DTCC

第十四届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA

数智赋能 共筑未来

IT168.com ChinaUnix ITPUB

北京国际会议中心 | 2023/8/16-18

一个兼容 Redis 协议的大容量 KV 数据库 Pika

360云平台基础架构组负责人 赵新

个人简介

- Pika 项目负责人
- apache/dubbo-go 项目创始人
- 前蚂蚁集团 seata 开源负责人

- 2021 年阿里开源先锋人物、阿里开源大使
- 2022 开放原子开源基金会 年度开源贡献之星
- 2022 信通院 OSCAR 尖峰开源人物
- dubbo-go: 中国科学技术协会 2021年度优秀开源产品

目录

1 KV 数据库

开源 KV 数据库

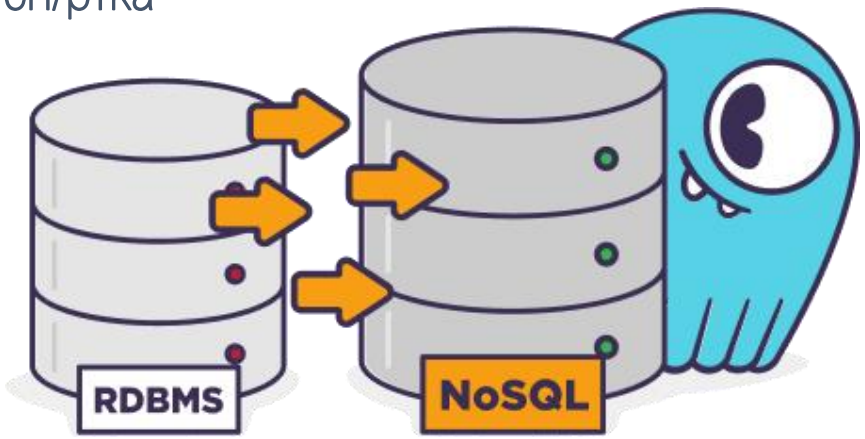
Database	Github	Stars(20230625)
Aerospike(DB)	https://github.com/aerospike/aerospike-server	867
Memcached	https://github.com/memcached/memcached	12.6k
Redis	https://github.com/redis/redis	60.4k
Riak KV	https://github.com/basho/riak_kv	626
Tair	https://github.com/alibaba/tair	2.1k
TiKV	https://github.com/tikv/tikv	13.3k
Voldemort	https://github.com/voldemort/voldemort	2.6k

兼容 Redis 协议的 KV 数据库

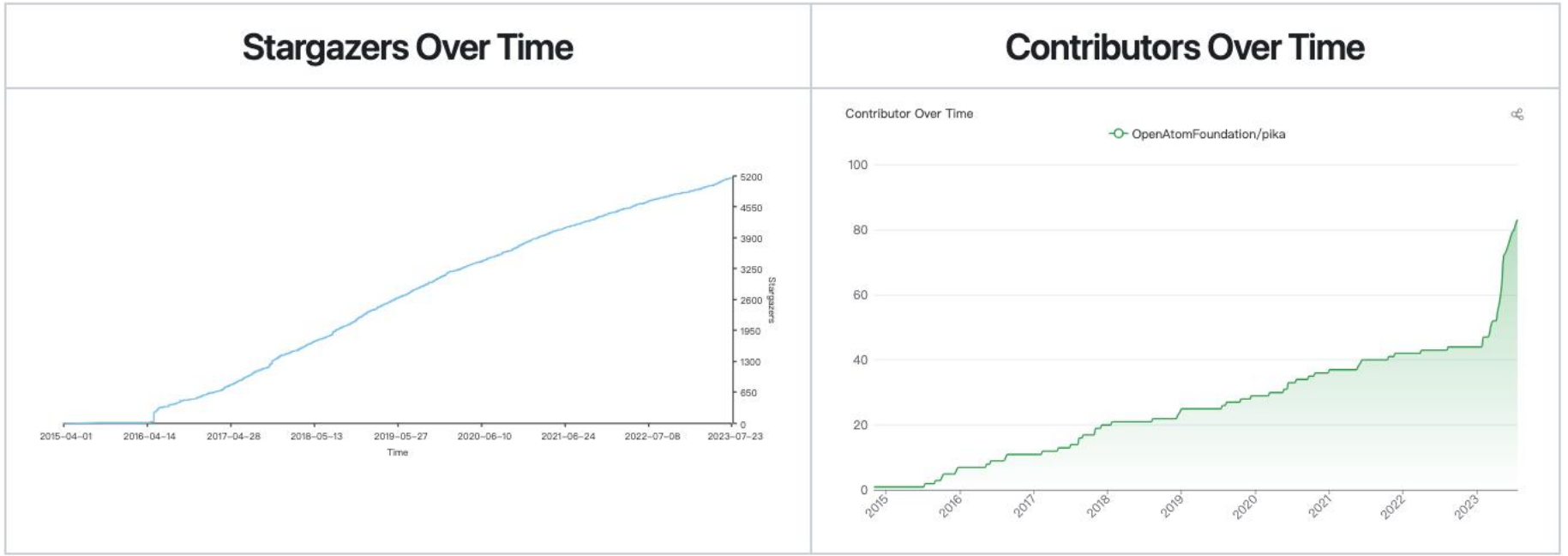
Database	Github	Stars(20230625)
SSDB	https://github.com/ideawu/ssdb	8.1k
Pika	https://github.com/openatomfoundation/pika	5.1k
Pegasus	https://github.com/apache/incubator-pegasus	1.9k
Kvrocks	https://github.com/apache/kvrocks	2.2k
Tedis	https://github.com/eleme/tedis	182
Tendis	https://github.com/Tencent/Tendis	2.6k
…		

