[19]中华人民共和国国家知识产权局

[51]Int. Cl[7]
H04L 12/44
H04L 29/06

[12]发明专利申请公开说明书

[21]申请号 00807858.0

[43]公开日 2002年5月29日

[11]公开号 CN 1351791A

[22]申请日 2000.5.16 [21]申请号 00807858.0
[30]优先权
[32]1999.5.21 [33]US [31]09/316,699
[86]国际申请 PCT/US00/13467 2000.5.16
[87]国际公布 WO00/72530 英 2000.11.30
[85]进入国家阶段日期 2001.11.21
[71]申请人 艾维西系统公司
地址 美国马萨诸塞州
[72]发明人 威廉·J·达利 菲利普·P·卡维
P·阿伦·金 威廉·F·曼恩
拉里·R·丹尼桑

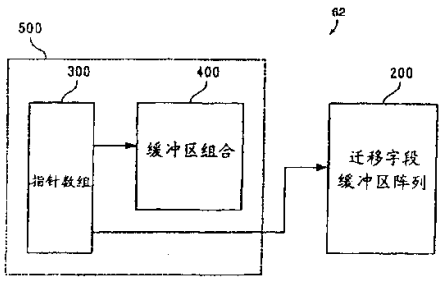
[74]专利代理机构 永新专利商标代理有限公司
代理人 过晓东

权利要求书6页 说明书19页 附图页数10页

[54]发明名称 带迁移字段缓冲的结构路由器

[57]摘要

在结构路由器中,迁移字段被储存在芯片上的第一套可快 速存取的迁移字段缓冲区中,而来自第一套迁移字段缓冲区的 溢出被储存在存取速度比第一套缓慢的在芯片外的第二套迁移 字段缓冲区中。迁移字段缓冲区可以包括通过指针数组访问的 缓冲区组合即成组相联的高速缓冲存储器。网络节点之间的流 量控制在第一套缓冲区和第二套缓冲区之间转移迁移字段时阻 止新的迁移字段到达。



ISSN1008-4274

01.11.21

权 利 要 求 书

1. 一种用来通过路由器转移信息单位集的包括缓冲区的路由器，该路由器包括：

用于信息单位集的第一套快速存取缓冲区；以及

对信息单位集的存取速度比第一套缓慢的第二套缓冲区。

2. 根据权利要求 1 的路由器，其中所述的路由器是在一个或多个的集成电路芯片上实现的；第一套缓冲区位于路由器的集成电路芯片上；第二套缓冲区位于与路由器的集成电路芯片分开的存储芯片上。

3. 根据权利要求 1 的路由器，其中第二套缓冲区为一整套虚拟信道保存信息单位集。

4. 根据权利要求 1 的路由器，其中第一套缓冲区包括缓冲区组合和指针数组。

5. 根据权利要求 1 的路由器，其中第一套缓冲区是作为成组相联的高速缓冲存储器组织起来的。

6. 根据权利要求 5 的路由器，其中成组相联的高速缓冲存储器中每个项目都包含一个信息单位集。

7. 根据权利要求 5 的路由器，其中成组相联的高速缓冲存储器中每个项目都包含数个缓冲区和适合整个虚拟信道的状态。

8. 根据权利要求 1 的路由器，进一步包括在第一套缓冲区和第二套缓冲区之间转移信息单位集期间阻止新的信息单位集到达的流量控制。

9. 根据权利要求 8 的路由器，其中所述的流量控制是阻断。

10. 根据权利要求 8 的路由器，其中流量控制是以信用为基础的。

11. 根据权利要求 1 的路由器，进一步包括过错状态寄存器，以便保存等待访问第二套缓冲区的信息单位集。

12. 根据权利要求 1 的路由器，进一步包括逐出缓冲区，以便保存为从第一套缓冲区转移到第二套缓冲区准备的项目。

13. 根据权利要求 1 的路由器在多计算机互连网络中的应用。

14. 根据权利要求 1 的路由器在网络开关或路由器中的应用。

15. 根据权利要求 1 的路由器，其中所述的路由器是结构路由器，而信息单位是迁移字段。

16. 在路由器中缓存信息单位集的方法，该方法包括：

把信息单位集储存在第一套快速存取缓冲区中；以及

把来自第一套缓冲区的溢出储存在存取速度比第一套缓慢的第二套缓冲区中。

17. 根据权利要求 16 的方法，其中路由器是在一个或多个集成电路芯片上实现的；第一套缓冲区位于路由器的集成电路芯片上；而第二套缓冲区位于与路由器的集成电路芯片分开的存储芯片上。

18. 根据权利要求 16 的方法，其中第二套缓冲区为一整套虚拟信道保存信息单位集。

19. 根据权利要求 16 的方法，进一步包括用第一套缓冲区把信息单位集储存在诸信道共享的缓冲区组合中，并且由与各个信道相关联的指针数组指出缓冲区组合内的信息单位集。

20. 根据权利要求 16 的方法，其中第一套缓冲区是作为成组相联的高速缓冲存储器组织起来的。

21. 根据权利要求 20 的方法，其中成组相联的高速缓冲存储器中每个项目都包含一个信息单位。

22. 根据权利要求 20 的方法，其中成组相联的高速缓冲存储器中每个项目都包含用于整个虚拟信道的数个信息单位缓冲区和状态。

23. 根据权利要求 16 的方法，进一步包括流量控制，以便在第一套缓冲区和第二套缓冲区之间转移迁移字段时阻止新的信息单位集到达。

24. 根据权利要求 23 的方法，其中流量控制是阻断。

25. 根据权利要求 23 的方法，其中流量控制是以信用为基础的。

26. 根据权利要求 16 的方法，进一步包括把等待访问第二套迁移字段缓冲区的信息单位集储存在过错状态寄存器中。

27. 根据权利要求 16 的方法，进一步包括把准备从第一套缓冲区转移到第二套缓冲区的信息单位集储存在逐出缓冲区中。

28. 根据权利要求 16 的方法，其中路由器在多计算机互连网络中。

29. 根据权利要求 16 的方法，其中结构路由器在网络开关或路由器中。

30. 根据权利要求 16 的方法，其中路由器是结构路由器，而信息单位是迁移字段。

31. 一种包括众多互连路由器的网络，每个路由器都包括数个信息单位缓冲区，其中包括：

第一套快速存取的信息单位缓冲区；以及

存取速度比第一套缓慢的第二套信息单位缓冲区。

32. 根据权利要求 31 的网络，其中路由器是在一个或多个集成电路芯片上实现的；第一套缓冲区位于路由器的集成电路芯片上；第二套缓冲区位于与路由器的集成电路芯片分开的存储芯片上。

33. 根据权利要求 31 的网络，其中第二套缓冲区为一整套虚拟信道保存信息单位集。

34. 根据权利要求 31 的网络，其中第一套缓冲区包括缓冲区组合和指针数组。

35. 根据权利要求 31 的网络，其中第一套缓冲区是作为成组相联的高速缓冲存储器组织起来的。

36. 根据权利要求 31 的网络，进一步包括流量控制，以便在第一套缓冲区和第二套缓冲区之间转移信息单位集时阻止新的信息单位集到达。

37. 根据权利要求 31 的网络，进一步包括流量控制，以便在第一套缓冲区和第二套缓冲区之间转移迁移字段时阻止新的信息单位集到达。

38. 根据权利要求 31 的网络在网络开关或路由器中的应用。

39. 根据权利要求 31 的网络，其中路由器是结构路由器，而信息单位是迁移字段。

40. 一种路由器，其中包括:

用来把信息单位集储存在第一套快速存取缓冲区中的装置；以及

用来把信息单位集储存在存取速度比第一套缓慢的第二套缓冲区中的装置。

41. 根据权利要求 40 的路由器，其中第二套缓冲区为一整套虚拟信道保存信息单位集。

42. 根据权利要求 40 的路由器，其中第一套缓冲区包括诸信道共享的缓冲区组合和在缓冲区组合中为各个信道指出项目的装置。

43. 根据权利要求 40 的路由器，其中第一套缓冲区是作为成组相联的高速缓冲存储器组织起来的。

44. 根据权利要求 40 的路由器，进一步包括用来提供流量控制的装置，以便在第一套缓冲区和第二套缓冲区之间转移信息单位集时阻止新的信息单位集到达。

45. 根据权利要求 40 的路由器，其中路由器是结构路由器，而信息单位是迁移字段。

说 明 书

带迁移字段缓冲的结构路由器

本发明的现有技术

互连网络由一套通过信道连接的节点构成。这样的网络被用来在节点之间传送数据包。例如，它们被用在多计算机系统、多处理机系统、网络开关系统和路由器系统中。在多计算机系统中，它们在处理节点之间传送信息。在多处理机系统中，它们把来自处理节点的存储器请求传送给存储器节点并且在相反的方向上传送应答。在网络开关系统和路由器系统中，它们把数据包从输入线路卡传送到输出线路卡。例如，William J. Dally、Philip P. Carvey、Larry R.、Dennison 和 P. Allen King 以 “Router With Virtual Channel Allocation（带虚拟信道分配的路由器）” 为题的国际专利申请公开第 PCT/US98/16762 （W099/11033）号介绍了为互连网路由器提供切换结构的三维环面互连网络的应用，在此通过引证将其全部教导并入。

互连网络设计中的关键问题是在构成互连结构的节点或结构路由器中缓冲存储的管理。包括在著名的未审 PCT 专利申请中描述的那种路由器在内的许多早期的路由器都利用 William J. Dally 于 1992 年 3 月在 IEEE Transactions on Parallel and Distributed Systems 的第 3 卷第 2 期第 194 至 205 页发表的“Virtual Channel Flow Control（虚拟信道流量控制）” 中介绍的虚拟信道流量控制来管理这些缓冲区。采用这种方法，与每个信道相关联的缓冲空间都被分割，而且每个分区都与一个虚拟信道相关

