Klustron 多IDC高可用方案及容器化展望

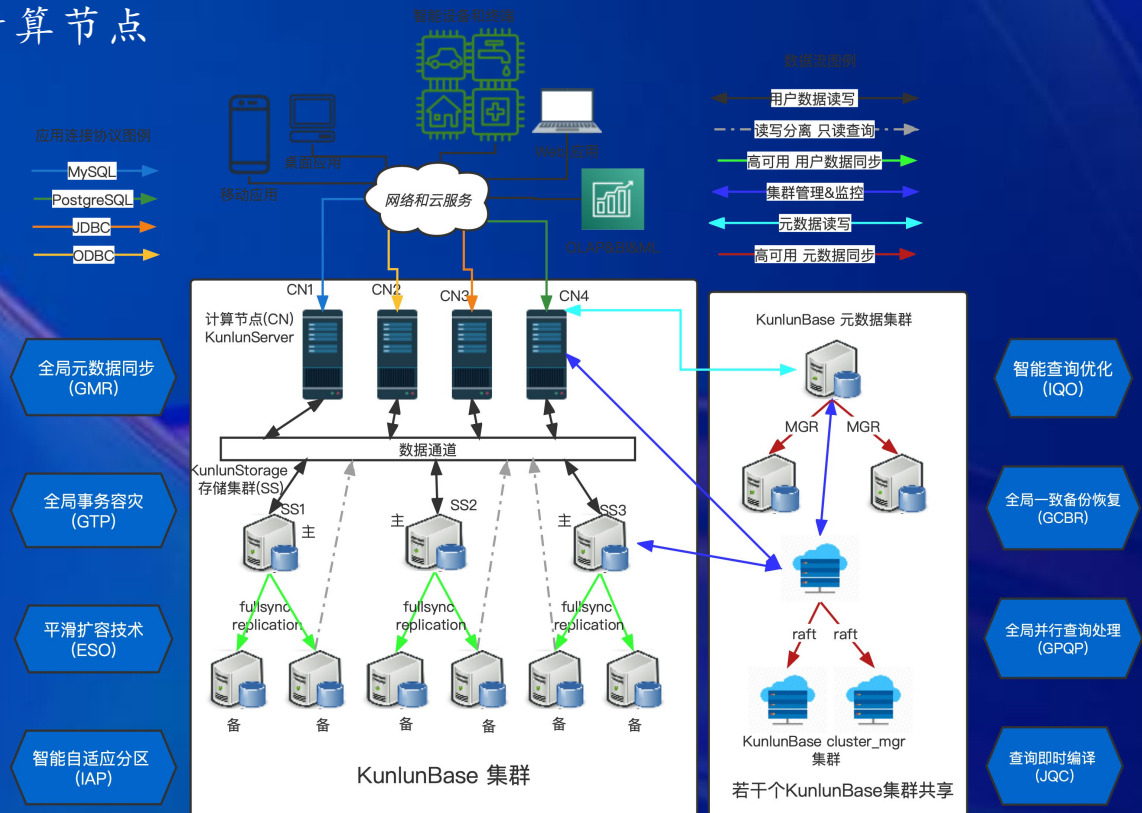
泽拓科技

www.klustron.com

目录

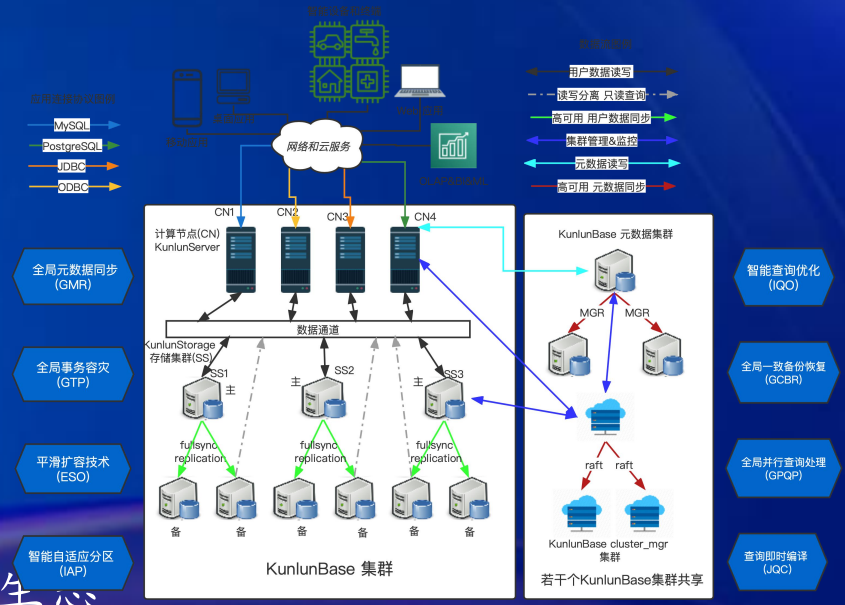
Klustron 架构与核心技术简介

- **弹性伸缩的计算和存储能力**
  - 多种可定制策略自动完成数据拆分和分布，实现最佳性能
  - 单台服务器CPU和内存有限，需要利用大量服务器的CPU和内存
  - 扩缩容：自动、柔性、不停服、无业务侵入、终端用户无感知
  - 存储和计算分离，多点读写，按需增减存储和/或计算节点
- **金融级高可用& 高可靠**
  - 高可用机制基于本机文件系统做数据存储
  - 自动处理软硬件故障、网络故障、机房故障
    - 数据不丢不乱，服务持续在线
  - 自动发现主节点故障并选主和主备切换
  - 多机房高可用和同城/异地双活
  - 确保RTO < 30秒 & RPO=0
- 极致的数据安全保障