关于 Pika

- 项目范围：NoSQL 数据库
- 创始团队：奇虎 360 云平台部基础架构组
- 开源时间：2016年2月
- 项目地址： https://github.com/OpenAtomFoundation/pika
- 许　可　证：BSD-3-Clause



发展历程



2015.04 项目启动 → 2015.11 发布 1.0 → 2016.02 开源 → 2016.04 发布 2.0 → 2018.08 发布 3.0 → 2020.08 申请加入 OpenAtom → 2021.03 孵化运营

解决的问题

Pika 的出现并不是为了替代 Redis，而是 Redis 的场景补充。Pika 力求在完全兼容Redis 协议、继承 Redis 便捷运维设计的前提下，通过持久化存储的方式解决 Redis在大容量场景下的问题，如：

- 单线程易阻塞
- 容量有限
- 加载数据慢
- 故障切换代价高

应用场景

key-string
高性能 KV
搜索推荐、机器学习

key-hash
复杂在线业务
用户信息、好友关系、对象存储元数据

key-list
简单高效的消息中间件
分布式任务系统

项目特点

- ✔ 在完全兼容 Redis 协议的前提下追求 极大容量 与 极致性能
- ✔ 支持 Redis 的常用数据结构 bitmap、string、hash、list、set、zset、geo、hyperloglog、pubsub
- ✔ 持久化存储到 RocksDB
- ✔ 单机主从、codis 集群两种方式部署
- ✔ 相比于 Redis 的内存存储方式，能极大减少服务器资源的占用，增强数据的可靠性

