[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
H04L 12/56 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200780026281.6

[43] 公开日 2009 年 7 月 22 日

[11] 公开号 CN 101491028A

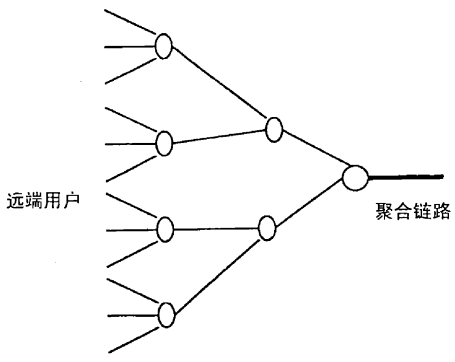
[22] 申请日 2007.7.12
[21] 申请号 200780026281.6
[30] 优先权
[32] 2006. 7.12 [33] US [31] 60/807,088
[86] 国际申请 PCT/CN2007/070283 2007.7.12
[87] 国际公布 WO2008/009235 英 2008.1.24
[85] 进入国家阶段日期 2009.1.12
[71] 申请人 华为技术有限公司
地址 中国广东省深圳市龙岗区坂田华为总部办公楼
[72] 发明人 露西·雍 琳达·邓巴
[74] 专利代理机构 永新专利商标代理有限公司
代理人 林锦辉
权利要求书 3 页 说明书 6 页 附图 1 页

[54] 发明名称

用于控制拥塞的方法

[57] 摘要

一种用于控制聚合链路中的拥塞的方法，包括：分析来自多个远端站点中的每个的数据传输率；确定来自所述多个远端站点的数据传输率是否超过阈值；响应于来自所述多个远端站点的数据传输率超过阈值的确定结果，访问拥塞控制策略；以及根据所述拥塞控制策略，通知所述多个远端站点中的至少一个远端站点降低数据传输率。 本发明的实施例实现了积极主动的拥塞控制管理，并区分了由超额流量导致的拥塞以及由一些链路故障导致的拥塞。 与 PAUSE 机制或丢弃尽力型机制相比，本发明的实施例还为运营商提供了管理网络的灵活性。



