

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.

H04L 12/56 (2006.01)

H04L 12/24 (2006.01)



[12] 发明专利申请公布说明书

[21] 申请号 200710103013.1

[43] 公开日 2008 年 10 月 29 日

[11] 公开号 CN 101296177A

[22] 申请日 2007.4.29

[21] 申请号 200710103013.1

[71] 申请人 华为技术有限公司

地址 518129 广东省深圳市龙岗区坂田华为
总部办公楼

[72] 发明人 郑波 周青 廖洪波 施有铸

[74] 专利代理机构 北京德琦知识产权代理有限公司

代理人 罗正云 宋志强

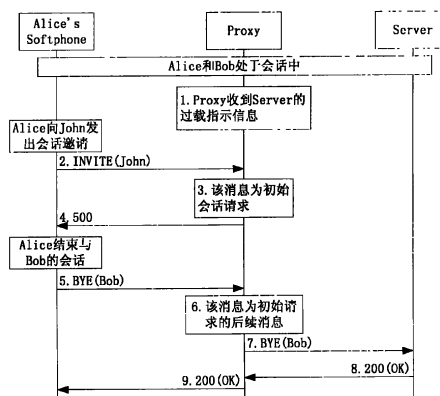
权利要求书 3 页 说明书 8 页 附图 2 页

[54] 发明名称

在分组网络中实现过载控制的方法、系统和装置

[57] 摘要

本发明公开了一种在分组网络中实现过载控制的方法，包括：当服务器过载时，客户端判断接收到的当前服务请求是否满足预先设置的服务请求路由条件，如果满足，则客户端将该当前服务请求路由到服务器。本发明还公开了一种在分组网络中实现过载控制的系统和装置。应用本发明以后，并不是简单地不转发任何请求到过载服务器，而是根据请求的特性进行相应处理，从而提高了分组网络的可用性。另外，本发明还保证了紧急优先的会话能够得到执行。



1、一种在分组网络中实现过载控制的方法，其特征在于，该方法包括：

当服务器过载时，客户端判断接收到的当前服务请求是否满足预先设置的服务请求路由条件，如果满足，则所述客户端将该当前服务请求路由到所述服务器。

2、根据权利要求1所述的在分组网络中实现过载控制的方法，其特征在于，所述预先设置的服务请求路由条件为：当前服务请求为初始请求的后续消息。

3、根据权利要求2所述的在分组网络中实现过载控制的方法，其特征在于，所述初始请求的后续消息包括以下请求中的至少一个：重新邀请 Invite 请求、取消 Cancel 请求、断线 Bye 请求、证实 ACK 请求、查询 OPTIONS 请求、会话初始化协议 SIP 响应消息。

4、根据权利要求1所述的在分组网络中实现过载控制的方法，其特征在于，所述服务请求路由条件为：当前服务请求为一个紧急优先的初始会话请求；

该方法包括：

当服务器过载时，客户端判断接收到的当前服务请求是否为一个紧急优先的初始会话请求，如果是，则所述客户端将该当前服务请求路由到所述服务器。

5、根据权利要求1所述的在分组网络中实现过载控制的方法，其特征在于，所述预先设置的服务请求路由条件为：

预先在客户端设置的服务请求路由条件；或

预先在服务器设置的服务请求路由条件，其中所述服务请求路由条件携带在过载指示信息中被发送到所述客户端。

6、根据权利要求5所述的在分组网络中实现过载控制的方法，其特征在于，所述服务请求路由条件携带在过载指示信息中发送到客户端包括：

服务器在过载后主动发送所述过载指示信息到客户端；或者

服务器在收到客户端的消息后返回所述过载指示信息到客户端。

7、根据权利要求6所述的在分组网络中实现过载控制的方法，其特征在于，

所述客户端的消息包括：客户端主动查询服务器负载情况消息、或者由客户端转发的用户会话请求消息。

8、一种在分组网络中实现过载控制的系统，其特征在于，包括客户端和服务端，其中所述服务器处于过载状态，所述客户端中预先设置有服务请求路由条件；

客户端，用于判断接收到的当前服务请求是否满足所述服务请求路由条件，如果满足，则所述客户端将该当前服务请求路由到所述服务器。

9、根据权利要求8所述的在分组网络中实现过载控制的系统，其特征在于，客户端中配置的所述服务请求路由条件为：当前服务请求为初始请求的后续消息，或当前服务请求为一个紧急优先的初始会话请求。。

