[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 13/14 (2006.01)
H04Q 11/04 (2006.01)

[12] 发明专利说明书

专利号 ZL 01820868.1

[45] 授权公告日 2008 年 6 月 25 日

[11] 授权公告号 CN 100397373C

[22] 申请日 2001.10.19 [21] 申请号 01820868.1

[30] 优先权

[32] 2000.10.19 [33] US [31] 09/693,357

[86] 国际申请 PCT/US2001/050544 2001.10.19

[87] 国际公布 WO2002/034004 英 2002.4.25

[85] 进入国家阶段日期 2003.6.19

[73] 专利权人 英特拉克蒂克控股公司

地址 美国新泽西州

[72] 发明人 科克·S·里德 约翰·赫斯

[56] 参考文献

US5649108A 1997.7.15

US5859981A 1999.1.12

US5408231A 1995.4.18

US5917426A 1999.6.29

EP0868104A2 1998.9.30

US5535197A 1996.7.9

WO97/30555A2 1997.8.21

审查员 江 红

[74] 专利代理机构 北京市柳沈律师事务所

代理人 马 莹 邵亚丽

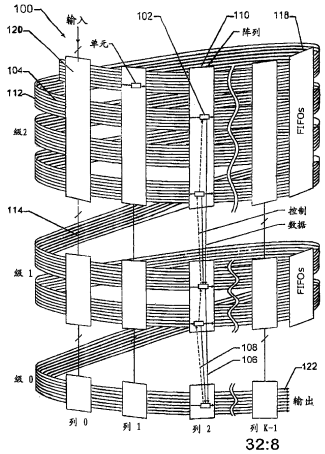
权利要求书 4 页 说明书 16 页 附图 10 页

[54] 发明名称

可伸缩蠕虫洞路由选择集中器

[57] 摘要

一种互连结构(100)通过在控制单元之间利用控制信号进行单比特路由选择，实质性地改进了信息集中器(700)的运行。 该互连结构和运行技术支持蠕虫洞路由选择和消息的流动。 消息分组总是缓存在结构内，且不会被丢弃，从而保证了所有进入该结构的分组可以退出。 在一个例子中，该互连结构包括一个连接在不相交路径上的多个节点的互连线条带(112)。 互连线条带(112)从源级至目标级的缠绕过多级。 绕匝的数目从源级至目标级减少。该互连结构还包括多个列，这些列由耦合通过各级的缠绕交叉区横跨条带的节点的互连线构成。



32:8

1. 一种互连结构，包括:

一个并行数据路径的集合S;

耦合在该并行数据路径的集合中的多个节点，这些并行数据路径被设置在从源级至目标级的多个级中，交叉源级中数据路径的节点数大于交叉目标级中数据路径的节点数，使得所述互连结构是一个集中器，节点的级完全由该节点在所述结构中的位置确定;

所述并行数据路径的集合S,使得以分组形式的数据可以从该互连结构外的一个或多个设备进入到所述数据路径，并在集合 S 的一个或多个数据路径上的以蠕虫洞传输方式移动到一个数据输出端口；其中，

对于在集合S的数据路径P上的节点A:

数据可以从所述节点A移动到在数据路径P上该节点A的直接后继节点B，使得数据从节点A至节点B的移动是更接近数据路径P数据输出端口的一个移动步骤；或者

数据可以从所述节点A移动到一个集合S中数据路径Q上的节点C，所述节点C比节点B相对数据路径P输出端口更靠近数据路径Q的输出端口。

2. 按照权利要求1所述的互连结构，还包括:

一在所述数据路径Q上的节点D，该节点D是所述节点C的直接前任节点，其中，数据从节点D传送到节点C优先于数据从节点A传送到节点C。

3. 按照权利要求1所述的互连结构，其中，在该互连结构中的业务不会造成数据被丢弃。

4. 一种互连结构，包括:

设置在一结构中的多个节点（102），该结构包括:

从源级至目标级的级层;

沿级展开的不相交路径中的多个节点；和

在级的不相交路径的交叉区中的多个节点，源级交叉区中的节点数大于目标级交叉区中的节点数，使得所述互连结构是一个集中器，节点的级完全由该节点在所述结构中的位置确定；和

多条将节点耦合在所述结构中的互连线（104），包括用于级 L 的路径 P 上节点 N 的互连线；包括：

一消息输入互连线，与级 L 的路径 P 上的第一相临节点相耦合；

一消息输出互连线，与级 L 的路径 P 上的第二相临节点相耦合；

至少一条消息互连线，与一个或多个节点 N 的朝向源的节点相耦合，用于从在分层中朝向源的节点接收数据，和/或与一个或多个节点 N 的朝向目标的节点相耦合，用于从在分层中朝向目标的节点发送数据；和

至少一条控制互连线，与一个节点 N 的朝向源的节点相耦合，用于向朝向源的节点发送控制信号，和/或与一个节点 N 的朝向目标的节点相耦合，用于从在分层中朝向目标的节点接收控制信号。

