



(12) 发明专利申请

(10) 申请公布号 CN 105847171 A

(43) 申请公布日 2016. 08. 10

(21) 申请号 201610184538. 1

(22) 申请日 2016. 03. 28

(71) 申请人 乐视控股(北京)有限公司

地址 100025 北京市朝阳区姚家园路 105 号
3 号楼 10 层 1102

申请人 乐视云计算有限公司

(72) 发明人 李洪福 刘斌

(74) 专利代理机构 北京商专永信知识产权代理
事务所(普通合伙) 11400

代理人 方挺 黄谦

(51) Int. Cl.

H04L 12/801(2013. 01)

H04L 12/803(2013. 01)

H04L 12/26(2006. 01)

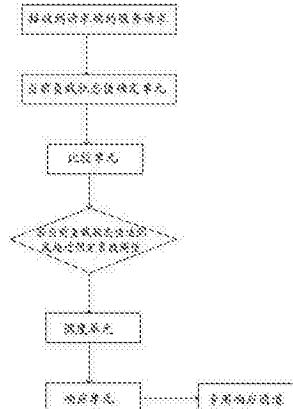
权利要求书1页 说明书3页 附图2页

(54) 发明名称

网络设备过载保护方法

(57) 摘要

本发明实施例提供一种网络设备过载保护方法，包括：在接收到请求端的服务请求时，比较网络设备的当前负载状态值与预定负载阈值的大小，其中所述预定负载阈值小于所述网络设备的最大负载能力值；若所述当前负载状态值大于所述预定负载阈值，则经由专用响应通道向所述请求端返回重定向报文。本发明实施例还提供了一种相应的网络设备过载保护系统。本发明通过设定负载阈值，为后续的重定向报文的发送预留了系统资源；另外，将重定向报文经由专用响应通道发送，一定程度上避免了重定向报文发送时的通信拥塞。另外，通过向请求端发送重定向报文而非错误报文，使得请求端在不接收错误提醒的情况下直接与其他网络设备进行再请求，不影响用户的体验。



1.一种网络设备过载保护方法,包括:

在接收到请求端的服务请求时,比较网络设备的当前负载状态值与预定负载阈值的大小,其中所述预定负载阈值小于所述网络设备的最大负载能力值;

若所述当前负载状态值达到或超过所述预定负载阈值,则经由专用响应通道向所述请求端返回重定向报文。

2.根据权利要求1所述的方法,其中,所述预定负载阈值根据所述网络设备的额定工作参数确定。

3.根据权利要求2所述的方法,其中,所述额定工作参数至少包括:CPU负荷和/或内存和/或带宽。

4.根据权利要求1所述的方法,其中,所述网络设备的当前负载状态值根据所述网络设备的当前负载测量值和上一次负载测量值确定。

5.根据权利要求4所述的方法,其中,所述网络设备的当前负载状态值等于所述当前负载测量值和所述上一次负载测量值分别加权后之和,其中,所述上一次负载测量值的权重大于所述当前负载测量值的权重。

6.根据权利要求1所述的方法,其中,当所述网络设备需要执行多个任务时,优先执行所述多个任务中向所述请求端返回重定向报文的任务。

7.根据权利要求1-6中任一项所述的方法,其中,当所述服务请求的协议为HTTP时,所述重定向报文为302返回码。

8.一种网络设备过载保护系统,包括:

比较单元,用于在接收到请求端的服务请求时,比较网络设备的当前负载状态值与预定负载阈值的大小,其中所述预定负载阈值小于所述网络设备的最大负载能力值;

响应单元,用于当所述当前负载状态值达到或超过所述预定负载阈值时经由专用响应通道向所述请求端返回重定向报文。

9.根据权利要求8所述的系统,其中,所述预定负载阈值根据所述网络设备的额定工作参数确定。

10.根据权利要求9所述的系统,其中,所述额定工作参数至少包括:CPU负荷和/或内存和/或带宽。

11.根据权利要求8所述的系统,其中,所述系统包括:

当前负载状态值确定单元,用于根据所述网络设备的当前负载测量值和上一次负载测量值确定所述网络设备的当前负载状态值。

12.根据权利要求11所述的系统,其中,所述网络设备的当前负载状态值等于所述当前负载测量值和所述上一次负载测量值分别加权后之和,其中,所述上一次负载测量值的权重大于所述当前负载测量值的权重。

