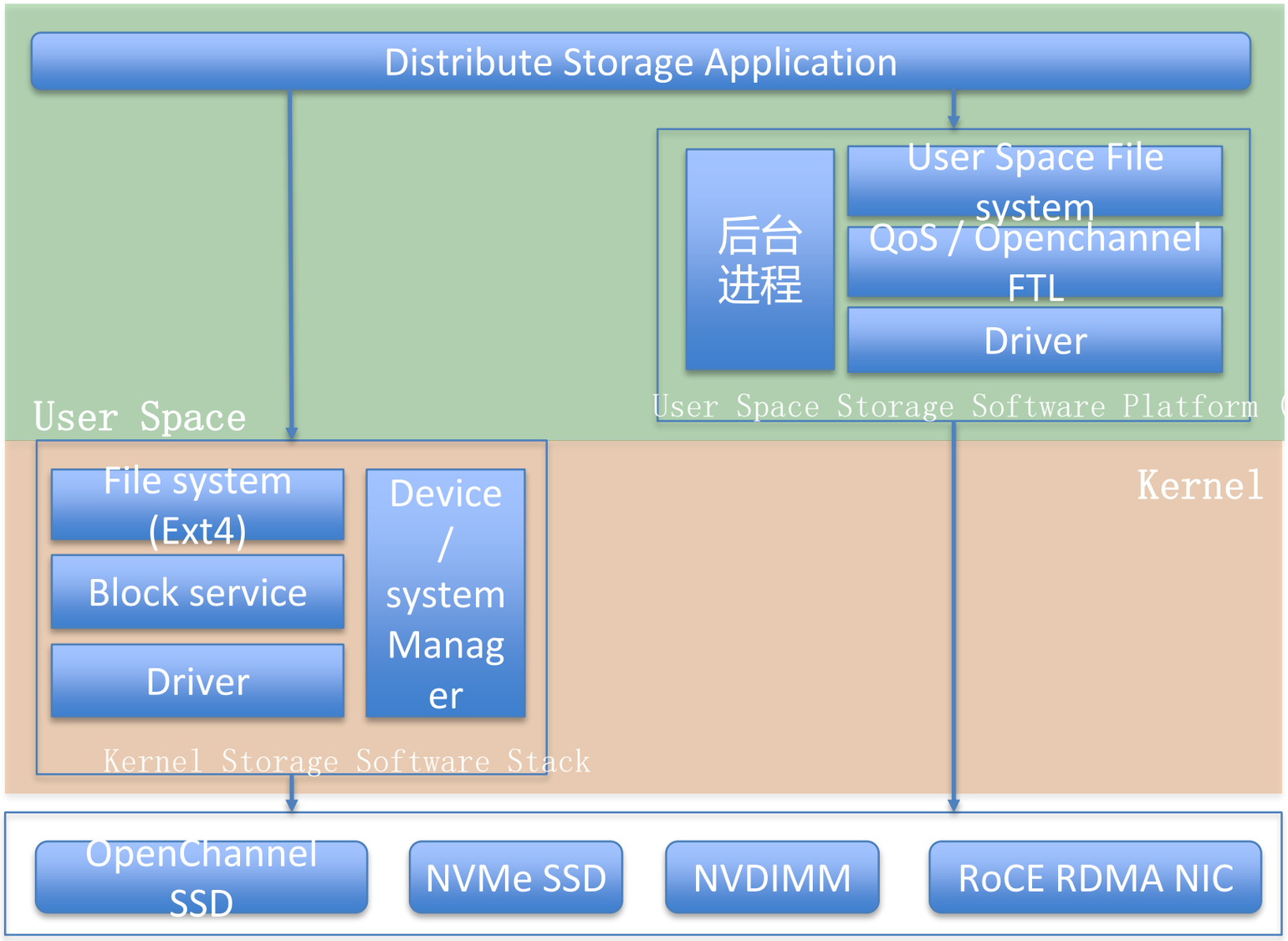
- NVME SSD带来的挑战
  - 以NVMe SSD为代表的新硬件介质的发展推动了存储软件栈的变革。网络和处理器逐渐成为严重的性能瓶颈点。
  - 旧的存储软件栈的设计围绕磁盘的特性，设计NVME SSD存储软件栈时需要采用不同的技术手段。
- 传统内核栈的问题
  - 线上运维对于云存储而言非常重要。
  - 传统分布式存储往往架构在Ext4之类的本地文件系统之上
  - 定位问题难度大。内核软件栈中问题的定位是一个非常困难和漫长的过程。
  - 解决问题风险大。在实际应用中，即使root cause的问题都有可能不会去修复。