联。每个穿越信道的数据包都被分配给特定的虚拟信道，而且不与穿越其它虚拟信道的数据包争夺缓冲空间。

最近，为了在被引导到不同的目的地的不同类别的通信量之间提供隔离，带数百个虚拟信道的路由器已被构成。采用这些路由器，保存虚拟信道缓冲区所需要的空间数量成为问题。为了在许多通信量类别之间提供隔离，缓冲区的数目（N）必须相当大。同时，为了把良好的吞吐量提供给单一的虚拟信道，每个缓冲区的规模（S）都必须相当大。总的输入缓冲区规模是这两项的乘积，$T = N \times S$。就 $N = 512$ 个虚拟信道和 $S = 4$ 个迁移字段（每个迁移字段 576 位）的路由器而言，所需要的总缓冲区空间是 2048 个迁移字段即 1,179,648 位。就上述专利申请中所描述的那种类型的路由器上的 7 个输入信道而言，必不可少的缓冲区将占用 7 兆位以上的存储器。

这些存储器需求使采用受每个路由器 ASIC 芯片大约 1Mbit 限制的现有的超大规模集成电路技术实现具有大量的大规模虚拟信道缓冲区的单片结构路由器变成不可行的。过去，这是通过下述方法寻址的，即具有数量比较少的虚拟信道，这可能导致在不同的通信量类别之间的缓冲区干扰；或者使每个虚拟信道都比较小（往往是一个迁移字段大小），这将导致在单一的虚拟信道上性能不良；或者把路由器分到几个 ASIC 芯片上，这将增加成本和复杂性。

本发明的概述

现在的发明通过提供作为高速缓冲存储器使用的芯片上的迁移字段缓冲区的小型组合和使任何不适合这种组合的迁移字段都溢出到芯片外的存储器中克服了现有技术的路由器的存储器局限性。我们的模拟研究表明为了保证通信量隔离需要缓冲

区隔离时实际上通常只占用缓冲区的一个微小的部分。因此，大部分时间所有活跃的虚拟信道都与高速缓冲存储器相适应，而外部的存储器难得被访问。

因此，按照现在的发明，路由器包括为通过路由器转移的诸如迁移字段之类的信息单位集准备的缓冲区。缓冲区包括用于信息单位集的第一套可快速存取的缓冲区和存取速度比第一套缓慢的用于信息单位集的第二套缓冲区。

在优选的实施方案中，结构路由器是在一个或多个集成电路芯片上实现的。第一套缓冲区被置于路由器的集成电路芯片上，而第二套缓冲区被置于与路由器的集成电路船分开的存储芯片上。第二套缓冲区可以为一整套虚拟单元保存信息单位集。

在一个实施方案中，第一套缓冲区包括缓冲区组合和指针数组。缓冲区组合是被各个虚拟信道共享的，而指针数组指出在缓冲区组合范围内与各个信道相关联的信息单位集。

在另一个实施方案中，第一套缓冲区是作为成组相联的高速缓冲存储器组织起来的。具体地说，在成组相联的高速缓冲存储器中每个项目都可以包含一个信息单位集，或者它可以包含数个缓冲区和适合某个完整的虚拟信道的状态。

流量控制可以被提供，以便在第一套缓冲区和第二套缓冲区之间转移信息单位集期间阻止新的信息单位集到达。流量控制可以是阻断型的或者是以信用为基础的。

过错状态寄存器可以保存等待访问第二套缓冲区的信息单位集。逐出缓冲区可以保存为从第一套缓冲区到第二套缓冲区的转移准备的项目。

本发明的应用包括多计算机互连网络、网络开关或路由器、以及互连网路由器内的结构路由器。

附图简要说明

本发明的上述和其它目的、特点和优点通过下面更具体地介绍用附图图解说明的本发明的优选实施方案将变得十分明显，在不同的附图中同样的参考符号始终指的是同一部分。这些附图不必按比例绘制，而是把重点放在图解说明本发明的原则上。

图 1 图解说明说明可以采用现在的发明的路由器的互连网配置。

图 2 图解说明形成图 1 的路由器的三维结构。

图 3 图解说明在图 2 的实施方案中使用的结构路由器。

图 4 图解说明一套先前在图 3 的结构路由器中提供的输入缓冲区。

图 5 图解说明按照本发明的一个实施方案的输入缓冲区阵列。

图 6 更详细地图解说明图 5 的指针数组和缓冲区组合。

图 7 图解说明在图 6 的指针数组中特定的项目。

图 8 图解说明本发明的另一个实施方案，其中输入缓冲区阵列是作为成组相联的高速缓冲存储器组织起来的，每个项目包含一个迁移字段。

图 9 图解说明本发明的第三个实施方案，其中输入缓冲区阵列是作为每个项目都包含迁移字段缓冲区和适合某个完整的虚拟信道的状态的成组相联的高速缓冲存储器组织起来的。

图 10 给出模拟信道占有直方图。

本发明的详细说明

虽然现在的发明适用于包括在多计算机系统、多处理机系统、网络开关和路由器中的那些路由器在内的任何路由器应用，但是与互连网路由器内的结构路由器相关的部分将予以描述。这样的路由器是在上述的 PCT 申请中提出的。

如同用图 1 图解说明的那样，因特网被安排成网络的分级结构。典型的末端用户具有接在局域网（LAN）24 上的工作站 22。为了允许 LAN 上的用户访问互连网的其余部分，LAN 经路由器 R 接到由整个地区的网络供应商（RNP）维护和操作的地区网络 26 上。这种连接是由因特网服务提供者（ISP）完成的。为了访问其它地区，地区网络在网络访问点（NAP）接到主干网络 28 上。这些 NAP 通常仅仅被放在主要的城市中。

网络是由链路和路由器构成的。在网络主干中，链路通常是使用 SONET（同步的光学网络）协议操作的光纤通信信道。SONET 链路以在 OC-3（155Mb/s）和 OC-192（9.9Gb/s）之间变动的各种数据传输速率操作。这些链路（有时也叫做中继线）往往越过相当大的距离把数据从一个点移动到另一个点。

路由器一组链路连接在一起，并且完成两种功能：转发和路径选择。在路由器的一条链路上到达的数据包通过把它在不同的链路上（取决于其最后的目的地和输出链路的状态）发送

出去被转发。为了计算给定的数据包的输出链路，路由器参与路径选择协议，在那里所有的路由器都在因特网上交换与网络的连接性有关的信息并且根据这个信息计算的路径选择表。

大多数现有技术的因特网路由器是以公用总线或纵横制跨过（crossbar switch）为基础的。通常，给定的 SONET 链路 30 接在线路接口模块上。这个模块从进入的 SONET 数据流中提取数据包。就每个进入的数据包而言，线路接口读出数据包的标题并且利用这个信息确定转发该数据包的输出口（或多个口）。为了转发该数据包，路由器内的通信路径被判优，而数据包被转移到输出线路接口模块。随后，该模块把数据包在出网 SONET 链路上转移到通向其目的地的路程上的下一跳。

上述的国际专利申请的路由器通过使用多级跳（multi-hop）互连网络（尤其是用图 2 图解说明的三维的环面网络）作为路由器克服了基于总线和纵横制的现有技术的路由器的带宽和缩放能力的限制。采用这种安排，在广域主干网络中每个路由器都有效地包含小型的柜内网络。为了避免混淆，我们将把每个路由器内部的小型网络称作切换结构，把这个小型网络内的路由器和链路称作结构路由器和结构链路。

在用图 2 图解说明的节点的三维环面切换结构中，每个节点 N 都包括接在 SONET 互连网的入网链路和出网链路上的线路接口模块。在这些线路接口节点中每个节点都包含一个开关结构路由器，这个开关结构路由器包括在环面中连接它的 6 个相邻节点的结构链路。在一个 SONET 链路上到达的数个 IP 包在节点 A 上说是为了确定它们应该在哪个 SONET 链路上离开互连网路由器而被核查的，而在节点 B 上说是随后经三维环面从 A 转发给 B 的开关结构。

通过互连网转发的典型的数据包在 50 字节和 1.5 千字节之间变动。为了通过现在的发明的互连网路由器的结构网络转移，数据包被分为若干个部分即迁移字段，每个部分 36 个字节。为了控制数据通过路由器结构的转移，至少修改包括在数据包的第一个迁移字段中的标题。在优选的路由器中，数据是按照蠕虫式路径选择协议（wormhole routing protocol）通过结构转移的。

数据包的迁移字段在包括一套缓冲区的虚拟网络中提供该结构流动。用于每个虚拟网络的一个或多个缓冲区都是在结构中的每个节点上提供的。每个缓冲区都是按保存至少一个流量控制数字或至少一个迁移字段的信息的规模制作的。所有的虚拟网络共享在真实的结构网络的节点之间的一套物理信道，并且为了在竞争的虚拟网络上多路复用各个物理信道采用公正的判优政策。

在开关结构上把数据包从与输入链路相关联的模块转发到与其输出链路相关联的模块所使用的结构路由器是用图 3 予以图解说明的。这种路由器有 7 条输入链路 58 和 7 条输出链路 60。其中 6 条链路与图 2 所示的三维环面网络中的毗邻节点相连接。第七条输入链路接收来自转发引擎 50 的数据包，而第七条输出链路把数据包发送给在这个路由器的线路接口模块中的数据包输出缓冲区 52。每条输入链路 58 都与输入缓冲区阵列 62 相关联，而每条输出链路 60 都与输出寄存器 64 相关联。输入缓冲区和输出寄存器通过 7×7 的纵横开关 66 被连接在一起。虚拟网络是为每对输出节点准备的，而七个输入缓冲区一组的阵列 62 中每个阵列都包括为机器中的每个虚拟网络准备的例如四个迁移字段缓冲区。

如果结构路由器中预定到输出节点的虚拟信道在数据包的标题迁移字段到达那个虚拟信道时是空闲的，那么该信道在数据包的持续时间内被分配给那个数据包，即直到数据包的结尾迁移字段通过为止。但是，多个数据包可以是在路由器处通过多个输入端为同一虚拟信道接收的。如果虚拟信道已经被分配，那么新的标题迁移字段必须在其迁移字段缓冲区中等待。如果信道虽然未被分配但是用于那个信道的两个标题迁移字段一起到达，那么公正的判优必须发生。迁移字段在公正的判优过程中被选定之前留在输入缓冲区中，向上游施加反压力。

