(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 101159694 B
(45)授权公告日 2011.04.06

(21)申请号 200710187234.1

(22)申请日 2007.11.16

(73)专利权人 中兴通讯股份有限公司
地址 518057 广东省深圳市南山区高新技术产业园科技南路中兴通讯大厦法务部

(72)发明人 刘明 秦春华 杨长江

(74)专利代理机构 信息产业部电子专利中心 11010
代理人 梁军

(51)Int.Cl.
H04L 12/56(2006.01)
H04L 29/12(2006.01)
H04L 29/06(2006.01)

(56)对比文件
CN 101035082 A,2007.09.12, 说明书全文 .
CN 100420238 A,2006.11.08, 说明书全文 .
US 2003200328 A1,2003.10.23, 说明书全文 .
CN 100448225 A,2007.04.04, 说明书全文 .
CN 1592215 A,2005.03.09, 说明书全文 .

审查员 梁年顺

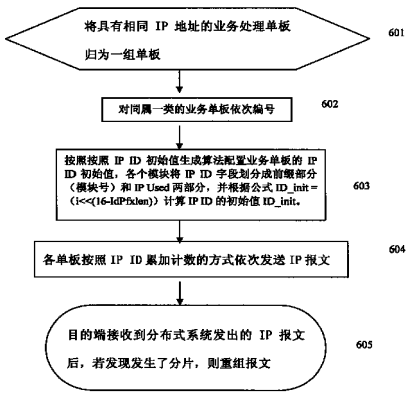
权利要求书 1 页 说明书 4 页 附图 3 页

(54)发明名称

一种IP共享的分布式系统避免分片重组失败的方法

(57)摘要

本发明公开了一种成本低、易于实现的IP共享的分布式系统避免分片重组失败的方法，包括以下步骤：将具有相同IP地址的业务处理单板归为一组单板；将属于同一组的单板依次编号，设为模块号i，经过编号的单板设为模块；将IP ID字段划分为模块号i和可变部分IP Used，计算IP ID的初始值ID_init = (i << (16-IdPfxlen))，其中IdPfxlen为IP ID字段的前缀长度；各单板按照IP ID累加计数的方式依次发送IP报文；目的端接收到IP报文后若发现发生了分片则重组报文。本发明方法对报文传送路径上的转发设备和应用层以及终端应用层没有特别的要求，因而成本较低，易于实现。



CN 101159694 B

1. 一种 IP 共享的分布式系统避免分片重组失败的方法，其特征在于包括如下步骤：

步骤一，将具有相同 IP 地址的业务处理单板归为一组单板；

步骤二，将所有属于同一组的单板依次编号，设为模块号 i，i 为大于等于 0 的整数，经过编号的单板设为模块；

步骤三，各个模块将 IP ID 字段划分为前缀部分和可变部分 IP Used，其中前缀部分用于描述模块号 i；如果 IP Used 字段值溢出，则 IP Used 字段值重新从 0 开始计算；

根据公式 ID_init = (i << (16-IdPfxlen)) 计算 IP ID 的初始值 ID_init，式中 IdPfxlen 为 IP ID 字段的前缀长度，i 的取值应满足 i <= 2^IdPfxlen；式中符号“<<”指按位左移；

步骤四，各单板按照 IP ID 累加计数的方式依次发送 IP 报文；

步骤五，目的端接收到所述 IP 报文后，若发现其发生了分片，则对其进行重组。

2. 根据权利要求 1 所述的 IP 共享的分布式系统避免分片重组失败的方法，其特征在于：在所述步骤一中，分布式系统分配多个对外 IP 地址；各个 IP 地址分别被多个业务处理单板共享。

3. 根据权利要求 1 所述的 IP 共享的分布式系统避免分片重组失败的方法，其特征在于：在所述步骤一中，分布式系统分配多个对外 IP 地址；各个 IP 地址分别被分配至一个业务处理单板。

4. 根据权利要求 1 所述的 IP 共享的分布式系统避免分片重组失败的方法，其特征在于：在所述步骤一中，分布式系统分配一个对外 IP 地址，该 IP 地址被多个业务处理单板共享。