10、根据权利要求8所述的分组网络中实现过载控制的系统，其特征在于，所述客户端为用户代理客户端 UAC 或代理服务器 Proxy。

11、一种在分组网络中实现过载控制的系统，其特征在于，包括客户端和服务端，其中所述服务器处于过载状态，

服务器，设置有所述服务请求路由条件，并用于将所述服务请求路由条件携带在过载指示信息中发送到所述客户端；

客户端，用于从过载指示信息中获取服务请求路由条件，并判断接收到的当前服务请求是否满足所述服务请求路由条件，如果满足，则将该当前服务请求路由到所述服务器。

12、根据权利要求11所述的分组网络中实现过载控制的系统，其特征在于，所述客户端包括：

上行单元过载信息处理模块，用于从服务器接收过载指示信息，从所收到的过载信息中获取服务请求路由条件，并将所述服务请求路由条件发送到消息处理模块；

消息处理模块，用于判断接收到的当前服务请求是否满足所述服务请求路由条件，如果满足，则将该当前服务请求路由到所述服务器。

13、一种在分组网络中实现过载控制的装置，其特征在于，该装置包括上

行单元过载信息处理模块和消息处理模块，其中：

上行单元过载信息处理模块，用于从服务器接收过载指示信息，从所收到的过载信息中获取服务请求路由条件，并将所述服务请求路由条件发送到消息处理模块；

消息处理模块，用于判断接收到的当前服务请求是否满足所述服务请求路由条件，如果满足，则将该当前服务请求路由到所述服务器。

在分组网络中实现过载控制的方法、系统和装置

技术领域

本发明涉及下一代网络（NGN）技术领域，更具体地说，本发明涉及一种在分组网络中实现过载控制的方法、系统和装置。

背景技术

随着分组技术的不断成熟，基于电路交换的传统电信网络正在向基于分组交换的宽带电信网发展，而使用会话初始化协议（Session Initiation Protocol, SIP）作为分组电信核心网的呼叫控制信令则是当前的技术发展趋势之一。比如：在国际电信同盟（ITU-T）和欧洲电信标准协会（European Telecommunications Standards Institute, ETSI）等标准组织对下一代网络（Next Generation Network, NGN）的研究过程中，都采用了第3代合作伙伴项目（The third Generation Partnership Project, 3GPP）标准组织所定义的IP多媒体子系统（IP Multimedia Subsystem, IMS）网络架构作为NGN的核心网。

IMS是由3GPP提出的支持IP多媒体业务的子系统。IMS是一个在分组域上的多媒体控制/呼叫控制平台，可以支持会话类和非会话类多媒体业务。IMS的核心特点是采用SIP协议管理多媒体会话。

当网络中的某个网元的资源不足以接纳新的服务请求时，该网元便进入过载状态。网络的规划、不可预料和日益增长的用户服务请求和雪崩效应等，都可能导致网络中某一个或多个网元在某时间进入过载状态。网元的资源可以是该网元用于处理消息的所有能力，包括CPU处理能力、内存、输入/输出资源、磁盘容量等等。上述资源同样包括扩展的资源，比如一个数据库或者DNS服务器。

过载控制就是在网元进入过载状态时进行处理,从而使得网络不瘫痪并能够继续提供服务。显然,过载控制是一个网络商用必须解决的问题。IMS网络采用 SIP 协议管理会话,过载控制也采用了 SIP 协议定义的过载控制机制。

在现有技术的过载控制机制中,定义有 503 (Service Unavailable) 响应码,该响应码由过载的网元发送给客户端。当由于临时过载或者服务器管理导致的服务器暂时不可用时,该服务器在 503 应答中增加一个 Retry-After 指示来让客户端重试这个请求。如果没有 Retry-After 指示,客户端就像收到了一个 500 (Server Internal Error) 应答一样处理。

客户端收到 503 (Service Unavailable) 响应码后,应当尝试转发这个请求到另外一个服务器处理。在 503 响应码的 Retry-After 头域中指定的时间内,客户端不应当转发其他任何请求到这个服务器。作为 503 (Service Unavailable) 的替代,服务器也可以拒绝连接或者把该请求抛弃。