360公司内部部署使用规模 7000+ 实例，单实例数据量 1.8TB；

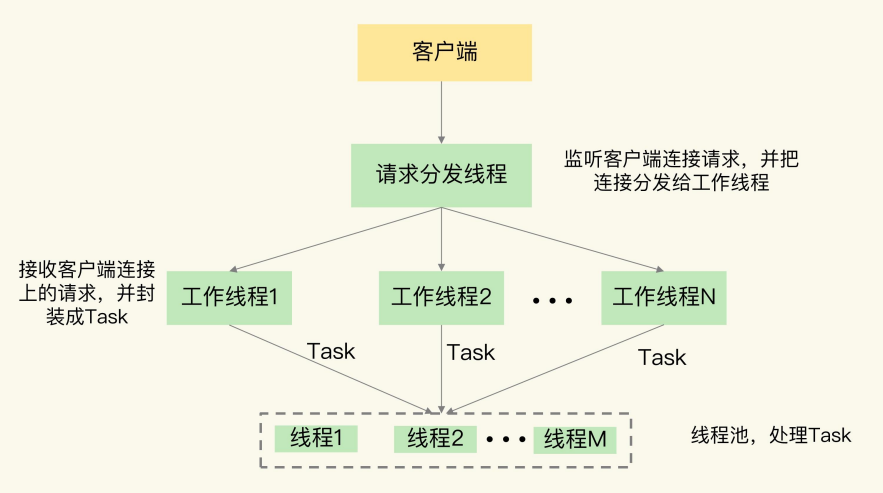
微博公司内部部署实例 10000+；

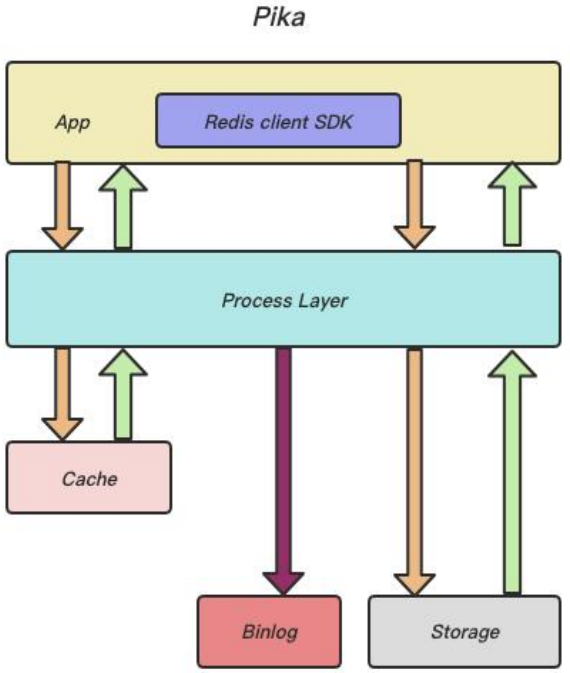
喜马拉雅(X Cache)实例数量 2500+，数据量 120TB+ ；

2 Pika 架构

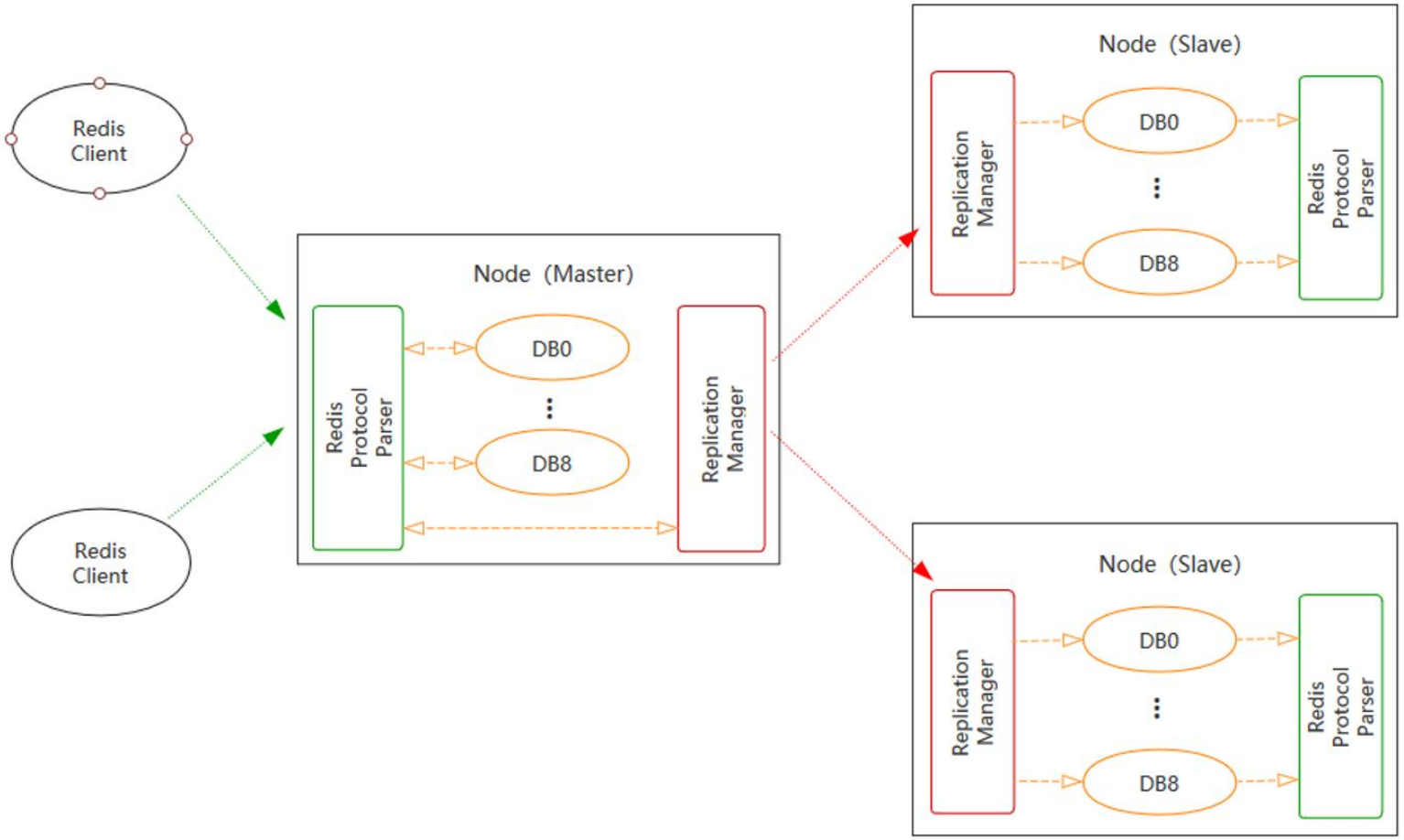
存储引擎

- 支持多平台 centos、ubuntu、macOS
- 多线程模型
- 基于 RocksDB 的存储引擎
- 多粒度数据缓存模型
- ……