5. 根据权利要求 1、2、3 或 4 所述的 IP 共享的分布式系统避免分片重组失败的方法，其特征在于：在所述步骤二中，同一组单板依次从 1 开始编号，不同组的单板允许具有相同的编号。

一种 IP 共享的分布式系统避免分片重组失败的方法

技术领域

[0001] 本发明涉及 IP 共享的分布式系统领域，尤其涉及一种 IP 共享的分布式系统避免分片重组失败的方法。

背景技术

[0002] 网络与通信是当今最热和发展最快的领域之一，为了应付日益繁忙的信息流，网络的速度由几年前的低速链路发展到目前的 10Gb/s 以上。同时通讯设备的处理能力不断提高，各种大容量节点设备不断出现，单 CPU、集中式处理事务已经越来越不能适应市场的需求。如今，各制造商普遍采用分布式处理技术，将通讯业务的处理分散在多个模块、多个 CPU 上加以处理。

[0003] 然而，为了给运营商和最终用户提供方便，同时也为了与现有的设备更好的兼容，制造商们仍然希望其分布式处理系统能够只对外暴露少数几个或者仅一个业务 IP 地址，对于客户端而言，只能见到这些对外暴露的 IP 地址，或者在客户端看来与之交互的是一台性能十分强劲的服务器。

[0004] 通常，分布式处理系统可以同时处理多个业务流，每块处理单板预先约定好处理的业务范围；比如通过报文的端口范围来区分每块单板处理的业务；这样，分布式系统发送的报文即使源 IP 地址相同，也可以通过端口来区分；对于报文的接收处理，我们在分布式系统的外围接口单板上配置业务流分发表，根据 IP 地址、端口号（有时包括协议字段）来区分业务流，按照分发规则将报文分发到特定的处理单板上。

[0005] 对于上述的共享 IP 的分布式处理系统，报文发送所至目的端可能是个纯粹的处理单元也可能是另一个分布式系统；这两种情况下都可能存在这样的问题：接收端接收到的报文可能来自分布式系统的不同业务流，这些来自分布式系统的不同业务流的报文有着相同的 IP 地址、不同的端口号，它们在源端或者路由转发至接收端的中途被分片，而分片报文除了第一个分片会包含端口信息外，后续分片都没有。因此这些拥有相同 IP 的分片报文无法在接收端正确重组。

[0006] 为了避免和解决这个问题，人们常常在发送报文的时候先确定至目的端整条路径的最小 mtu 即路径 mtu 值，发送的报文长度小于该 mtu 值，从而避免发送报文在路径上被分片， 这种方式的缺点是，沿该条路径上的所有转发设备都要求支持路径 mtu 机制，并且要求应用层支持报文按路径 mtu 封装、发送报文；此外，路由的变化会导致路径 mtu 的变化，这些变化必须让终端的应用层感知，实现起来会有一定的复杂度并且成本较高。

发明内容

[0007] 本发明要解决的技术问题是提供一种成本低、易于实现的 IP 共享的分布式系统避免分片重组失败的方法。

[0008] 为解决上述技术问题，本发明方法包括如下步骤：

[0009] 步骤一，将具有相同 IP 地址的业务处理单板归为一组单板；

[0010] 步骤二，将所有属于同一组的单板依次编号，设为模块号 i，i 为大于等于 0 的整数，经过编号的单板设为模块；

[0011] 步骤三，各个模块将 IP ID字段划分为前缀部分和可变部分，其中其中前缀部分描述模块号 i；如果 IP Used 字段值溢出，则可变部分 IP Used 字段值重新从 0 开始计算；

[0012] 根据公式 ID_init = (i<<(16-IdPfxlen)) 计算 IP ID 的初始值 ID_init，式中 IdPfxlen 为 IP ID 字段的前缀长度，i 的取值满足 i< = 2^IdPfxlen；式中符号“<<”表示按位左移；

[0013] 步骤四，各单板按照 IP ID 累加计数的方式依次发送 IP 报文；

[0014] 步骤五，目的端接收到所述 IP 报文后，若发现其已发生分片，则对其进行重组。

[0015] 在所述步骤一中，分布式系统可以分配多个对外 IP 地址，也可以只有一个 IP 地址；各个 IP 地址被多个业务处理单板共享或者被分配至一个业务处理单板。