1、一种用于控制聚合链路中的拥塞的方法，包括：

分析来自多个远端站点中的每个的数据传输率；

确定来自所述多个远端站点的数据传输率是否超过阈值；

响应于来自所述多个远端站点的数据传输率超过阈值的确定结果，访问拥塞控制策略；以及

根据所述拥塞控制策略，通知所述多个远端站点中的至少一个远端站点降低数据传输率。

2、根据权利要求 1 所述的方法，其中，确定来自所述多个远端站点的数据传输率是否超过阈值包括：

获得所述多个远端站点的超额信息速率（EIR）之和；

确定所述 EIR 之和是否大于所述阈值。

3、根据权利要求 2 所述的方法，其中，所述阈值包括所述聚合链路的带宽。

4、根据权利要求 1 所述的方法，其中，确定来自所述多个远端站点的数据传输率是否超过阈值包括：

确定正在向所述聚合链路移送流量的受影响的远端站点；

获得所述受影响的远端站点的承诺信息速率（CIR）之和；

确定所述 CIR 之和是否大于所述阈值。

5、根据权利要求 4 所述的方法，其中，所述阈值包括：λ%*G，

其中 G 为所述聚合链路的带宽，λ 为 1 到 100 之间的数值。

6、根据权利要求 4 所述的方法，其中，确定所述受影响的远端站点包括：

如果包连接是由信令或人工设置预先确定的，则确定所述受影响的远

端站点；

由监测业务实例的中间节点确定所述受影响的远端站点；或

如果所述聚合链路的链路利用率超过预定阈值，则将所有所述远端站点确定为受影响的远端站点。

7、根据权利要求 2 到 6 中的任一项所述的方法，其中，所述拥塞控制策略包括：

如果来自所述多个远端站点的数据传输率未超过所述阈值，则所述多个远端站点将它们的数据传输率控制在其 EIR 以下；

如果来自所述多个远端站点的数据传输率超过所述阈值，则基于定义的策略计算调节后的 EIR（aEIR）。

8、根据权利要求 7 所述的方法，其中，所述定义的策略包括：

$aEIRi = (\lambda\% * G - (\sum CIRi)) / n + CIRi$

其中 G 为所述聚合链路的带宽，λ 为 1 到 100 之间的数值，n 为远端站点的数量，i 为 1 到 n 之间的常数。

9、根据权利要求 7 所述的方法，其中，所述定义的策略包括：

$aEIRi = (((\lambda\% * G) / \sum CIRi) - 1) * CIRi$

其中 G 为所述聚合链路的带宽，λ 为 0 到 100 之间的数值，i 为 1 到所述远端站点数量之间的常数。

10、根据权利要求 4 或 6 所述的方法，其中，所述拥塞控制策略包括：

如果所述 CIR 之和小于或等于所述阈值，则当前数据传输率在其 CIR 之上的远端站点将其数据传输率控制在所述当前数据传输率以下；

如果所述 CIR 之和大于所述阈值，则为当前数据传输速率在其 CIR 之上的远端站点赋予所计算的调节后的 EIR。

11、根据权利要求 4 或 6 所述的方法，其中，所述拥塞控制策略包括：

如果所述 CIR 之和小于或等于所述阈值，则当前数据传输率在其 CIR

之上的远端站点将其数据传输率控制在所述当前数据传输率以下；

如果所述 CIR 之和大于所述阈值，则当前数据传输率在其 CIR 之上的远端站点降低其数据传输率。

12、根据权利要求 10 或 11 所述的方法，其中，所述拥塞控制策略还包括：

如果大于所述阈值的所述 CIR 之和变为小于或等于所述阈值，则当前数据传输率在其 CIR 之上的远端站点停止控制其数据传输率。

用于控制拥塞的方法

技术领域

本发明通常涉及网络技术，尤其涉及一种用于控制拥塞的方法。

背景技术

拥塞控制是诸如 IP 或以太网网络的非宽带专用传输网的难题。在 IEEE 802.1 和 IEEE 802.3 中讨论了拥塞管理。当前典型的方案是当端口发生拥塞时，首先丢弃所有尽力型包（best effort packet），然后丢弃使其拥塞的其他包，或者向其相邻节点发送 PAUSE 消息，以请求降低流量发送速率。

以太网交换机具有在它们的端口上控制承诺信息速率（CIR）和超额信息速率（EIR）的能力。提供商为用户保证 CIR，允许用户以比 CIR 更高的流量速率进行发送，直至 EIR。如果发送速率超过该 CIR，一些包将被标记为尽力型包，如果在经过路径（traverse path）上没有拥塞，则将传输尽力型包。如果在经过路径上存在拥塞，则将首先丢弃这些包。

有些客户希望利用直到 EIR 的速率发送流量，但希望网络告知他们网络中何时发生拥塞，从而他们能够降低发送速率。这些客户不希望网络在拥塞时简单地丢弃它们的超额流量。

图 1 示出了聚合的接入网。所有用户流量被配置来通过右侧的聚合链路。每个用户端口都具有 CIR 和 EIR。当很多用户以超过 CIR 的流量速率发送时，可能会导致该聚合链路拥塞。一种控制拥塞的方法是在聚合端口处丢弃所有尽力型包。

另一种方法是使用 PAUSE 帧。如果使用 PAUSE 帧来通知远端端点降低速率，则时延可能会导致在拥塞链路上丢弃更多包。当每个远端节点都接收到 PAUSE 帧并降低流量速率时，拥塞点处的速率可能会降得过低。

此外，PAUSE 帧不会在远端节点之间进行协调并协调每个远端节点需要减少多少流量来停止拥塞。

发明内容

本发明的实施例提供了一种防止网络中发生拥塞的拥塞控制机制和协议。该拥塞可能是由远端位置输送的太多流量导致的。本发明的机制将允许网络入口端口在网络中存在潜在拥塞时通知该入口端口的相邻用户降低其输出速率。

一种用于控制聚合链路中的拥塞的方法，包括：

分析来自多个远端站点中的每个的数据传输率；

确定来自所述多个远端站点的数据传输率是否超过阈值；

响应于来自所述多个远端站点的数据传输率超过所述阈值的确定结果，访问拥塞控制策略；以及

根据所述拥塞控制策略，通知所述多个远端站点中的至少一个远端站点降低数据传输率。

本发明创建了一种智能拥塞控制机制，允许通过网络拓扑和带宽利用率来激活拥塞控制。利用链路利用率数据，本发明实现了积极主动的拥塞控制管理。该机制还能够区分由超额流量导致的拥塞和由一些链路故障导致的拥塞。与 PAUSE 机制或丢弃尽力型机制相比，本发明还为运营商提供了管理网络的灵活性。

附图说明

为了更完整地理解本公开及其优点，现在结合附图参考以下描述，在附图中类似的附图标记表示类似的部件：

图 1 示出了聚合以太网接入网的实例。

具体实施方式

本发明的实施例提供了一种所有远端站点之间的机制和协议，以控制来自所有远端站点的流量，使得在聚合所有远端流量的链路上不超过可用带宽（如上图所示）。本发明的一个部分涉及到算法，该算法基于聚合链路利用率和可用带宽计算每个受影响的入口端口处的最大允许速率，以防止拥塞造成的随机包丢弃。本发明的第二部分涉及通知入口端口相应控制其入口速率的信令和消息。

