

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织  
国际局

(43) 国际公布日  
2012年11月8日 (08.11.2012)



(10) 国际公布号  
WO 2012/149857 A1

- (51) 国际专利分类号:  
H04L 12/56 (2006.01)
- (21) 国际申请号: PCT/CN2012/073735
- (22) 国际申请日: 2012年4月10日 (10.04.2012)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:  
201110115794.2 2011年5月5日 (05.05.2011) CN
- (71) 申请人 (对除美国外的所有指定国): **中兴通讯股份有限公司 (ZTE CORPORATION)** [CN/CN]; 中国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, Guangdong 518057 (CN)。
- (72) 发明人; 及
- (75) 发明人/申请人 (仅对美国): **孙延涛 (SUN, Yantao)** [CN/CN]; 中国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, Guangdong 518057 (CN)。 **刘真 (LIU, Zhen)** [CN/CN]; 中国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, Guangdong 518057 (CN)。 **方维维 (FANG, Wei-**

wei) [CN/CN]; 中国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, Guangdong 518057 (CN)。 **刘强 (LIU, Qiang)** [CN/CN]; 中国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, Guangdong 518057 (CN)。

(74) 代理人: **北京派特恩知识产权代理事务所 (普通合伙) (CHINA PAT INTELLECTUAL PROPERTY OFFICE)**; 中国北京市海淀区海淀南路21号中关村知识产权大厦B座2层, Beijing 100080 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA,

[见续页]

(54) Title: ROUTING METHOD FOR DATA CENTER NETWORK SYSTEM

(54) 发明名称: 数据中心网络系统的路由方法

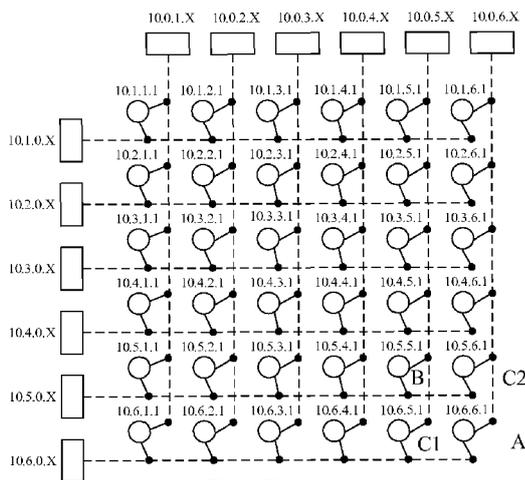


图 1 / Fig.1

(57) Abstract: The present invention relates to a routing method for a data center network system. The data center network system comprises row header switches, column header switches, and access switches arranged in an array. The row header of each row is provided with at least one row header switch. The column header of each column is provided with at least one column header switch. Servers are connected to the access switches. The access switches are connected to all row header switches of the row where the access switches are located and to all column header switches of the column where the access switches are located. In the present invention, the servers of a same subnet communicate therebetween via access switches connected thereto, and servers of different subnets need to communicate therebetween via the access switches, the row header switches, and the column header switches. The routing method employed in the present invention is simplified, efficient, allows for convenient implementation via hardware, and for great routing speed.

(57) 摘要:

[见续页]

WO 2012/149857 A1



RW, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF,

CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

**本国际公布:**

— 包括国际检索报告(条约第 21 条(3))。

---

本发明涉及一种数据中心网络系统的路由方法。该数据中心网络系统包括行首交换机、列首交换机和呈矩阵排列的接入交换机，每行的行首至少部署一个行首交换机，每列的列首至少部署一个列首交换机，服务器与接入交换机相连接，接入交换机与其所在行的所有行首交换机及其所在列的所有列首交换机相连接。在本发明中，同一子网的服务器之间通过与其相连接的接入交换机进行通信，不同子网的服务器之间需要通过接入交换机、行首交换机和列首交换机进行通信。本发明采用的路由方法更为简单、高效，便于通过硬件实现，路由速度快。

## 数据中心网络系统的路由方法

### 技术领域

本发明涉及一种路由技术领域，尤其涉及一种数据中心网络系统的路由方法。

### 5 背景技术

数据中心是企业各种应用服务的提供中心，是数据运算、交换和存储的中心。它结合了先进的网络技术和存储技术，承载了网络中 80% 以上的服务请求和数据存储量，为客户业务体系的健康运转提供服务和运行平台。

数据中心最早出现在 20 世纪 60 年代初。随着互联网的快速建设和信息技术的迅猛发展，到 20 世纪 90 年代中后期，数据中心进入了蓬勃发展期，建设规模和服务器数量每年都以惊人的速度增长。互联网技术的蓬勃发展掀起了建设数据中心的高潮，不但政府机构和金融电信等大型企业扩建自己的数据中心，中小企业也纷纷构建数据中心。自 2006 年 Google 公司提出云计算以来，在亚马逊、微软、雅虎、IBM 等 IT 公司的大力推动下，云计算技术得到长足发展，美国、韩国、日本政府都宣布了国家云计算发展战略。云计算的发展进一步带动数据中心的迅速发展，数据中心网络规模不断扩大，目前一个大型数据中心可能包含数万台服务器。