由于初始请求的后续消息并不需要建立新的会话,也不会带来新的负荷,因此如果当前服务请求为初始请求的后续消息,这些后续消息就不应该被拒绝。然而,在现有技术中,当使用了 Retry-After 消息头后,客户端在 Retry-After 头域中指定的时间内,不转发其他任何请求到这个服务器,因此这种处理方式就降低了分组网络的可用性。实际上,某些请求还能够直接释放服务器上的会话,从而节省服务器的资源,因此此时更是不应该被拒绝。

另外,在实际应用中,诸如报警呼叫等紧急的请求应该需要尽量处理,而不应该如现有技术这般简单地予以拒绝。

发明内容

有鉴于此,本发明实施例的主要目的是提出一种在分组网络中实现过载控制的方法,以增强分组网络的可用性。

本发明实施例的另一目的是提出一种在分组网络中实现过载控制的系统,以增强分组网络的可用性。

本发明实施例的再一目的是提出一种在分组网络中实现过载控制的装置，以增强分组网络的可用性。

为达到上述目的，本发明的技术方案是这样实现的：

一种在分组网络中实现过载控制的方法，预先设置服务请求路由条件，该方法包括：

当服务器过载时，客户端判断接收到的当前服务请求是否满足预先设置的服务请求路由条件，如果满足，则所述客户端将该当前服务请求路由到所述服务器。

一种在分组网络中实现过载控制的系统，包括客户端和服务器，其中所述服务器处于过载状态，所述客户端中预先设置有服务请求路由条件；

客户端，用于判断接收到的当前服务请求是否满足所述服务请求路由条件，如果满足，则所述客户端将该当前服务请求路由到所述服务器。

一种在分组网络中实现过载控制的系统，包括客户端和服务器，其中所述服务器处于过载状态，

服务器，设置有所述服务请求路由条件，并用于将所述服务请求路由条件携带在过载指示信息中发送到所述客户端；

客户端，用于从过载指示信息中获取服务请求路由条件，并判断接收到的当前服务请求是否满足所述服务请求路由条件，如果满足，则将该当前服务请求路由到所述服务器。

一种在分组网络中实现过载控制的装置，包括上行单元过载信息处理模块和消息处理模块，其中：

上行单元过载信息处理模块，用于从服务器接收过载指示信息，从所收到的过载信息中获取服务请求路由条件，并将所述服务请求路由条件发送到消息处理模块；

消息处理模块，用于判断接收到的当前服务请求是否满足所述服务请求路由条件，如果满足，则将该当前服务请求路由到服务器。

从上述技术方案中可以看出，在本发明中，当服务器过载时，客户端判

断接收到的当前服务请求是否满足服务请求路由条件,如果满足,则客户端将该当前服务请求路由到所述服务器。比如:如果当前服务请求是一个初始请求的后续消息,客户端将该请求路由到服务器;如果当前服务请求是一个新的会话请求,且该会话请求是一个紧急优先的会话请求,客户端将该请求路由到所述服务器;如果当前服务请求是一个非紧急优先的初始会话请求,客户端可以根据事先收到的响应决定是否将请求路由到服务器。因此当服务器发生过载时,本发明并不是简单地不转发其他任何请求到过载服务器,而是根据请求的特性进行相应处理,从而提高了分组网络的可用性。

另外,在本发明中,如果当前服务请求为紧急优先的初始会话请求,那么依然将该当前服务请求路由到服务器,从而保证了紧急优先的会话能够得到执行。

附图说明

图 1 是根据本发明实施例在分组网络中实现过载控制的示范性方法流程图;

图 2 是根据本发明实施例在分组网络中实现过载控制的方法流程示意图;

图 3 是根据本发明实施例在分组网络中实现过载控制的系统结构示意图;

图 4 是根据本发明实施例在分组网络中实现过载控制的装置结构示意图。

具体实施方式

为使本发明的目的、技术方案和优点表达得更加清楚明白,下面结合附图及具体实施例对本发明再作进一步详细的说明。

图 1 是根据本发明实施例在分组网络中实现过载控制的示范性方法流程图。

如图 1 所示, 该方法包括:

