DTCC2013

百度大规模存储系统

钱一峰

qianyifeng@baidu.com

提纲

- 百度的数据
- 百度的存储系统
- 面临的挑战
- 新存储体系
- 经验教训

百度的数据

百PB级:

- 网页 & 超链
- 日志 + 数据仓库
- 广告
- UGC
- 个人云

百度的数据——特点

- 性能
  - 高吞吐 VS 低延迟、高并发
- 规模
  - 10PB级 VS T级
- 时效性
  - 非实时 VS 实时
- 读写
  - 易变 VS 静态
- 大小
  - 小记录 VS 大记录
- 数据组织
  - 无序 VS 有序
- 一致性
  - 弱 VS 强
- 处理方式
  - 批量(顺序) VS 单个(随机)

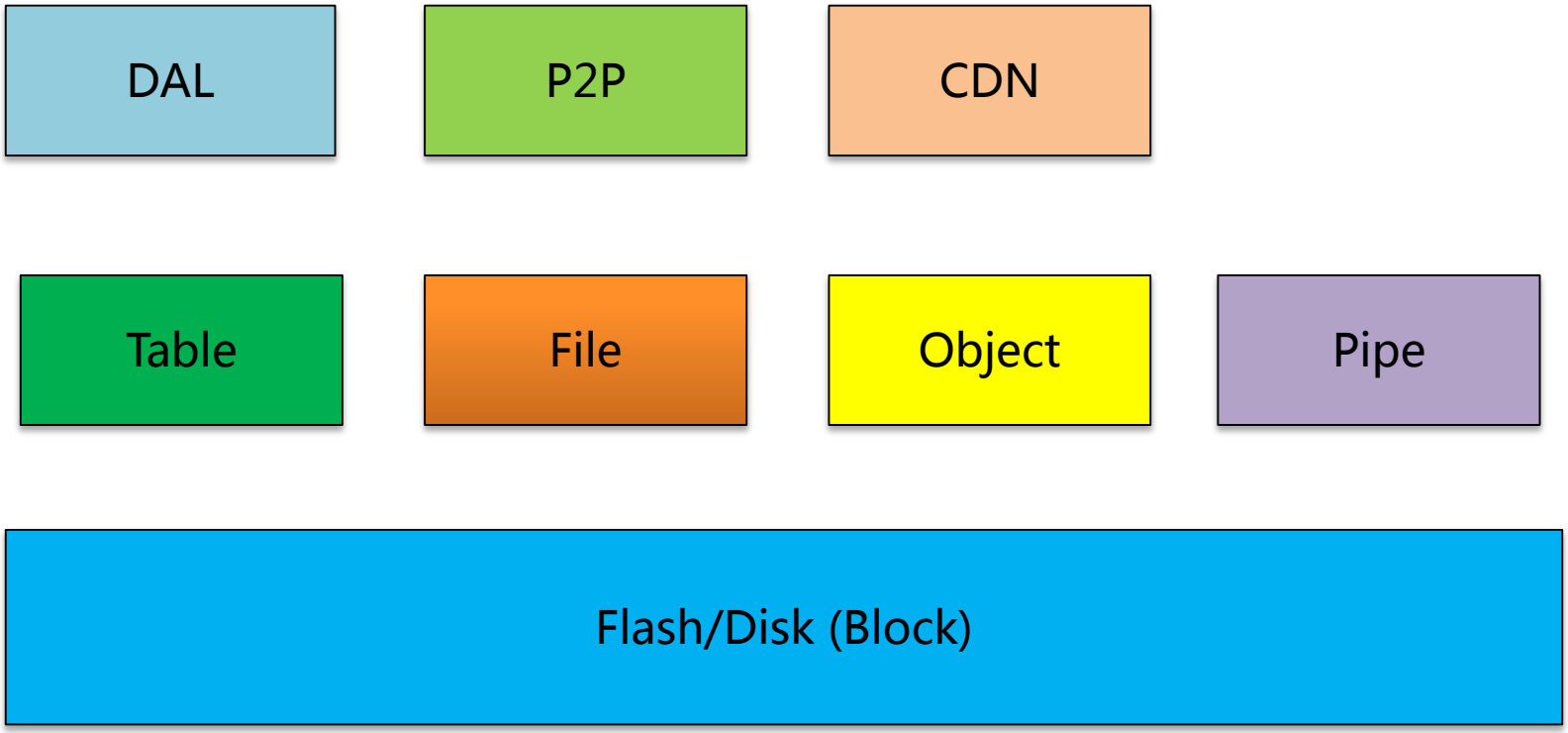
百度的存储系统——从前

- Bailing（网页库）
  - 海量、高吞吐
- Mola（Key-Value存储）
  - 低延迟、高并发
- Peta（HDFS2）
  - 无序、大数据
- DDBS（分布式数据库）
  - 复杂关系、强一致

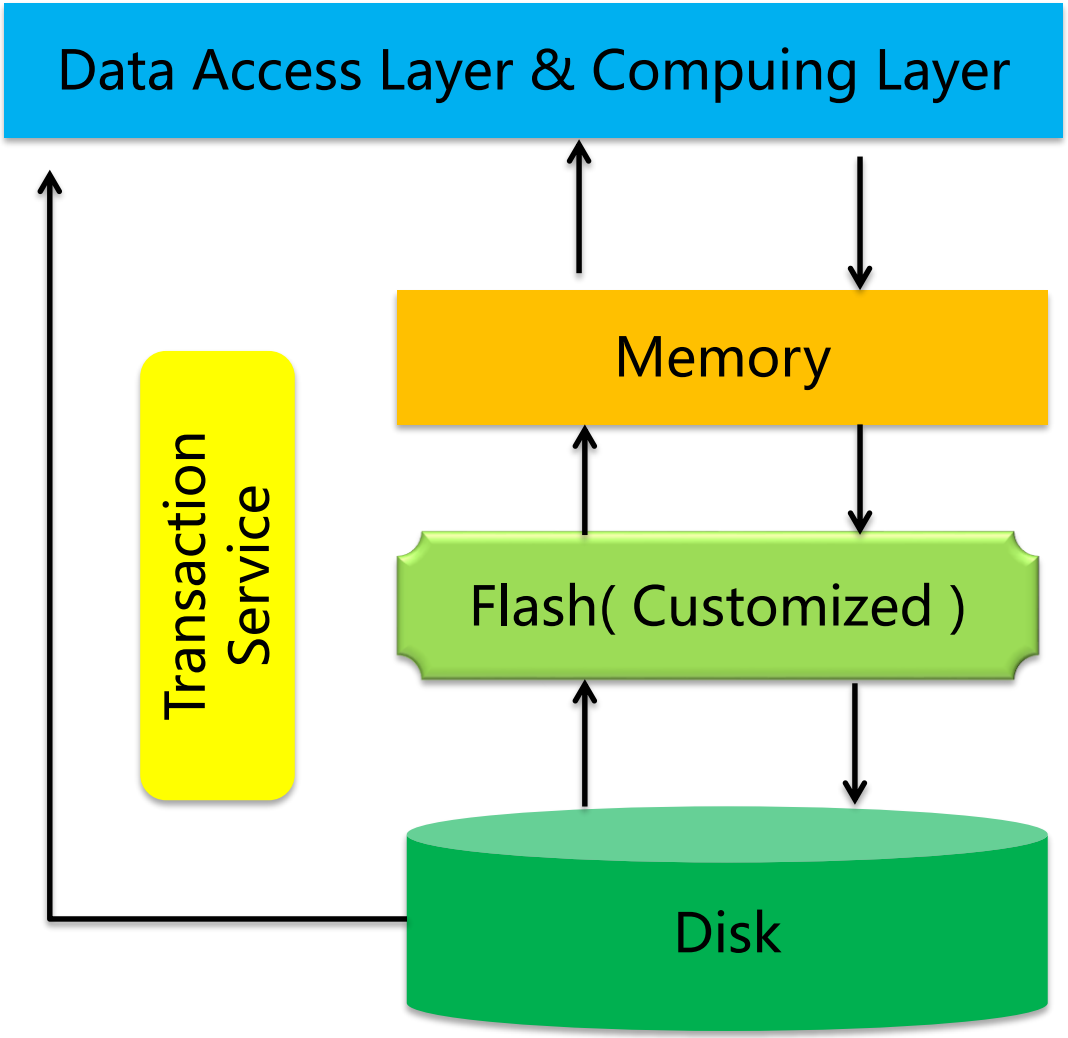
面临的挑战

- One Baidu One Storage
- 海量与实时
- 高吞吐与低延迟、高并发
- 一致性
- 可扩展性
- 可用性与可靠性
- 新硬件（SSD，ARM，etc）
- 平台化（服务化）