[0016] 所述步骤二中，同一组单板依次从 1 开始编号，不同组的单板允许具有相同的编号。

[0017] 所述步骤三中，IP ID 生成算法可以确保相同 IP 的不同模块发送的 IP 报文其 IP ID 值不会重叠。

[0018] 本发明 IP 共享的分布式系统避免分片重组失败的方法，对报文传送路径上的转发设备和应用层以及终端应用层没有特别的要求，因而成本较低，易于实现。

附图说明

[0019] 图 1 是本发明应用的分布式系统共享 IP 模块示意图；

[0020] 图 2 是本发明中 IP ID 字段划分示意图；

[0021] 图 3 是本发明中 IP ID 初始值生成算法流程图；

[0022] 图 4 是本发明中分布式系统报文分片示意图；

[0023] 图 5 是本发明中分布式系统报文重组示意图；

[0024] 图 6 是本发明 IP 共享的分布式系统避免分片重组失败的方法流程图。

具体实施方式

[0025] 下面结合附图对本发明的技术方案进行详细说明。

[0026] 图 1 是本发明应用的分布式系统共享 IP 模块示意图，左图部分 A1、A2、A3 共享一个 IP 地址，模块编号依次是 1、2、3；B1、B2 共享一个 IP，模块编号依次是 1、2。

[0027] 图 2 是本发明中 IP ID 字段划分示意图，协议中 IP 报头的 ID 字段共占用 16 位，IP ID 被划分成两部分，即前缀部分和可变部分，其中高位段是前缀部分，前缀部分描述该报文所属模块号 i，低位段是 ID 字段可变部分。

[0028] 图 3 是本发明中 IP ID 初始值生成算法流程图，该算法中的 i、IdPfxlen 预先配置给各个模块，包括以下步骤：

[0029] 步骤 301，流程开始；

[0030] 步骤 302，读取所属模块号；

[0031] 步骤 303，读取配置的模块前缀长度；

[0032] 步骤 304，根据计算公式进行计算；

[0033] 步骤 305，获得 IP ID 初始值。

[0034] 图 4 和图 5 是本发明中分布式系统报文分片和重组示意图；图 4 的 B1、B2 共享一个 IP，编号分别是 1、2；B1 和 B2 发送的 IP 分片经过路由转发发送至 M；虽然分片的源 IP、目的 IP 相同，M 仍然能够根据报文 ID 正确重组。

[0035] 本发明主要内容是在共享 IP 的分布式系统中提供有效的 IP ID 生成策略，解决多处理板的分片在接收端重组不正确的问题；该 ID 生成策略包括如下：

[0036] 1) 将分布式系统中共享 IP 的处理板归为一组单板，并依次编号，设为模块号 i；每一组单板编号都从 1 开始；

[0037] 2) 每一组单板配置一个 IP ID 前缀长度 IdPfxlen，IdPfxlen 需要满足如下约束条件：

[0038] 组中单板的模块号 i 不大于 2 的 IdPfxlen 次方，即 i< = 2^IdPfxlen；

[0039] 3) 单板在 IP 模块初始化时，按照图 3 的 IP ID 初始值生成算法计算出单板的 IP ID 初始值 ID_init，该算法公式为：ID_init = (i<<(16-IdPfxlen))；

[0040] 4) IP 模块每发送一个 IP 报文，IP ID 的 IP Used 字段值加 1，该 IPUsed 字段的初始值为 0；

[0041] 5) 如果 IP Used 字段值溢出，则 IP Used 字段值重新从 0 开始计算；

[0042] 该策略能保证分布式系统共享 IP 的多块处理单板发出的 IP 报文不会有 IP ID 值的重叠。

[0043] 下面描述 IP 模块采用本发明的 IP ID 生成策略后报文的分片重组过程：

[0044] 1) 假设 B1 和 B2 是分布式系统中共享 IP 的两块单板，M 是与之通讯的一台主机；分布式系统与 M 之间的路径 mtu 值是 1500 字节，如图 4；

