面向大规模应用实例的配置分发系统

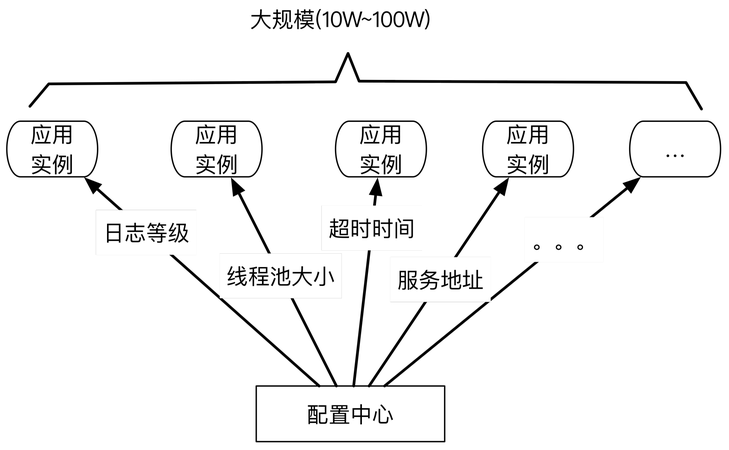
唐震，邵佳琦，王伟，许利杰
软件工程技术研究开发中心
wangwei@otcaix.iscas.ac.cn xulijie@iscas.ac.cn

工作介绍

场景：随着业务复杂度、容量的增加，云服务提供商的应用实例规模急剧增长，部署在云端的应用面临着其实例的维护与管理等业务需求，如配置管理需求。

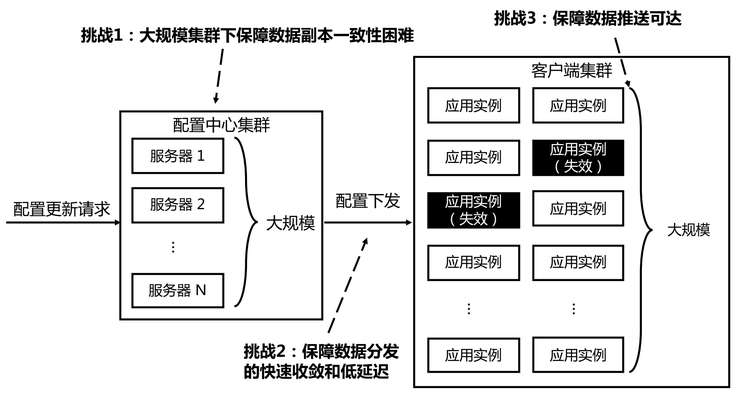
挑战：配置管理系统面临着大规模服务器集群之间以及服务器向实例集群进行配置分发过程的收敛速度慢，可靠性难以保障等挑战，传统星型通知与raft协议难以用于大规模集群的配置分发场景。

工作内容：本课题对现有问题进行建模与分析，提出了可以充分利用集群各节点传播能力的基于树模型与图模型的数据分发机制。实验结果表明，本课题提出的数据分发机制比传统算法收敛时间降低了50%以上，并可保障在复杂网络环境下数据分发的可靠性。



