QCon 全球软件开发大会
INTERNATIONAL SOFTWARE DEVELOPMENT CONFERENCE

BEIJING 2017

OpenStack Swift海量小文件优化之路

爱奇艺 技术产品中心 李杰辉

主办方 Geekbang. 极客邦科技 InfoQ

关于我

- 盛大云/VMware/爱奇艺
- Cloud Foundry/vNAS/vBlob/Swift
- Cloud Computing/Distributed Storage/DevOps/Docker
- @blue-salt

Agenda

- 背景
- Swift海量小文件问题以及应对之法
- Swift合并存储设计和实现
- 性能数据对比
- 展望

爱奇艺Swift实践

- 爱奇艺12年开始调研使用Swift
- 视频、图片、文本
- 海量小文件需求

分布式存储系统3个基本问题

- 副本定位
- 数据持久化
- 副本一致性

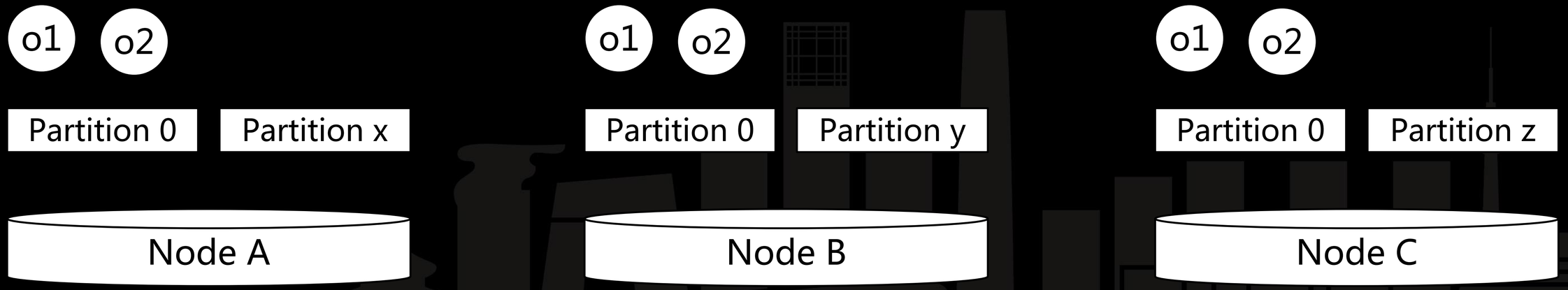
Swift介绍

- 高可用容错分布式对象存储系统
- 本质是一个键值系统

```
curl -T unix.png -X PUT http://172.28.128.82:8080/v1/tenant000000/container0001/unix.png
```

Swift副本定位

- 类一致性哈希
- 虚拟节点
- 副本以虚拟节点为单位



Swift数据持久化

- Replication
  - 一个对象保存成一个POSIX文件
  - 元数据保存在POSIX文件的扩展属性上

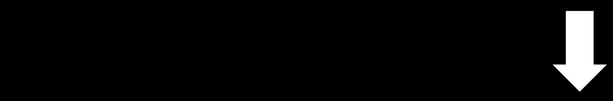
```
/srv/node/sdb/objects/3/63c/3e19cafe6fc6d71c6ee3fe814ef4d63c/1491478264.82777.data

/srv/node/sdb/objects/3/63c/3e19cafe6fc6d71c6ee3fe814ef4d63c/1491478264.82777.data
```

- Erasure Coding

Swift副本一致性

- 最终一致性
- 哈希列表

```
/srv/node/sdb/objects/3/63c/3e19cafe6fc6d71c6ee3fe814ef4d63c/1491478264.82777.data
```