随着数据中心规模的日益扩大，数据中心容纳的服务器数量也越来越多，因此需要巨大的上层网络带宽支持。数据中心网络的典型拓扑结构是由路由和交换单元组成的类似树形的网络结构，其上层网络为了支持大量的带宽需求不得不采用昂贵的专用设备。问题是，即使采用最高端的 IP 交换机或路由器，核心层也是只能支持到 50% 的边缘网络汇集的带宽，而且需要巨大的费用开销。因此树形拓扑结构不可避免地会在上层核心网络产

生通信瓶颈，从而导致网络在传输延迟、传输效率等方面整体性能的下降。另外，在数据中心，这种非对称的网络带宽还会为应用设计带来复杂性。

为了解决上层核心网络带宽不足带来的通信瓶颈问题，目前通过检索到的方法来看，一般采用 Clos 网络或者胖树 (Fat Tree) 拓扑组成无阻塞网络，并根据拓扑结构的特点，提出相应的路由/交换方法，比如 Fat Tree 结构、Clos 网络、多分支胖树网络拓扑结构。另外，还有些方法利用服务器的多网卡技术，把服务器同时连接到多个交换机/路由器上，从而增加服务器之间的连接数量，来解决上层核心网络带宽不足的问题。这些方法和传统的树形结构相比，具有较高的二分带宽 (bisection bandwidth)，并且每一层路由/交换设备的超额订购 (Oversubscription) 比例都可以达到 1:1，因此可以有效消除顶层带宽不足带来的网络瓶颈问题。

上述网络结构虽然解决了构建大规模数据中心网络的上层带宽瓶颈问题，但仍然存在如下缺点：(1) 由于其拓扑结构的限制，网络流量的超额订购 (Oversubscription) 比例很难按照实际需求进行灵活的调整，网络配置的灵活性差；(2) 其网络拓扑结构比较复杂，设备编址需要遵循严格复杂的规则，设备端口之间按照严格的顺序进行连接，这些问题导致数据中心在网络布线和设备部署方面比较繁琐，运行过程中网络维护也会相对比较困难；(3) 由于拓扑结构复杂，导致其路由算法也相对比较复杂；(4) 在构建规模较小的数据中心网络时，会存在端口空余浪费的情况，网络伸缩性较差。

## 发明内容

本发明提出了一种数据中心网络系统的路由方法，该数据中心网络系统采用的是称为交换式矩阵的网络结构。该方法可以充分发挥交换式矩阵网络拓扑结构的特点和优势，解决数据中心网络中的通信瓶颈问题。在该路由方法中，构建路由表只需要网络设备交换很少量信息，构建方法简单

易行，所构建的路由表规模小，路由速度快。此外，该路由方法可以充分利用源和目的节点之间的多条并行链路实现无阻塞路由和负载均衡。

实现本发明中的路由方法所基于的数据中心网络系统包括交换机和服务  
5 器两类设备，所述交换机包括行首交换机、列首交换机和呈矩阵排列的  
接入交换机，矩阵的每行行首至少部署一个行首交换机，每列列首至少部  
署一个列首交换机，服务器与接入交换机相连接，每个接入交换机与其所  
在行的所有行首交换机及其所在列的所有列首交换机相连接，任意行首交  
换机和列首交换机之间、以及各接入交换机之间不直接相连。在该网络结  
构中，交换机和服务器采用内部网络 IP 地址，并按照如下规则进行编址：

10 行首交换机的 IP 地址配置为 10.Row.0.X；列首交换机的 IP 地址配置为  
10.0.Col.X；接入交换机的 IP 地址配置为 10.Row.Col.1；服务器的 IP 地址  
配置为 10.Row.Col.X。其中 Row 为交换机或服务器所在行的行号，Col 为  
交换机或服务器所在列的列号；对于行首/列首交换机， $0 < X \leq 255$ ，对于  
服务器， $1 < X \leq 255$ 。所有设备的子网掩码设为 255.255.255.0。

15 本发明中的数据中心网络系统的路由方法为：同一子网的服务器之间  
通过与其相连接的接入交换机进行通信，不同子网的同行的服务器之间通  
过与其相连接的接入交换机和位于该行的行首交换机进行通信，同列的服  
务器之间通过与其相连接的接入交换机和位于该列的列首交换机进行通  
信，不同行列的服务器之间通过接入交换机、行首交换机和列首交换机进  
20 行通信。

