(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102387210 A
(43)申请公布日 2012.03.21

(21)申请号 201110325988.5

(22)申请日 2011.10.25

(71)申请人 曙光信息产业(北京)有限公司
地址 100084 北京市海淀区水磨西街 64 号

(72)发明人 张攀勇 袁重桥 赵力 邵宗有
刘新春 苗艳超 王勇

(74)专利代理机构 北京安博达知识产权代理有限公司 11271
代理人 徐国文

(51)Int.Cl.
*H04L 29/08*(2006.01)

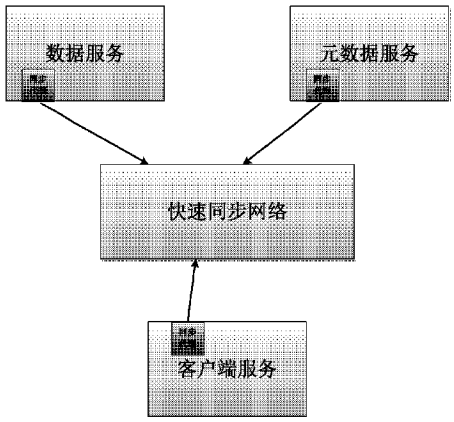
权利要求书 1 页 说明书 3 页 附图 2 页

(54)发明名称

一种基于快速同步网络的分布式文件系统监控方法

(57)摘要

本发明提出的一种基于快速同步网络的分布式文件系统监控机制，能够快速同步系统中的每一个节点上服务的状态，对于 N 个节点，全系统只需要发送 3N 个消息即可完成整个系统的状态监控和同步，因此具有较好的可扩展性，能够随着系统规模进行扩展。同时采用了基于选举的动态换主方法，不存在单点故障。同时根据系统与其他节点的通信状态，进一步压缩了同步消息的数量，降低了监控的开销。




CN 102387210 A

1. 一种基于快速同步网络的分布式文件系统监控方法，其特征在于：

服务器节点通过快速同步网络相互连接组成一个同步组，每个服务器节点在快速同步网络中有独立的编号；

在一个同步组中，设置一个主节点用于周期性发起状态收集命令，向所有节点报告整个系统的运行状态；

其他从节点负责处理主节点发出的状态收集命令，收集本地节点服务状态，并向主节点应答本节点的状态；

同时从节点负责接收来自主节点的系统状态报告，根据该状态报告进行故障处理。

2. 如权利要求 1 所述的方法，其特征在于：所述同步组中存在一个同步代理模块，用于获取节点上运行服务的状态，和同步组中间的其他节点之间进行消息交互，包括报告自身的服务状态，获取整个系统的节点服务状态。

3. 如权利要求 1 所述的方法，其特征在于：所述同步组可以分层，每一层内部为一个同步组，内部完成同步之后，由每一层提供的同步组的头结点构成的高一层同步组之间再进行同步操作。

4. 如权利要求 1 所述的方法，其特征在于：所述主节点在出现故障时，同步组的换主方法为：

S1，对于每一个从节点，系统周期性的检测是否存在来自主节点的系统状态报告消息；

S2，如果发现主节点未报告，则增加主节点失败计数；如果主节点的失败记录超过阈值，则从节点向同步组中间上次活动的序号最小的从节点发出接管命令，等待接管完成命令；

S3，该最小节点在接收到接管命令之后，向主节点发出状态询问命令，如果收到主节点的应答命令，则无操作；否则，认为主节点死机，进行主节点接管操作，在接管操作完成之后，向所有节点发出接管完成命令；

S4，如果其他从节点在几个周期内未收到接管完成命令，则认为该最小节点死机，继续步骤 S2，选出下一个可用的主节点为止，

S5，新接管完成的主节点定期发起请求状态查询。

# 一种基于快速同步网络的分布式文件系统监控方法

## 技术领域

[0001] 本发明涉及并行文件系统的通信系统，特别涉及一种基于快速同步网络的分布式文件系统监控方法。

## 背景技术

[0002] 分布式系统中通信系统作为系统各节点之间的通信基础，具有重要的意义。由于分布式系统中的每一个节点均可能出现故障，而为了提高系统的可用性，需要在节点和服务出现故障之后，能够快速的进行故障检测和定位，以便于系统的后续故障恢复和处理。因此分布式系统的监控和检测的速度和准确率变得重要。传统的系统监控方法，通过节点之间相互发出服务状态监控消息，这种方法能够检测出系统故障，但是随着系统规模的扩大，其对网络中正常消息的干扰会急剧上升。同时故障检测和通知的速度均存在问题。在这种背景下，本专利提出了一种利用快速同步网络进行分布式文件系统监控的机制，有效的保证了可扩展性和检测通知的速度。

