KEYSPACE

2025 / 北京

基于Valkey构建超大规模的集群服务

下一代集群管理架构展望

郑晓茵
字节跳动研发

Valkey

01.

为什么需要超大规模的Valkey集群

02.

Valkey集群架构的现状

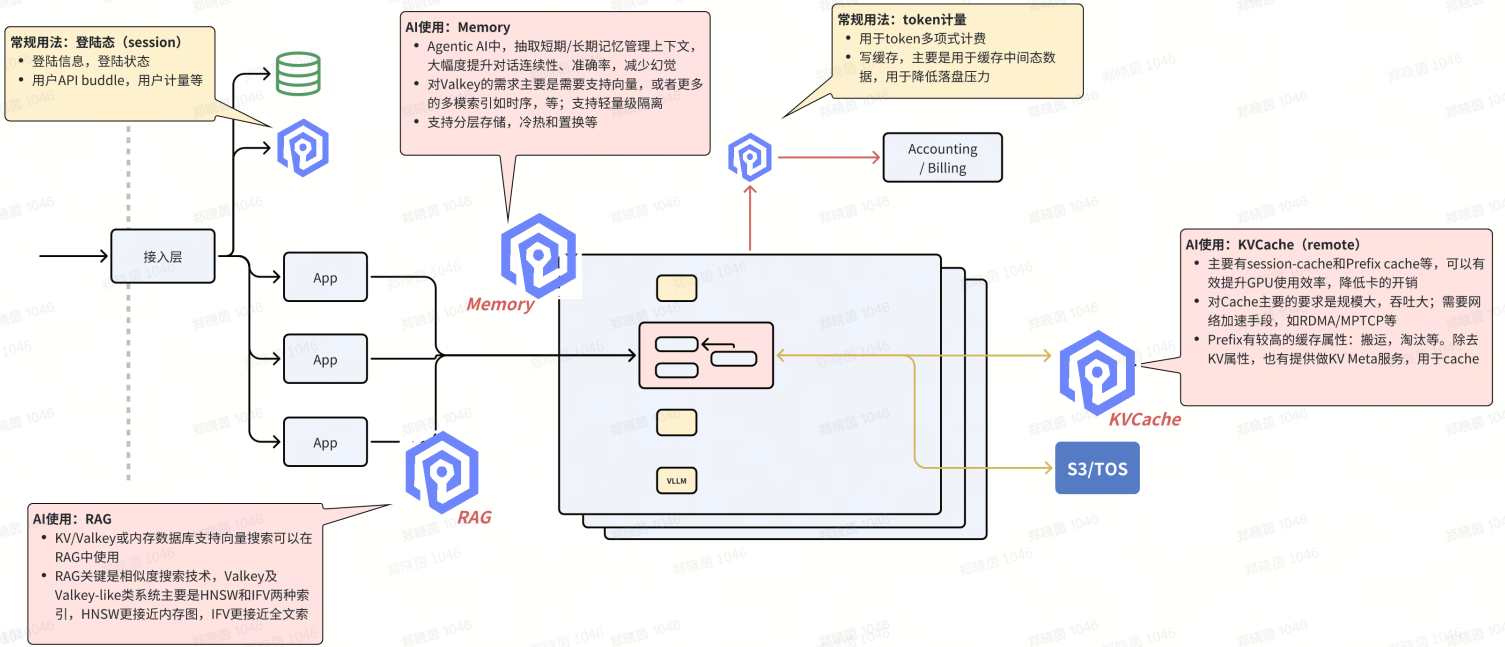
03.

下一代Valkey集群管理架构

04.