同一行内的服务器 A 和服务器 B 进行通信时，服务器 A 先和与其相连  
的接入交换机 A 进行通信，接入交换机 A 再通过位于该行的行首交换机  
与和服务器 B 相连接的接入交换机 B 进行通信，接入交换机 B 再与服务器  
B 进行通信。

25 同一列内的服务器 A 和服务器 B 进行通信时，服务器 A 先和与其相连

接的接入交换机 A 进行通信，接入交换机 A 再通过位于该列的列首交换机和与服务器 B 相连接的接入交换机 B 进行通信，接入交换机 B 再与服务器 B 进行通信。

5 位于不同行列的服务器 A 和服务器 B 进行通信时，服务器 A 先和与其相连接的接入交换机 A 进行通信，接入交换机 A 再通过与其同行的行首交换机与位于该行的且与服务器 B 同列的接入交换机 C1 进行通信，接入交换机 C1 再通过其所在列的列首交换机和与服务器 B 相连接的接入交换机 B 进行通信，接入交换机 B 再与服务器 B 进行通信。

10 位于不同行列的服务器 A 和服务器 B 进行通信时，服务器 A 先和与其相连接的接入交换机 A 进行通信，接入交换机 A 再通过与其同列的列首交换机与位于该列的且与服务器 B 同行的接入交换机 C2 进行通信，接入交换机 C2 再通过其所在行的行首交换机和与服务器 B 相连接的接入交换机 B 进行通信，接入交换机 B 再与服务器 B 进行通信。

15 与现有路由方法相比，本发明充分发挥了交换式矩阵网络拓扑结构的特点和优势，具有以下优点：

1) 相邻交换机之间只需要定期交换本机的 IP 和 MAC 地址信息就可以建立构造出路由表，交换的数据量非常少。

2) 路由表的构造方法非常简单，构造速度快，对链路失效的反应速度快，不存在路由收敛问题。

20 3) 路由表规模小，路由速度快，可以通过硬件设备利用交换技术实现路由。

4) 网络可靠性高，任意两个终端之间存在多条等价路径，本路由算法支持等价多路径路由 ECMP (Equal-Cost Multipath Routing) 技术，具备负载均衡能力。

## 附图说明

图 1 为由 6 个端口的交换机组成的数据中心网络系统。

## 具体实施方式

为使本发明的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本发明作进一步详细的说明。

本发明提供的数据中心网络系统中，网络采用规则化的拓扑结构（交换式矩阵拓扑），并按照一定的规则进行编址。数据中心网络系统由一组行首交换机 10.Row.0.X、列首交换机 10.0.Col.X、接入交换机 10.Row.Col.1 和连接到接入交换机上的各种服务器 10.Row.Col.X 构成。其中 Row 为交换机或服务器所在行的行号，Col 为交换机或服务器所在列的列号；对于行首/列首交换机， $0 < X \leq 255$ ，对于服务器， $1 < X \leq 255$ 。所有设备的子网掩码均设为 255.255.255.0。

服务器之间通信通过接入交换机、行首交换机和列首交换机的路由转发功能完成。行首交换机负责把本行的所有接入交换机连接在一起，列首交换机负责把本列的所有接入交换机连接到一起。每一个接入交换机同时连接到所在行的所有行首交换机和所在列的所有列首交换机上面，行首交换机和列首交换机之间、各接入交换机之间没有直接连接关系。每个服务器都连接到一个接入交换机上。为了完成路由转发功能，在每个行首/列首交换机和接入交换机上都维护一张路由表，数据分组根据路由表进行转发。相连的交换机通过互相交换信息（包括本机 IP 地址和 MAC 地址）学习之间的连接关系，并根据连接关系生成路由表。

### 1、交换式矩阵拓扑结构

本实施例中，数据中心网络系统由交换机和服务器两类设备组成。交换机设备提供二层（链路层）和三层（网络层）网络交换功能，服务器设备提供数据运算和存储服务。其中交换机又分为三种类型，称为行首交换

机、列首交换机和接入交换机。行首交换机和列首交换机属于网络核心层，具有三层交换/路由能力，负责把接入交换机连接在一起；接入交换机属于网络接入层，具有二层交换和三层交换/路由能力，负责把服务器接入到网络中。本实施例中的行首交换机、列首交换机和接入交换机都可以采用高  
5 性价比的普通交换机。行首交换机、列首交换机和接入交换机连接在一起构成了交换式矩阵拓扑结构。