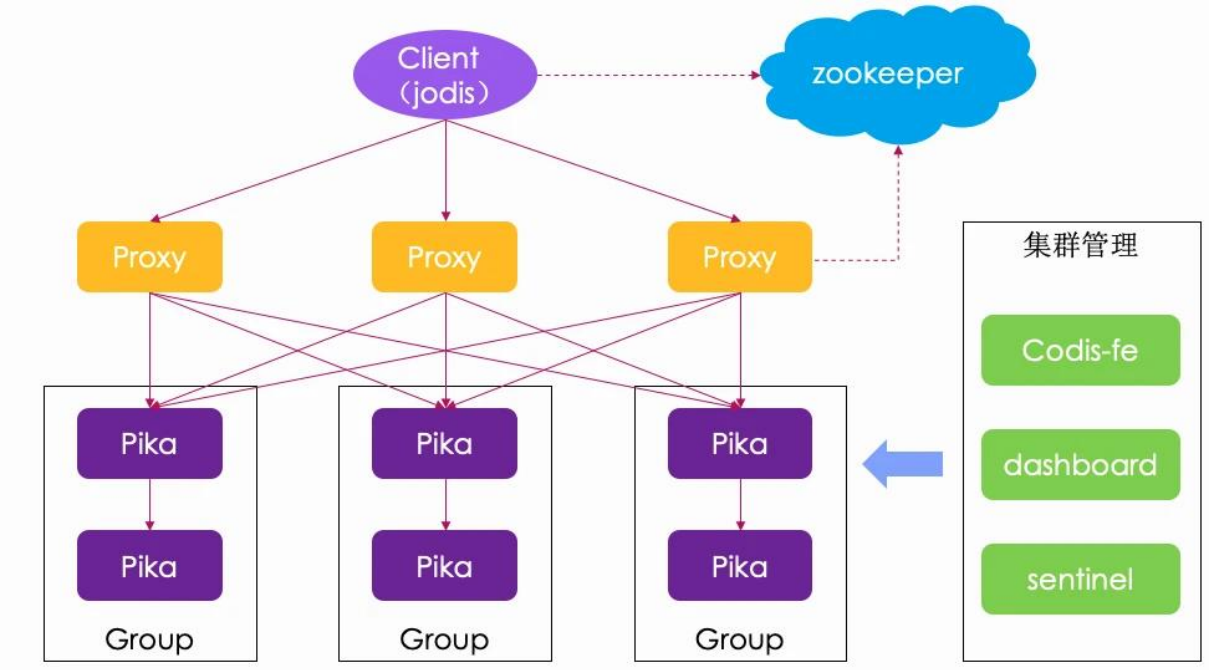


主从集群



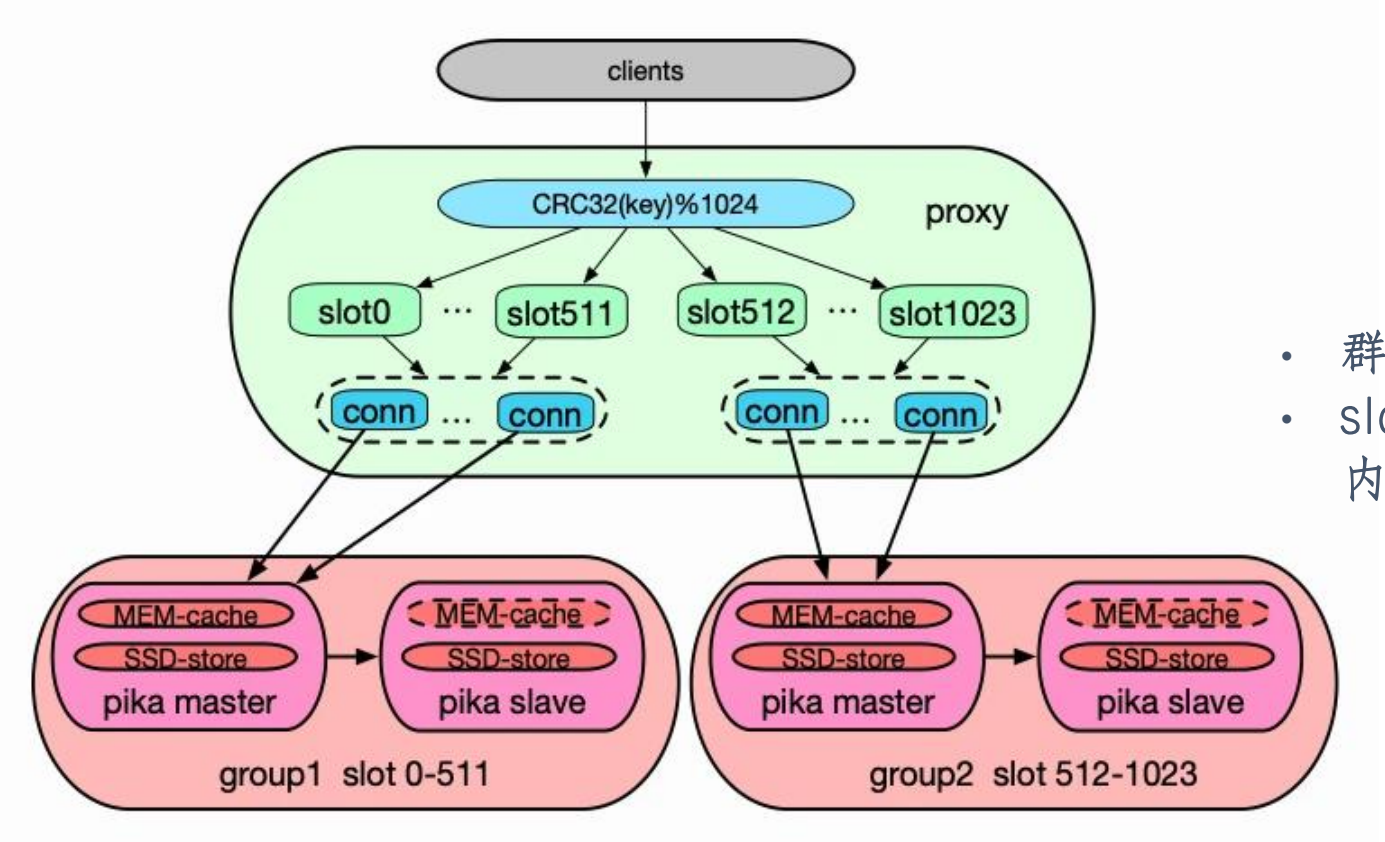
- 架构与 Redis 类似
- 与 Redis 协议和数据结构兼容性好
- 每种数据结构使用一个 RocksDB 实例
- 主从采用 Binlog 异步复制方式

群集架构(1)



- 采用 Codis 架构，支持多 group
- 单 group 内是一个主从集群
- 以 group 为单位进行弹性伸缩

群集架构(2)



- 群集内共 1024 个 slot
- slot 与 group 映射关系存于 zk/etcd 内

3.5 新特性

done

- 1 跨平台： Centos、Ubuntu、MacOS
- 2 RocksDB 能力升级： BlobDB KV 分离、LSM 分级压缩、Compaction 限速、动态更新 sst 文件 size
- 3 基于 C++17 重构 Pika，降低内存泄漏风险
- 4 统一经典模式和 Sharding 模式，完全兼容 Codis
- 5 去除 Rsync，全新的主从全量数据同步机制

doing

- 6 Lua 引擎
- 7 可观测性增强：支持采集 Pika 网络请求、主从同步、RocksDB 等更多 Metrics
- 8 测试集支持了 Redis TCL 单测、E2E测试（主从数据同步、主从切换、故障自愈）
- 9 Pika Operator 支持主从实例、Codis 集群部署
- 10 事务与锁