如果分配了输出虚拟信道，在收到来自下游节点的表明在那个节点的输入缓冲区可供那个虚拟信道使用的信号之前，迁移字段不被允许穿过链路转移。

早期的路由器已经使用图 4 图解说明的缓冲区组织，其中每个迁移字段缓冲区被分配给一个特定的虚拟信道并且不能被用于保存与任何其它虚拟信道相关联的迁移字段。图 4 展示到达的迁移字段 100 和用于路由器 200 的一个输入口的迁移字段缓冲区阵列。缓冲区阵列 200 包含一行个个都得到路由器支持的虚拟信道。每行都包括 S 个（S = 2）迁移字段缓冲区 204 和 205、列（F）中的第一迁移字段的指针 201、列（L）中的最后一个迁移字段的指针 202 和指出何时该行中没有迁移字段的空位（E）203。

当迁移字段 100 到达输入信道时，它被储存在由迁移字段的虚拟信道标识符（VCID）字段 101 和迁移字段缓冲区中选定的行的 L-字段确定位置的迁移字段缓冲区中。首先，VCID 被用来确定缓冲区阵列 200 的地址，以便选择与这个虚拟信道相关联的行 210。用于这行的 L-字段指向置于缓冲区中的最后一个迁移字段。为了识别把到达的迁移字段储存到其中的下一个

开着的迁移字段缓冲区，它被加 1。当输入信道的某个特定的虚拟信道被选定用来输出迁移字段时，打算读出的迁移字段缓冲区是利用该虚拟信道的 VCID 字段和缓冲区阵列中相应的行的 F 字段以类似的方法选定的。

模拟实验业已表明，就典型的通信量而言，在给定的时刻，特定的输项目只有一小部分虚拟信道被占用。我们通过仅仅提供小型的片上缓冲区组合而用比较缓慢但比较便宜的片外存储器提供完整的缓冲区阵列可以开拓这种行为。组合中的缓冲区被动态地分配给各个虚拟信道，以致在任何瞬间这些缓冲区都为一部分正在使用的虚拟信道保存迁移字段。缓冲区组合实际上是在芯片外存储器中用于完整的迁移字段缓冲区的高速缓冲存储器。带廉价的外部动态随机存取存储器的简单的路由器 ASIC 可以同时支持大量的虚拟信道和用于每个虚拟信道的大缓冲区。在优选的实施方案中，有 V 个虚拟信道（V = 2560），每个虚拟信道有 S 个迁移字段（S = 4），每个迁移字段在存储器的诸迁移字段中占用 F 位（F = 576）。

现在的发明的第一个优选的实施方案是用图 5 图解说明的。迁移字段缓冲区 400 的比较小的组合被用来把当前活跃的一套迁移字段保存在路由器芯片 500 上。完整的迁移字段缓冲区阵列 200 被置于廉价的芯片外存储器上，并且保存超过组合容量的任何迁移字段。指针数组 300 作为目录保存每个虚拟信道的状态并指出与该虚拟信道相关联的每个迁移字段被放在哪里。

指针数组 300 和缓冲区组合 400 的更详细的视图是用图 6 展示的。指针数组 300 是每个虚拟信道一行。每行都包含 3 个状态字段和 S 个指针字段。在这个实例中，S = 4。F-字段 301 指出 S 个指针字段（S = 4）中哪一个与信道上的第一迁移字段相对应。L-字段 302 指出哪个指针字段与该信道上最后到达的

迁移字段相对应，，而 E-字段 303 如果设置的话则指出该信道空闲并且没有迁移字段出现。

每个指针字段如果在用则规定对应的迁移字段的位置。如果指针字段的数值（P）在[0，B-1]范围内，其中 B 是在缓冲区组合中缓冲区编号，那么迁移字段被放在缓冲区组合 400 的缓冲区 P 中。另一方面，如果 P = B，那么该指针表明对应的迁移字段被放在芯片外的迁移字段缓冲区阵列中。在芯片外的迁移字段缓冲区阵列中的迁移字段是按 VCID 和迁移字段编号放置的，而不是按指针放置的。

这种结构的运用是用图 7 所示的实例予以图解说明的。这张图展示在单一的虚拟信道（VCID = 4）的指针数组 300、缓冲区组合 400 和晶片外的缓冲区阵列 200 中包含 3 个迁移字段的项目。这些迁移字段当中的两个迁移字段（第一个和最后一个）在缓冲区组合中，而中间的迁移字段在芯片外的缓冲区阵列中。为了确定特定的迁移字段的位置，VCID 被用来在指针数组 300 范围内选择行 310。在被选定的行的范围内状态字段详细说明三个迁移字段都在公用缓冲区中，第一个迁移字段用指针 1（F=1）识别，而最后一个迁移字段用指针 3（L=3）识别。指针 2 识别中间的迁移字段。指针 1 包含说明第一个迁移字段在缓冲区组合 400 的位置 5 中的数值 5。类似地，指针 3 说明最后一个迁移字段在缓冲区组合 400 的位置 3 中。但是，指针 2 包含数值“零”，这表明中间的迁移字段在芯片外的缓冲区阵列 200 中，在第 4 行第 2 列。行是由 VCID 确定的，在这种情况下是 4，而列是由指针的位置确定的，就这个例子来说是指针 2。在优选的实施方案中，缓冲区组合包含标号为 0 到 2n-2 的 2n-1 个缓冲区，其中一些 n 和值 2n-1 表示没有指针。

为了看到迁移字段缓冲区高速缓存的优点，考虑我们的示范系统，该系统有 V = 2560 个虚拟信道，每个虚拟信道有 S = 4 个容纳 F = 576 位的迁移字段的迁移字段缓冲区。就路由器的每个输入信道而言，传统的缓冲区组织（图 4）需要 V × S ×F = 5,898,240 位的存储器。另一方面，采用具有 P = 255 个缓冲区的芯片上的缓冲区组合与芯片外的缓冲区阵列相结合（图 5），就每个输入信道 241,600 位芯片上的总存储量而言，需要 P × F = 146,880 位的缓冲区组合和 37 × V = 94,720 位的指针数组。另外，5,898,240 位的芯片外的缓冲区阵列也是必不可少的。这意味着在芯片上的存储器需求方面减少 24 倍。

输入信道控制器采用双重临界值的缓冲区管理算法来保证组合中缓冲区的适当的供应。无论何时只要该组合中空闲缓冲区的数目落在某个临界值（例如，32）以下，信道控制器就开始把迁移字段从缓冲区组合 400 驱逐到芯片外的缓冲区阵列 200 中，并且更新指针数组 300 以指出这种变化。一旦开始逐出迁移字段，该逐出过程就将持续到组合中空闲缓冲区的数目超过第二个临界值（例如，64）。在逐出期间里，被驱逐的迁移字段可以用任何算法（随机、可利用的第一个、或者最近用过的最后一个）选定。尽管 LRU 算法给出稍高的性能，但是逐出事件如此稀少，以致可能的最简单的算法就足够了。为了制止上游的控制器无限期地等待迁移字段离开缓冲区，逐出过程是必不可少的。如果不逐出，这种等待将在无关的虚拟信道之间产生依赖作用，因此有可能导致树饱和或僵持。象在上述的未审的 PCT 申请中介绍的那种路由器那样的早期的路由器采用基于信用的流量控制，在这种情况下上游的控制器为每个下游的虚拟信道保持信用计数 73（图 3），空闲迁移字段缓冲区的数目的计数。输出控制器仅仅在它对某个特定的虚拟信道具有表明下游有可利用的迁移字段缓冲区的非零的信用计数时才把迁移字段

转发给那个虚拟信道。无论何时只要上游的控制器转发迁移字段，它就将用于对应的 VC 的信用计数减 1。当下游的控制器倒空迁移字段缓冲区时，它向上游转移信用。上游的控制器在收到每个信用时将用于对应的虚拟信道的信用计数加 1。

采用迁移字段缓冲区高速缓存，为了避免在未必存在的缓冲区组合溢出事件中超额订购这个带宽，上游输出信道控制器的流量控制战略必须被修改。这种附加的流量控制是必要的，因为芯片外的缓冲区阵列具有比芯片上的缓冲区组合以及信道本身低得多的带宽。如果该组合被充满，迁移字段必须被逐出到芯片外的缓冲区阵列中，而且来自所有的虚拟信道的转移在组合中有空间可以利用之前都必须被阻断。为了处理迁移字段缓冲区的高速缓存，早期的基于信用的流量控制机制通过增添反映在下游的缓冲区组合中空闲的缓冲区数目的缓冲区组合的信用计数得以扩展。上游的控制器在它可以向下游转发迁移字段之前必须有两个非零的信用计数，即用于虚拟信道的信用计数 73 和缓冲区组合的信用计数 75。这保证在缓冲区组合中有供所有到达的迁移字段使用的空间。开始时，为所有的虚拟信道和所有的虚拟信道共享的缓冲区组合设置最大的计数。每逢上游的控制器转发迁移字段时，它就将相应的虚拟信道的信用计数和共享的缓冲区组合的信用计数都减 1。当下游的控制器把信用向上游发送给任何虚拟信道时，如果得到信用的迁移字段是从缓冲区组合中发送出来的，那么它将把某个组合位设置在信用中。当上游的控制器收到设置了组合位的信用时，它把缓冲区组合的信用计数以及虚拟信道的信用计数都加 1。在把迁移字段从缓冲区组合逐出到芯片外的缓冲区阵列时，组合专用的信用被发送到上游的控制器，以便更新其反映这种变化的信用。因此，新迁移字段的转移仅仅在缓冲区组合和芯片外的缓冲区阵列之间转移迁移字段期间才被停止。