63c : 12c76fd39c8a409049d30ec30d1c2c21

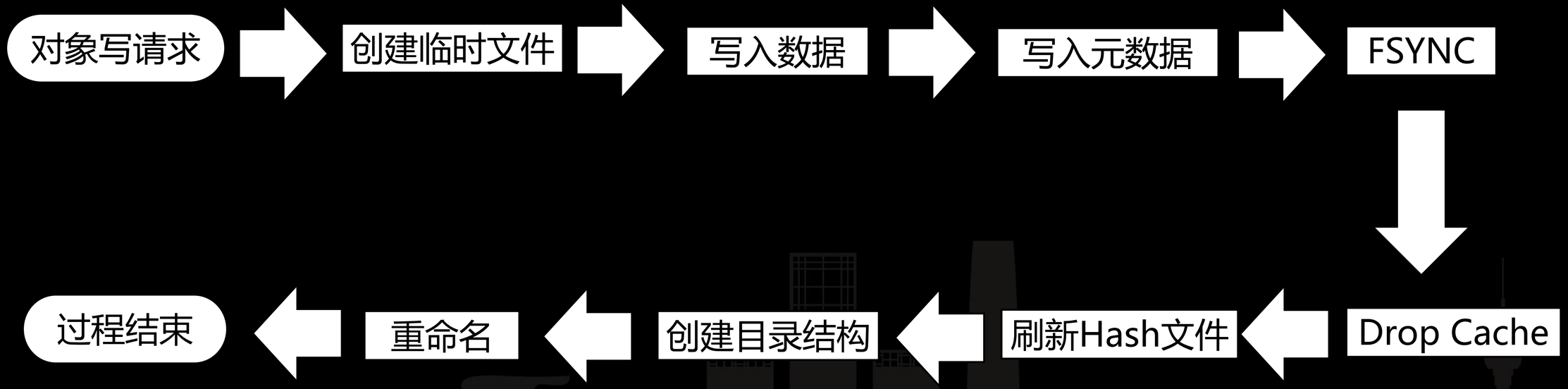
Agenda

- 背景
- Swift海量小文件问题以及应对之法
- Swift合并存储设计和实现
- 性能数据对比
- 展望

性能问题定位

- 阅读源码
- 原因猜测
- 代码性能分解

Replication 引擎写流程



Replication 存储引擎的问题

- 一个对象多层目录
- Inode占用过多
- Random IO操作
- 同步操作
- 文件锁

应对方法

- 加机器
- PyPy
- Go Swift
- 合并存储

合并存储介绍

- 源于Log Structured File System
- 核心思想：将Random IO变为Sequential IO
- FastDFS， Haystack，SeaweedFS，Ambry， TFS，BFS等

合并存储架构

- 副本定位
  - 中心服务器 + 文件Handle
- 副本一致性
  - 异步同步
- 数据持久化
  - 小文件合并存储到大文件
  - 索引文件 + 文件Handle

Agenda

- 背景
- Swift 海量小文件问题及应对方法
- Swift合并存储设计和实现
- 性能数据对比
- 展望

Swift合并存储的难点

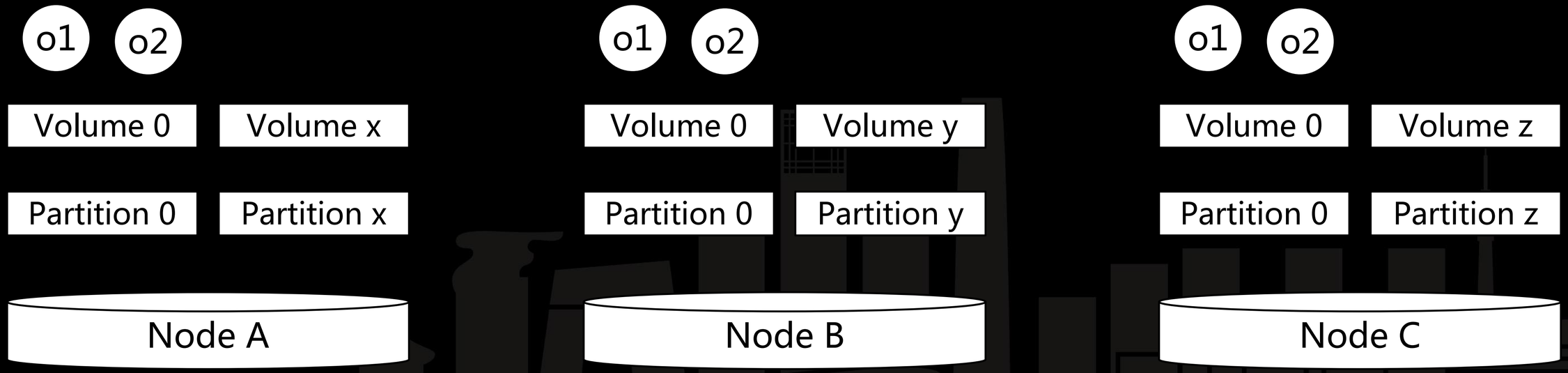
- Python WSGI 并发模式
- 去中心化副本定位
- 自定义元数据
- 基于目录哈希列表同步