项目背景 — 2018年1月9日，上线ESSD公有云项目



USSOS: User Space Storage Operating System

- Bypass kernel提升处理器效率
- 简化存储软件栈提升IO性能
- 建立统一的用户态设备管理体系
- 切合高性能存储介质的发展趋势
- 全用户态软件栈，简化运维

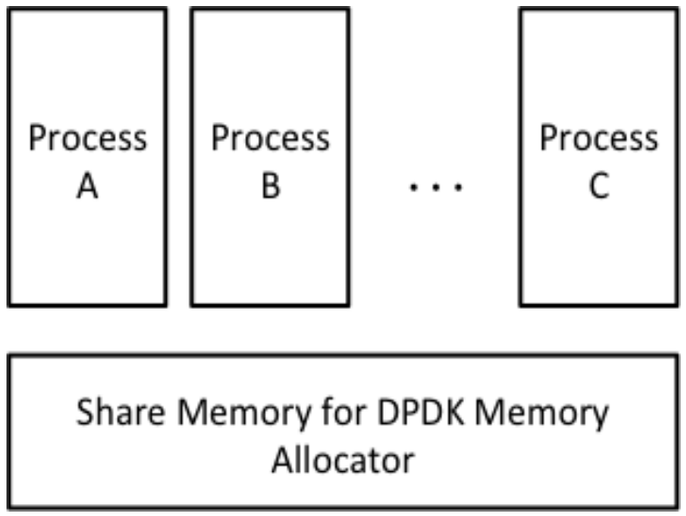


与原生SPDK的关系

- DPDK
  - USSOS会依赖于DPDK库中的内存管理实现
    - 共享内存的实现
    - 内存分配器的实现，包括rte_malloc，mempool以及hashtable
- SPDK(17.07)
  - USSOS在SPDK NVMe驱动的基础上对用户态驱动进行了改造及增强

垃圾内存回收机制

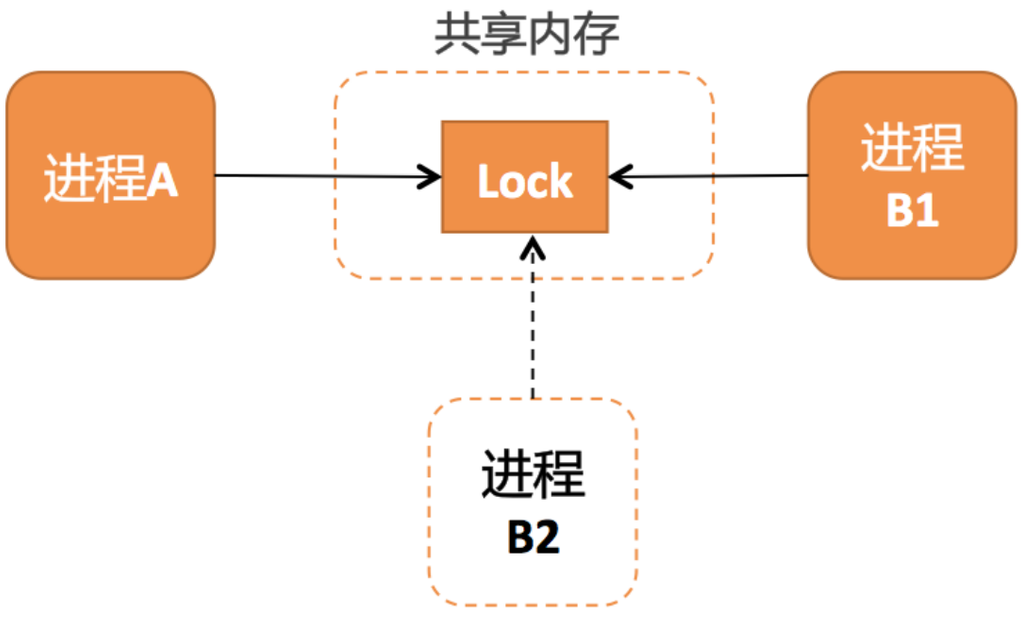
在多进程共享Memory分配器的情况下，当一个进程异常重启，该进程所占Memory将会发生泄漏