作为替代，为了防止组合超出限度可以使用阻断型流量控制而不是基于信用的流量控制。采用这种方法，当缓冲区组合中空闲的缓冲区的数目落在临界值以下时，在所有上游的信用中都设置阻断位。当这个阻断位被设定时，上游的输出控制器被禁止向下游发送任何迁移字段。一旦空闲缓冲区的数目增加到临界值以上，该阻断位就被清除，于是上游的控制器可以重新开始发送迁移字段。阻断型流量控制是有利的，因为它在上游的控制器中不需要组合的信用计数，而且可以为了其它理由被用于禁止迁移字段的传输。

如同在图 8 中图解说明的那样，优选的替代实施方案不用指针数组，而是使用成组相联的高速缓冲存储器组织。这种组织是由状态数组 500、一个或多个高速缓冲存储器阵列 600 和逐出 FIFO 800 组成的。芯片外的缓冲区阵列 200（未示出）也被使用，以便支援高速缓冲存储器阵列。与特定的虚拟信道 V 的特定的缓冲区 B 相关联的迁移字段以类似于传统的成组相联的高速缓冲存储器的方法被编入每个高速缓冲存储器阵列中可能的位置（例如，参阅 Hennessey 和 Patterson 的 Computer Architecture: A Quantitative Approach（计算机体系结构：一种定量的方法），第二版，摩根考夫曼出版社，1996，第 5 章）。然后，储存在每个位置的迁移字段都按相关联的高速缓冲存储器部件予以记录。如果在每个高速缓冲存储器位置的有效位所包含的项目代表有效的数据，那么它将发出信号。

把高速缓冲存储器中的特定的迁移字段 F= {V:B}考虑进去的位置是由 F 的低阶位确定的。例如，考虑这样的情况：有 V = 2560 个虚拟信道、每个虚拟信道有 S= 4 个缓冲区，而且在 A = 2 个高速缓存存储器阵列中每个阵列有 C = 128 个项目。在这种情况下，迁移字段标识符 F 是 14 位，12 位的 VCID 和 2 位的缓冲

区标识符 B。7 位的高速缓冲存储器阵列索引是通过把 B 附加到 V 的 5 个低阶位上构成的，I = {V [4:0]:B。然后，V 剩余的 7 个高阶位被用于高速缓冲存储器的标记，T = V [11:5]。

当迁移字段在信道上到达时，其 VCID 被用于检索状态数组 500，以便读出 L 字段。然后，这个字段被加 1，以便确定打算存入该迁移字段的缓冲区在虚拟信道内的位置 B。然后，通过把 B 与 VCID 的 5 个低位连接起来形成阵列索引（I），并且用这个索引访问高速缓冲存储器阵列 600。尽管展示的是两个高速缓冲存储器阵列，但是应该理解成可以使用的阵列数目是任意的。然后，高速缓冲存储器阵列之一被选定利用选择算法接收这个迁移字段。优先选择可以被赋予在这个位置包含无效的项目的阵列，如果所有的项目都是有效的，则赋予包含最近用过的最后一个位置的阵列。如果被选定的高速缓冲存储器阵列已经包含有效的迁移字段，那么这个迁移字段首先被读出并且与其标识符（VCID 和 B）一起被放进逐出 FIFO 800。然后，到达的迁移字段被写入腾出的位置，那个位置上的标记与 VCID 的高位一起被更新，于是那个位置被注明是有效的。

当读出下一个迁移字段的请求来自某个特定的虚拟信道时，VCID 再一次被用来检索状态数组 500，并且把 F 字段读出，以便确定打算读出的缓冲区位置 B。然后，F 字段被加 1 并且被写入该状态数组。然后，VCID 和缓冲区编号（B）被用来在三个位置中搜寻迁移字段。首先，象前面介绍的 I = {V [4:0],B}那样形成索引，并且访问高速缓冲存储器阵列。利用比较器 701 将从每个阵列中访问到的标记与 V [11:5]进行比较。如果有某种匹配并且有效位已设定，那么迁移字段已被定位并且已借助三态缓冲区 702 从相应的阵列中读出。然后，将包含迁移字段的项目的有效位清除，以便为较晚的使用释放这个项目。

如果被请求的迁移字段在任何高速缓冲存储器阵列中都没有找到，则搜索逐出 FIFO，以确定它是否包含与 VCID 和 B 相匹配的有效的项目。如果匹配被找到，则将迁移字段从逐出 FIFO 中读出，并且将项目的有效位清除，以便释放那个位置。最后，如果迁移字段在高速缓冲存储器阵列和逐出 FIFO 中找不到，则在位置{V[11:0],B}读芯片外的迁移字段数组 200，以便从备份存储器中取回该迁移字段。

如同第一个优选实施方案那样，为了确保来自上游的控制器的请求不超出逐出 FIFO 的限度，必须使用流量控制和完成逐出操作。如上所述，为了在逐出 FIFO 中的自由空间落在临界值以下时阻止上游的控制器发送迁移，既可以使用基于信用的流量控制也可以使用阻断型的流量控制。在超过这个临界值时，来自逐出 FIFO 的迁移字段也被写入芯片外的迁移字段数组 200。

与第一个优选实施方案相比较，成组相联的组织需要更少的芯片上的存储器位就能实现，但是有可能由于冲突失误具有比较低的命中率。就上述的示范数字而言，状态数组需要 2560 × 5 = 12,800 位的存储器，具有规模为 256 个迁移字段的项目的高速缓存储器阵列需要 256 × (576 + 1 + 7) = 149,504 位，具有 16 个项目的逐出 FIFO 需要 16 × (576 + 14) = 9,440 位，总共 171,744 位，与第一个优选实施方案的 241,600 位不相上下。

因为迁移字段不能驻留在任何迁移字段缓冲区中，而宁可说是仅仅驻留在每个阵列数组（一套缓冲区）的单一位置，所以发生冲突失误。因此，活跃的迁移字段可能由于几个其它的迁移字段被安排在同一位置而在全部缓冲区被充满之前被逐出该阵列。但是，这些冲突失误的影响由于联合搜索起牺牲品高速缓存器作用的逐出 FIFO 多少得以减轻（见 Jouppi 发表在

Proceedings of 17th Annual International Symposium on Computer Architecture, 1990, pp 364-375（关于计算机体系结构的第 17 次年会的会议录，1990 年，第 364 页到 375 页）上的“Improving Direct Mapped Cache Performance by Addition of a Small Fully-Associasive Cache and Prefetch Buffers（通过添加小型的全面联合的高速缓存器和预取缓冲区提高直接映射的高速缓存器的性能）”）。

通过使用图 9 图解说明的第三个优选实施方案可以进一步降低对迁移字段高速缓存器的存储要求。这个实施方案也使用成组相联的高速缓存器组织。但是，不同于图 8 所示的实施方案，它把与给定的虚拟信道相关联的全部状态和缓冲区都放进阵列 900 的单一的高速缓存器项目。尽管展示的是单一的高速缓存器阵列 900，但是熟悉这项技术的人将理解可以使用的高速缓存器阵列的数目是任意的。把所有的状态都放在高速缓存器阵列中可以取消对状态数组 500 的需求。用于这个组织的逐出 FIFO1000 也包含用于虚拟信道的全部状态和缓冲区（但不需要字段 B）。类似地，为了把用于每个虚拟信道的状态字段（F、L 和 E）包括在内，芯片外的迁移字段阵列 200（未示出）也被扩展。

当迁移字段到达缓冲区时，用于该迁移字段的虚拟信道的项目在芯片上定位，如果它已经不在那里则把它带到芯片上。然后，该项目被更新，以便插入新的迁移字段和更新 L 字段。具体地说，到达的迁移字段的 VCID 字段被用来对三个位置中的虚拟信道项目进行搜索。首先，利用 VCID 的低位（例如，V[6:0]）作为索引和利用 VCID 的高位（例如,V[11:7]）作为标记搜索高速缓冲存储器阵列。如果储存的标记与提交的标记相匹配并且有效位已设定，那么项目就已经被找到。在这种情况下，相匹配的项目的 L 字段被读出并且被用来在项目范围内选

择储存到达的迁移字段的位置。然后，L 项被加 1 并且被写入该项。

如上所述，在高速缓冲存储器失误的情况下，在高速缓冲存储器阵列中没有匹配，高速缓冲存储器阵列之一被选定，以便接收必要的项目。如果被选定的项目当前是有效的，保存着不同的项目，那么那个项目被逐出 FIFO 1000。然后，就必要的项目对逐出 FIFO 进行搜索。如果被找到，该项目将从缓冲区加载到选定的高速缓冲存储器阵列并且象前面介绍的那样被更新。如果在逐出 FIFO 中没找到项目，那么将从芯片外的缓冲区阵列中提取该项目。为了在等待项目从芯片外装入时允许处理其它到达的迁移字段，悬而未决的迁移字段被暂时储存在保存过错的寄存器（1002）中，直到芯片外的索引完整为止。一旦项目已从芯片外装入，更新就象前面介绍的那样继续进行。

为了从给定 VCID 的高速缓冲存储器中读出迁移字段，搜索以类似于写入的方式继续进行。虚拟信道项目如果尚未在那里则从逐出 FIFO 或芯片外的缓冲区阵列装入高速缓冲存储器。该项目的 F 字段被用来在用于读出的项目内选择迁移字段。最后，F 字段被加 1，并且被重新写入该项目。

如同其它的优选实施方案那样，阻断型的或基于信用的流量控制对于制止上游的控制器在逐出 FIFO 或保存过错的寄存器中空闲位置的数目落在门限值以下时发送迁移字段是必不可少的。

第三个优选实施方案的优点是在虚拟信道的数目非常大时其芯片上的存储器阵列的规模却相当小。因为芯片上没有每个虚拟信道的阵列，所以规模在很大程度上独立于虚拟信道的数目（标记字段的规模确实随着虚拟信道的数目以对数方式增

加）。例如，就我们的示范参数而言，V=2560、S=4、F=576、具有 64 个项目（256 个迁移字段）的芯片上的阵列包含 64(((S(F) + 12) = 148,224 位。逐出 FIFO 和有 16 个项目的保存过错的数组各自分别添加额外的 37,152 位和 9,328 位。芯片上存储器的总量（194,704）略微大于第二个优选实施方案的 171,744 ，但是当虚拟信道的数目增加到 2560 以上时该总量基本上保持恒定不变。