展望

为什么需要超大规模的Valkey集群



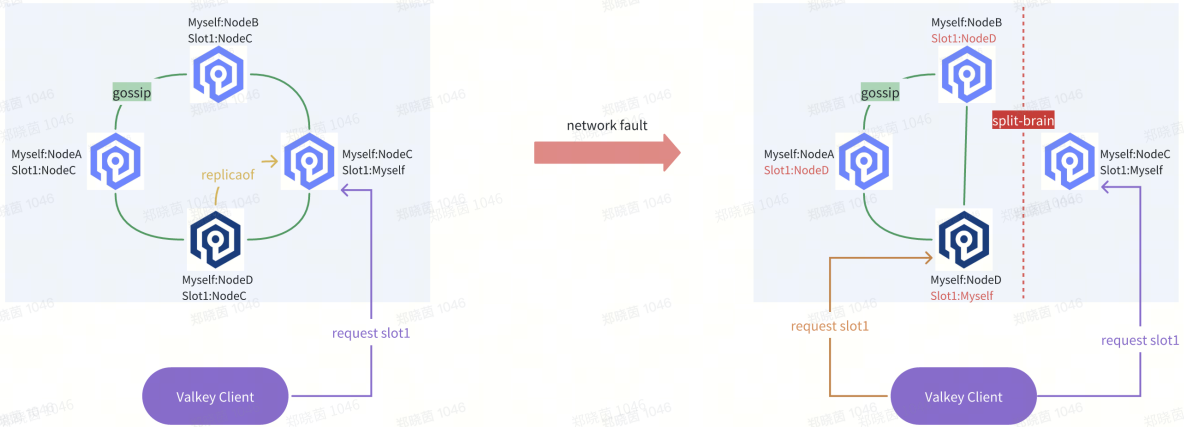
为什么需要超大规模的Valkey集群

规模与性能示例（字节内部）

4096	256T	400GB/s
Shards	Memory	Throughput
分片规模	总内存容量	集群总带宽

Valkey集群架构的现状

- gossip协议通信指数增长
- 分布式协商故障场景收敛时间长
- 网络脑裂拓扑分布不一致



下一代Valkey集群管理架构

社区对Valkey集群演进的讨论

https://github.com/valkey-io/valkey/issues/384

内嵌控制节点：
从Valkey Cluster中选举控制节点

维持sentinel：
组建外部控制组件

下一代Valkey集群管理架构

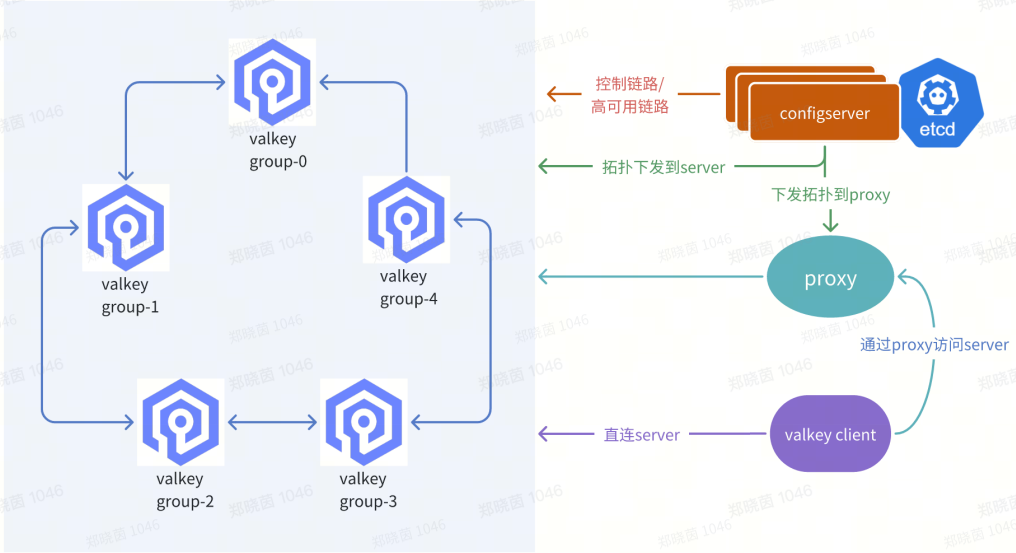
configserver中控集群管理架构