本发明提出的交换式矩阵拓扑结构要求行首交换机、列首交换机和接入交换机的参与交换/路由的端口数（活动端口数）最好相同，设端口数为  $N$  ( $N > 3$ )。接入交换机的端口分为三部分，其中第一部分端口用来连接服  
10 务器，第二部分端口连接行首交换机，剩下的一部分端口用来连接列首交换机。完整的拓扑结构总共有  $N$  行 $\times$  $N$  列个接入交换机，每一行的行首部署多个行首交换机，每一列的列首部署多个列首交换机。任一个接入交换机需要连接其所在行和列的全部行首交换机和列首交换机。本发明允许服务器和接入交换机之间、接入交换机和所在行首/列首交换机之间通过任意端  
15 口进行连接。每一个接入交换机和所在行的每个行首交换机之间都有一条单独的连接，和所在列的每个列首交换机也都有一条单独的连接。图 1 是一个交换机的端口数  $N = 6$  的交换式矩阵的例子，为清晰起见，图中没有画出服务器，并且在图中用一条虚横线表示一行内的所有接入交换机和该行的所有行首交换机之间的连接，用一条虚竖线表示一列内的所有接入交换机和该列的所有列首交换机之间的连接。  
20

接入交换机的端口分配比例可以根据实际需要分配，典型分法是把端口分成 3 等份， $1/3$  的端口用于连接行首交换机， $1/3$  的端口用于连接列首交换机，剩下  $1/3$  的端口用于连接服务器设备。这样每行的行首交换机和每列的列首交换机数量为  $N/3$ 。这种配置方式可以保证每层设备的超额订  
25 购比例达到 1:1。如果需要行首交换机或列首交换机参与转发的通信量不是

很多，也可以根据需求适当减少行首交换机和列首交换机的数量，从而降低建网成本。比如把接入交换机的 1/2 端口分配给服务器，1/4 的端口连接行首交换机，剩下 1/4 的端口用于连接列首交换机。这样每行的行首交换机和每列的列首交换机数量可以减少到  $N/4$ 。这种情况下，行首/列首交换机的超额订购比例为 1:2。

上面描述的是一个完整的交换式矩阵拓扑。在某些情况下，可以针对实际需要对网络拓扑进行调整。比如对于网络规模较小，服务器数量不多的数据中心，也可以构建不完全的交换式矩阵网络。完整的交换式矩阵网络拥有  $N$  行 $\times$  $N$  列个接入交换机，如果服务器的数量达不到  $N^3/3$ ，可以按照自右向左，自下向上的顺序减少接入交换机的数量。对于不完整的行或列，可以相应地按照比例减少行首或列首交换机的数量。行首/列首交换机上的多余空闲端口通过端口汇聚（Trunk）技术合并到其他端口上。比如交换机的端口数为 12，则可以构成最大为 12 行\* 12 列的网络拓扑。每行的行首交换机和列首交换机的数量均为  $12/3=4$  个。如果是不完全的拓扑结构，比如只有 6 行\*12 列的接入交换机，则列首交换机的数量即可减少一半，为 2 个。此时列首交换机的连接方案为：首先每个列首交换机用 6 个端口连接本列内的 6 个接入交换机，然后剩下的端口按照顺序逐次平均地汇聚到这些端口上。

## 2、网络编址方案

本数据中心网络系统内的各种交换机和服务器采用内部网络 IP 地址 10.X.X.X ( $0 < X \leq 255$ ) 进行编址，需要和外部网络通信时采用网络地址转换（NAT）技术转换成外部网络地址。

行首交换机的 IP 地址配置为 10.Row.0.X，其中 Row 为行首交换机所在的行号， $0 < Row \leq N$ ， $0 < X \leq 255$ ，在这个规定范围内，行首交换机的地址可以任意配置（X 可以任意指定）。

列首交换机的 IP 地址配置为 10.0.Col.X，其中 Col 为列首交换机所在的列号， $0 < Col \leq N$ ， $0 < X \leq 255$ ，在这个规定范围内，列首交换机的地址可以任意配置（X 可以任意指定）。

5 接入交换机的 IP 地址配置为 10.Row.Col.1，其中 Row 为接入交换机所在的行号，Col 为接入交换机所在的列号， $0 < Row \leq N$ ， $0 < Col \leq N$ 。

服务器的 IP 地址配置为 10.Row.Col.X，其中 Row 为该服务器所在的行号，Col 为该服务器所在的列号， $0 < Row \leq N$ ， $0 < Col \leq N$ ， $1 < X \leq 255$ ，在这个规定范围内，服务器的地址可以任意配置（X 可以任意指定）。

10 在上面的编址方案中，我们可以根据设备的 IP 地址区分出其设备类型，以及该设备在网络中所处的位置，这有助于确定设备连接关系，简化路由方案。根据编址方案和设备连接关系，可以看出每一个接入交换机和其连接的全部服务器构成一个物理子网，子网掩码为 255.255.255.0。同一行的行首交换机或同一列的列首交换机虽然其网络地址前缀相同，但是没有直接连接关系。

### 15 3、路由方法

在本实施例提出的交换式矩阵网络中，行和列是一个对称的结构，因此，不同行列内的服务器之间进行通信，可以先经过行首交换机再经过列首交换机，或者先经过列首交换机再经过行首交换机，比如有一台设备 10.2.2.X 和 10.4.4.X 通信，先经过列首交换机的路径如下：

