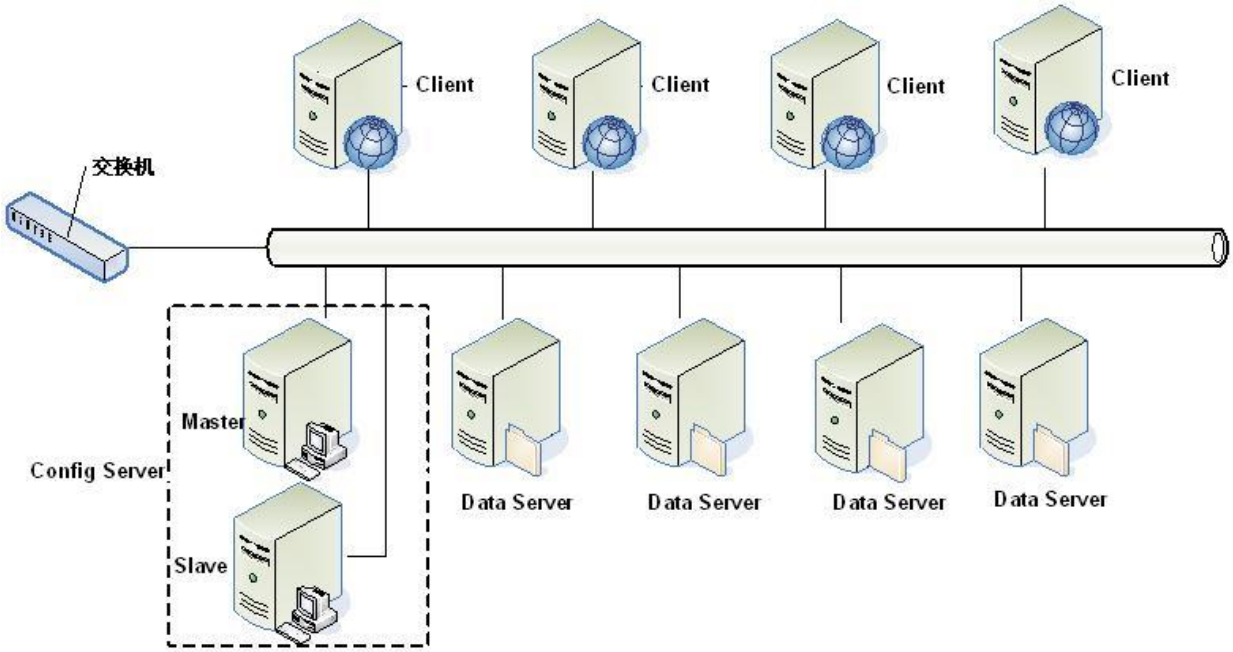
简介

tair 是淘宝自己开发的一个分布式 key/value 存储引擎. tair 分为持久化和非持久化两种使用方式. 非持久化的 tair 可以看成是一个分布式缓存. 持久化的 tair 将数据存放于磁盘中. 为了解决磁盘损坏导致数据丢失, tair 可以配置数据的备份数目, tair 自动将一份数据的不同备份放到不同的主机上, 当有主机发生异常, 无法正常提供服务的时候, 其于的备份会继续提供服务.