[0045] 2) B1 和 B2 配置的 IP 地址是 201.1.1.1，模块号分别是 1、2，两块单板的 IP ID 前缀长度配成 5；M 的 IP 地址是 201.1.2.1；

[0046] 3) 按照上述的 IP ID 生成策略，B1 发送的 IP 报文 ID 范围是 (2048 ～ 4095)，B2 的 IP ID 范围是 (4096 ～ 6143)；

[0047] 4) B1 和 B2 向 M 发送 IP 报文，报文长度都是 4000 字节；B1 发出的报文被分成 3 个分片，B2 发出的报文同样被分成 3 片；

[0048] 5) M 收到 B1 和 B2 发来的分片报文，由于 B1 和 B2 发出的报文 IP ID 不会出现相同值；按照通用的 IP 重组算法，B1 和 B2 的分片报文不会在一个重组队列重装；

[0049] 6) 经过重装后，B1 和 B2 的 IP 报文交由相应模块处理。

[0050] 图 6 是本发明 IP 共享的分布式系统避免分片重组失败的方法流程图，如图所示，本发明方法包括以下步骤：

[0051] 步骤 601，对共享 IP 的业务单板进行归类，将具有相同 IP 地址的业务处理单板归为一组单板；

[0052] 步骤 602，将所有属于同一组的单板依次编号，设为模块号 i，i 为大于等于 0 的整数，经过编号的单板设为模块；

[0053] 步骤 603，各个模块将 IP ID 字段划分成前缀部分 (模块号 i) 和可变部分 (IP Used) 两部分，并根据公式 ID_init = (i<<(16-IdPfxlen)) 计算 IP ID 的初始值 ID_init，式中 IdPfxlen 为 IP ID 字段的前缀长度，IdPfxlen 的取值满足 i< = 2^IdPfxlen；式中符

号“<<”指按位左移；

[0054] 步骤604，IP ID初始值配置结束，各单板按照IP ID累加计数的方式依次发送IP报文；

[0055] 步骤605，目的端接收到分布式系统发出的IP报文后，若发现IP报文发生了分片，则对该IP报文进行重组。

[0056] 本发明所提出的共享IP分布式系统中IP ID分片策略，根据每块单板所属模块号和配置的IP ID前缀长度来计算IP ID的初始值，通过划分IP ID范围有效地避免单板发送的分片报文重组失败；这种IP ID生成策略只是调整IP ID，不需要改动其他协议模块的流程，并且对于报文接收端完全透明，具有很强的实用性。

[0057] 以上所述的实施例只是本发明方法的一个实现方式描述，该部分的功能完全可以在其他物理实体中实现。在不脱离本发明的精神和范围的情况下，所有的变化和修改都在本发明的保护范围之内。