20 10.2.2.X→10.2.2.1→10.0.2.X→10.4.2.1→10.4.0.X→10.4.4.1→10.4.4.X

先经过行首交换机的路径如下：

10.2.2.X→10.2.2.1→10.2.0.X→10.2.4.1→10.0.4.X→10.4.4.1→10.4.4.X

25 本实施例规定：同一行内的设备之间进行通信，只通过行首交换机进行转发，同一列内的设备之间进行通信，只通过列首交换机进行转发。不同行列之间的设备通信，采用先经过列首交换机的路径。

#### 3.1 路由表的结构

本发明提出的数据中心网络系统具有规则的拓扑结构，因此路由方法可以设计的非常简单。考虑到灵活性和扩展性，本发明采用基于路由表的路由转发方法。路由表的结构如下：

目标子网	子网掩码	下一跳 IP 地址	下一跳 MAC 地址	出端口	时间戳
10.1.0.0	255.255.0.0	10.1.0.1	XX-XX-XX-XX-XX-XX	1	
10.2.0.0	255.255.0.0	10.2.0.1	XX-XX-XX-XX-XX-XX	2	

5 说明：

(1) 时间戳用来记录本条路由表项的创建或更新时间。

(2) 后面描述路由表时，省略了下一跳 IP 地址、下一跳 MAC 地址和时间戳信息。

行首交换机 10.Row.0.X 的路由表（采用非连续子网掩码）

10	目标子网/子网掩码	出端口
	10.0.1.0/255.0.255.0	P1(10.Row.1.1 对应的端口)
	10.0.2.0/255.0.255.0	P2(10.Row.2.1 对应的端口)
	10.0.3.0/255.0.255.0	P3(10.Row.3.1 对应的端口)
	10.0.4.0/255.0.255.0	P4(10.Row.4.1 对应的端口)
15	10.0.5.0/255.0.255.0	P5(10.Row.5.1 对应的端口)
	10.0.6.0/255.0.255.0	P6(10.Row.6.1 对应的端口)

列首交换机 10.0.Col.X 的路由表

20	目标子网/子网掩码	出端口
	10.1.0.0/255.255.0.0	P1(10.1.Col.1 对应的端口)
	10.2.0.0/255.255.0.0	P2(10.2.Col.1 对应的端口)
	10.3.0.0/255.255.0.0	P3(10.3.Col.1 对应的端口)
	10.4.0.0/255.255.0.0	P4(10.4.Col.1 对应的端口)
	10.5.0.0/255.255.0.0	P5(10.5.Col.1 对应的端口)

10.6.0.0/255.255.0.0 P6(10.6.Col.1 对应的端口)

接入交换机 10.Row.Col.1 的路由表

接入交换机的路由转发规则有如下三条： 1) 对于本子网的通信，采用传统的二层交换进行转发，此处不作说明； 2) 对于目的地址是本行内的服务器（但位于不同子网），转发至行首交换机； 3) 对于目的地址是不同行的服务器（位于不同子网）的通信，转发至列首交换机。对于情况（2）和（3），需要进行三层路由转发，路由表如下：

	目标子网/子网掩码	出端口	
	10.Row.0.0/255.255.0.0	10.Row.0.X <sub>1</sub> 对应的端口	} 转发给行首交换机，共 K 个等价路径
10	10.Row.0.0/255.255.0.0	10.Row.0.X <sub>2</sub> 对应的端口	
	.....		
	10.Row.0.0/255.255.0.0	10.Row.0.X <sub>K</sub> 对应的端口	
	10.0.0.0/255.0.0.0	10.0.Col. X <sub>1</sub> 对应的端口	} 转发给列首交换机，共 M 条等价路径
	10.0.0.0/255.0.0.0	10.0.Col. X <sub>2</sub> 对应的端口	
15	.....		
	10.0.0.0/255.0.0.0	10.0.Col. X <sub>M</sub> 对应的端口	

说明：

1) 10.Row.0.X<sub>i</sub> 是第 Row 行的第 i 个行首交换机的 IP 地址(1≤i≤K, K 为第 Row 行的行首交换机的数量), 10.0.Col.X<sub>j</sub> 是第 Col 列的第 j 个列首交换机的 IP 地址 (1≤j≤M, M 为列首交换机的数量)。

2) 对于同一目标子网，接入交换机的路由表里存在多条等价的路由路径，本发明采用等价多路径路由 ECMP (Equal-Cost Multipath Routing) 技术，实现从多条重复的等价路径中随机选择一条路径。

### 3.2 路由表构造方法

通过自动学习行首/列首交换机和接入交换机之间的连接关系，可以非常容易地把路由表构造出来。为了学习交换机之间的连接关系，需要所有

交换机定期向所有活动端口发送 PDU (协议数据单元), 包含内容为本机 IP 地址和本机 MAC 地址。对于任意交换机来说, 每个端口最多对应一条路由表项, 因此路由表的条目数最多等于交换机的端口数。