## 发明内容

[0003] 本发明的目的是解决随着系统规模扩到导致的文件系统监控的可扩展性和速度的问题，基于快速同步网络，提供了一种分布式系统监控机制。

[0004] 本发明采用了一套快速同步网络，该快速同步网络与每一个服务节点相互连接，采用树状或者其他拓扑形状，连接起来，构成一个同步组，每一个服务节点在快速同步网络中间有一个独立的编号。

[0005] 对于同步组中间的节点上，存在一个同步代理模块，用于获取节点上运行服务的状态，和同步组中间的其他节点之间进行消息交互，包括报告自身的服务状态，获取整个系统的节点服务状态。

[0006] 在一个同步组中间，存在一个主节点，用以周期性的发起状态收集命令，向所有节点报告整个系统的运行状态。其他从节点负责处理主节点发出的状态收集命令，收集本地节点服务状态，并向主节点应答本节点的状态。同时从节点负责接收来自主节点的系统状态报告，根据该状态报告进行故障处理。当系统规模扩大之后，同步组可以分层，每一层内部为一个同步组，内部完成同步之后，由每一层提供的同步组的头结点构成的高一层同步组之间再进行同步操作。

[0007] 当主节点出现故障的时候，需要同步组能够迅速检查到主节点故障，并进行换主操作。本发明实现了一种动态换主方法。详细步骤为：

[0008] S1，对于每一个从节点，会周期性的检测是否存在来自主节点的系统状态报告消息；

[0009] S2，如果发现主节点未报告，则增加主节点失败计数；如果主节点的失败记录超过阈值，则从节点向同步组中间上次活动的序号最小的从节点发出接管命令，等待接管完成命令；

[0010] S3，该最小节点在接收到接管命令之后，向主节点发出状态询问命令，如果收到主节点的应答命令，则无操作；否则，认为主节点死机，进行主节点接管操作，在接管操作完成之后，向所有节点发出接管完成命令；

[0011] S4，如果其他从节点在几个周期内未收到接管完成命令，则认为该最小节点死机，继续步骤 S2，选出下一个可用的主节点为止；

[0012] S5，新接管完成的主节点定期发起请求状态查询。

[0013] 本发明提出的一种基于快速同步网络的分布式文件系统监控机制，能够快速同步系统中的每一个节点上服务的状态，对于 N 个节点，全系统只需要发送 3N 个消息即可完成整个系统的状态监控和同步，因此具有较好的可扩展性，能够随着系统规模进行扩展。同时采用了基于选举的动态换主方法，不存在单点故障。同时根据系统与其他节点的通信状态，进一步压缩了同步消息的数量，降低了监控的开销。

## 附图说明

[0014] 以下，结合附图来详细说明本发明的实施例，其中：

[0015] 图 1 为基于快读同步网络的监控系统示意图；

[0016] 图 2 为节点监控机制示意图；

[0017] 图 3 为多层次系统的同步方式示。

## 具体实施方式

[0018] 下面结合附图和具体实施方式对本发明的方法进行说明。

[0019] 快速同步网络的连接如图 1 所示，快速同步网络与每一套服务节点相互连接，同步网络的具体实现形式不限定，可以为物理的独立的管理网络，可以为和数据网络一样的物理网络。

[0020] 节点监控方式如图 2 所示，对于同步组中间的状态收集和同步方式分为如下几个步骤：

[0021] 步骤 S1，头节点根据当前节点和其他节点的连接状态，以及在监测时间间隔内是否发送过消息，构建状态收集组，如果时间间隔内发送过消息，则无需向该节点发送状态收集命令；如果未发送消息，则将对应节点加入到状态收集组中

[0022] 步骤 S2，头结点向其他节点以广播的方式发出状态收集请求，等待其他节点应答状态通知

[0023] 步骤 S3，同步组中其他节点在接收到状态收集消息之后，检测自己的服务状态，并向头结点发出应答通知。

[0024] 步骤 S4，头结点收集所有的应答通知，如果有节点在规定时间内未应答，则认为该节点上的服务未响应，如果死机次数超过规定阈值，则认为该节点死机，修改对应的节点状态。

[0025] 步骤 S5，头结点在完成收集应答通知步骤之后，向同步组中的所有有效节点发出系统状态通知。

[0026] 步骤 S6，其他节点在接收到系统状态通知之后，获知系统中整个节点状态，根据状态通知，进行相应的故障处理。

[0027] 多层次系统的同步方式示例如图 3 所示：

[0028] 步骤 S1，主节点 0 和次主节点 1，次主节点 2 构成一个同步组 0，主节点 0 首先以广播方式发起状态收集请求，等待同步组的应答

[0029] 步骤 S2 次主节点 1 和其内部的从节点构成一个次同步组 1，次主节点 1 在收到主节点 0 发出的状态收集请求，次主节点 1 以广播的方式发起状态收集请求，在内部收集完成之后，向主节点 0 发送同步组 0 的应答操作。同样的次主节点 2 在次同步组 2 内部发起状态收集请求，在次同步组 2 内部完成状态收集之后，向主节点 0 发送同步组 0 的应答操作。