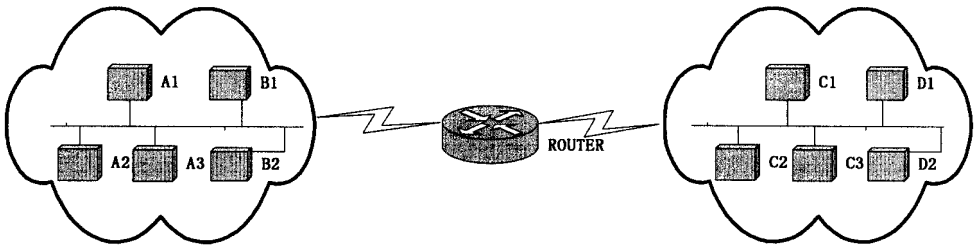


图 1

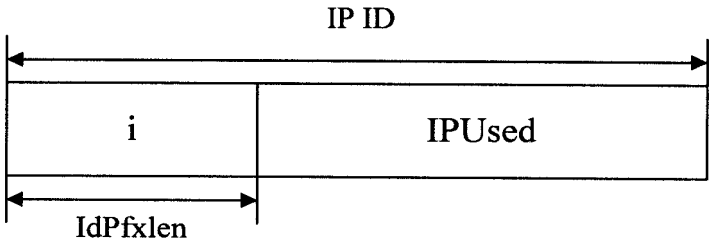


图 2

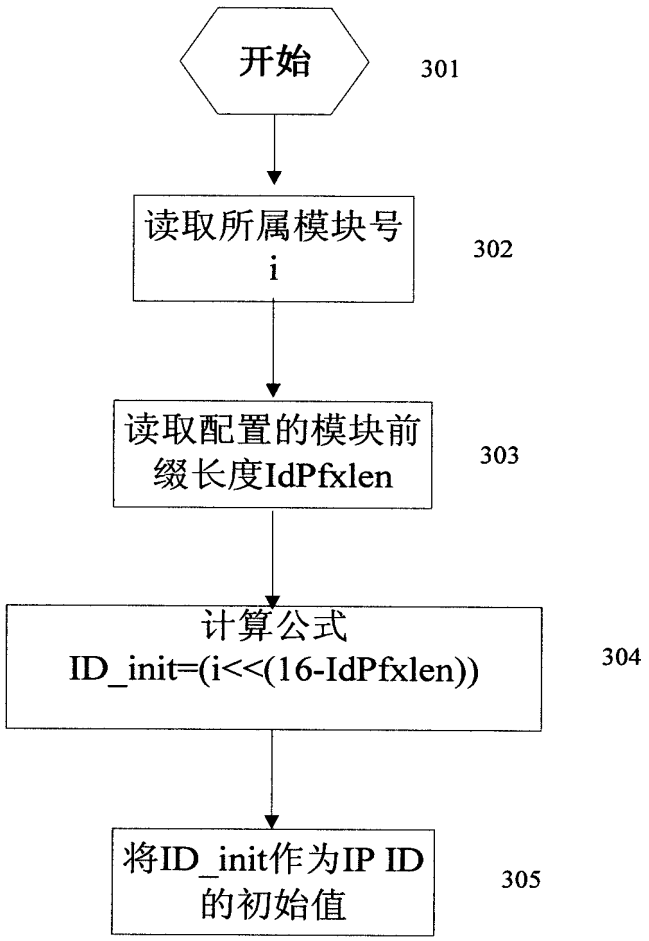

开始
301
读取所属模块号
i
302
读取配置的模块前
缀长度IdPfxlen
303
计算公式
ID_init=(i<<(16-IdPfxlen))
304
将ID_init作为IP ID
的初始值
305