多进程共享内存分配器导致的问题

共享内存中的竞争锁问题

- 多个进程共享内存，通过锁机制进行临界区保护。当一个进程持锁异常退出之后，整个系统将会面临死锁的问题。对外表现为进程Hang住。



共享内存竞争锁问题

内存管理优化

```
    68.39%    68.27%  pangu_chunkserver    [.] malloc_heap_alloc
    61.25%     0.00%  [unknown]            [.] 0000000000000000
    27.40%    27.29%  pangu_chunkserver    [.] malloc_elem_free
    23.74%     0.00%  [unknown]            [.] 0xec81485354415541
```

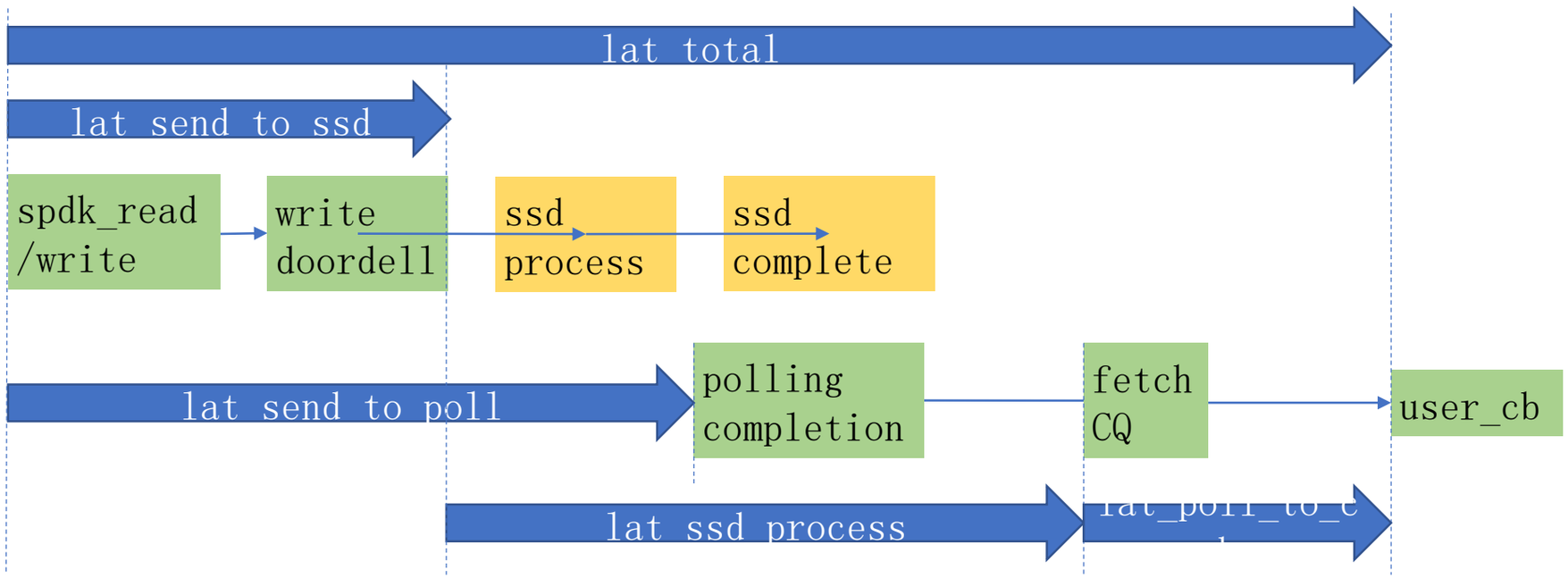
1. 问题的提出

IO有一段时间比较慢，或者压力很大的时候，发现cpu使用率飙高。然后使用perf top进行查看，发现，cpu的热点， 都发生在内存分配和释放的时候。经过代码分析，cpu利用率都在spinlock里面。原因是大压力下，dpdk分配效率过低。导致线程之间争抢spinlock，cpu消耗过大，然后IO更加慢。