图 10 图解说明使用三个优选实施方案中任何一个时迁移字段缓冲区的高速缓存效果。这张图展示模拟 512 个节点（8 × 8 × 8）的三维环面网络的实验结果，其中对每个目的地的通信量是在分开的虚拟信道上运行的。在这个模拟实验期间，虚拟信道的占有率是在每个点及时记录的。这张图展示这种信道占用率的两种直方图，分别对应于以其最大容量的 30 %（典型的负载）操作的网络和以其最大容量的 70 %（极重的负载）操作的网络。即使在 70 %的容量下，38 个以上虚拟信道缓冲区将在给定的点及时被占用的概率也在 10^{-5} 以下。这意味着容量为 38S 个迁移字段的迁移字段缓冲区的高速缓冲存储器应该有高于 99.999 %的命中率（或者反过来说，过错率在 0.001 %以下）。如果我们将这个结果外推，它暗示容量为 256 个迁移字段的迁移字段缓冲区的高速缓冲存储器将具有难以察觉地小的过错率。

尽管我们已经就三个优选实施方案介绍了迁移字段高速缓冲存储器的特定的安排，但是熟悉结构路由器设计的人将理解可能有许多替代安排和组织。例如，尽管我们已经介绍了基于指针的和成组相联的组织，但是人们还可能使用全面关联的组织（尤其就各种小型高速缓存器的容量而言）、混编表或树结构的高速缓冲存储器。尽管我们已经介绍了一个迁移字段和一个

虚拟信道的高速缓冲存储器的块规模，但是其它规模也是可能的。尽管我们已经介绍了采用字段F，L和E的虚拟信道状态的特定的编码，但是许多其它编码也是可能的。此外，尽管我们仅仅高速缓存了迁移字段和输入虚拟信道状态的内容，但是这种高速缓存可能被延伸到高速缓存与虚拟信道和输出虚拟信道状态相关联的输出口。

尽管这项发明已参照其优选的实施方案被具体地展示和介绍，但是熟悉这项技术的人将理解不脱离权利要求书定义的本发明的精神和范围可以在形式和细节上做出各种变化。例如，虽然优选的实施方案在结构路由器中提供迁移字段缓冲区，但是本发明可以延伸到在其它路由器中的信息单位集，例如数据包和信息。

说 明 书 附 图

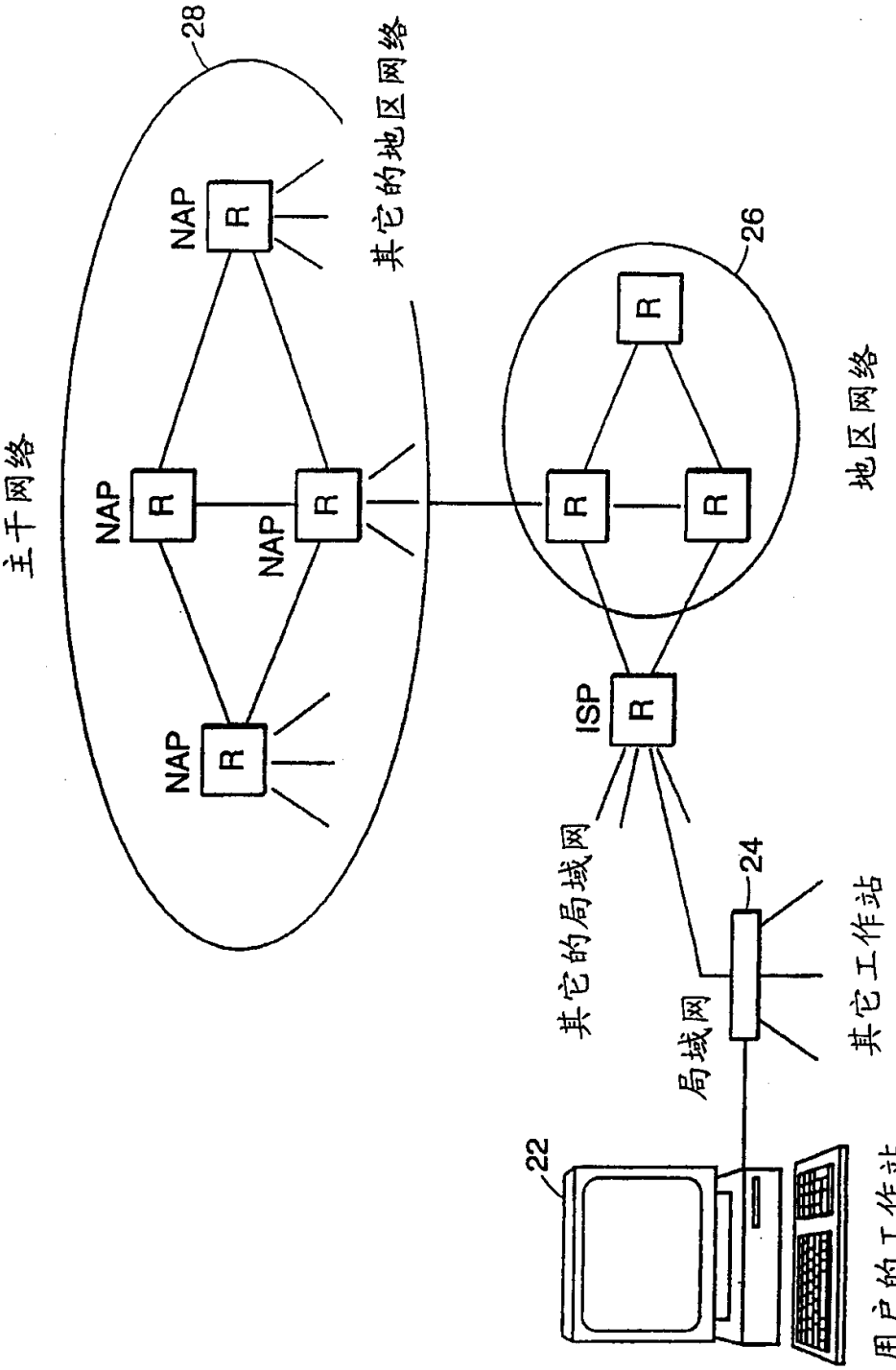

主干网络
28
其它的地区网络
NAP
R
NAP
R
NAP
R
NAP
R
26
R
R
R
地区网络
ISP
R
其它的局域网
24
局域网
其它工作站
22
用户的工作站