问题分析

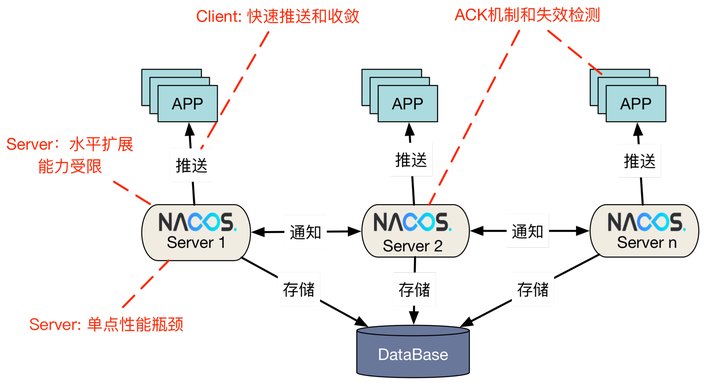
本课题关注大规模应用实例配置分发场景下面临的性能与可靠性问题。



数据副本快速达成一致困难
在应对洪峰流量时需要对微服务容器进行扩容，如果未能保障扩容配置数据的一致性，将导致部分节点配置数据过旧或错误；

对数据分发的快速收敛要求高
为应对业务和负载的快速变化，实现大规模应用实例配置快速变更，数据分发需要在限定时间（如10秒）内完成，否则可能带来严重后果；

大规模应用实例的失效检测困难
在大规模应用实例的环境下，由于硬件、网络等问题，极易出现少数应用实例失效的场景。



本课题基于Nacos(Dynamic Naming and Configuration Service)
提供服务注册与发现功能以及配置管理功能

具体问题

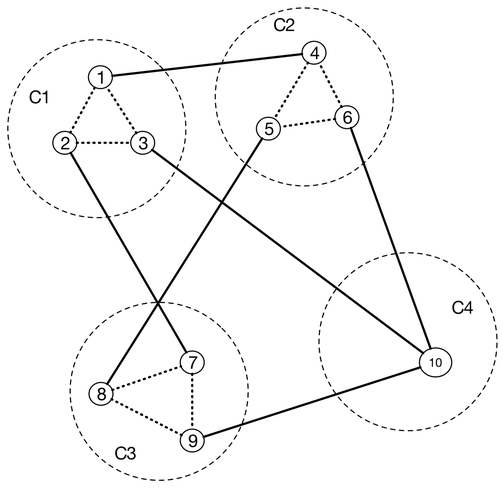
- 大规模服务器间的高效数据分发
- 服务器集群向大规模客户端的高效数据分发
- 配置分发的最终可靠性保障

算法模型

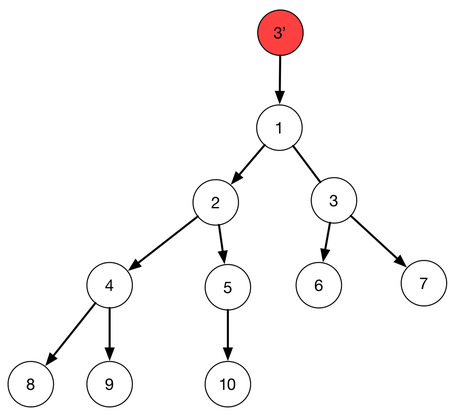
为了解决数据分发过程中的性能与可靠性问题，本课题提出以下两种算法模型

基于图模型的传播算法

算法概述
将整个集群拆分为若干小集群，分层传播。
组网方式
小集群内部采用全连接的结构，将小集群抽象为单节点后，小集群间为全连接结构，小集群间的度数由内部节点均分。
传播过程
更新消息首先在小集群内部传播，随后在小集群之间传播。
容错策略
根据层次拓扑结构，推算二跳连接节点，从而绕开失效节点。



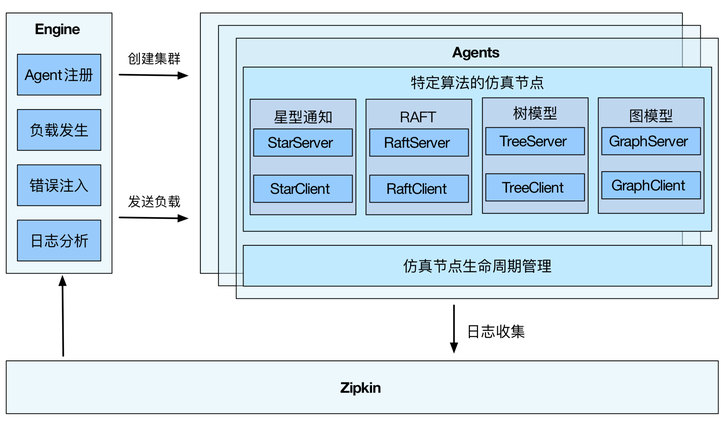
基于树模型的传播算法



算法概述
将集群节点组织成为满N叉树结构，采用多轮次传播解决单点性能瓶颈。
组网方式
节点列表以二维数组形式存储，根据数组下标，每个节点均个推算出其父节点与子节点的索引。
传播过程
消息。
容错策略
(1) 由根节点传播至前K层所有的非叶子节点，可以容忍K-1层节点全部失效；
(2) 根据满N叉树的数据结构，绕开失效节点，传播任务直达其子节点。