优势

完全命令式，无gossip协议无协商式
轻量部署，控制模块和数据模块拆分
灵活托管实例，可独享，可共享实例池

不足

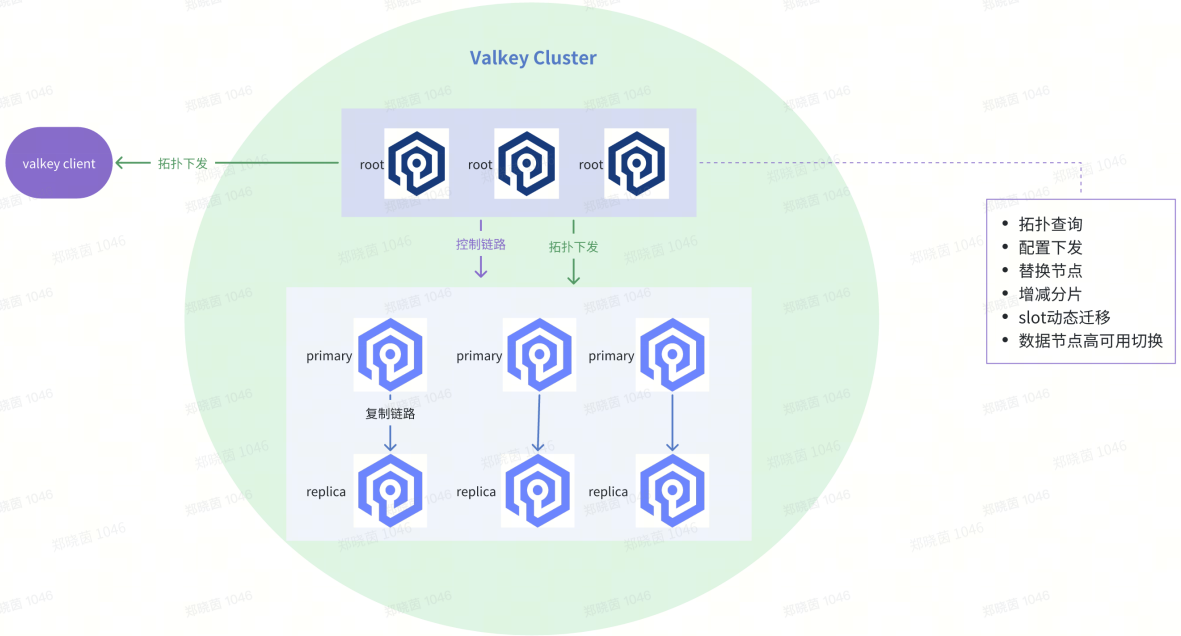
网络脑裂拓扑分布不一致
拓扑局部更新客户端全量拉取负载大
构建部署configserver架构维护成本高
valkey内核改造和维护成本高



下一代Valkey集群管理架构

raft一体化中控集群管理架构

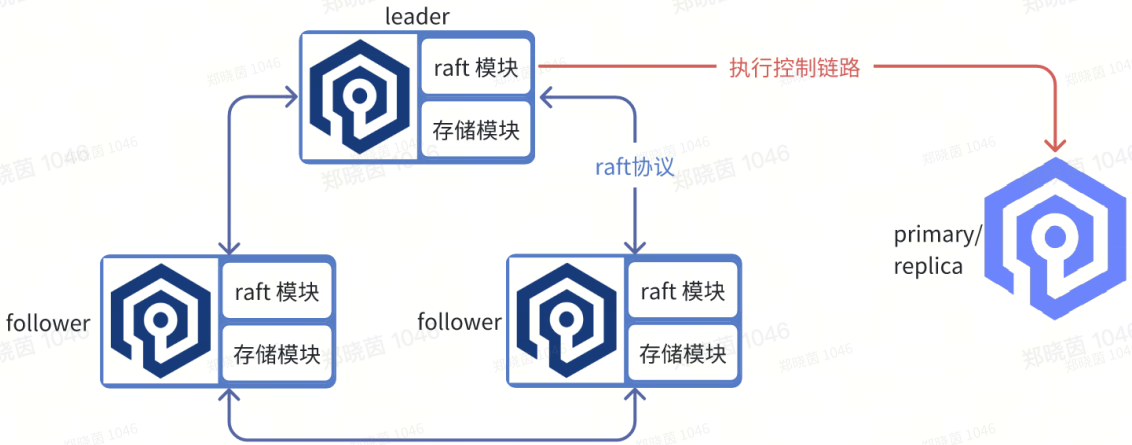
中控服务集成在Valkey cluster内，彻底去掉gossip协议
Valkey node 新加角色：root



下一代Valkey集群管理架构

raft-root

root由3个以上的奇数个节点组成，内部使用raft协议
root有独立的存储模块，兼容rdb/aof格式
结合管理metaspace，root可兼容数据节点能力，适用小规模集群