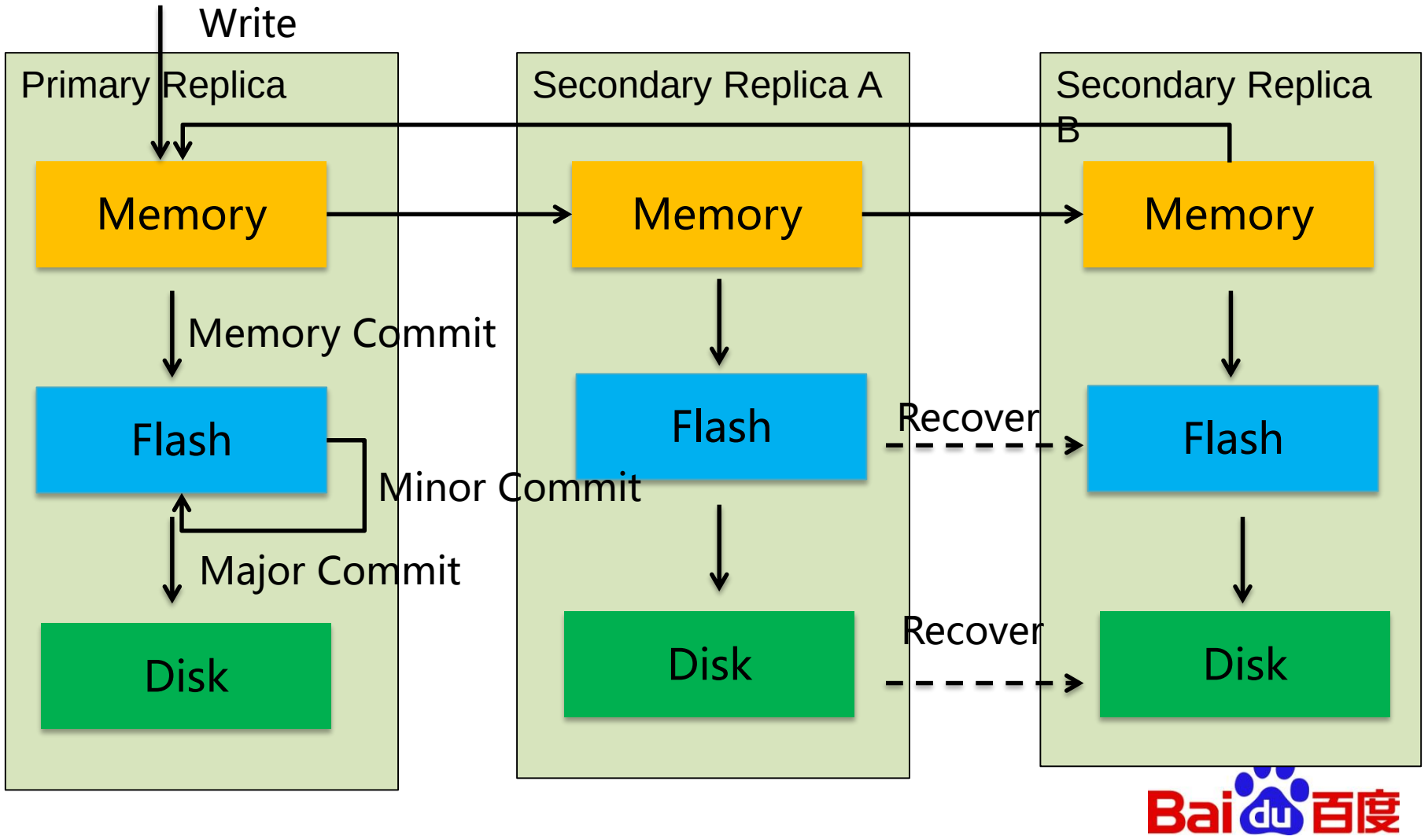
新存储体系



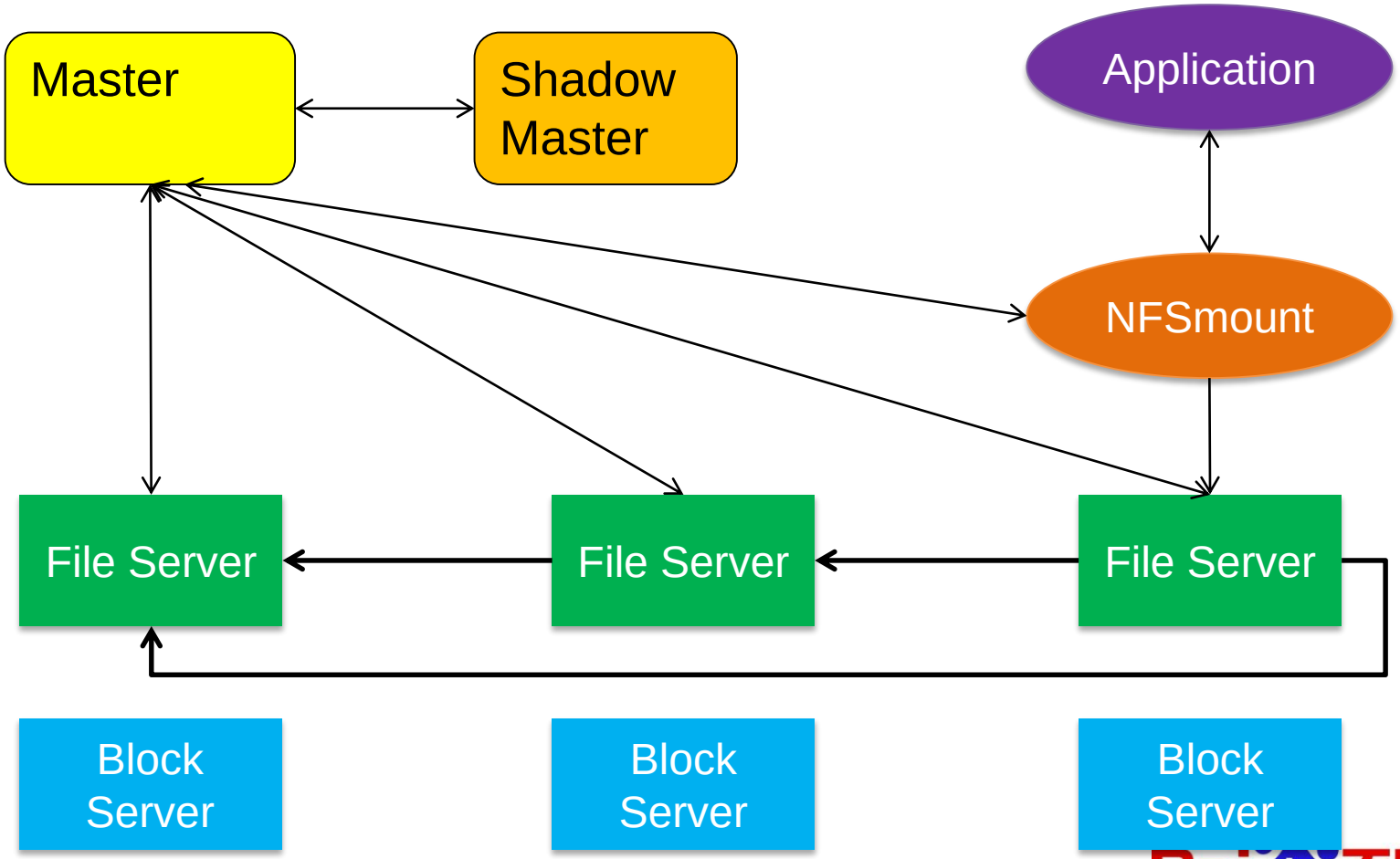
Table——Architecture