13.根据权利要求8所述的系统,其中,所述系统包括:

调度单元,用于当所述网络设备需要执行多个任务时,优先执行所述多个任务中所述响应单元向所述请求端返回重定向报文的任务。

14.根据权利要求8-13中任一项所述的系统,其中,当所述服务请求的协议为HTTP时,所述重定向报文为302返回码。

网络设备过载保护方法

技术领域

[0001] 本发明实施例涉及通信技术领域,尤其涉及一种网络设备过载保护方法。

背景技术

[0002] 现代网络随着业务量的提高、访问量和数据流量的快速增长,对于网络核心的处理能力和计算强度需求也急剧增大,这就使得单一设备根本无法承担。在此情况下,所以负载均衡机制应运而生。

[0003] 图1为现有技术中一种网络设备过载保护方法流程图,当接收到请求端的服务请求时,如果网络设备已经过载,网络设备处理超时会返回错误,请求端继续重试访问,这种情况下网络设备很容易出现问题并且具有破坏性。

发明内容

[0004] 针对现有技术存在的问题,本发明实施例一方面提供一种网络设备过载保护方法,包括:

[0005] 在接收到请求端的服务请求时,比较网络设备的当前负载状态值与预定负载阈值的大小,其中所述预定负载阈值小于所述网络设备的最大负载能力值;

[0006] 若当前负载状态值达到或超过预定负载阈值,则经由专用响应通道向请求端返回重定向报文。

[0007] 本发明实施例另一方面提供一种网络设备过载保护系统,包括:

[0008] 比较单元,用于在接收到请求端的服务请求时,比较网络设备的当前负载状态值与预定负载阈值的大小,其中所述预定负载阈值小于所述网络设备的最大负载能力值;

[0009] 响应单元,用于当当前负载状态值达到或超过预定负载阈值时经由专用响应通道向请求端返回重定向报文。

[0010] 本发明通过设定负载阈值,为后续的重定向报文的发送预留了系统资源;另外,将重定向报文经由专用响应通道发送,一定程度上避免了重定向报文发送时的通信拥塞。解决了现有技术中即使网络设备已经“满”负载却仍然需要请求端继续重试访问的情况,另外,通过网络设备向请求端发送重定向报文而非错误报文,使得请求端在不接收错误提醒的情况下直接与其他网络设备进行再请求,不影响用户的体验。

附图说明

[0011] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍,显而易见地,下面描述中的附图是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0012] 图1为现有技术中一种网络设备过载保护方法流程图;

[0013] 图2为本发明一种网络设备过载保护方法实施例流程图;

[0014] 图3为本发明一种网络设备过载保护系统实施例结构示意图。

具体实施方式

[0015] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0016] 如图2所示，一种网络设备过载保护方法，包括：

[0017] 在接收到请求端的服务请求时，比较网络设备的当前负载状态值与预定负载阈值的大小，其中所述预定负载阈值小于所述网络设备的最大负载能力值；

[0018] 若所述当前负载状态值达到或超过所述预定负载阈值，则经由专用响应通道向请求端返回重定向报文。

[0019] 这里假定服务器端分别为负载均衡设备1和负载均衡设备2，客户端可以是PC端；当PC端发送HTTP请求给负载均衡设备1时，负载均衡设备1会比较当前负载状态值与预定负载阈值的大小，如果当前负载状态值达到或超过预定负载阈值，说明负载均衡设备1超过设备负载，负载均衡设备1优先通过专用响应通道向PC端返回302返回码，PC端基于302返回码返回的重定向地址把HTTP请求重新发给负载均衡设备2，负载均衡设备2继续比较当前负载状态值与预定负载阈值的大小，如果当前负载状态值小于预定负载阈值，则负载均衡设备2做出响应，否则继续向PC端返回302返回码。

[0020] 如图2所示，通过网络设备的当前负载状态值与预定负载阈值作比较，来决定是否优先调用专用响应通道向请求端返回重定向报文，改变了现有技术中即使网络设备已经负载却仍然继续重试访问，从而导致访问容易出现问题且具有破坏性影响系统正常运行，本方法有效提高系统的安全性和可用性。