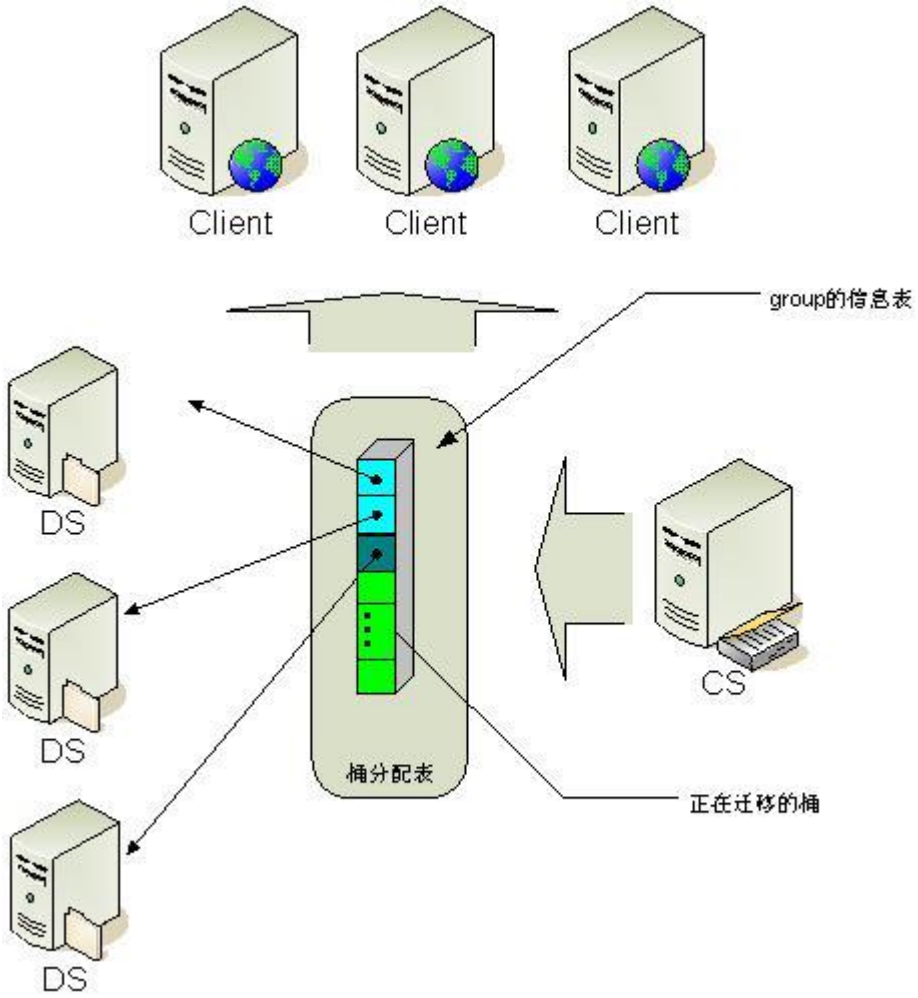
tair 的总体结构

tair 作为一个分布式系统, 是由一个中心控制节点和一系列的服务节点组成. 我们称中心控制节点为 config server. 服务节点是 data server. config server 负责管理所有的 data server, 维护 data server 的状态信息. data server 对外提供各种数据服务, 并以心跳的形式将自身状况汇报给 config server. config server 是控制点, 而且是单点, 目前采用一主一备的形式来保证其可靠性. 所有的 data server 地位都是等价的.



tair 的负载均衡算法是什么

tair 的分布采用的是一致性哈希算法, 对于所有的 key, 分到 Q 个桶中, 桶是负载均衡和数据迁移的基本单位. config server 根据一定的策略把每个桶指派到不同的 data server 上. 因为数据按照 key 做 hash 算法, 所以可以认为每个桶中的数据基本是平衡的. 保证了桶分布的均衡性, 就保证了数据分布的均衡性.