步骤 101: 设置服务请求路由条件。

其中, 本发明实施例可以在客户端中设置服务请求路由条件。本发明实施例也可以在服务器设置服务请求路由条件, 然后再由服务器将服务请求路由条件携带在过载指示信息中发送到客户端。

服务器可以在过载发生后主动发送过载指示信息到客户端, 或者服务器在收到客户端的消息后, 再返回过载指示信息到客户端。此处的客户端的消息可以包括: 客户端主动查询服务器负载情况消息、或者由客户端转发的用户会话请求消息, 等等。

步骤 102: 当服务器过载时, 客户端判断接收到的当前服务请求是否满足服务请求路由条件, 如果满足, 则客户端将该当前服务请求路由到所述服务器。

其中, 服务请求路由条件可以为: 当前服务请求为初始请求的后续消息; 此时, 当服务器过载时, 客户端判断接收到的当前服务请求是否为初始请求的后续消息, 如果是, 则客户端将该当前服务请求路由到所述服务器。

优选地, 这些初始请求的后续消息是指在初始请求之后的后续消息, 可以包括重新邀请 (Invite) 请求、取消 (Cancel) 请求、断线 (Bye) 请求、证实 (ACK) 请求、查询 (OPTIONS) 请求或 SIP 响应消息, 等等。

可选地, 服务请求路由条件还可以为: 当前服务请求为一个紧急优先的初始会话请求。此时, 当服务器过载时, 客户端判断接收到的当前服务请求是否为一个紧急优先的初始会话请求, 如果是, 则所述客户端将该当前服务请求路由到所述服务器, 从而保证了紧急优先的会话能够得到执行。如果当前服务请求是一个新的非紧急优先的会话请求, 客户端可以根据预先配置的服务请求路由条件, 或者服务器的预先过载指示, 决定是否将请求路由到服务器。

图 2 是根据本发明实施例在分组网络中实现过载控制的方法流程示意图。在该流程中, 服务器 (Server) 已经由于某种原因处于过载状态, 而且

艾丽斯（Alice）和鲍比（Bob）处于通话中。

如图 2 所示，方法包括：

步骤 1: Alice 和 Bob 位于会话中，代理服务器（Proxy）收到 Server 的过载指示信息，该过载指示信息可以由 Server 过载后主动发送给 Proxy，也可以是在 Server 收到 Proxy 的消息后返回给 Proxy。

如果 Server 收到 Proxy 的消息后返回过载指示信息，此时 Proxy 的消息可以是 Proxy 主动查询 Server 负载情况的消息，也可以是 Proxy 转发用户的会话请求消息。Proxy 记录下 Server 过载指示信息，便可以按照预先配置，或者 Server 在过载指示信息中携带的拒绝或接受初始会话请求设置执行相关操作。

步骤 2: Alice 向约翰（John）发出新的会话邀请。

步骤 3: Alice 向 John 发出的该初始会话请求到达 Proxy 后，Proxy 判断出该消息为初始会话请求，且该请求路由下一跳为 Server，然后 Proxy 便根据上述步骤中记录的 Server 过载指示信息及 Proxy 的预先配置，或者 Server 在过载指示信息中携带的拒绝或接受初始会话请求设置执行相关操作。

步骤 4: 在此处，Proxy 拒绝 Alice 所发出的新会话邀请，并向 Alice 的用户设备发送 500 消息，从而避免加剧 Server 的过载程度。

步骤 5: Alice 结束与 Bob 的会话，向 Proxy 发送 Bye 请求。

步骤 6-步骤 7: 该 Bye 请求到达 Proxy 后，Proxy 判断出该消息为初始请求的后续消息，且该请求路由下一跳为 Server，Proxy 将该请求转发给 Server。

步骤 8-步骤 9: 尽管 Server 当前处于过载状态，Server 依然处理该 Bye 请求，并最终发送给 Bob 的用户设备，Bob 响应该 Bye 请求，结束和 Alice 的会话。