当务之急：CPU load

在磁盘没瓶颈的情况下，实例数提升总是很容易到达瓶颈

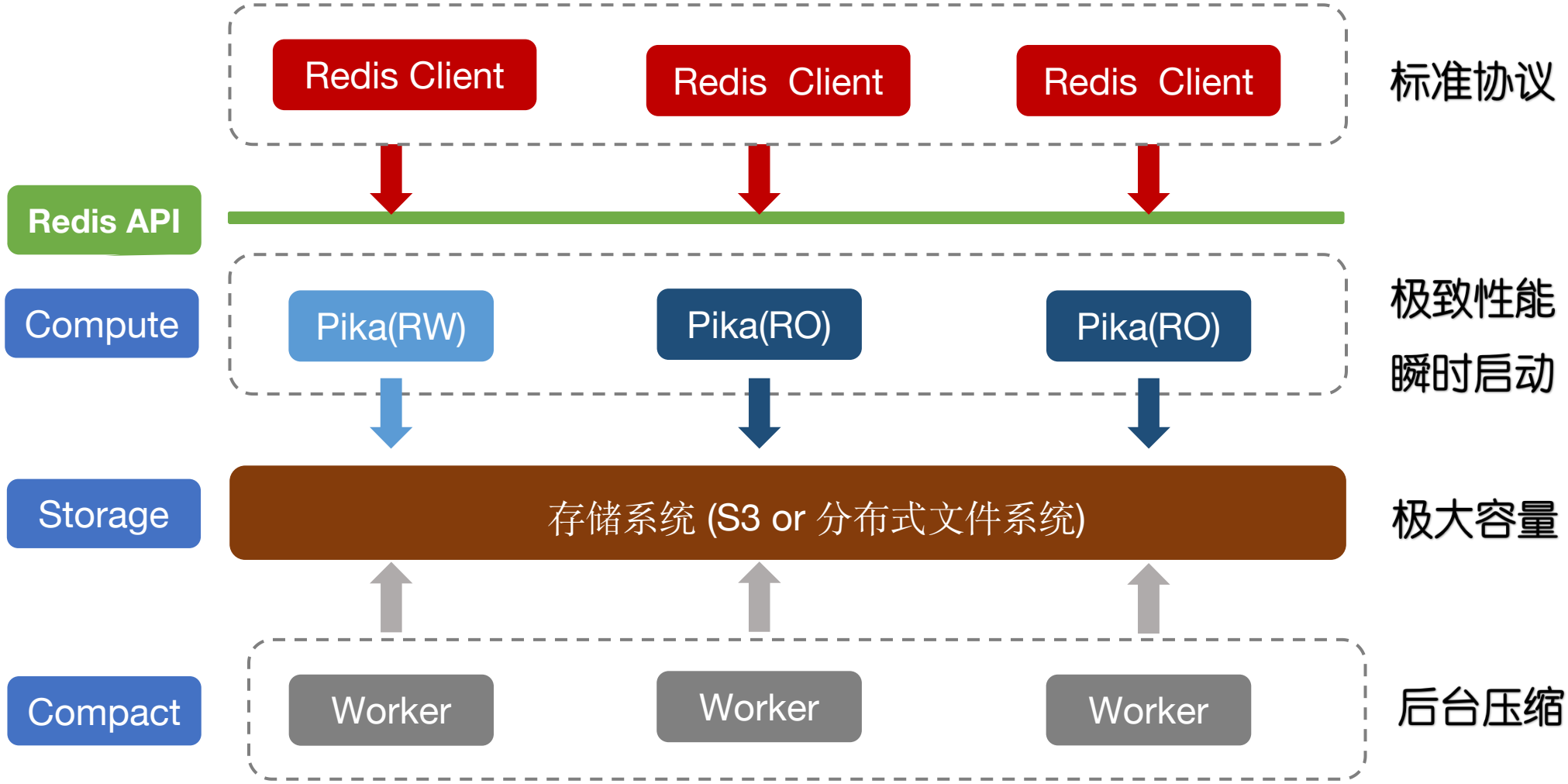
- 24 核常规服务器，2 个和 4 个实例的总 QPS 几乎相等
- 96 核超级服务器，7 个实例以后再增加实例，提升也很小