桶在 data server 上分布时候的策略

程序提供了两种生成分配表的策略，一种叫做负载均衡优先，一种叫做位置安全优先：我们先看负载优先策略．当采用负载优先策略的时候，config server 会尽量的把桶均匀的分布到各个 data server 上．所谓尽量是指在不违背下面的原则的条件下尽量负载均衡．1 每个桶必须有 COPY_COUNT 份数据 2 一个桶的各份数据不能在同一台主机上；位置安全优先原则是说，在不违背上面两个原则的条件下，还要满足位置安全条件，然后再考虑负载均衡．位置信息的获取是通过 _pos_mask(参见安装部署文档中关于配置项的解释)计算得到．一般我们通过控制 _pos_mask 来使得不同的机房具有不同的位置信息．那么在位置安全优先的时候，必须被满足的条件要增加一条，一个桶的各份数据不能都位于相同的一个位置(不在同一个机房)．这里有一个问题，假如只有两个机房，机房 1 中有 100 台 data server，机房 2 中只有 1 台 data server．这个时候，机房 2 中 data server 的压力必然会非常大．于是这里产生了一个控制参数 _build_diff_ratio(参见安装部署文档)．当机房差异比率大于这个配置值时，config server 也不再 build 新表．机房差异比率是如何计出来的呢？首先找到机器最多的机房，不妨设使 RA，data server 数量是 SA．那么其余的 data server 的数量记做 SB．则机房差异比率=|SA – SB|/SA．因为一般我们线上系统配置的 COPY_COUNT 是 3．在这个情况下，不妨设只有两个机房 RA 和 RB，那么两个机房什么样的 data server 数量是均衡的范围呢？当差异比率小于 0.5 的时候是可以做到各台 data server 负载都完全均衡的.这里有一点要注意，假设 RA 机房有机器 6 台,RB 有机器 3 台．那么差异比率 = 6 – 3 / 6 = 0.5．这个时候如果进行扩容，在机房 A 增加一台 data server，扩容后的差异比率 = 7 – 3 / 7 = 0.57．也就是说，只在机器数多的机房增加 data server 会扩大差异比率．如果我们的_build_diff_ratio 配置值是 0.5．那么进行这种扩容后，config server 会拒绝再继续 build 新表．

tair 的一致性和可靠性问题

分布式系统中的可靠性和一致性是无法同时保证的，因为我们必须允许网络错误的发生. tair 采用复制技术来提高可靠性，并且为了提高效率做了一些优化，事实上在没有错误发生的时候, tair 提供的是一种强一致性. 但是在有 data server 发生故障的时候，客户有可能在一定时间窗口内读不到最新的数据. 甚至发生最新数据丢失的情况.

tair 提供的客户端

tair 的 server 端是 C++写的，因为 server 和客户端之间使用 socket 通信，理论上只要可以实现 socket 操作的语言都可以直接实现成 tair 客户端. 目前实际提供的客户端有 java 和 C++. 客户端只需要知道 config server 的位置信息就可以享受 tair 集群提供的服务了.

Tair 的安装

1、确保安装了 automake autoconfig，使用 autumake --version 查看，一般情况下已安装，否则通过 apt-get install 安装

2、安装依赖库或软件

2.1 安装 libtool

sudo apt-get install libtool

2.2 安装 boost-devel

下载源码

解压 tar -zxvf boost...

cd boost

./booststrap.sh

./bjam

2.3 安装 zlib 库

下载源码

tar -xvf zlib

cd zlib

./configure

make

sudo make install

2.4 安装 tbsys 和 tbnet

svn co -r 18 http://code.taobao.org/svn/tb-common-utils/trunk tb-common-utils

mkdir lib

cd tb-common-utils

su

export TBLIB_ROOT="/home/fl/lib/"

sh build.sh

3 安装 Tair

svn checkout http://code.taobao.org/svn/tair/trunk/ tair

cd tair

sh bootstrap.sh

./configure

make

make install

比赛中 Tair 如何使用

存入 tair 的数据格式 key 字符串格式, value 是 number 类型。key 统一以“固定前缀_整分时间戳”方式命名的字符串，整分时间戳就是整分时刻对应的时间戳，可以表示该一分钟，例如 2015/11/11 08:11:00 分钟对应的时间戳为 1447200660（10 位数），可表示 2015/11/11 08:11:00 ~ 2015/11/11 08:12:00（不包含该时刻）这一分钟。

类型	key	vaule
淘宝每分钟的交易	platformTaobao_ 整分时间戳	Number 类型
天猫每分钟的交易	platformTmall_ 整分时间戳	Number 类型
整分时刻无线和 PC 历史交易比	ratio_整分时间戳	Number 类型，保留两位小数