仿真系统

为了验证上文提出算法的正确性与有效性，本课题将在小规模集群中，采用一定比例，在物理节点上使用线程模拟应用实例与服务器节点（例如一台物理机仿真一百个虚拟节点），使用线程休眠的方式模拟网络通信延迟与数据库访问延迟。



实验流程

- Engine根据集群规模、类型参数发出集群创建命令；
- 各个Agent接收命令本地创建集群，并向Engine注册；
- Engine生成不同比例的负载，模拟真实环境中的负载不均衡，仿真节点运行具体传播算法；
- 收集运行日志并分析。

Agent

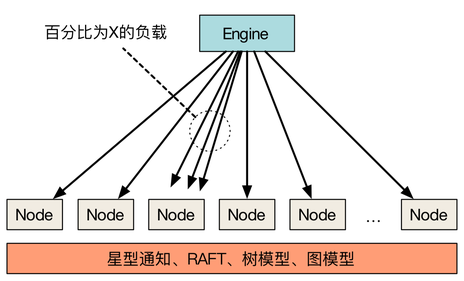
- 创建多个仿真节点；
- 运行具体数据传播算法。

Engine

- 发出创建节点命令并注册仿真节点；
- 向仿真节点发送负载；
- 错误注入以模拟复杂网络环境；
- 分析算法运行日志，图表分析。

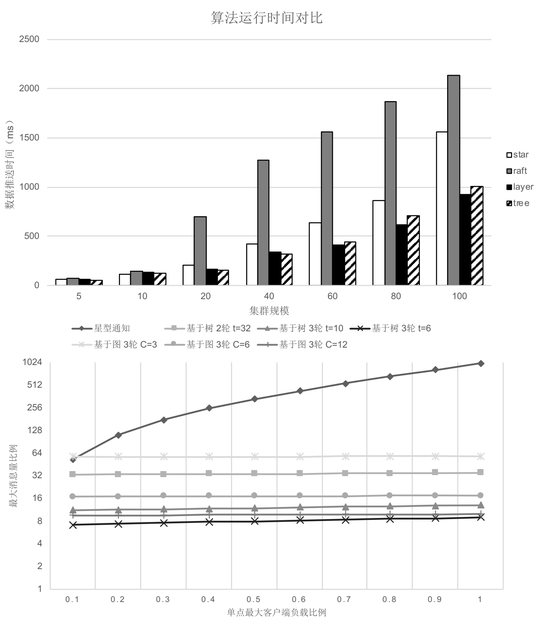
Zipkin

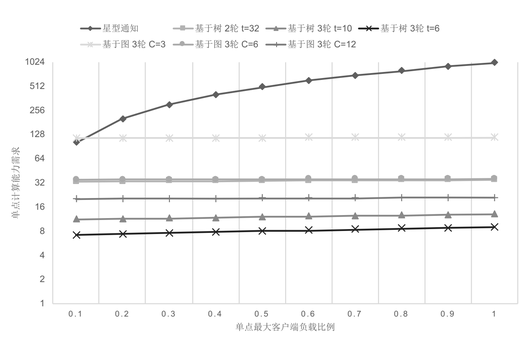
- 跟踪算法执行过程，收集日志。



实验结果

在由10个物理节点组成的集群中，按比例仿真出不同规模的节点。在仿真节点上分别进行星型通知算法，RAFT，基于树模型与基于图模型的实验，在不同规模仿真节点与不同单点最大客户端负载比例（来自客户端的更新请求由单节点承担一定百分比，剩余节点均分）下，计算统计各算法运行时间，单节点计算能力需求（网络中负载最大的节点所需的计算能力），最大消息量比例（网络中消息量最大的节点与消息量最小的节点的总消息量之比）。





实验结果：

1. 集群规模较小时（一般在个位数节点），几种算法性能差别较小，传统算法实现简单；

2. 在大规模集群下，本课题提出的基于树模型与图模型的传播算法降低了50%以上的运行时间，降低了97.7%的单节点计算能力需求与最大消息量比例。