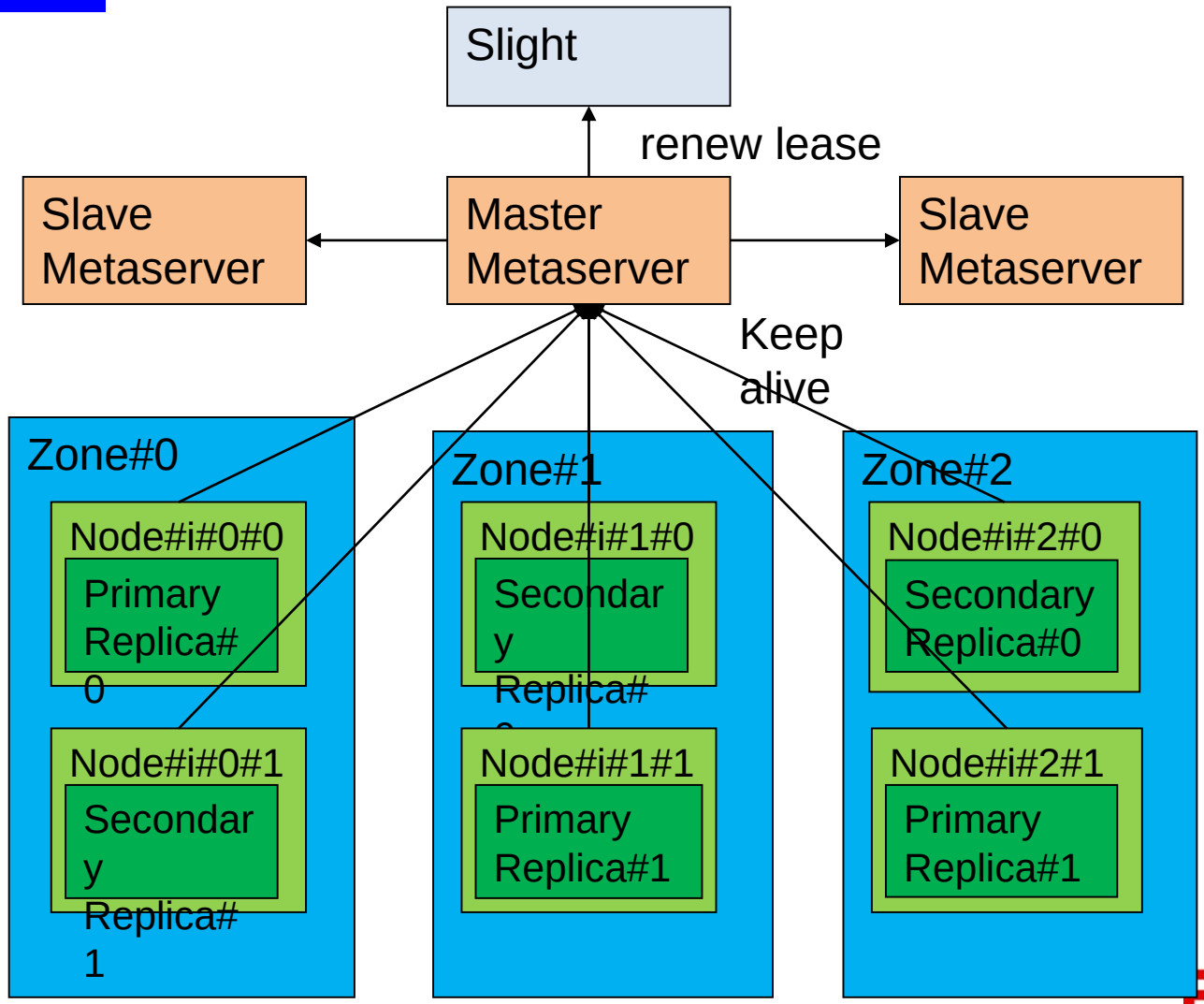
Table——Data Flow



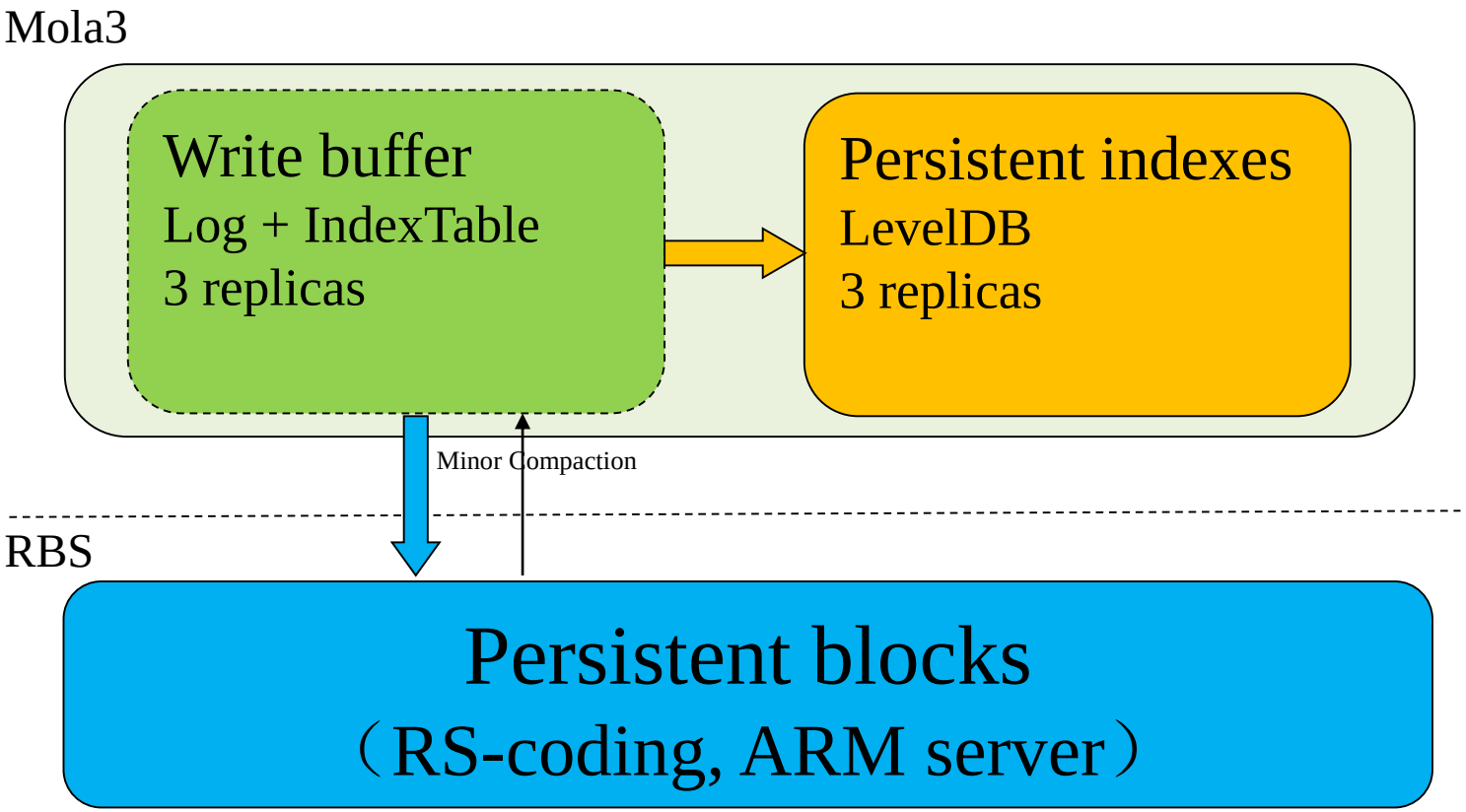
File——Architecture



Mola3——Architecture



Object——Architecture



经验教训

- KISS(Keep It Simple and Stupid)
- You built it you manage it
- Automate everything
- Layered design vs Vertical design
- 考虑到3~5年而不是10年或更久
- 平台化/服务化

Q&A

Thanks！ Questions?