[0021] 上述预定负载阈值根据网络设备的额定工作参数确定，其中额定工作参数至少包括：CPU负荷和/或内存和/或带宽。

[0022] 预定负载阈值可以由配置文件确定，例如，本机或交换机带宽的预定负载阈值可以按照当前带宽的90%或80%的比例来配置；同样CPU负荷、内存也是按比例来配置预定负载阈值的。这样专用响应通道仅仅占用了剩余的10%或20%的小容量空间专门用于向请求端返回重定向报文。

[0023] 网络设备的当前负载状态值根据网络设备的当前负载测量值和上一次负载测量值确定。网络设备的当前负载状态值等于当前负载测量值和上一次负载测量值分别加权后之和，其中，上一次负载测量值的权重大于当前负载测量值的权重。

[0024] 一般情况下，数据采集值会有抖动的情况，例如，宽带可能由原来的1.5G的采集值抖动成1.3G的采集值，这种情况下， $\text{当前负载状态值} = \text{上次测量值} \times X + \text{本次测量值} \times (1-X)$ ，X大于0.5，优选的，X=0.8。

[0025] 当网络设备需要执行多个任务时，优先执行多个任务中向请求端返回重定向报文的任务。

[0026] 即专用响应通道具有最高优先级别，若当前负载状态值超过预定负载阈值，则优先执行专用响应通道向请求端返回重定向报文，避免因重试访问出现的问题，实现了对网

络设备的负载保护,提高了访问效率。

[0027] 当服务请求的协议为HTTP时,重定向报文为302返回码。302返回码的优势主要在于客户端不需要动作,也不需要配置,只要客户端符合标准就可以了。

[0028] 如图3所示,一种网络设备过载保护系统,包括:

[0029] 比较单元,用于在接收到请求端的服务请求时,比较网络设备的当前负载状态值与预定负载阈值的大小,其中预定负载阈值小于所述网络设备的最大负载能力值;

[0030] 响应单元,用于当当前负载状态值达到或超过预定负载阈值时经由专用响应通道向请求端返回重定向报文。

[0031] 上述预定负载阈值根据网络设备的额定工作参数确定,其中额定工作参数至少包括:CPU负荷和/或内存和/或带宽。

[0032] 该网络设备过载保护系统还包括:当前负载状态值确定单元,用于根据网络设备的当前负载测量值和上一次负载测量值确定网络设备的当前负载状态值。网络设备的当前负载状态值等于当前负载测量值和上一次负载测量值分别加权后之和,其中,上一次负载测量值的权重大于当前负载测量值的权重。

[0033] 该网络设备过载保护系统还包括:调度单元,用于当网络设备需要执行多个任务时,优先执行所述多个任务中所述响应单元向所述请求端返回重定向报文的任务。

[0034] 当所述服务请求的协议为HTTP时,重定向报文为302返回码。

[0035] 以上所描述的装置实施例仅仅是示意性的,其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下,即可以理解并实施。

[0036] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到各实施方式可借助软件加必需的通用硬件平台的方式来实现,当然也可以通过硬件。基于这样的理解,上述技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品可以存储在计算机可读存储介质中,如ROM/RAM、磁碟、光盘等,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行各个实施例或者实施例的某些部分所述的方法。