静态闭环反馈控制

该机制使用内部链路利用率和可用带宽计算每个接入节点/端口处的允许入口速率。对于图 1 所示的拓扑而言，可以在所有接入节点/端口之间平均或按比例地划分聚合链路的带宽。可以有多种策略来确定如何由受影响的入口端口划分该可用带宽。

当网络中的一些链路出现故障时，例如在使用 802.3ad（链路聚合）时有些链路出现故障，在该聚合链路上可用的带宽要小得多。在这种情形下，用户可以具有被告知网络内速率下降的选项，以避免对在它们的协议 CIR 之下的流量的随机包丢弃。当然，网络提供商可以选择不通知他们的用户，因为与让其客户知道网络故障相比，他们宁愿在用户无法使用其协议 CIR 时缴纳罚款。

这里是如何使其工作的描述。

假设每个入口端口具有其自己的 CIR 和 EIR，其由 CIR1/CIR2.../CIRn 和 EIR1/EIR2/.../EIRn 表示。拥塞控制器知道每个端口的 CIR 和 EIR 信息。

如果 EIR1+EIR2+...EIRn<=聚合链路带宽（G），那么每个端口仅需要将它们所有的流量控制在其自己的 EIR 之下。

如果 EIR1+EIR2+..+EIRn>聚合链路带宽，则为每个端口分配调节后的 EIR（aEIR）。所述 aEIR 是基于所定义的策略来计算的。

对于如何在所有远端接入节点/端口之间划分聚合链路上的可用带宽，可以有多种策略。本专利公开描述了一些策略和它们的相关算法。

对于多条链路具有拥塞的情况，拥塞控制器将针对所有这些链路计算调节后的 EIR，并且最小的 aEIR 将被通知给受影响的入口端口。

对应于 aEIRi 的一个简单策略是在所有入口端口之间平均划分超过总 CIR 的额外带宽，这可以由下式来表示：

$$aEIRi = (\lambda\% * G(\text{聚合链路}) - (\sum CIRi)) / n + CIRi$$

其中 λ%是为聚合链路保留的阈值，以防止其过分接近整个带宽，λ 是 1 到 100 之间的数值，n 为远端入口用户/VLAN 端口的数量，i 为 1 到 n 之间的常数。

稍微复杂一些的策略可以是：允许每个入口端口超出它们的预定 CIR 速率固定百分比，这可以由下面的算术式表示：

$$aEIRi = ((\ (\lambda\% * G\ (\text{聚合链路})\ ) / (\sum CIRi)) - 1\ ) * CIRi$$

动态闭环反馈控制

在该模型中，流经聚合链路的流量是动态的；这意味着流经该链路的流量来自不同的端口。在上图中，流向树叶的流量可以源于各个非静态的远端节点。因此，该模型需要一种检测机制来检测哪个接入端口正在向该拥挤的链路输送流量。

有多种方式来检测正在向拥挤的链路输送流量的接入端口：

对于面向连接的数据路径，即包路由或路径是预定的

在通过信令（如在 MPLS 信令的情况下）或人工设置预先确定包连接（packer connection）时，每个链路已经具有正在向链路输送流量的网络入口端口的信息。

在这种情形中，可以使用静态闭环反馈控制机制。

真实的情况是并非所有包连接在任何时候都具有流量。静态闭环反馈控制针对其流量流经拥挤链路的所有端口计算调节后的 EIR，以防止可能的拥塞和包丢弃。有可能的是，即使在一个端口用尽超过其 EIR 时，也可能不丢弃包，这是因为某些其他端口未用尽它们指定的 EIR。然而，如果用户不希望丢弃任何包，该控制机制将给予用户选项来控制他们的输出速率。如果用户不关心他们的包丢弃，他们可以发送更多，并且超额的包将被标为“可丢弃”，在发生拥塞时网络可以丢弃它们。

基于业务实例的检测

为了实现该检测级，需要中间节点监测流经链路的业务实例，例如对于以太网流量而言跟踪最近的 SVLAN/BVLAN 或 I-TAG（802.1ah 的情况），对于 IP 网络而言跟踪 VPN-ID。

对于一些交换机（以太网或 IP），已经存在记录当前 VLAN 或 VPN-ID 的机制。对于这些交换机而言，监测流经链路的最近流动不是额外的工作。

该检测机制还需要示出了业务实例和它们的相关接入端口之间的关系的关联表。对于大部分提供商网络而言，在向个体用户销售业务时，该关联表是预先配置的。

所有接入端口的盲控制

对于一些网络运营商而言，与公平相比，简单更重要。运营商可以在一个接入端口的链路利用率超过预定阈值时，选择控制所有接入端口。这种简单控制不要求网络检测流经每条链路的业务实例。