  - 连接加密，数据和日志存储加密
  - 多层级访问控制，灵活配置规则



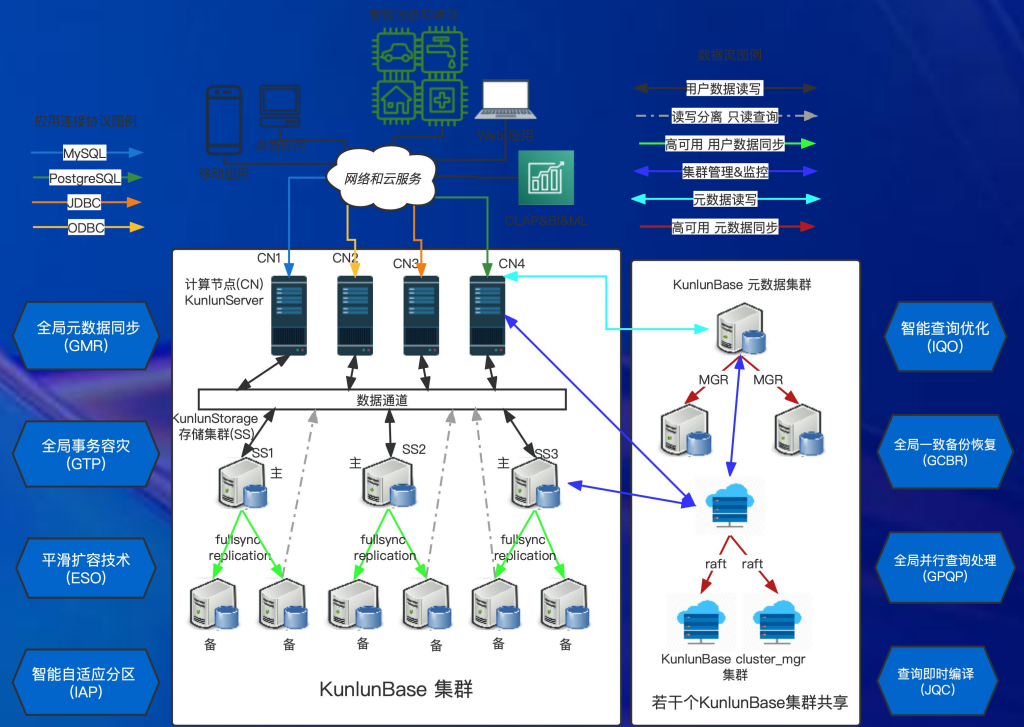
Klustron 架构与核心技术简介

- 支持K8S和容器化部署
  - 数据库系统实现云服务化的关键基础设施：分布式文件系统
  - 仅用于现有节点监控和拉起
- HTAP：OLTP & OLAP 一份数据，两类负载互不干扰
  - OLTP为主：对应用软件等价于使用MySQL或PostgreSQL
  - OLAP为辅：多层级并行查询实现高性能
    - 数据分析新场景：分析最新数据，捕获先机
    - 通过 TPC-H & TPC-DS
- 融合标准SQL、PostgreSQL 和 MySQL 的应用和工具软件生态
  - ✓ 支持PostgreSQL的DDL 和DML语法 和连接协议
  - ✓ 支持MySQL的DML语法 和连接协议
  - ✓ 支持JDBC，ODBC，所有常见编程语言的PostgreSQL和MySQL 客户端connector