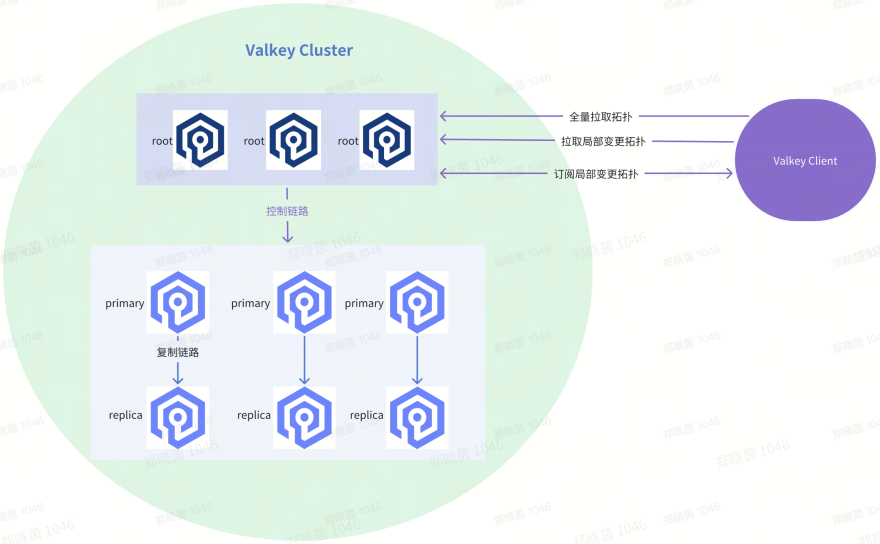


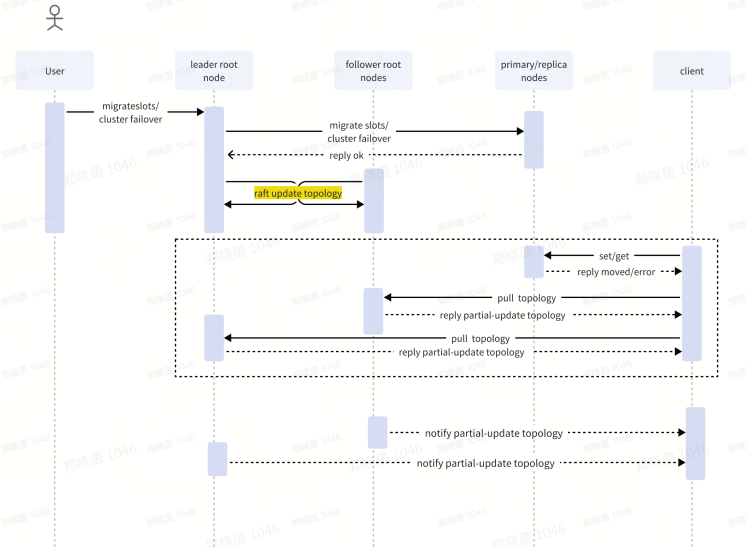
下一代Valkey集群管理架构

客户端-root

规避脑裂：客户端订阅root节点的拓扑（Linearizable read），脱离raft集群的root节点不对外提供服务

及时恢复：客户端访问data节点时收到moved或者报错，客户端到root节点重新拉取局部变更的拓扑





下一代Valkey集群管理架构

root-HA

独立cluster线程

cluster bus从主线程抽离出来，独立线程处理cluster port。
root节点控制模块采用独立的多线程实现。

多点探活

raft集群内超过一半以上root节点发起探活。
raft集群中的leader节点决策和执行切换。

多维度选主

root节点可根据更多维度的信息对候选节点列表进行优先级排序：

- 基于offset数据全局选主
- 区域亲和性选主
- 级联复制模式选主

展望

与社区共发展

字节跳动内部原有Redis也正在逐步向Valkey演进
字节也积极参与社区建设，已经向Valkey 8/9 提交诸如 RDMA，MPTCP，lttng可观测性等重量级feature和几十个bugfix等

对社区需求

随着复杂业务的发展，希望valkey engine能提供更丰富的特性

metaspace设计

可观测的资源统计
以支撑详细的资源规划、容量评估

可管理的资源
提供控制命令支持管理metaspace的布局

丰富的集群感知能力
为类valkey search module提供高效的集群基础协作框架

增强expire

引入Expire Bucket设计等弥补随机扫描过期删除的低效

支持Eviction bustable

平滑逐出能力实现可预测可管理的cpu开销

谢谢

本次分享回顾了社区集群架构的现状、业界通用成熟的集群架构设计，探讨了raft一体化集成在Valkey cluster内的新一代中控集群管理架构。
道阻且长，与社区共发展，同前进!