[0030] 步骤 S3，主节点 0 在收到次主节点 1 和次主节点 2 的应答操作之后，计算全系统的状态，向次主节点发送系统状态报告

[0031] 步骤 S4，次主节点在接收到系统状态报告之后，向各自的次同步组内部广播系统状态。

[0032] 步骤 S5，各个节点都接收到系统状态，完成一次同步操作。

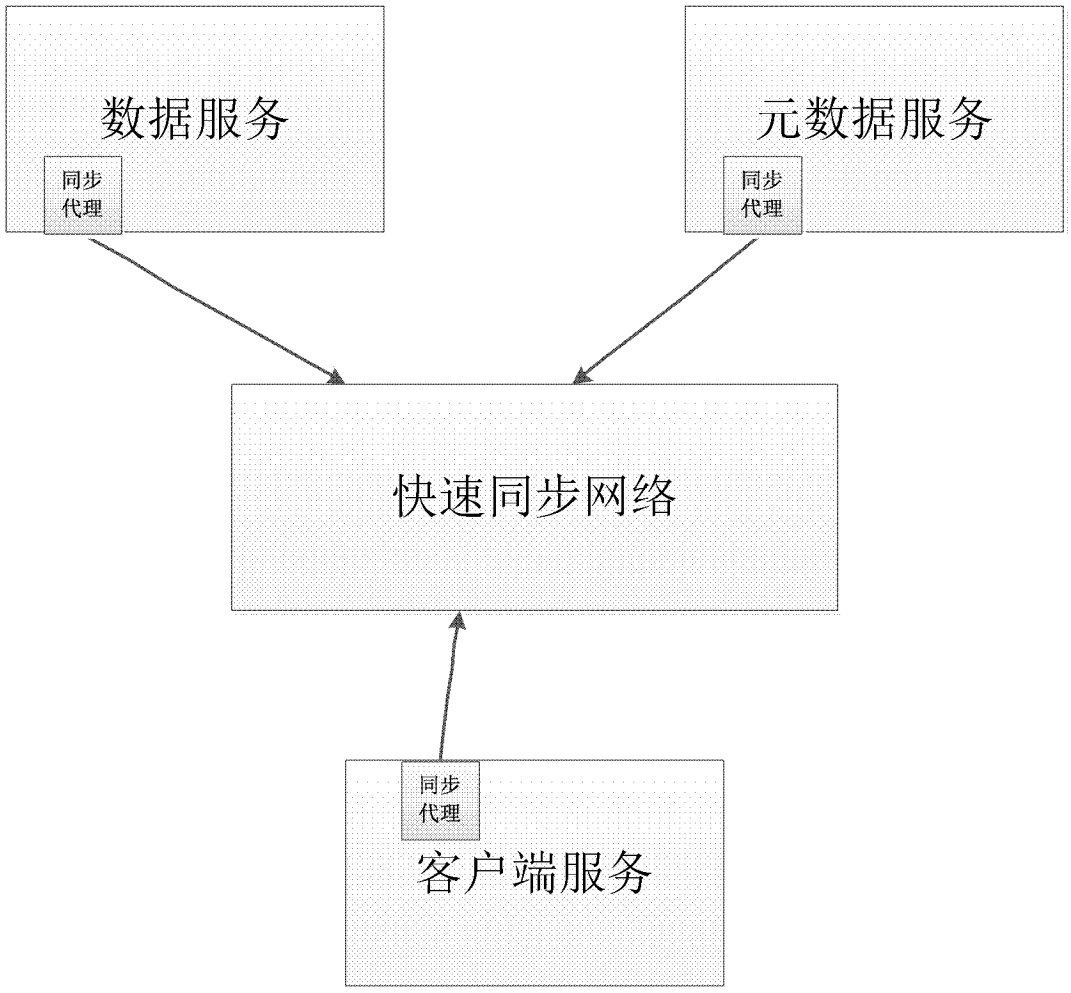