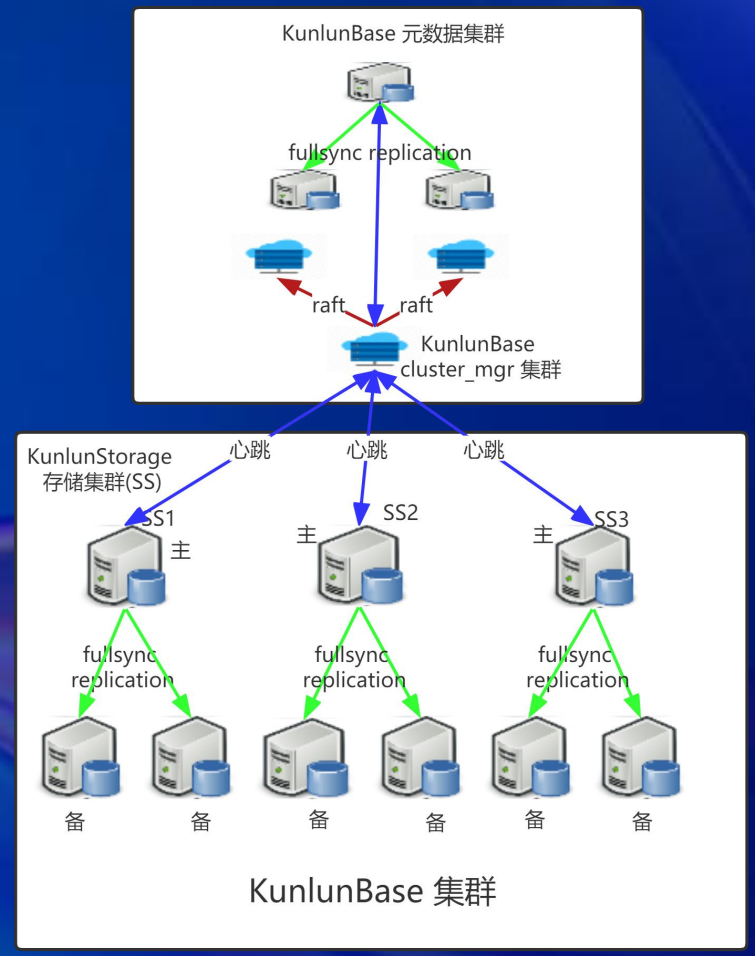
Klustron 资源管理和节点监控能力

- 元数据集群
  - 计算机服务器注册，供服务器资源分配
  - 集群元数据信息和拓扑结构
- node_mgr：部署在每台计算机服务器
  - 执行cluster_mgr下发的本地命令和脚本
  - 监控本机内节点运行状态，及时拉起节点
  - 监控服务器资源使用状况并上报
- cluster_mgr 集群：raft高可用
  - 主节点接收和执行集群管理命令 -- cluster_mgr API
  - 节点迁移和集群扩容：手动触发，自动完成
  - 集群管理命令下发到nodemgr
- bootstrap
  - 初始化计算机服务器：安装klustron组件
  - 初始化服务器后，使用XPanel或者cluster_mgr API做集群管理