图 1

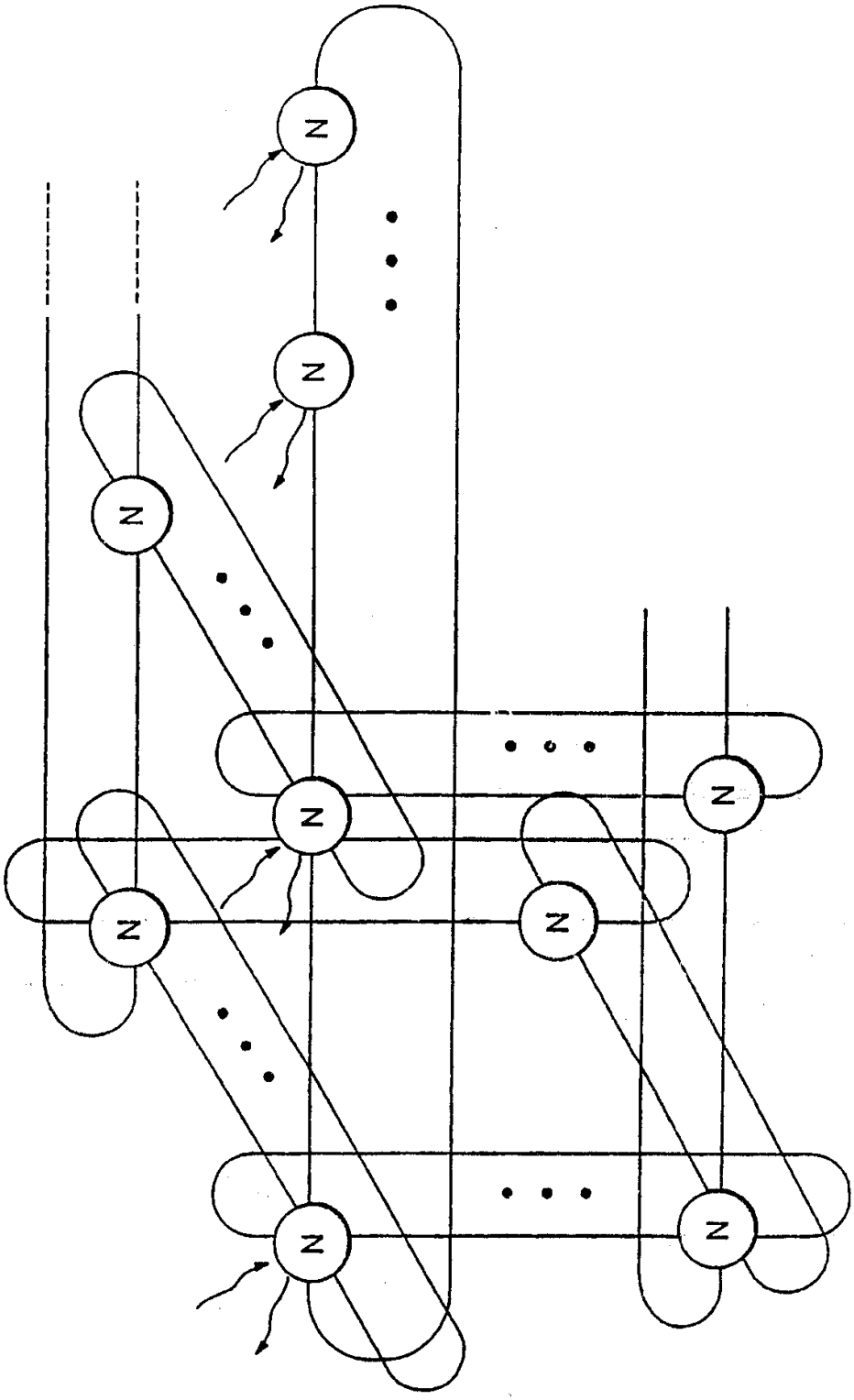


図2

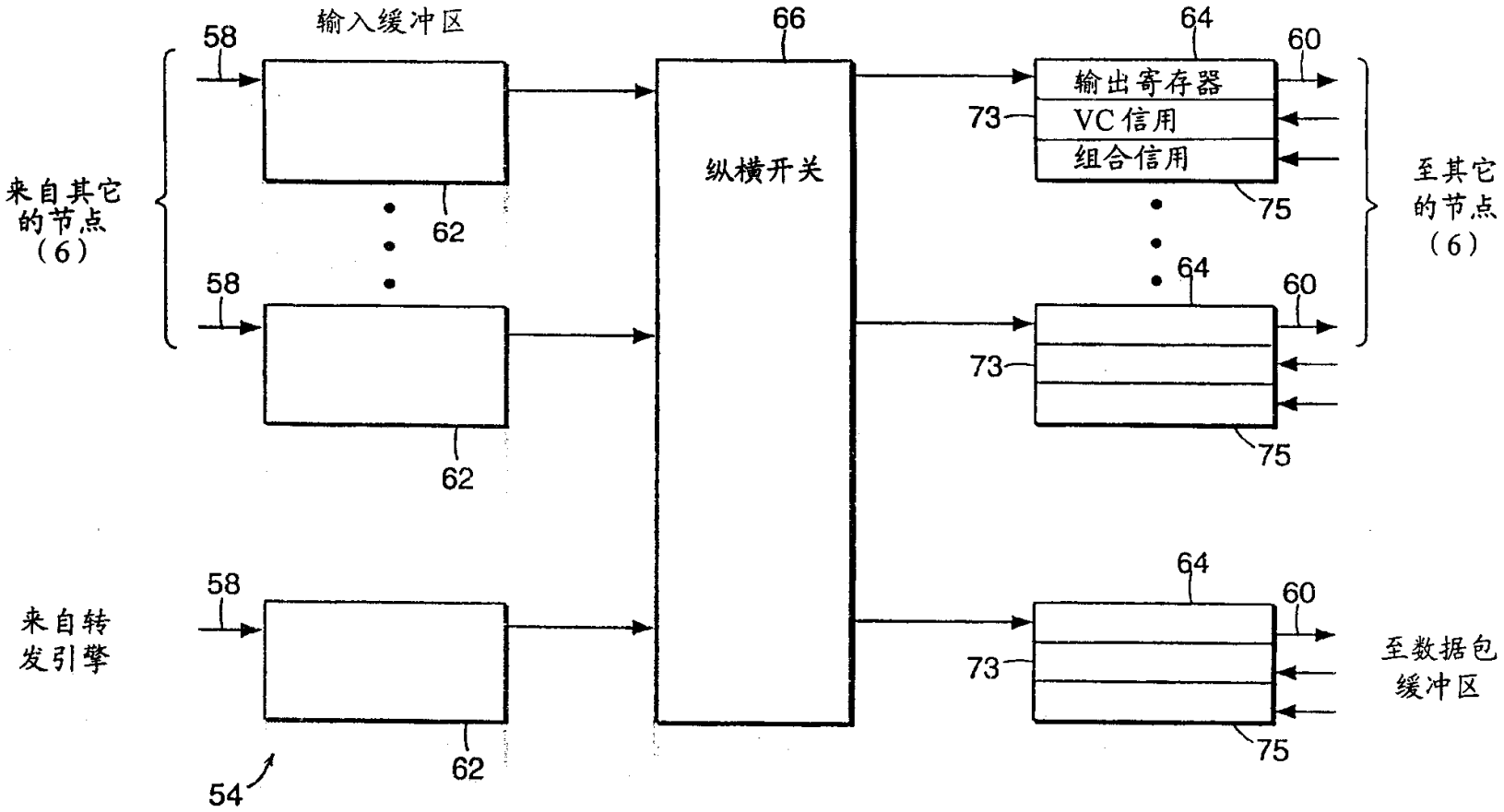

输入缓冲区
58
66
64
60
输出寄存器
VC 信用
组合信用
73
75
62
来自其它的节点（6）
至其它的节点（6）
纵横开关
来自转发引擎
54
至数据包缓冲区