Swift合并存储设计

- 副本定位
- 数据持久化
- 副本同步

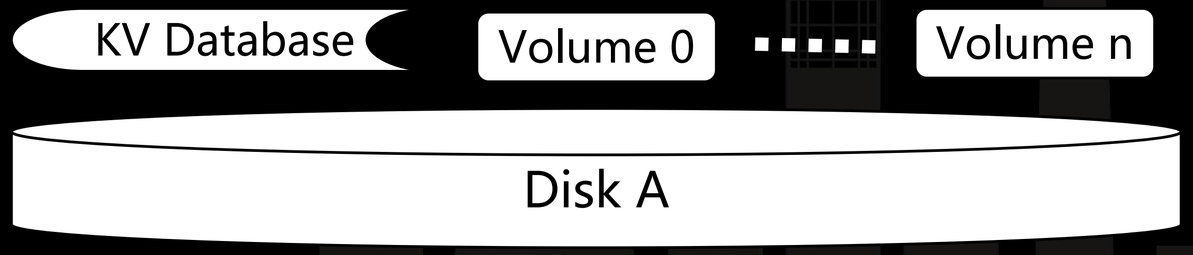
Swift合并存储副本定位

- 虚节点映射大文件
- 索引信息定位到偏移量



Swift合并存储数据持久化

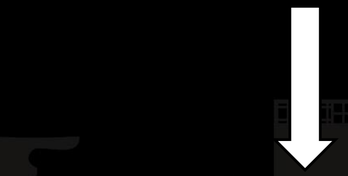
- Volume大文件
- 嵌入式键值数据库



Swift合并存储副本同步

- 基于哈希列表
- 模拟目录结构

/srv/node/sdb/objects/3/63c/3e19cafe6fc6d71c6ee3fe814ef4d63c/1491478264.82777.data