图 3

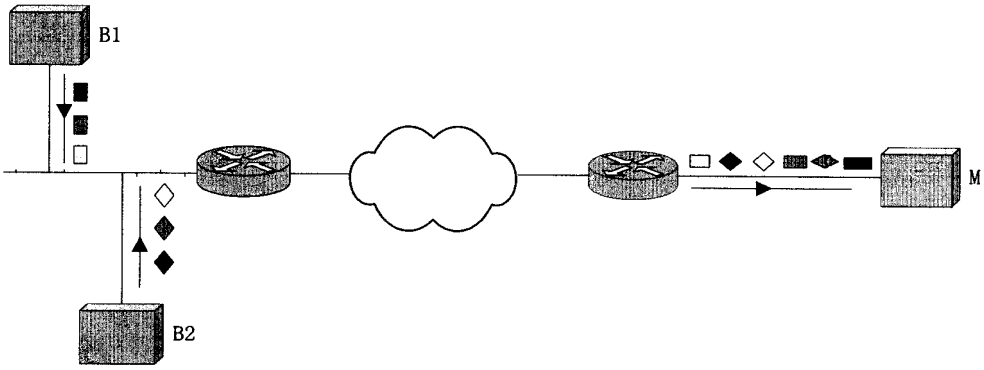

B1
B2
M

图 4

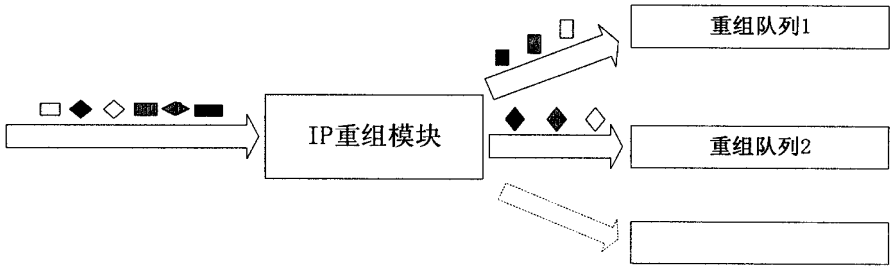

重组队列1
IP重组模块
重组队列2

图 5

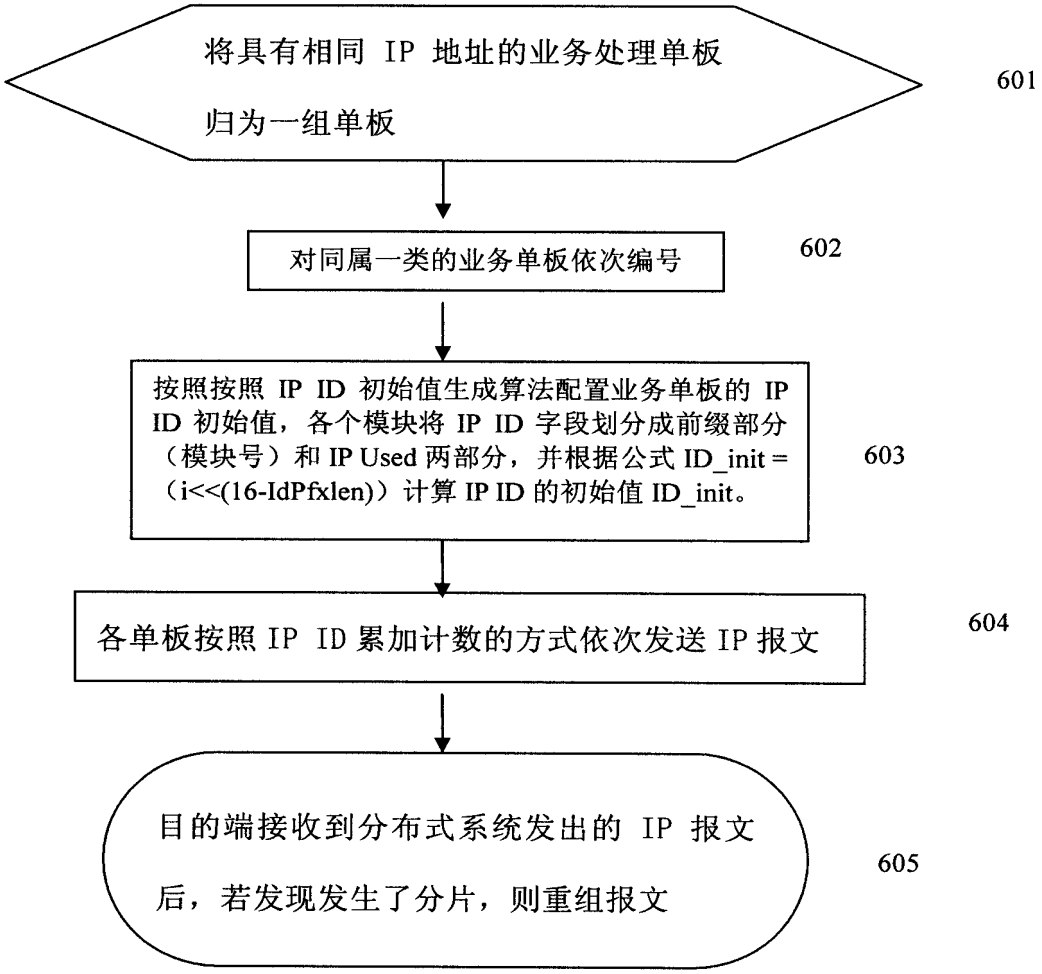

将具有相同 IP 地址的业务处理单板
归为一组单板
601
对同属一类的业务单板依次编号
602
按照按照 IP ID 初始值生成算法配置业务单板的 IP ID 初始值，各个模块将 IP ID 字段划分成前缀部分（模块号）和 IP Used 两部分，并根据公式 ID_init = （i<<(16-IdPfxlen)）计算 IP ID 的初始值 ID_init。
603
各单板按照 IP ID 累加计数的方式依次发送 IP 报文
604
目的端接收到分布式系统发出的 IP 报文后，若发现发生了分片，则重组报文
605

图 6