图 3

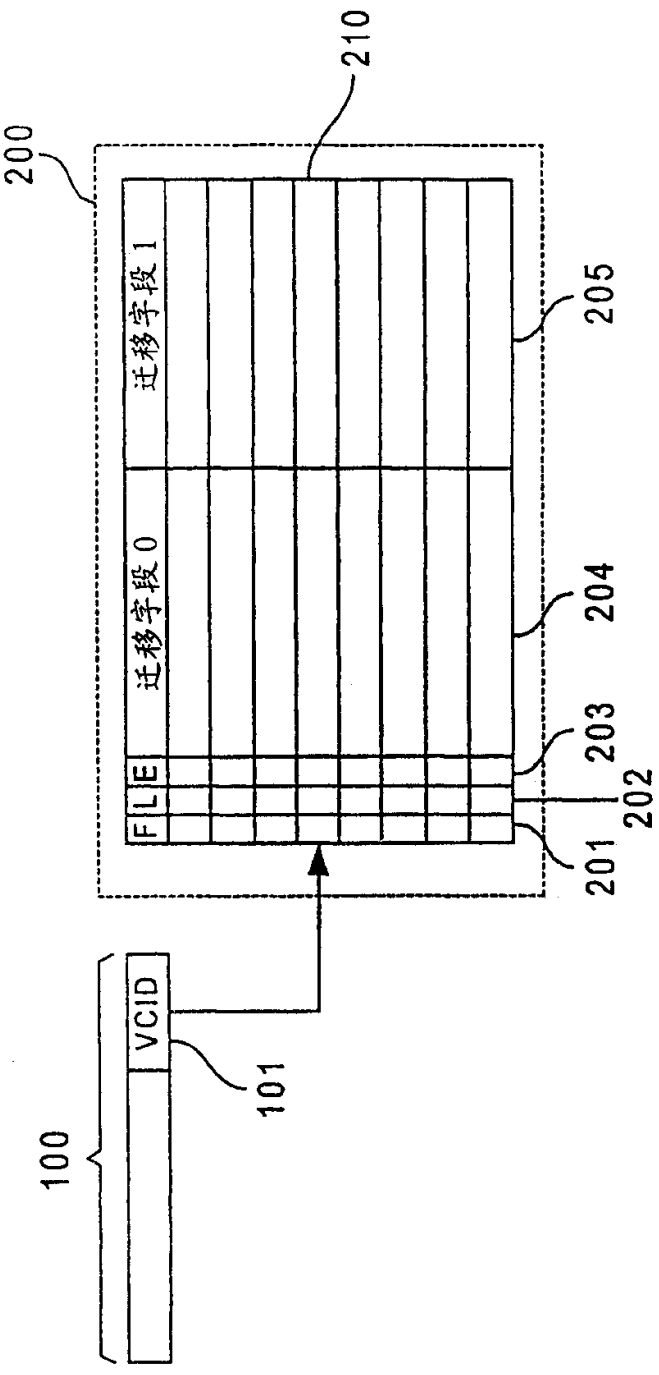


图 4

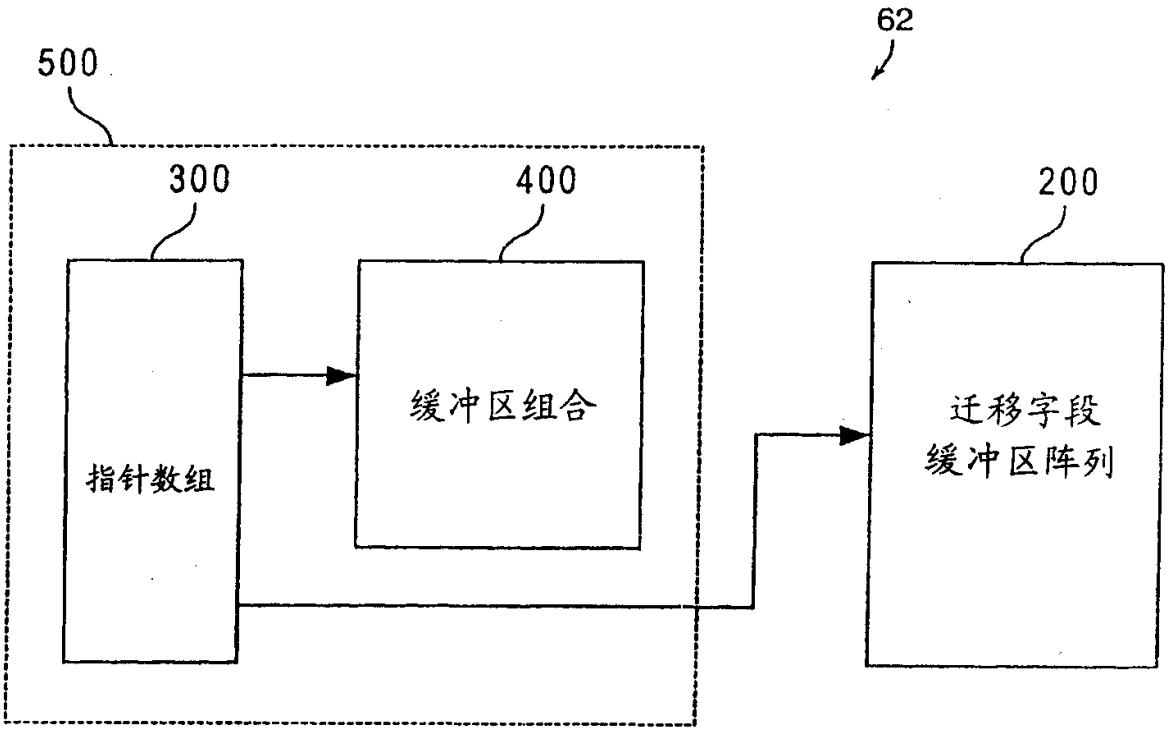


图 5

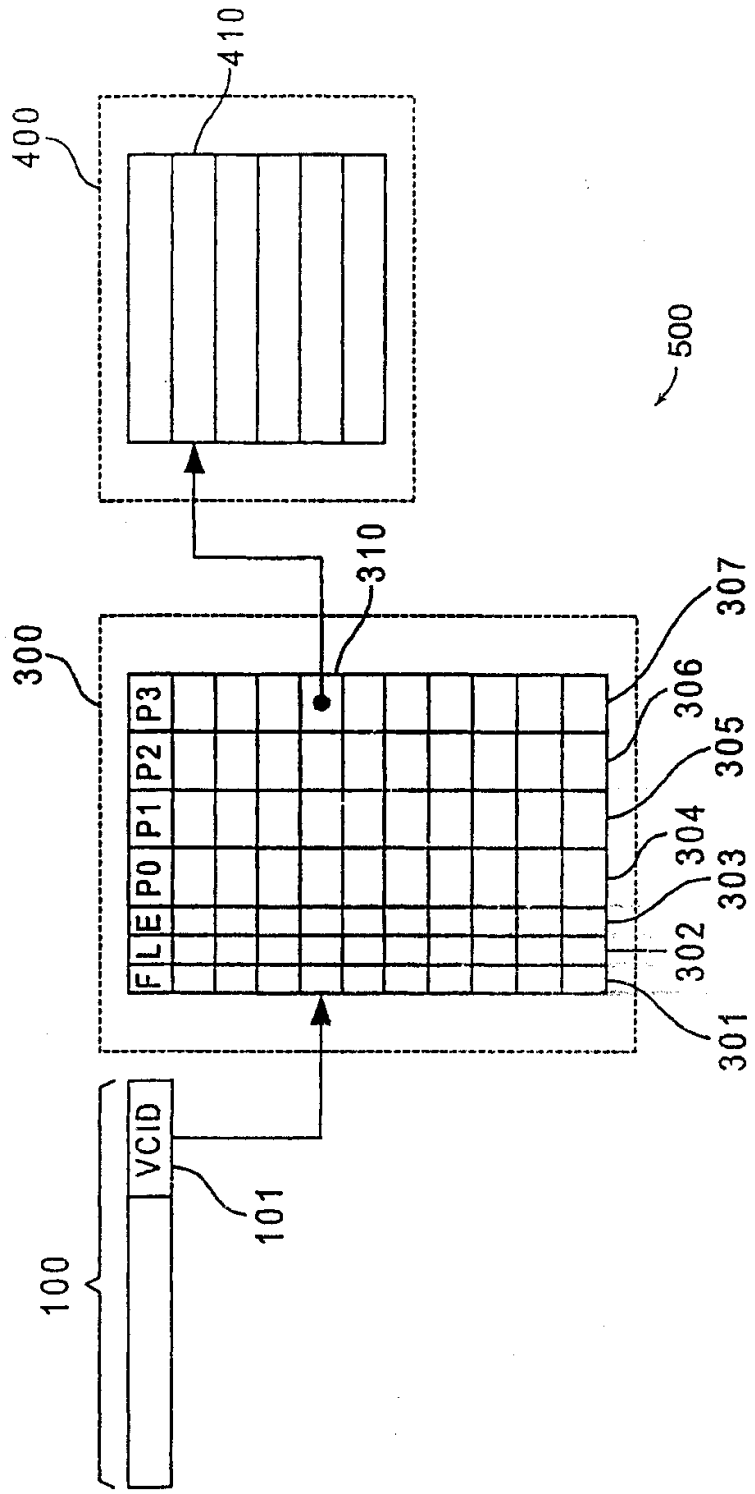

400
410
500
310
300
307
306
305
304
303
302
301
P3
P2
P1
P0
E
L
F
100
VCID
101

図 6

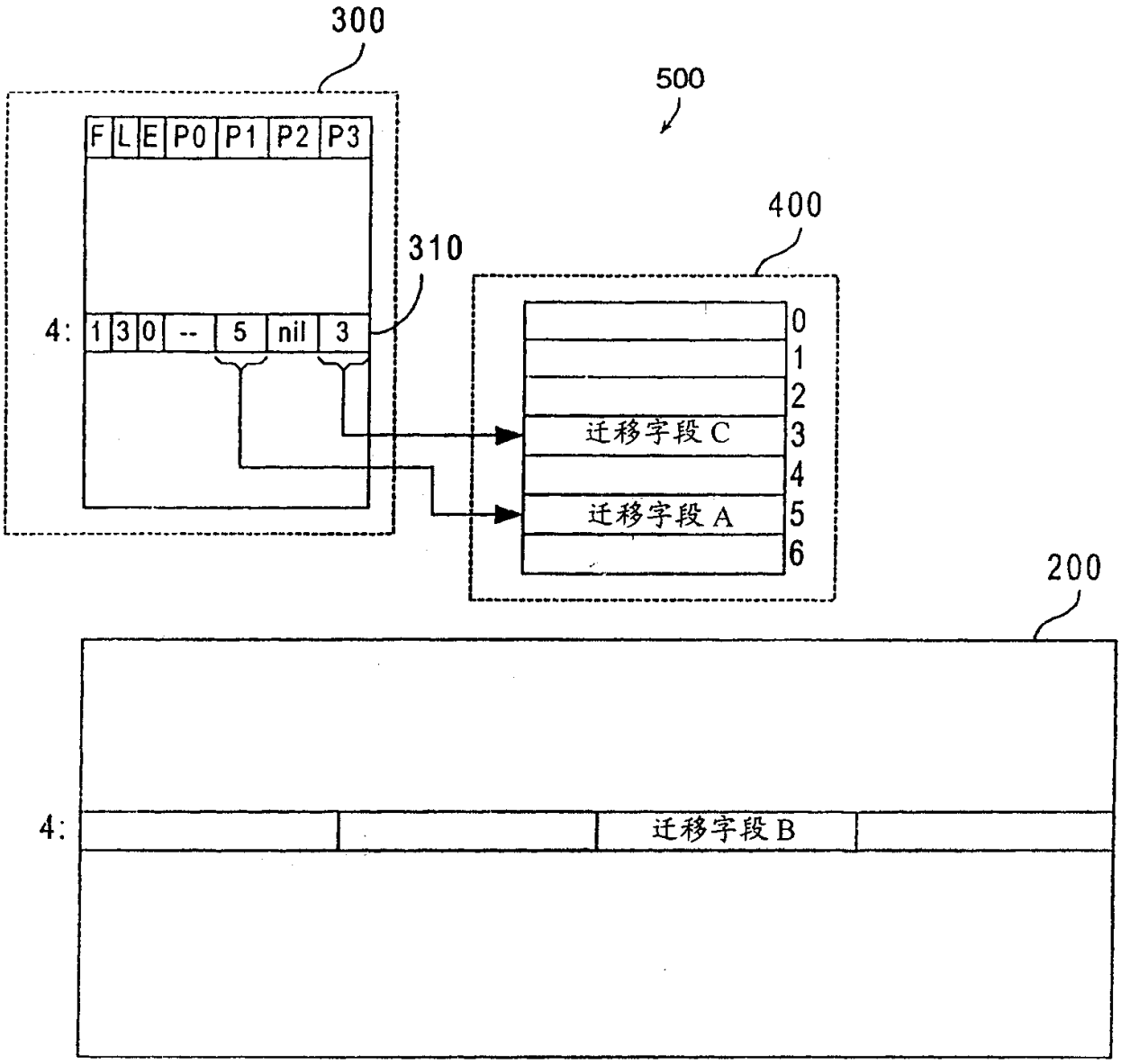

300
500
F L E P0 P1 P2 P3
400
310
4:
1 3 0 -- 5 nil 3
0
1
2
迁移字段 C
3
4
迁移字段 A
5
6
200
4:
迁移字段 B