基于上述描述，本发明实施例还提出了一种在分组网络中实现过载控制的系统。

图 3 是根据本发明实施例在分组网络中实现过载控制的系统结构示意图

图。

如图 3 所示,该系统 300 包括客户端 300 和服务器 301,其中服务器 301 处于过载状态,客户端 300 中预先设置有服务请求路由条件。

客户端 300,用于判断接收到的当前服务请求是否满足所述服务请求路由条件,如果满足,则客户端 300 将该当前服务请求路由到服务器 301。

具体地,客户端 300 中配置的服务请求路由条件可以为:当前服务请求为初始请求的后续消息。那么客户端 300,用于判断接收到的当前服务请求是否为初始请求的后续消息,如果是,则客户端 300 将该当前服务请求路由到服务器 301。

另外,客户端 301 中配置的服务请求路由条件还可以为:当前服务请求为一个紧急优先的初始会话请求。那么此时,客户端 300,用于判断接收到的当前服务请求是否为一个紧急优先的初始会话请求,如果是,则客户端 300 将该当前服务请求路由到所述服务器 301。

其中,客户端既可以是用户代理客户端(UAC),也可以是代理服务器(Proxy)。

可选地,可以在服务器 301 中设置有服务请求路由条件。服务器 301 将服务请求路由条件携带在过载指示信息中发送到客户端 300。此时,客户端 300,用于从过载指示信息中获取服务请求路由条件,并判断接收到的当前服务请求是否满足所述服务请求路由条件,如果满足,则客户端 300 将该当前服务请求路由到所述服务器 301。

图 4 是根据本发明实施例在分组网络中实现过载控制的装置结构示意图。

如图 4 所示,该装置 400 包括过载信息处理模块 401 和消息处理模块 402,其中:

上行单元过载信息处理模块 401,用于从服务器接收过载指示信息,从所收到的过载信息中获取服务请求路由条件,并将所述服务请求路由条件发送到消息处理模块 402;

消息处理模块 402，用于判断接收到的当前服务请求是否满足所述服务请求路由条件，如果满足，则将该当前服务请求路由到服务器。

其中，可以将所述装置 400 嵌入客户端 300 中，或者其它任意合适的网元中。

应用本发明实施例的方法和系统及装置，如果当前服务请求为初始请求的后续消息，那么无论服务器是否过载，都应当将该消息转发到服务器，这就和现有技术存在本质的不同。通过上述处理，服务器处理这种类型的消息能使当前已接纳的会话保持活跃状态而不致吊死。甚至，在某种情况下还能减轻服务器的过载程度。比如：如果服务器处理 Bye 请求最终释放服务器的会话，应用本发明实施例的处理方式就能有效减轻服务器的负载。

另外，如果当前服务请求为紧急优先的初始会话请求，那么无论服务器是否过载，都将该消息转发到服务器，从而可以保证诸如报警电话等紧急呼叫即使在过载的情况下也能够得到处理。

因此，本发明实施例较好地解决了基于 SIP 的分组网络中的过载控制问题，能够显著地增强网络的可用性，并且保证了紧急呼叫即使在过载的情况下也能够得到处理。

以上所述，仅为本发明的较佳实施例而已，并非用于限定本发明的保护范围。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