(1) 行首交换机路由表的构造

5 行首交换机 10.Row.0.X 按照如下规则构造路由表:

如果从端口 Port 收到 10.Row.Col.1 发来的 PDU, 向路由表内添加或更新路由表项:

10.0.Col.0/255.0.255.0 /10.Row.Col.1/MAC 地址/ Port/时间戳

如果规定时间内收不到更新 PDU, 则删除相应的路由条目 (已过期)。

10 (2) 列首交换机路由表的构造

列首交换机 10.0.Col.X 按照如下规则构造路由表:

如果从端口 Port 收到 10.Row.Col.1 发来的 PDU, 向路由表内添加或更新路由表项:

10.Row.0.0/255.255.0.0 /10.Row.Col.1/MAC 地址/ Port/时间戳

15 如果规定时间内收不到更新 PDU, 则删除相应的路由条目 (已过期)。

(3) 接入交换机路由表的构造

接入交换机 10.Row.Col.1 按照如下规则构造路由表:

a) 从端口 Port 收到本行的行首交换机 10.Row.0.X 的 PDU, 向路由表内添加或更新路由表项:

20 10.Row.0.0/255.255.0.0/10.Row.0.X / MAC 地址/ Port/时间戳

b) 从端口 Port 收到本列的列首交换机 10.0.Col.X 的 PDU, 向路由表内添加或更新路由表项:

10.0.0.0/255.0.0.0 /10.0.Col.X /MAC 地址/Port/时间戳

c) 如果规定时间内收不到更新 PDU, 则删除相应的路由条目 (已过期)。  
25

说明：对于同一子网内的数据通信，采用传统的二层交换技术进行数据转发，二层交换的地址转发表（AFT，Address Forwarding Table）的构造此处不做说明。

### 3.3 路由过程举例说明

5       （1）同一子网内设备的数据通信。假设有两台服务器 IP 地址分别为 10.1.1.2(源)和 10.1.1.3(目的)，则二者的通信直接通过接入交换机 10.1.1.1 进行转发。

      （2）同一行内设备的数据通信。假设有两台服务器 IP 地址分别为 10.1.3.2 和 10.1.5.2，数据分组要从 10.1.3.2 发送到 10.1.5.2，需要首先发往  
10 接入交换机 10.1.3.1，然后根据各交换机的路由表，路由过程如下：

10.1.3.2→10.1.3.1→10.1.0.X→10.1.5.1→10.1.5.2

      （3）同一列内设备的数据通信。假设有两台服务器 IP 地址分别为 10.2.2.2 和 10.4.2.2，数据分组要从 10.2.2.2 发送到 10.4.2.2，需要首先发往接入交换机 10.2.2.1，然后根据各交换机的路由表，路由过程如下：

15       10.2.2.2→10.2.2.1→10.0.2.X→10.4.2.1→10.4.2.2

      （4）不同行的设备的数据通信。假设有两台服务器 IP 地址分别为 10.2.2.2 和 10.4.4.2，数据分组要从 10.2.2.2 发送到 10.4.4.2，需要首先发往接入交换机 10.2.2.1，然后根据各交换机的路由表，路由过程如下：

10.2.2.2→10.2.2.1→10.0.2.X→10.4.2.1→10.4.0.X→10.4.4.1→10.4.4.2

20       以上对本发明所提供的数据中心网络系统及其路由方法进行详细介绍，本文中应用了具体实施例对本发明的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本发明的方法及其核心思想；同时，对于本领域的一般技术人员，依据本发明的思想，在具体实施方式及应用范围上均会有改变之处。综上所述，本说明书内容不应理解为对本发明的限制。  
25

### 工业实用性

本发明充分发挥了交换式矩阵网络拓扑结构的特点和优势，具有交换数据量少、路由表构造速度快、路由表规模小、路由速度快、网络可靠性高、具备负载均衡能力等优点。

## 权利要求书

1、一种数据中心网络系统的路由方法，该方法包括：

同一子网的服务器之间通过与其相连接的接入交换机进行通信，不同子网的同行的服务器之间通过与其相连接的接入交换机和位于该行的行首  
5 交换机进行通信，同列的服务器之间通过与其相连接的接入交换机和位于该列的列首交换机进行通信，不同行列的服务器之间通过接入交换机、行首交换机和列首交换机进行通信。

2、根据权利要求 1 所述的一种数据中心网络系统的路由方法，其中，同一行内的服务器 A 和服务器 B 进行通信时，服务器 A 先和与其相连接的  
10 接入交换机 A 进行通信，接入交换机 A 再通过位于该行的行首交换机与和服务器 B 相连接的接入交换机 B 进行通信，接入交换机 B 再与服务器 B 进行通信。