图 7

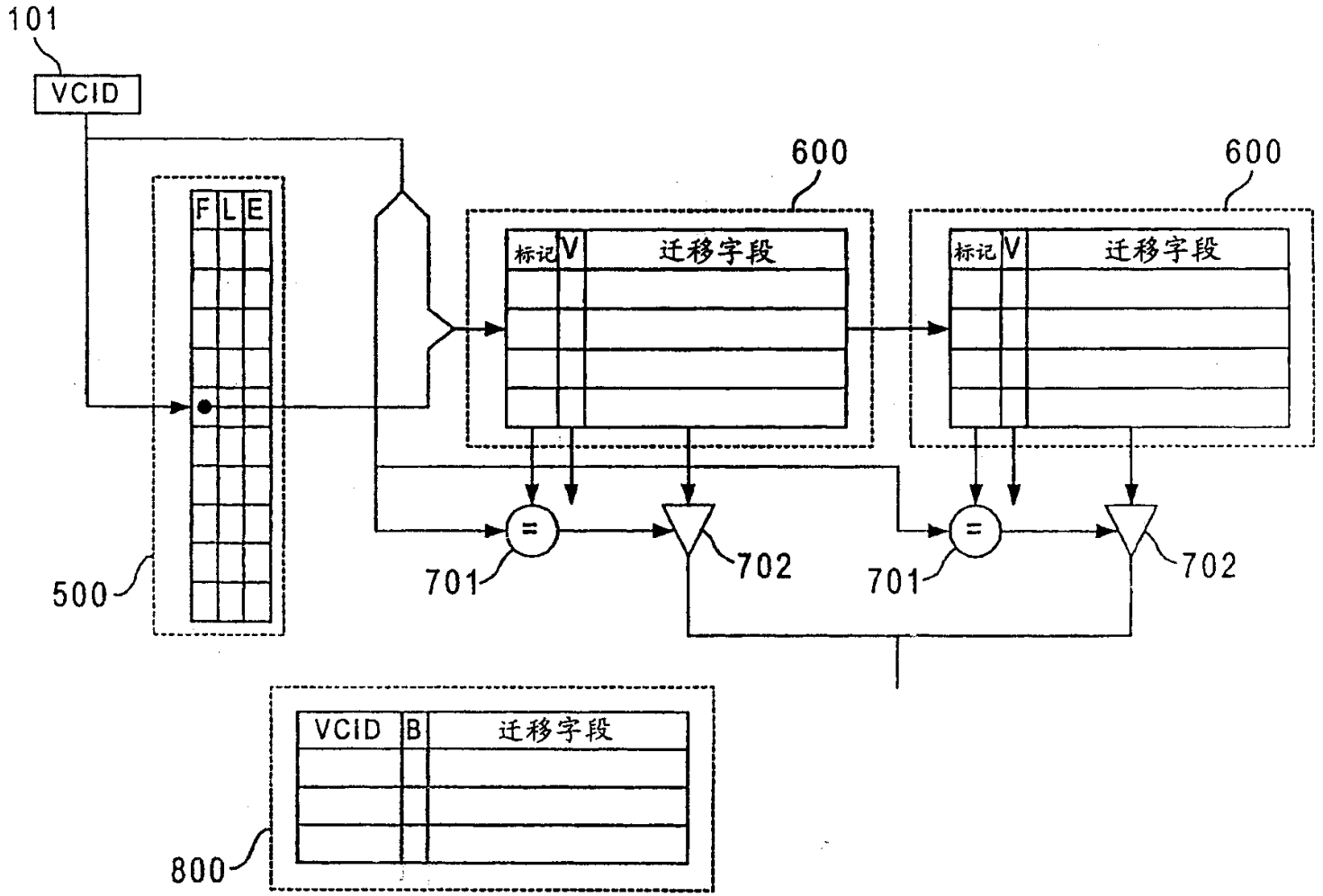

101
VCID
F L E
500
600
标记 V 迁移字段
600
标记 V 迁移字段
=
701
702
=
701
702
VCID B 迁移字段
800

图 8

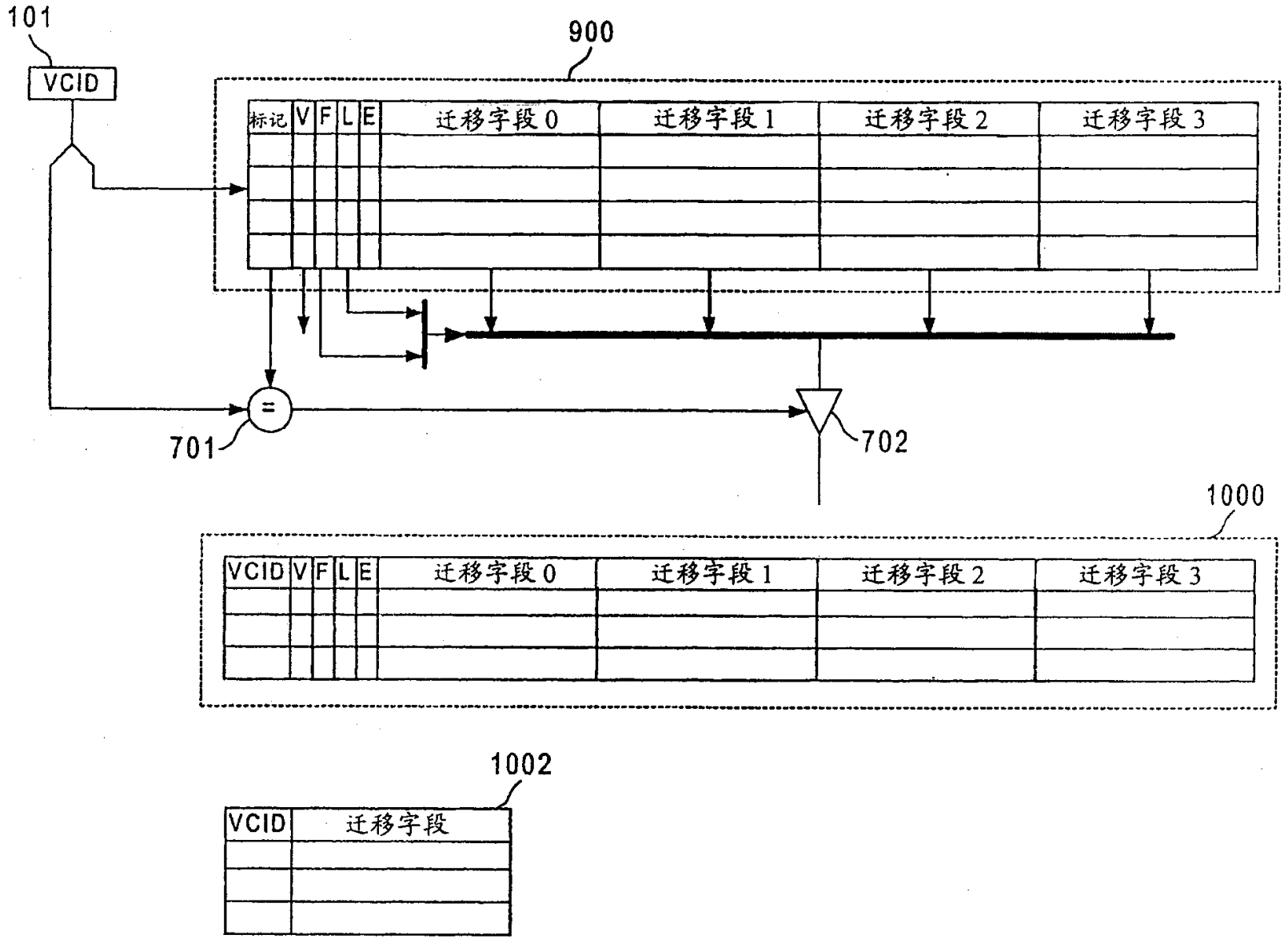

101
VCID
900
标记
V
F
L
E
迁移字段 0
迁移字段 1
迁移字段 2
迁移字段 3
=
701
702
1000
VCID
V
F
L
E
迁移字段 0
迁移字段 1
迁移字段 2
迁移字段 3
1002
VCID
迁移字段

图 9

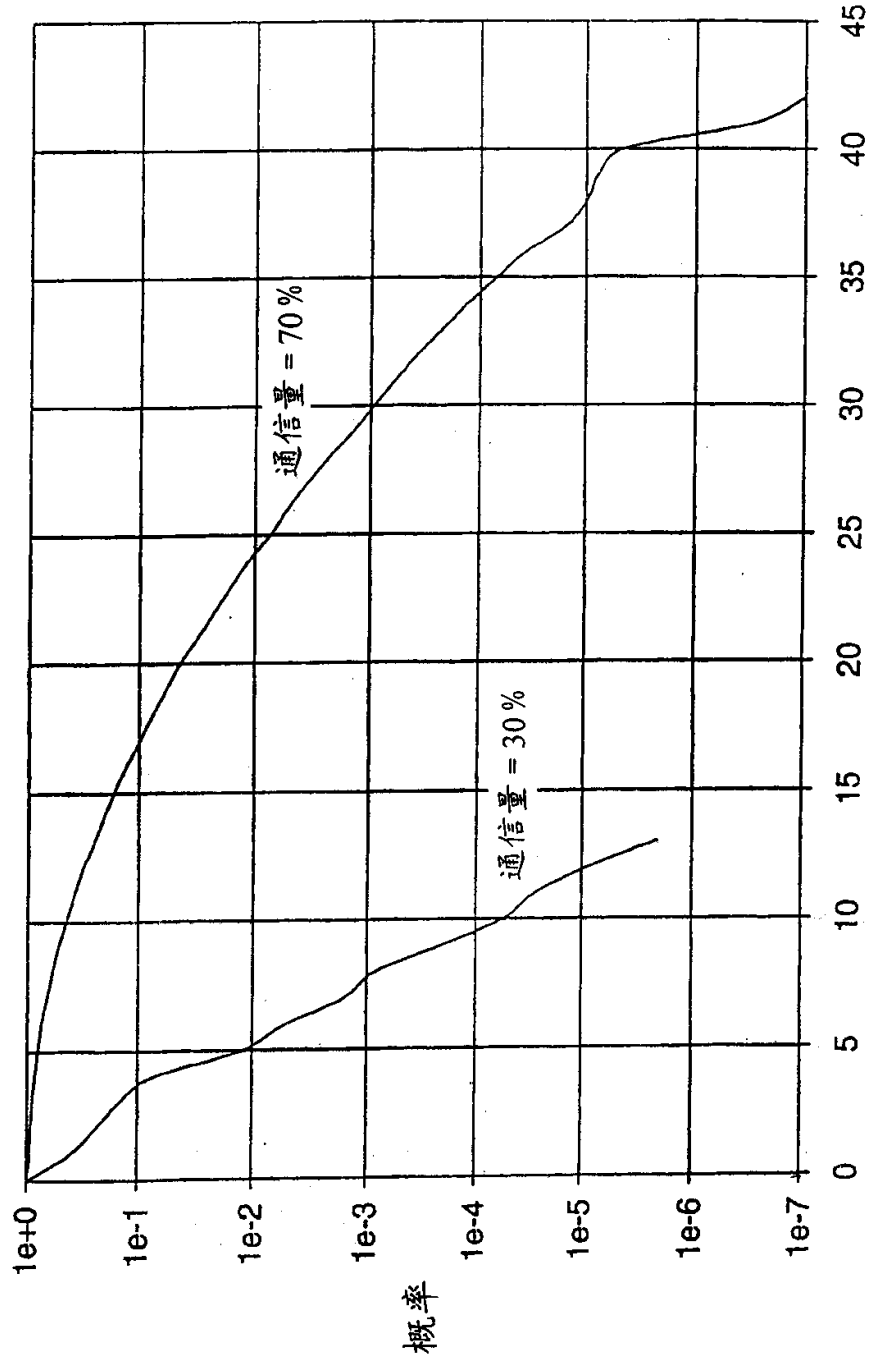

概率
1e+0
1e-1
1e-2
1e-3
1e-4
1e-5
1e-6
1e-7
0
5
10
15
20
25
30
35
40
45
通信量 = 70 %
通信量 = 30 %
被占用的虚拟信道数

图 10