DB Key: /3/63c/3e19cafe6fc6d71c6ee3fe814ef4d63c/

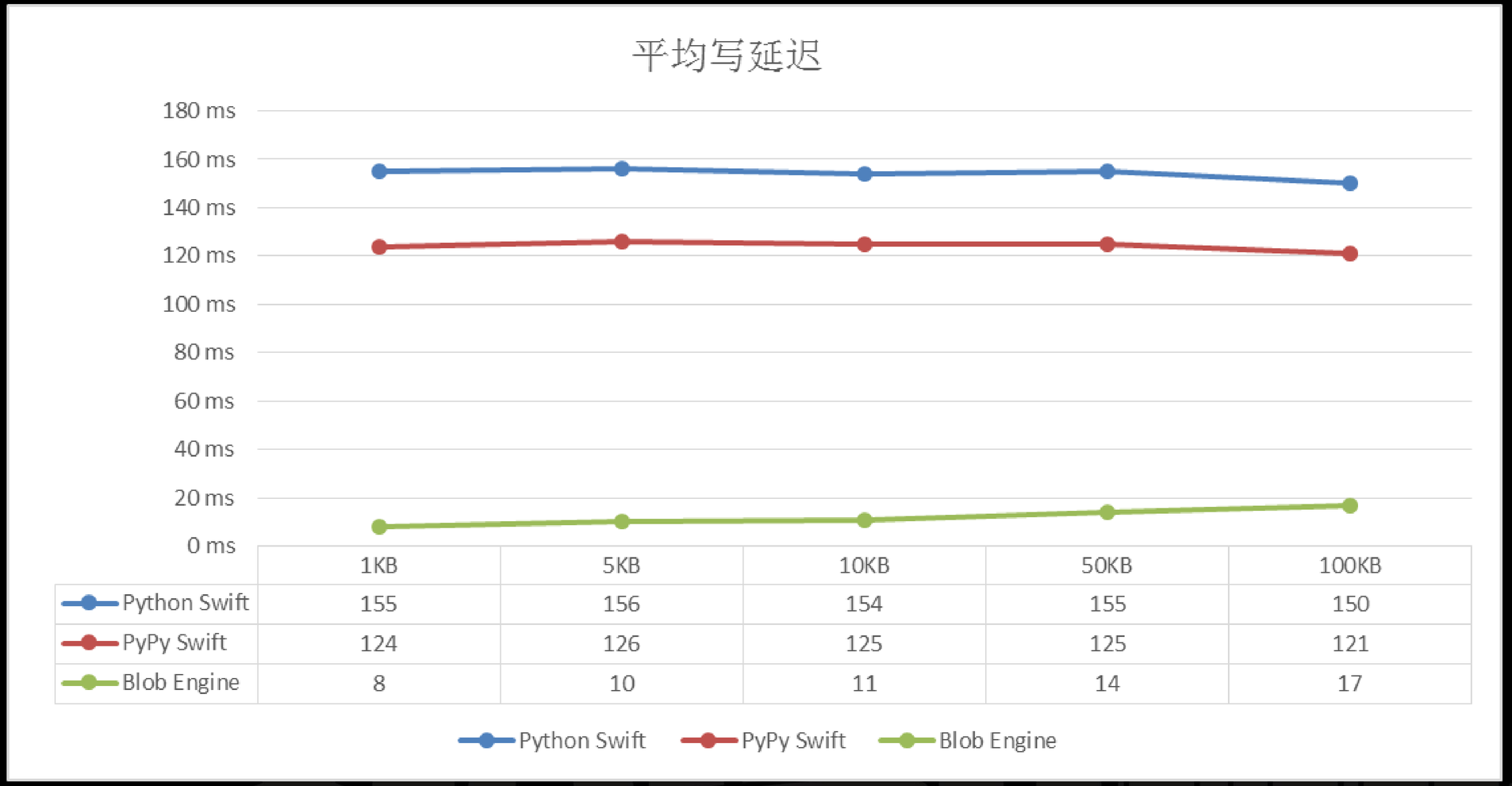
Swift合并存储实现

- Go语言为主
- RocksDB作为元数据库