2. 解决方法

dpdk现有的内存池mempool，只能是等大小的object。而IO的大小是不定的。然后mempool无法动态扩充和回收内存。

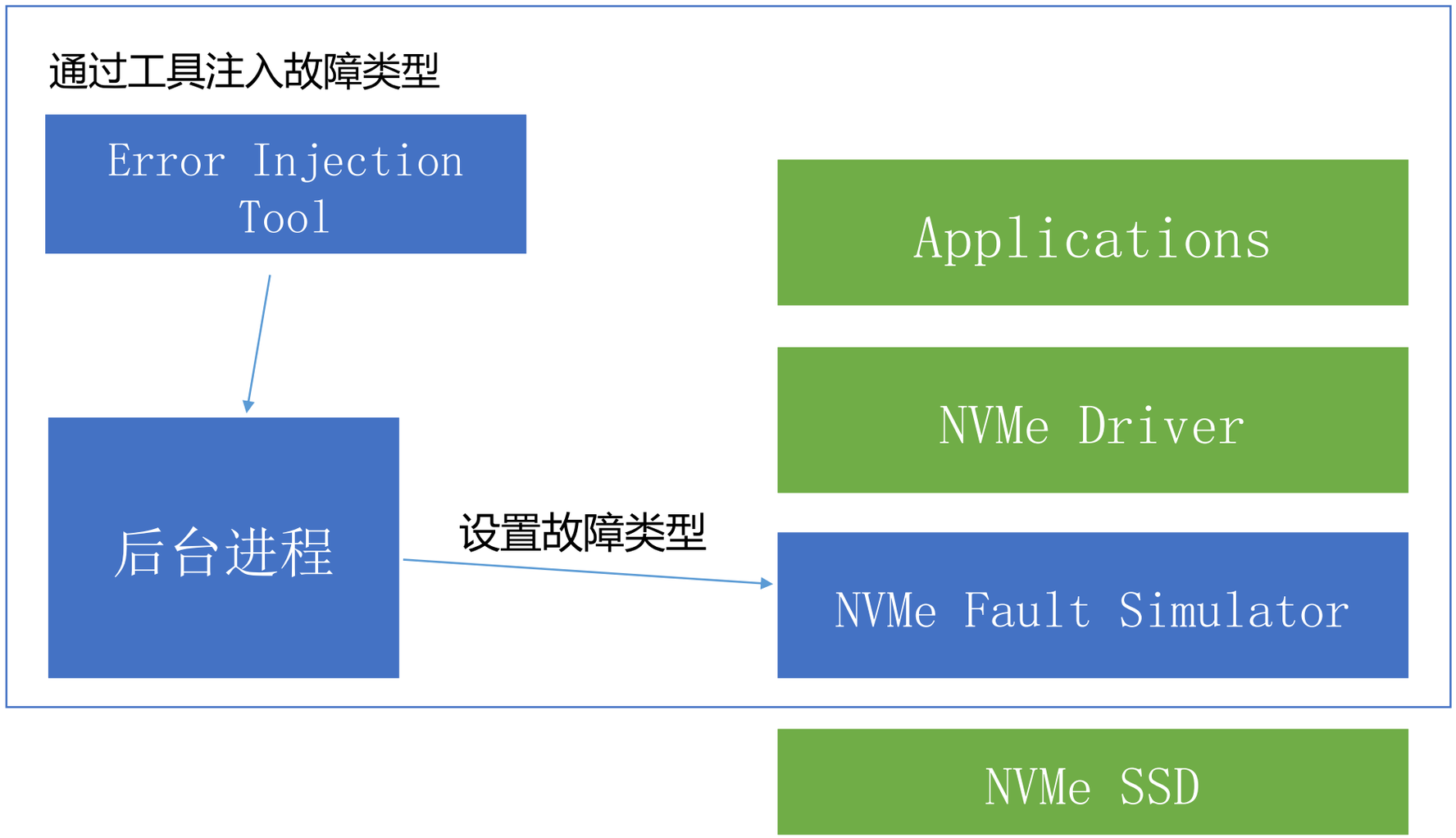
性能统计



- 在遇到IOComplete时间比较长的时候，如何定位
- IOComplete时间长一般要么是盘的处理速度慢，要么就是Polling调用慢；在polling模式下盘对于命令的处理速度并不能精确统计。