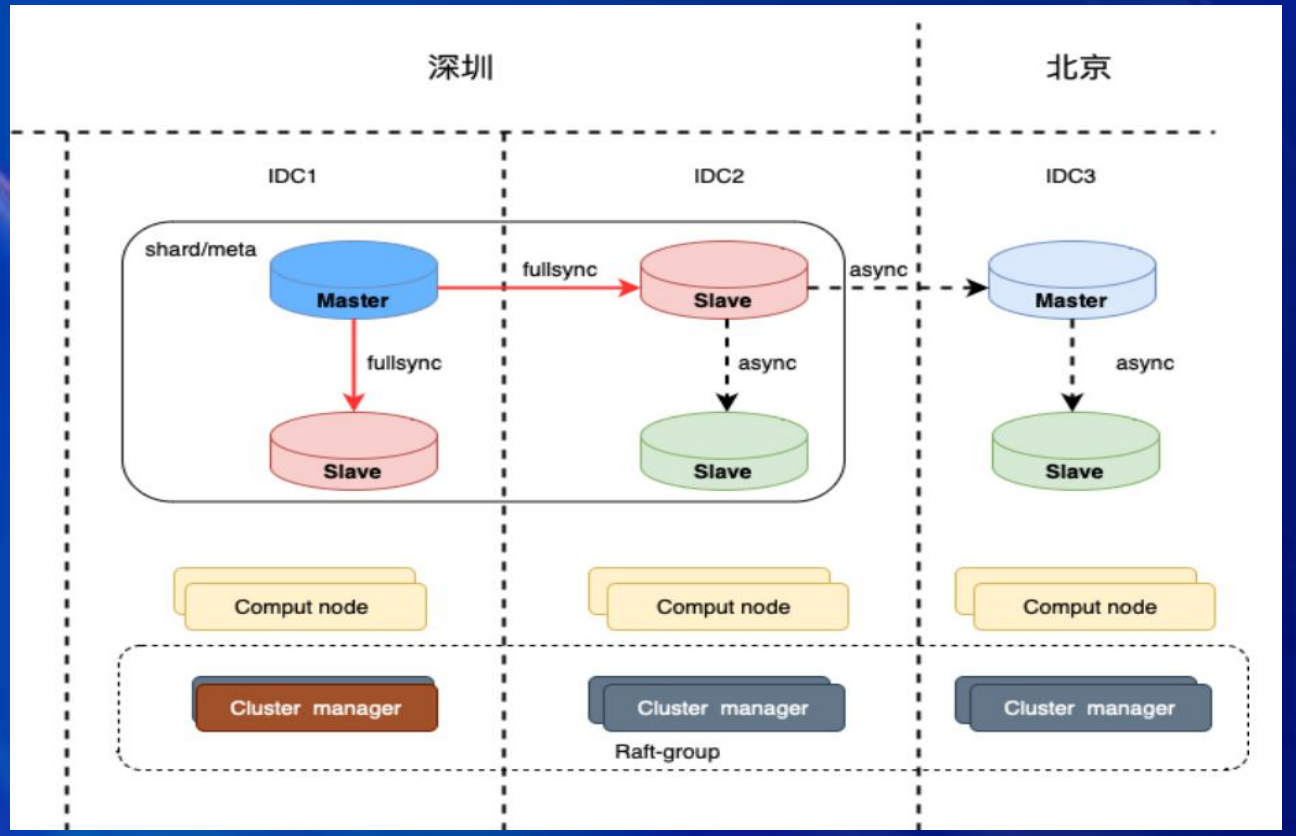
Klustron的集群高可用机制

- fullsync：高可用的基础技术
  - binlog replication，兼容所有事务存储引擎(innodb + rocksdb)
  - 确保主节点的数据更新事件被备机收到
  - 等待的方法和时机
  - 主备同时宕机？
  - 性能开销约等于0， how?
  - 多于2个备机：consistency_level
  - 与MySQL semisync和MGR 的区别
- fullsync HA
  - 主节点监测
  - 选主
  - 主备切换
  - 闪回