图 1

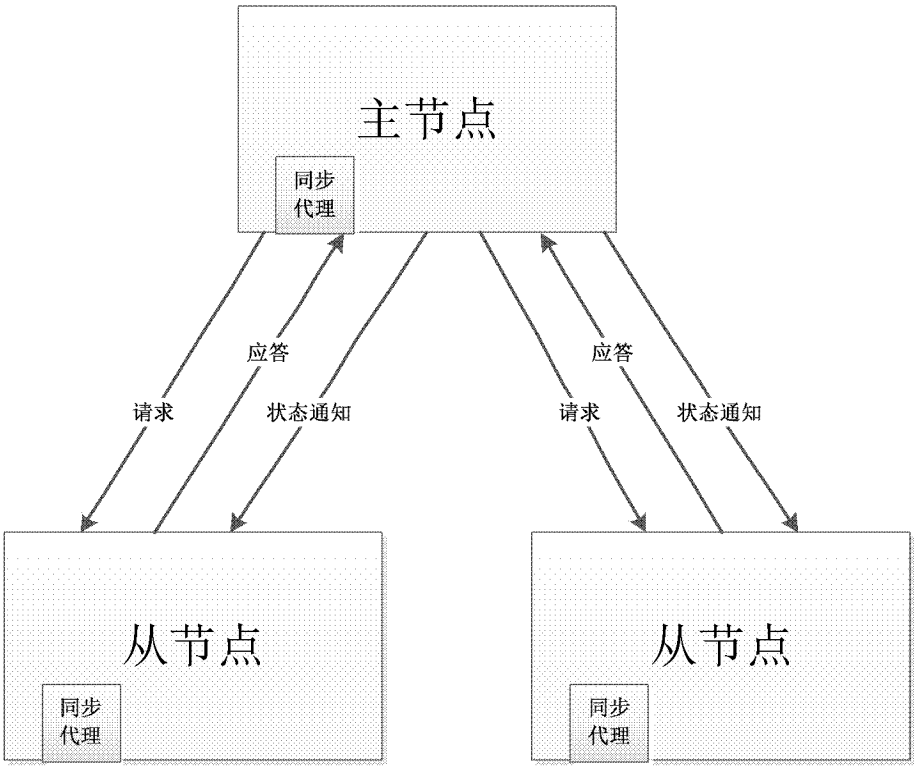

主节点
同步代理
请求
应答
状态通知
请求
应答
状态通知
从节点
同步代理
从节点
同步代理

图 2

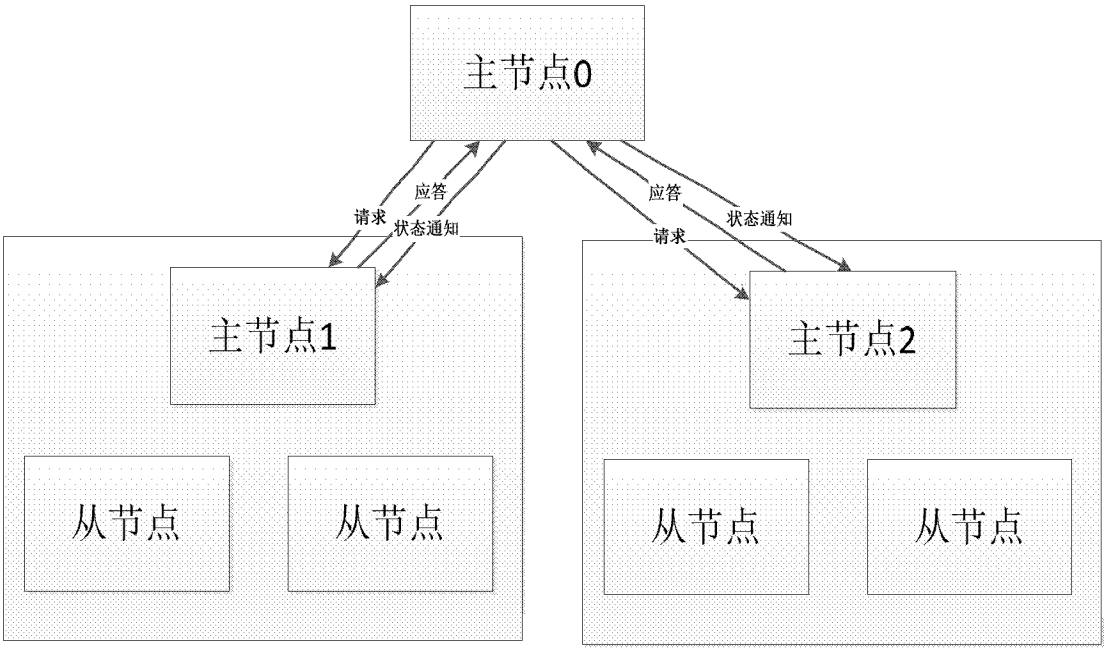

主节点0
请求
应答
状态通知
请求
应答
状态通知
主节点1
主节点2
从节点
从节点
从节点
从节点

图 3