工具支持 — 故障模拟器

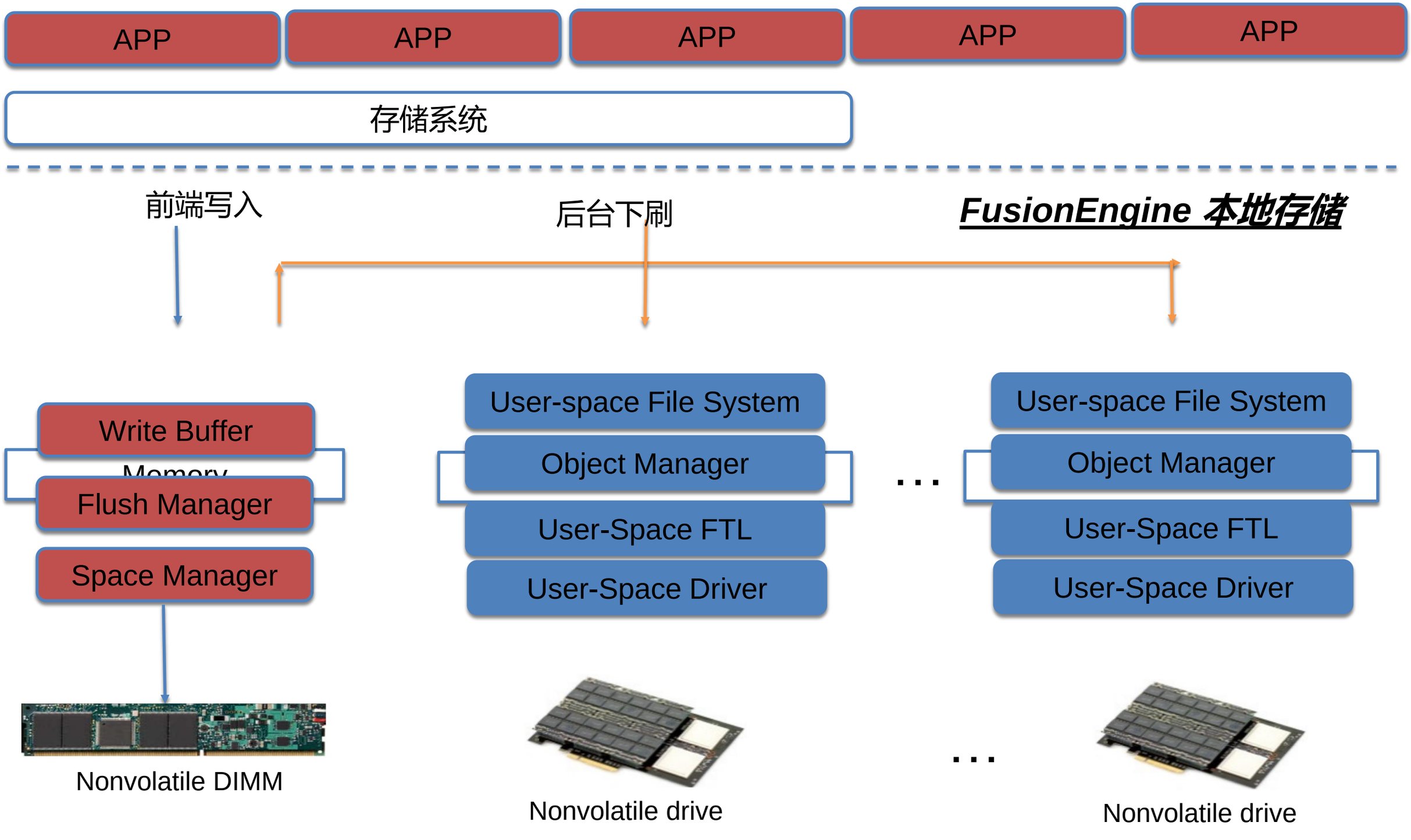
通过故障模拟器增强存储软件栈测试



通过模拟器实现IO故障模拟，实现对存储软件栈Error Handling code path测试

后续计划 — 支持高性能混合存储

✈ 优势：低延时，低成本，高效IO聚合，非易失，高吞吐



Alibaba Group 阿里巴巴集团 | 基础架构事业群 Alibaba Infrastructure Service

感谢

wanjun.lp@alibaba-inc.com