[0037] 最后应说明的是:以上实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

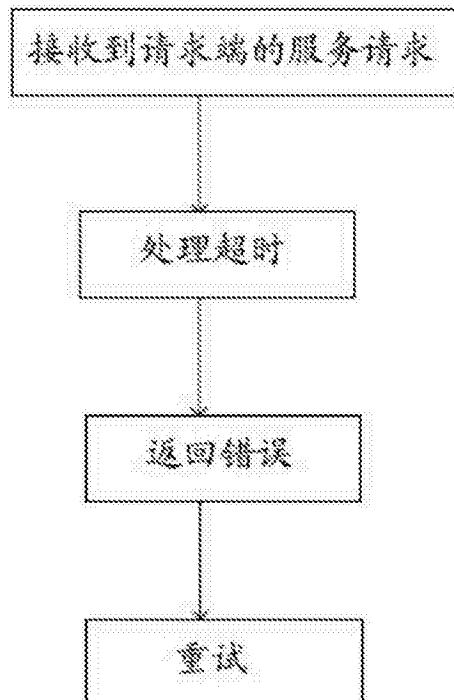


图1

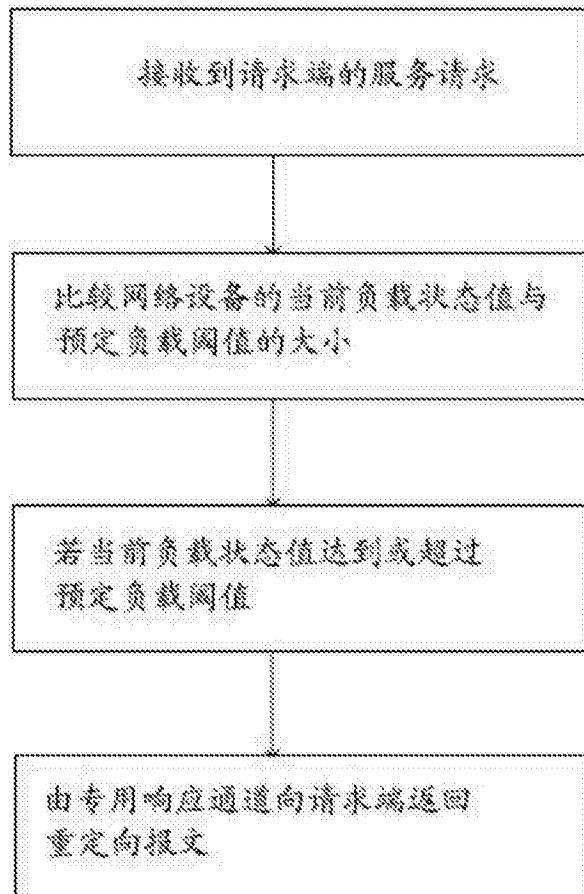


图2

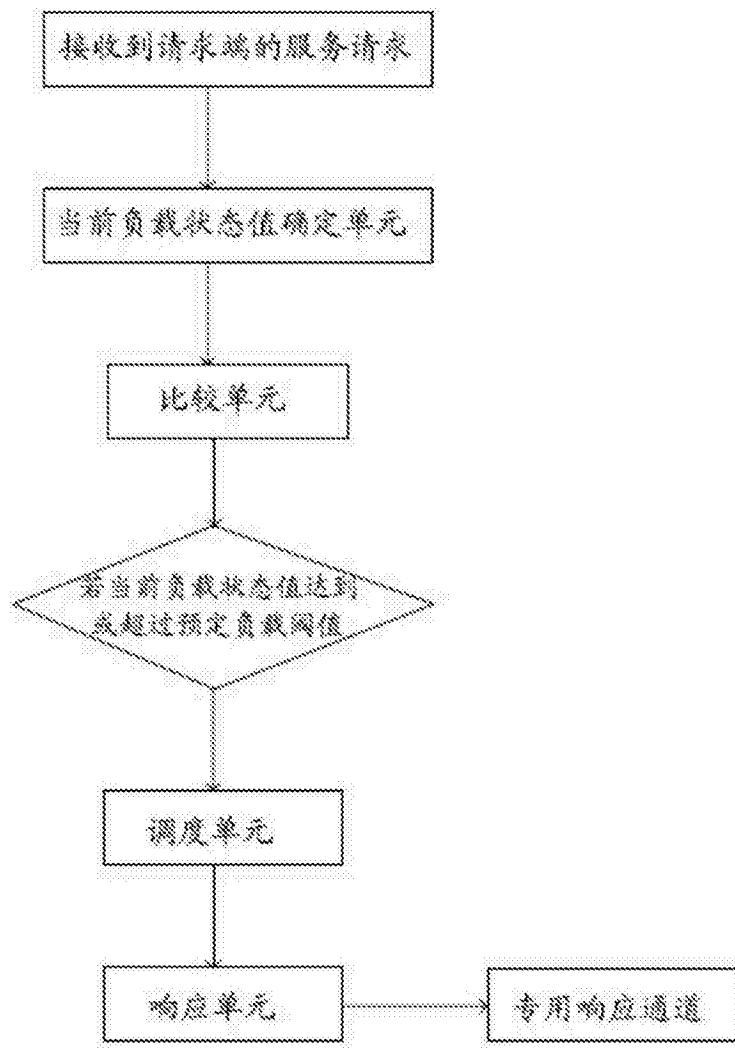


图3