- 复用Python Replicator和Auditor代码
- gRPC

Agenda

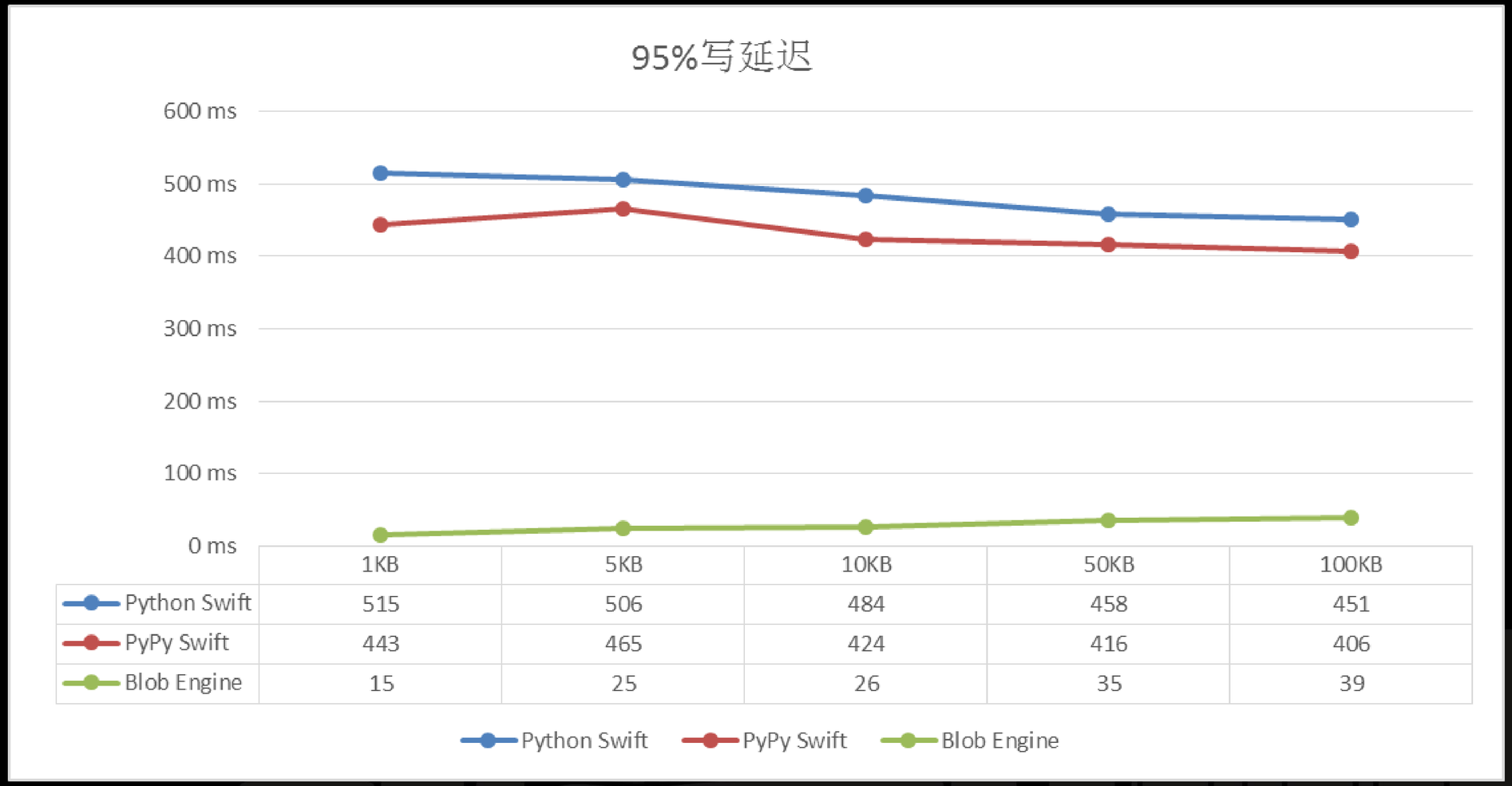
- 背景
- Swift 海量小文件问题及应对方法
- Swift合并存储设计和实现
- 性能数据对比
- 展望