极致性能

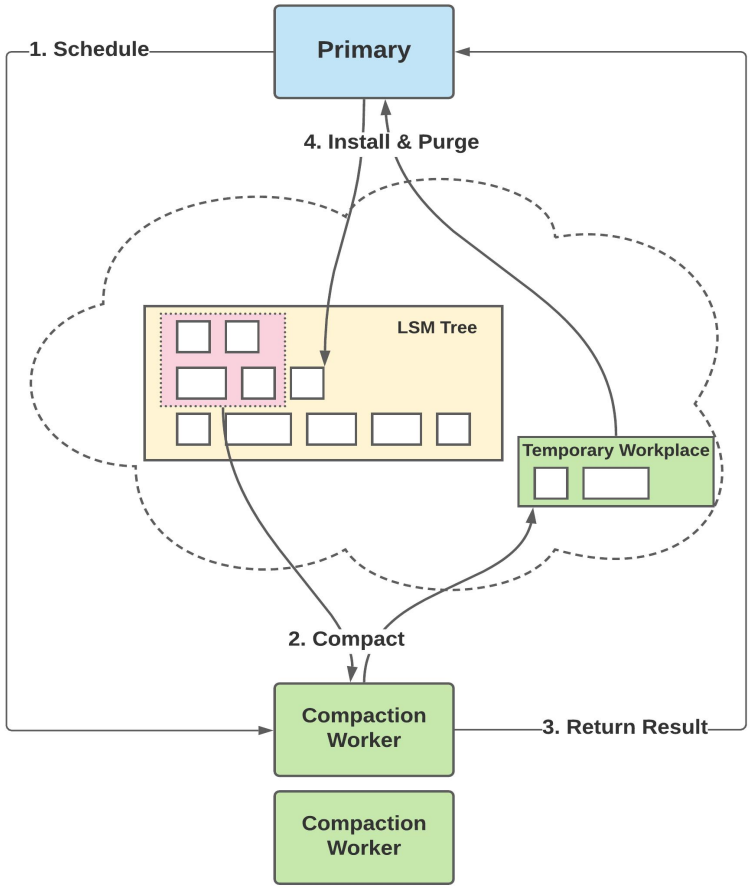
- 1 对 Pika 的 dispatcher-worker 架构进行升级，平衡 CPU 负担
- 2 基于 RocksDB io_uring 最新特性，对 Pika 进行协程化改造
- 3 使用 DPU 技术，减轻 CPU load
- 4 借助 SPDK/XDP 等较新的 IO 硬件技术，落地零拷贝技术，减轻 Pika 网络负担