3、根据权利要求 1 所述的一种数据中心网络系统的路由方法，其中，同一列内的服务器 A 和服务器 B 进行通信时，服务器 A 先和与其相连接的  
15 接入交换机 A 进行通信，接入交换机 A 再通过位于该列的列首交换机和与服务器 B 相连接的接入交换机 B 进行通信，接入交换机 B 再与服务器 B 进行通信。

4、根据权利要求 1 所述的一种数据中心网络系统的路由方法，其中，位于不同行列的服务器 A 和服务器 B 进行通信时，服务器 A 先和与其相连接  
20 的接入交换机 A 进行通信，接入交换机 A 再通过与其同行的行首交换机与位于该行的且与服务器 B 同列的接入交换机 C1 进行通信，接入交换机 C1 再通过其所在列的列首交换机和与服务器 B 相连接的接入交换机 B 进行通信，接入交换器 B 再与服务器 B 进行通信。

5、根据权利要求 1 所述的一种数据中心网络系统的路由方法，其中，  
25 位于不同行列的服务器 A 和服务器 B 进行通信时，服务器 A 先和与其相连

接的接入交换机 A 进行通信，接入交换机 A 再通过与其同列的列首交换机与位于该列的且与服务器 B 同行的接入交换机 C2 进行通信，接入交换机 C2 再通过其所在行的行首交换机和与服务器 B 相连接的接入交换机 B 进行通信，接入交换机 B 再与服务器 B 进行通信。

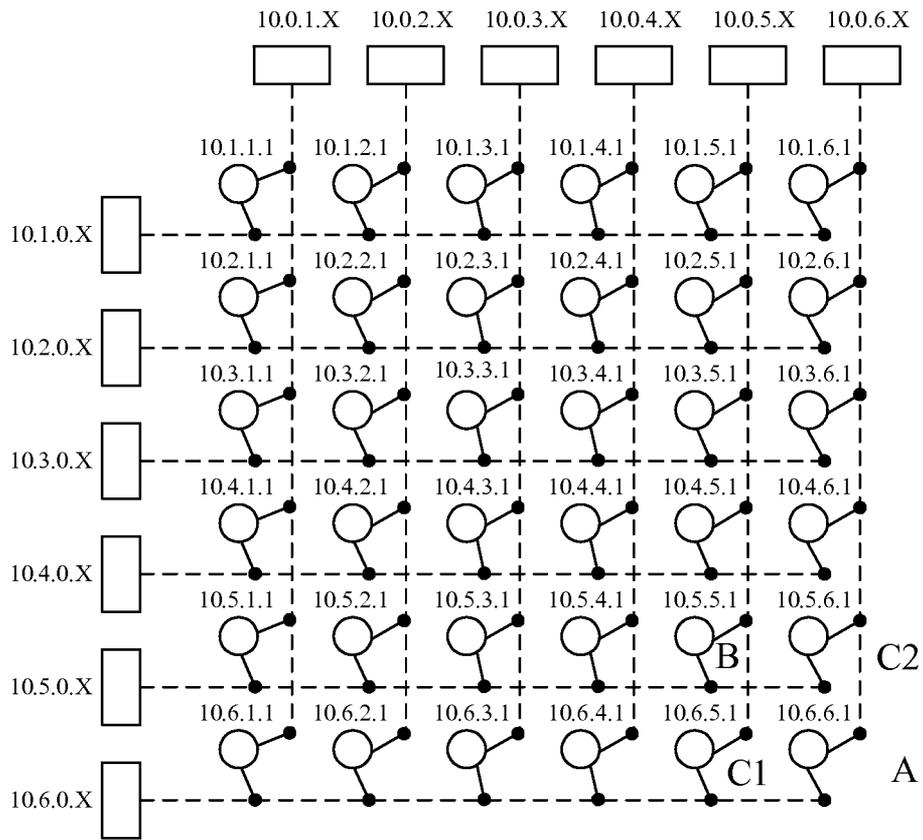


图 1

# INTERNATIONAL SEARCH REPORT

International application No.

**PCT/CN2012/073735**

## A. CLASSIFICATION OF SUBJECT MATTER

H04L 12/56 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: H04L; H04W

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CPRSABS, CNTXT, CNKI, VEN: structure, clos, switch+, rout+, data w center, server, network, topology, matrix, row, column, same, subnet+, fat w tree

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 102185772 A (BEIJING JIAOTONG UNIVERSITY), 14 September 2011 (14.09.2011), claims 1-5	1-5
PX	CN 102164088 A (BEIJING JIAOTONG UNIVERSITY), 24 August 2011 (24.08.2011), description, paragraphs [0018]-[0084]	1-5
A	US 2009303880 A1 (MICROSOFT CORP.), 10 December 2009 (10.12.2009), the whole document	1-5
A	CN 101485156 A (LEVEL 3 COMMUNICATIONS INC.), 15 July 2009 (15.07.2009), the whole document	1-5
A	CN 101383778 A (HANGZHOU H3C TECHNOLOGIES CO., LTD.), 11 March 2009 (11.03.2009), the whole document	1-5

Further documents are listed in the continuation of Box C.