写延迟



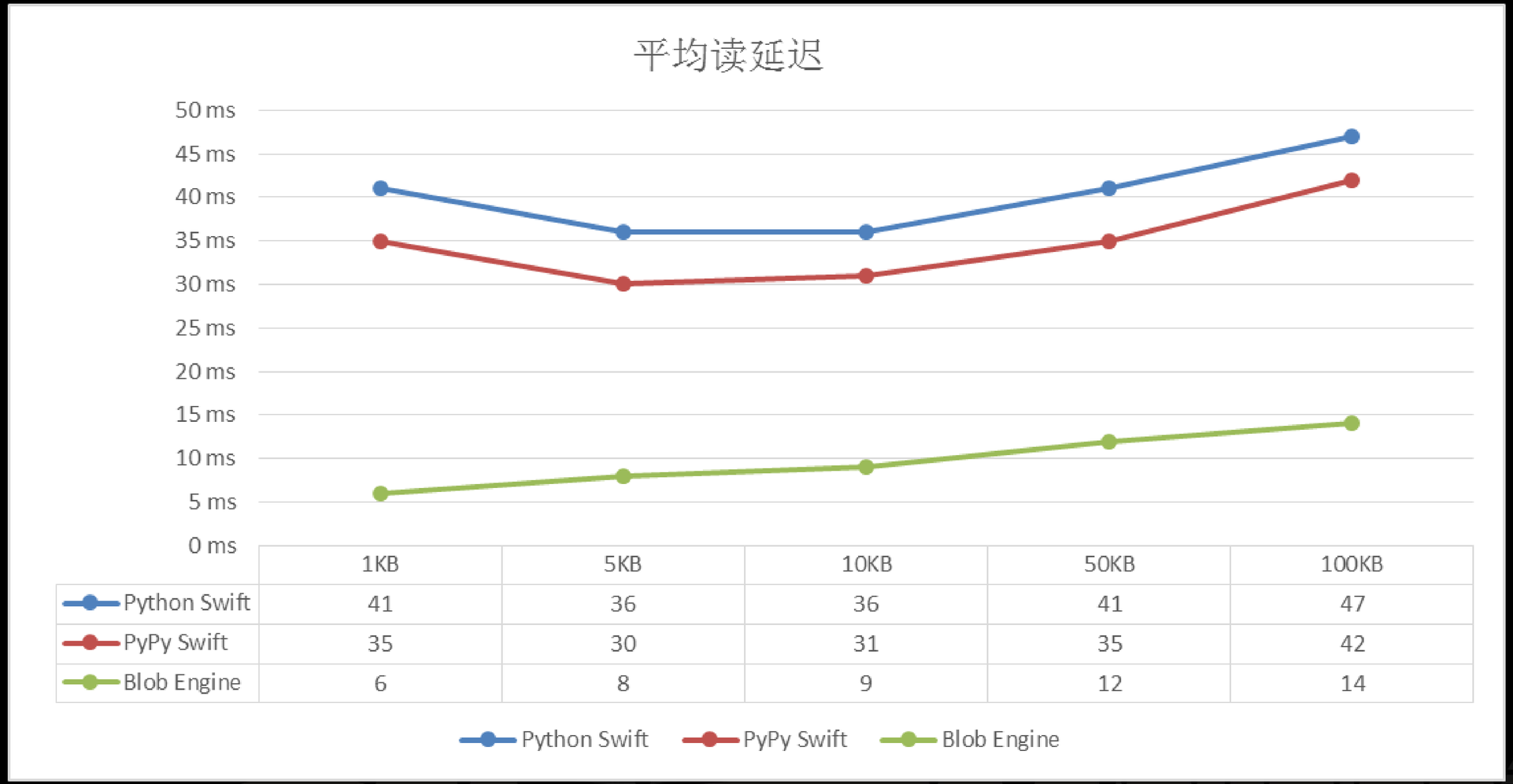
	1KB	5KB	10KB	50KB	100KB
Python Swift	155	156	154	155	150
PyPy Swift	124	126	125	125	121
Blob Engine	8	10	11	14	17