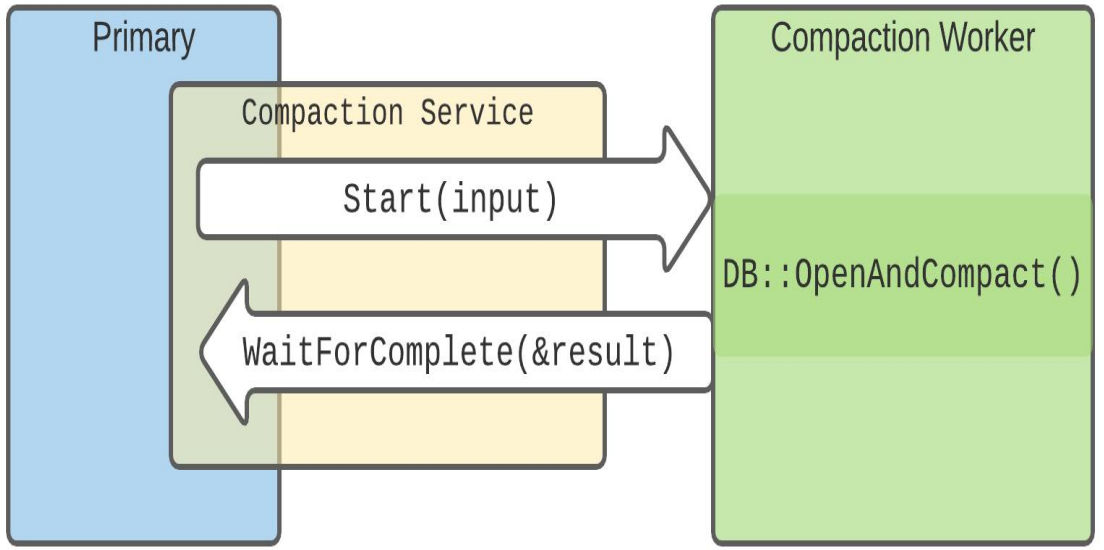
3 Pika Cloud

Pika Serverless



RocksDB Remote Compaction



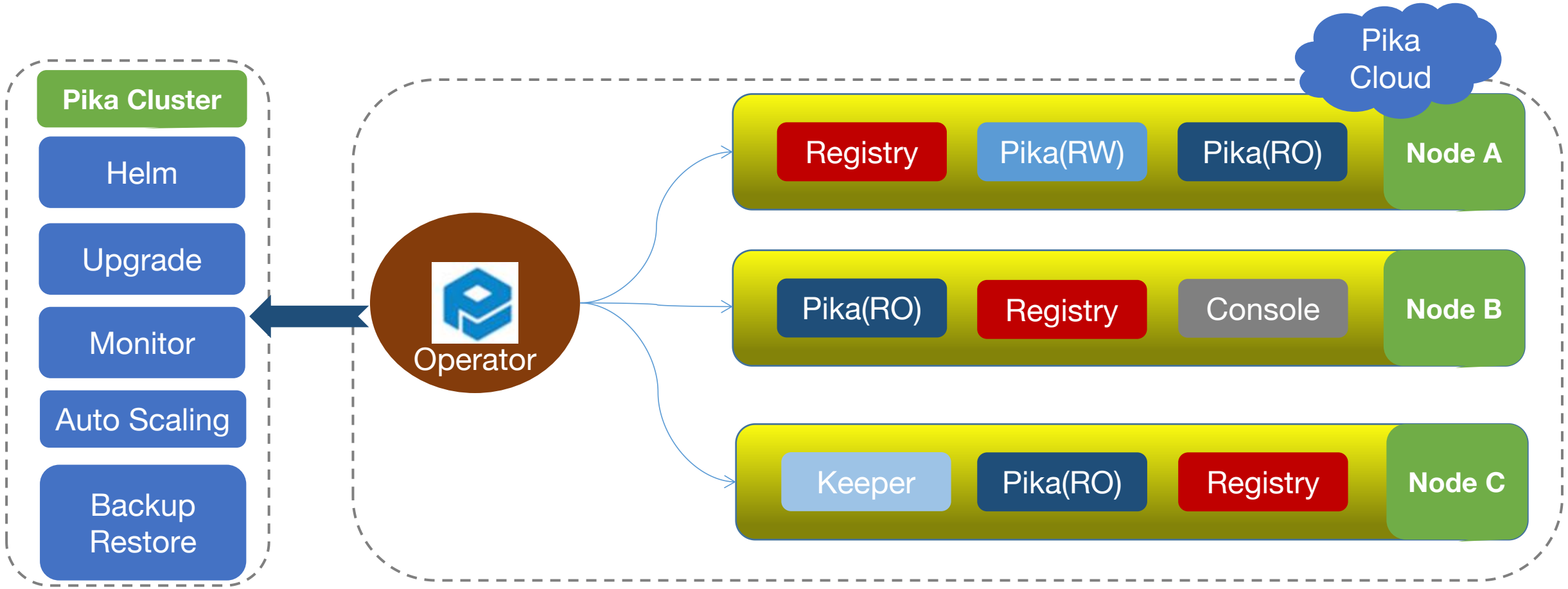


- **1 Primary** 调用 **CompactionService** 后，在 **WaitForComplate(&result)** 结果上同步等待
- **2 Worker** 实现 **DB::OpenAndCompact()**，以 **ReadOnly** 方式读取 **LSM Tree**，然后把压缩结果存入 **Temporary Workplace**
- **3 Worker** 把压缩后的 **SST** 的 **metadata** 等信息做为 **RPC Result** 返回给 **Primary**

Remote Compaction 收益

- **1 Pika(Compute)** 执行 **R/W** 任务时非常轻量，**P99** 延时平稳，几乎无毛刺
- **2 LSM Tree** 相对于 **Worker** 只读，对于 **Pika(Compute)** 是 **Append Only**，**unmutable**
- **3 Pika(Compute)** 瞬时启动，**Worker** 节点可弹性伸缩
- **4 Compaction** 任务分离后，**Storage** 层可使用廉价硬件实现 **Share-storage**，还可以通过 **EC** 方式实现极大容量

Pika Operator



Operator Workers

done

- **1 Pika Cluster** 各个组件已经可以在 **K8s** 成功部署
- **2** 支持扩容和缩容单个 **Pika** 节点

doing

- **3** 单批次启动一个 **Group**
- **4** 部署可观测性组件
- **5** 存储通过创建 **PV** 形式支持使用持久存储卷
- **6** 自动化扩缩容

4 未来展望

Pika Cloud

- 1 多租户（资源管控、数据隔离）
- 2 存算分离
- 3 弹性伸缩（扩缩容）
- 4 故障自愈（Operator、Sentinel）
- 5 异地多活（跨机房部署）

详细规划

强内核

- 向Redis接口靠拢
- 单机性能提升
- 多租户
- 更多平台支持
- 支持 CP

高质量

- 单测
- 集成测试
- 混沌压测
- ChatGPT Code Review

工具集

- 基于 Prometheus 的 exporter 接口
- Redis 与 Pika 迁移

大社区

- 开源大赛
- 和 OpenAtom 协作，提升项目知名度

云原生

- 存算分离：S3
- 弹性： Operator
- 故障自愈
- 多租户
- 资源隔离

最终形态


分布式
高性能流量转发
多租户隔离
负载均衡
资源管控
数据分片
集群监控
云原生
故障自愈
弹性伸缩
云化存储
Pika 核心功能
存算分离
Raft 共识算法
兼容 Redis 协议
无侵入观测
事务
共享存储
大容量存储
一主多从
Data Stream
ACL
持久化 KV
硬件加速
生态工具
数据迁移工具
第三方扩展
Lua 脚本
质量工程
混沌测试
集成测试
单元测试

THANKS