基于入口端口地址的检测

对于一些网络运营商而言，公平更重要。这些运营商可以选择在其流量流经拥挤链路的那些接入端口之间控制速率。为了实现该目标，网络运营商不仅必须监测业务实例，而且还要监测接入端口地址。对于 IEEE 802.1ah 的环境而言，这可能不太困难，因为仅存在很少数量的提供商骨干网地址。

一旦检测到这些接入端口，拥塞控制器将基于链路的链路利用率和可用带宽计算这些端口的调节后的 EIR。同样，不同的策略将规定如何计算每个受影响端口的调节后的 EIR。一种策略实例可以是，如果聚合链路超过预定阈值，则将每个受影响的入口端口控制在其当前速率之下，或者至少基于当前的使用情况来降低速率。

在该模型中，当拥挤链路的利用率正在越过预定阈值时，拥塞控制器将查询所有受影响的端口处的入口速率。假设处于其 CIR 之下的端口将用尽其 CIR，控制器首先将计算拥挤链路是否能够处理在其 CIR 以上的端口之间的当前入口速率。如果答案为是，那么控制器将通知所有受影响的入口端口控制进入速率，以不超过当前水平。如果答案为否，那么控制器将为当前正使用超过其 CIR 的速率的那些端口计算调节后的 EIR 并向它们通知该调节后的 EIR。这里仅是控制策略的一个实例。其他策略可以是在所有受影响的入口端口之间均分可用带宽。

假设允许目前使用低于其 CIR 的速率的端口具有整个 CIR，下面是确定聚合链路是否能够处理当前使用率的公式：

$TOTAL = \sum Ri + \sum CIRj,$

其中 Ri 表示其当前速率>CIRi 的端口的当前使用率，CIRj 表示其当前速率<CIRi 的端口的当前使用率。

如果 TOTAL <=λ% * G (聚合链路)

这表示拥挤链路上的可用带宽能够处理当前的使用率，拥塞控制器将通知当前入口速率高于其 CIR 的所有入口端口将其当前速率用作 aEIR，即，不允许额外的入口速率。

否则，拥塞控制器将通知其当前速率高于其 CIR 的所有入口端口将其当前速率降低一个百分比。

当链路利用率降到预定阈值之下时，拥塞控制器将向所有入口端口发送消息以停止速率控制。

同样的机制也可以应用于以太网虚拟 LAN。假设多个用户利用同样的 S-VLAN 形成以太网虚拟 LAN。网络能够监测每个 S-VLAN 的性能数据。控制器能够收集每个 S-VLAN 的每个入口的实际用户速率。那么，该计算将基于总的允许虚拟 LAN 带宽和每 S-VLAN 的实际总速率。

本发明创建了一种智能拥塞控制机制，并允许通过网络拓扑和带宽利用率来激活拥塞控制。利用链路利用率数据，本发明实现了积极主动的拥塞控制管理。该机制还能够区分由超额流量导致的拥塞和由一些链路故障导致的拥塞。与 PAUSE 机制或丢弃尽力型机制相比，本发明还为运营商提供了管理网络的灵活性。

尽管在此已经展示和描述了本发明的一些优选实施例，但在不脱离本发明的精神和教导的情况下，本领域的技术人员可以对其做出修改。这里所述的实施例仅仅为示范性的，并非意在限制。这里披露的本发明的很多变化、组合和修改都是可能的，且在发明范围之内。因此，保护范围不受上面给出的描述限制，而是由以下权利要求界定，该范围包括权利要求主题的所有等价物。

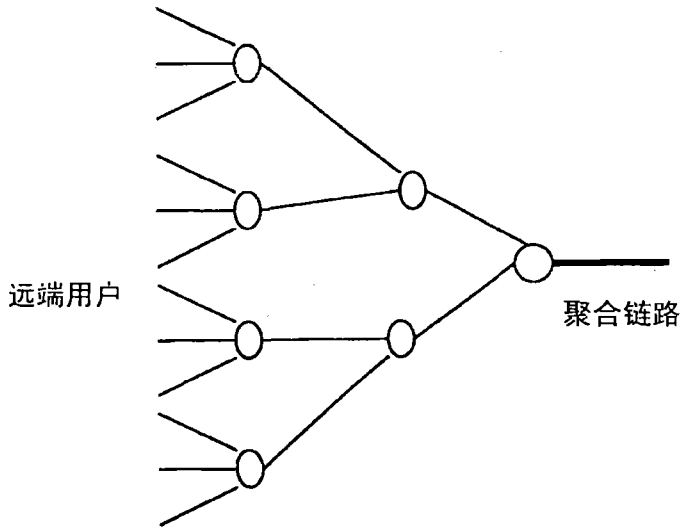


图1