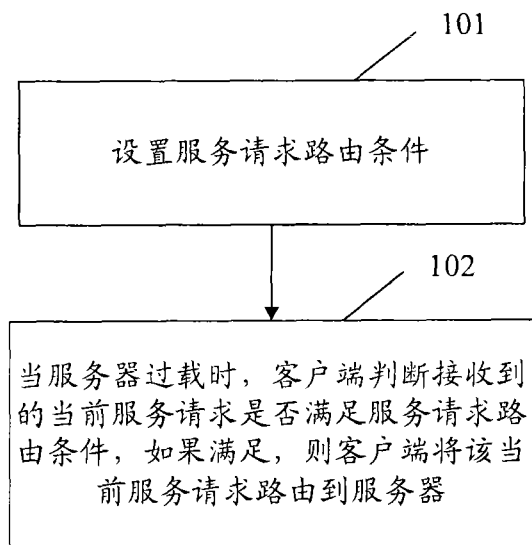


图 1

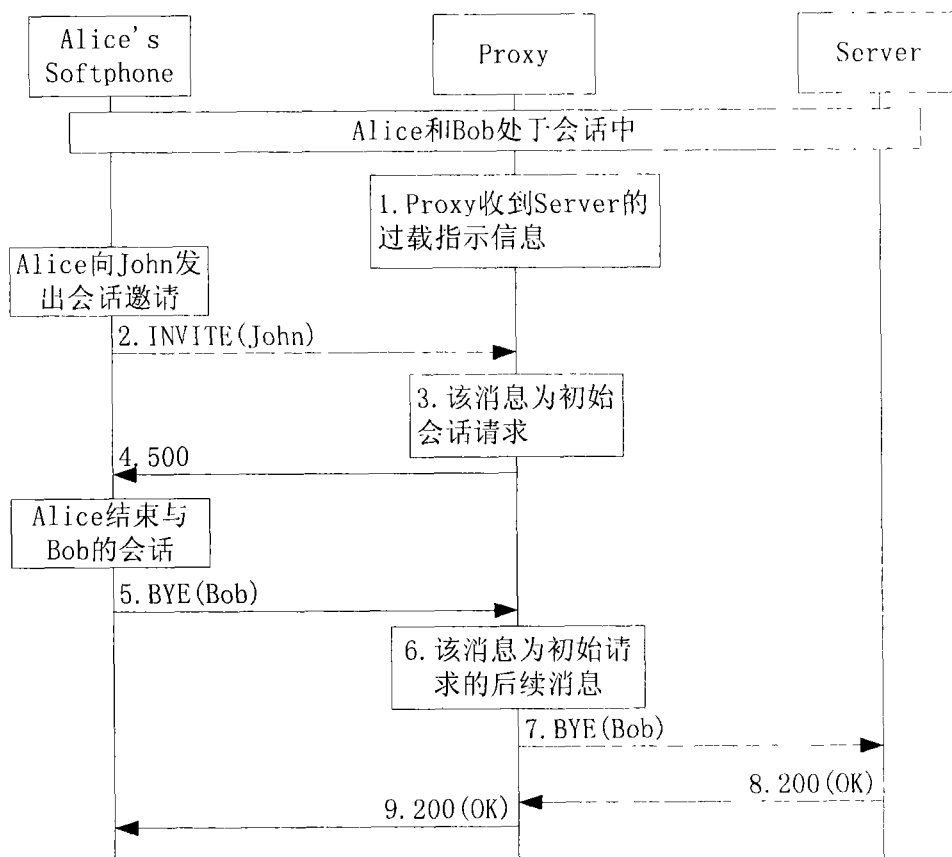


图 2

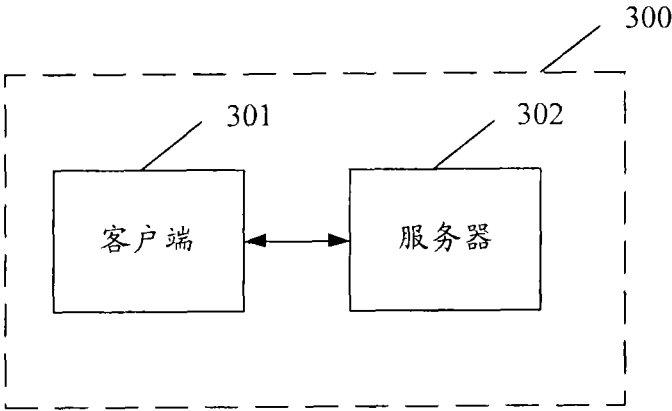


图 3

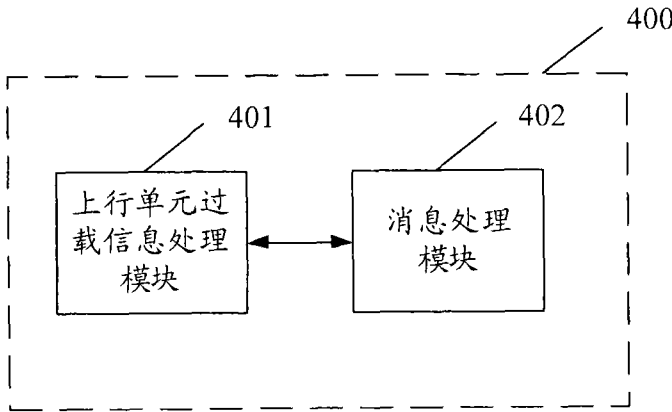


图 4