95%写延迟



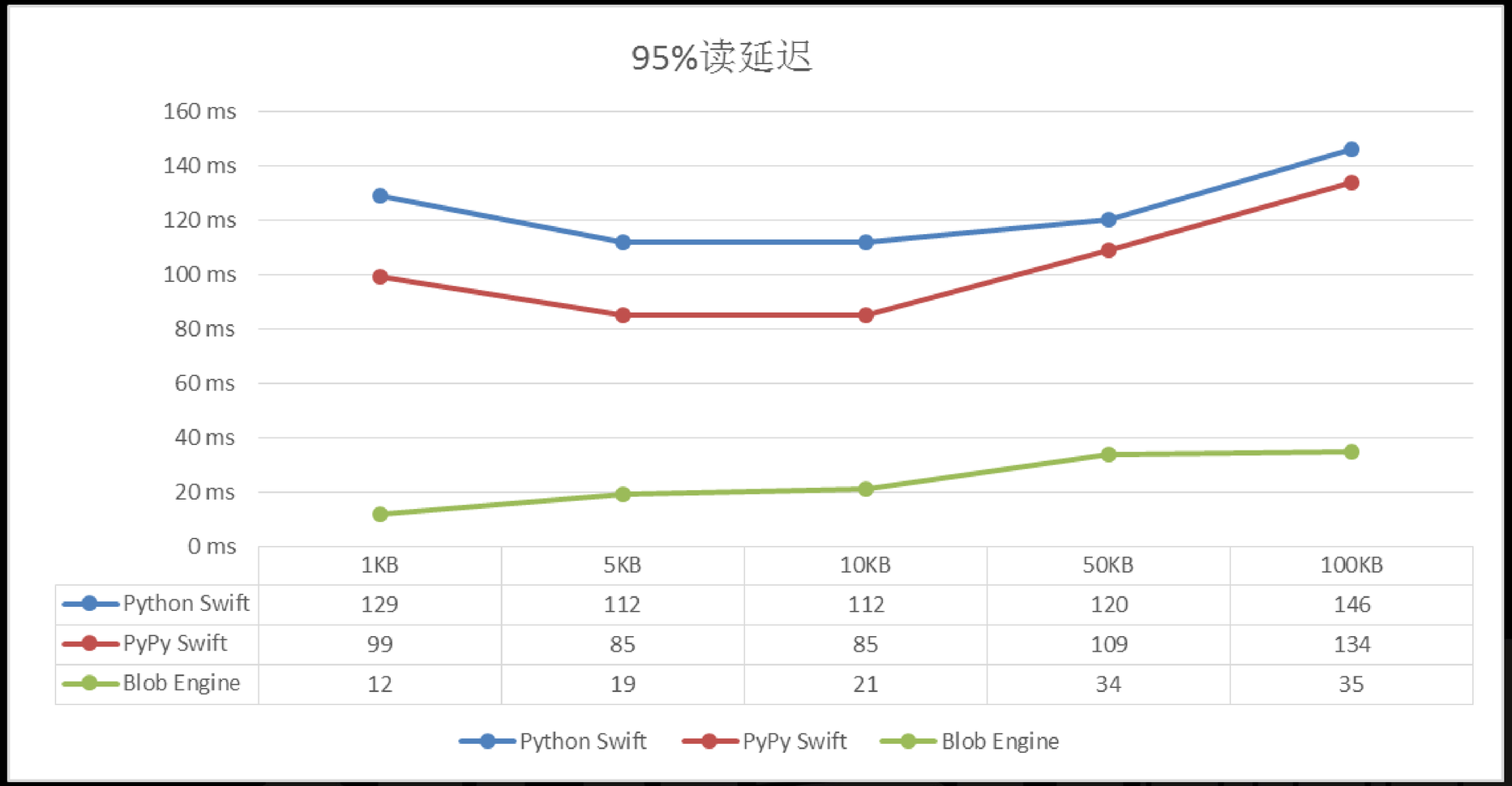
	1KB	5KB	10KB	50KB	100KB
Python Swift	515	506	484	458	451
PyPy Swift	443	465	424	416	406
Blob Engine	15	25	26	35	39

读延迟



	1KB	5KB	10KB	50KB	100KB
Python Swift	41	36	36	41	47
PyPy Swift	35	30	31	35	42
Blob Engine	6	8	9	12	14

95%读延迟



	1KB	5KB	10KB	50KB	100KB
Python Swift	129	112	112	120	146
PyPy Swift	99	85	85	109	134
Blob Engine	12	19	21	34	35

Agenda

- 背景
- Swift 海量小文件问题及应对方法
- Swift合并存储设计和实现
- 性能数据对比
- 展望

后续工作

- Full Go Stack
  - Swift 组件
  - 原生Go Key Value Database
- 大文件优化
- 性能监测支持

总结

- 分布式存储系统3个基本问题
- Swift 架构介绍以及Replication存储引擎剖析
- Swift合并存储设计与实现

关注QCon微信公众号，
获得更多干货！

Thanks!

INTERNATIONAL SOFTWARE DEVELOPMENT CONFERENCE

主办方 Geekbang. 极客邦科技 InfoQ