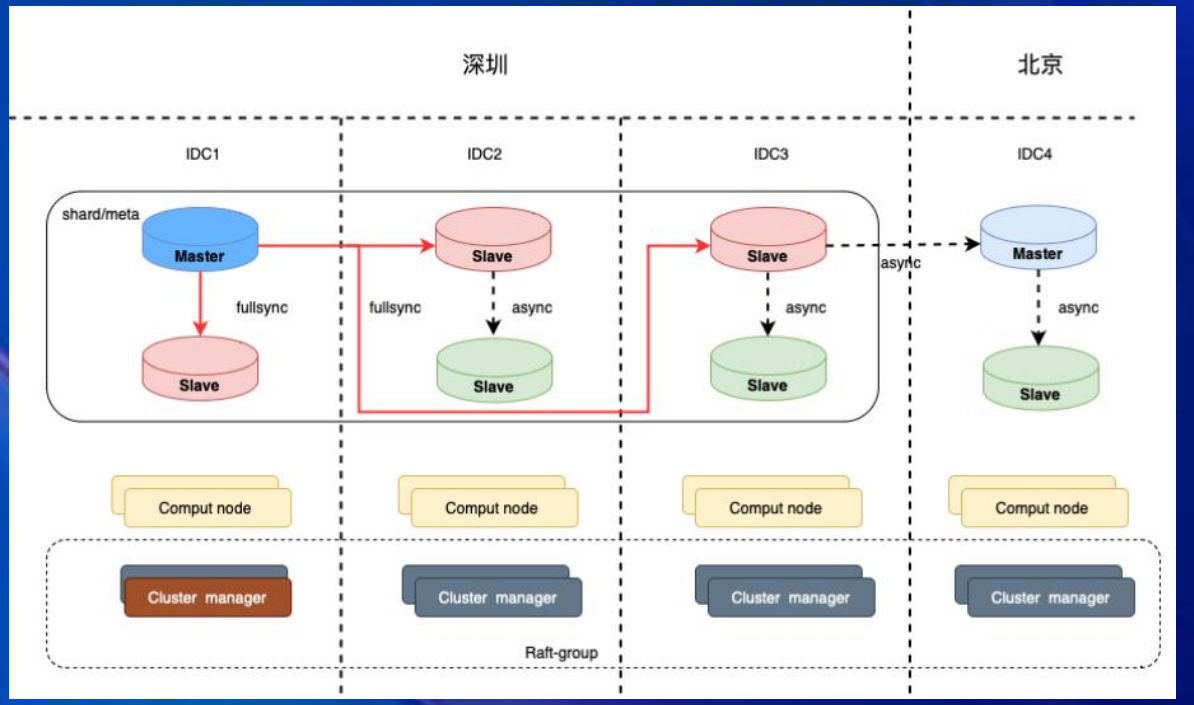
Klustron 集群多机房高可用机制

- 一个集群每个shard部署在多个机房
  - 主城主机房：shard主节点M+1个fullsync 备 MS
  - 主城备机房：M的一个fullsync备 Sx作为主 + Sx的1个二级async 备
    - 可以按需复制多个
  - 备城机房：主城备机房 Sx做async复制的主 + 二级async备
- 主机房内主备切换：不丢数据
  - MS提升为新的主节点M，M作为其备
  - 其他机房的Sx 从新主复制
- 主城机房间切换：不丢数据
  - Sx提升为主
  - 其他机房的Sx 从新主复制



Klustron 集群多机房高可用机制

- 备城机房提升为主机房
  - 主城所有机房全部失效时
  - 可能会丢数据
  - 可能无法完成自动主备切换从而需要人工介入
- 计算节点
  - 每个机房若干个
  - 主城所有计算节点都可以使用，首选同机房
  - 备城机房：只有升为主才用
- cluster_mgr
  - 所有机房的节点组成同一个集群
  - raft一致性：多个IDC失效需手工修改配置文件



Klustron集群多机房容器化高可用机制展望

- Klustron 资源管理和监控能力的不足
  - 节点动态迁移慢：重做存储节点需要copy大量数据
    - 原因：没有分布式文件系统，假设使用本地磁盘
    - 公有云上部署没有发挥出分布式文件系统的多副本能力
  - 资源隔离依赖于独立设置cgroup
  - 存储空间开销大
  - 截止1.2版本使用k8s+容器 不具备跨机房高可用能力
- 开发维护工作量比较大

Klustron集群多机房容器化高可用机制展望

- Klustron 在公有云上优化部署展望
  - 使用k8s+容器 做集群管理、监控和迁移、扩容
  - 完全基于分布式文件系统做数据存储和高可用
    - 每个shard的主备节点使用相同的一份（默认3副本）用户数据
    - 每个shard 拥有和读写一部分数据分片
- 技术挑战
  - 把主备切换融入k8s的节点监控管理流程
  - 介入k8s的节点迁移和扩容流程 -- 无需搬迁数据
  - 备机读取数据的一致性
  - k8s跨机房操作klustron节点