See patent family annex.

<p>* Special categories of cited documents:</p> <p>“A” document defining the general state of the art which is not considered to be of particular relevance</p> <p>“E” earlier application or patent but published on or after the international filing date</p> <p>“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)</p> <p>“O” document referring to an oral disclosure, use, exhibition or other means</p> <p>“P” document published prior to the international filing date but later than the priority date claimed</p>	<p>“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention</p> <p>“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone</p> <p>“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art</p> <p>“&amp;” document member of the same patent family</p>
---	---

Date of the actual completion of the international search 26 June 2012 (26.06.2012)	Date of mailing of the international search report <b>05 July 2012 (05.07.2012)</b>
--	--

<p>Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451</p>	<p>Authorized officer  <b>LI, Yanxin</b>  Telephone No.: (86-10) <b>62411241</b></p>
---	--

**INTERNATIONAL SEARCH REPORT**  
Information on patent family members

International application No.

**PCT/CN2012/073735**

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102185772 A	14.09.2011	None	
CN 102164088 A	24.08.2011	None	
US 2009303880 A1	10.12.2009	CN 102057631 A	11.05.2011
		US 8160063 B2	17.04.2012
		WO 2009151985 A2	17.12.2009
		WO 2009151985 A3	25.03.2010
		EP 2289206 A2	02.03.2011
		KR 20110027682 A	16.03.2011
		CN 101485156 A	15.07.2009
		WO 2008067493 A3	17.07.2008
		WO 2008067493 A2	05.06.2008
		CA 2655984 A1	05.06.2008
		EP 2087657 A2	12.08.2009
		EP 2087657 B1	25.01.2012
CN 101383778 A	11.03.2009	CN 101383778 B	13.04.2011

国际检索报告

国际申请号  
PCT/CN2012/073735

<b>A. 主题的分类</b>		
H04L 12/56(2006.01)i		
按照国际专利分类(IPC)或者同时按照国家分类和 IPC 两种分类		
<b>B. 检索领域</b>		
检索的最低限度文献(标明分类系统和分类号)		
IPC: H04L; H04W		
包含在检索领域中的除最低限度文献以外的检索文献		
在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))		
CPRSABS, CNTXT, CNKI, VEN: 交换机, 路由, 数据中心, 服务器, 网络, 拓扑, 结构, 矩阵, 行, 列, 同, 子网, 胖树, clos, switch+, rout+, data w center, server, network, topology, matrix, row, same, subnet+, fat w tree		
<b>C. 相关文件</b>		
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN102185772A (北京交通大学) 14.9 月 2011(14.09.2011) 权利要求 1-5	1-5
PX	CN102164088A (北京交通大学) 24.8 月 2011(24.08.2011) 说明书第[0018]-[0084]段	1-5
A	US2009303880A1 (MICROSOFT CORP) 10.12 月 2009(10.12.2009) 全文	1-5
A	CN101485156A (第三级通讯公司) 15.7 月 2009(15.07.2009) 全文	1-5
A	CN101383778A (杭州华三通信技术有限公司) 11.3 月 2009(11.03.2009) 全文	1-5
<input type="checkbox"/> 其余文件在 C 栏的续页中列出。 <input checked="" type="checkbox"/> 见同族专利附件。		
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件		“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件
国际检索实际完成的日期 26.6 月 2012(26.06.2012)		国际检索报告邮寄日期 05.7 月 2012 (05.07.2012)
ISA/CN 的名称和邮寄地址: 中华人民共和国国家知识产权局 中国北京市海淀区蓟门桥西土城路 6 号 100088 传真号: (86-10)62019451		受权官员  李彦欣  电话号码: (86-10) 62411241

国际检索报告  
关于同族专利的信息

国际申请号  
**PCT/CN2012/073735**

检索报告中引用的 专利文件	公布日期	同族专利	公布日期
CN102185772A	14.09.2011	无	
CN102164088A	24.08.2011	无	
US2009303880A1	10.12.2009	CN102057631A	11.05.2011
		US8160063B2	17.04.2012
		WO2009151985A2	17.12.2009
		WO2009151985A3	25.03.2010
		EP2289206A2	02.03.2011
		KR20110027682A	16.03.2011
CN101485156A	15.07.2009	US2008151863A1	26.06.2008
		WO2008067493A3	17.07.2008
		WO2008067493A2	05.06.2008
		CA2655984A1	05.06.2008
		EP2087657A2	12.08.2009
		EP2087657B1	25.01.2012
CN101383778A	11.03.2009	CN101383778B	13.04.2011