5. 按照权利要求 4 所述的互连结构，还包括：

一种与节点 N 关联的逻辑，该逻辑能够判断节点 N 是否被级 L 的路径 P 上的消息所占据，并在该判断的基础上将控制信号发送至朝向源的节点、加速消息在朝向源的节点的前进。

6. 按照权利要求 4 所述的互连结构，还包括：

多个列，每个列将一个级中不相交路径交叉区中的多个节点互连，所述列包括节点之间的互连，该互连包括在至少一条消息互连线和至少一条控制互连线上的朝向源和朝向目标的耦合。

7. 按照权利要求 4 所述的互连结构，还包括：

多个 FIFO 缓存器（118），分别与沿级展开的不相交路径相耦合。

8. 按照权利要求 4 所述的互连结构，其中：

所述多条耦合所述结构中节点的互连线，包括用于级 L 的路径 P 上的节点 N 的互连线，还包括：

一第一控制输出互连线，与路径 P 朝向源的一个节点相耦合并在所述级 L 上；和

一第二控制输出互连线，与所述级 L 的一朝向源的级的节点相耦合。

9. 按照权利要求 4 所述的互连结构，其中，

所述多条耦合所述结构中节点的互连线，包括用于级 L 的路径 P 上节点 N 的互连线，还包括：

一第一消息输入互连线，与路径 P 朝向源的一个节点相耦合并在所述级 L 上；

一第二消息输入互连线，与所述级 L 的一朝向源的级的节点耦合；和

一控制输出互连线，与路径 P 的一个朝向源的节点相耦合并在所述级 L 上。

10. 按照权利要求 4 所述的互连结构，其中，

所述多条耦合所述结构中节点的互连线，包括用于一个级 L 的路径 P 上节点 N 的互连线，还包括：

一第一控制输出互连线，与路径 P 朝向源的一个节点相耦合并在所述级 L 上；

一第二控制输出互连线，与所述级 L 的朝向源的级的节点耦合；和

一消息输入互连线，与路径 P 的一个朝向源的节点耦合。

11. 按照权利要求 4 所述的互连结构，还包括：

在所述多个节点中有优先权节点，而其它节点是非优先权节点，所述优先权节点和非优先权节点有选择地互连。

12. 按照权利要求 4 所述的互连结构，还包括：

一种逻辑，将节点互连成组以将在朝向源路径上的 n 个节点集中到朝向目标路径上的 m 个节点，其中，n 大于 m 而 n:m 是集中率。

13. 按照权利要求 4 所述的互连结构，还包括：

一个在不相交路径中连接多个节点的互连线条带（112），该互连线条带从源级至目标级缠绕过各级，其中，绕匝数从源级至目标级减少；和

多个列，耦合通过级的所有缠匝的所述条带的交叉区中的节点。

14. 按照权利要求 13 所述的互连结构，其中，

所述互连条带的缠匝数从源级至目标级在每级减少一半。

15. 一种系统，包括：

多个开关（702）；和

多个集中器（700），分别与所述多个开关相耦合，所述集中器是按照权利要求 4 所述的互连结构。

16. 一种互连结构，包括：

一连接不相交路径中多个节点的互连线条带（112），该互连线条带从源级至目标级绕过多个级，绕匝数从源级至目标级减少；和

多个列，其由耦合通过各级绕匝交叉区横跨条带节点的互连线形成。

17. 按照权利要求 16 所述的互连结构，还包括：

多个输入端口，与源级第一列中的节点相耦合。

18. 按照权利要求 16 所述的互连结构，还包括:

多个输出端口，与目标级最后一列中的节点相耦合。

19. 按照权利要求 16 所述的互连结构，还包括:

多个 FIFO 缓存器，分别与沿级展开的不相交路径相耦合。

20. 按照权利要求 16 所述的互连结构，还包括:

所述互连线条带的绕匝数从源级至目标级在每级减少一半。

21. 按照权利要求 16 所述的互连结构，还包括:

从所述互连结构内节点至一个或多个该互连结构外设备的控制线，用于控制消息进入该互连结构。

22. 按照权利要求 16 所述的互连结构，还包括:

一种逻辑，将节点互连成组以将朝向源路径上的 n 个节点集中到朝向目标路径上的 m 个节点，其中，n 大于 m 而 n:m 是集中率。

23. 按照权利要求 16 所述的互连结构，还包括:

一种逻辑，该逻辑能够判断一个节点是否被一级的路径上的消息所占据，并在该判断的基础上将控制信号发送至一朝向源的节点，加速消息在该朝向源节点的前进。

24. 按照权利要求 16 所述的互连结构，还包括:

在所述多个节点中有优先权节点，而其它节点是非优先权节点，所述优先权节点和非优先权节点有选择地互连。

25. 一种系统，包括:

多个开关（702）；和

多个集中器（700），分别与所述多个开关耦合，所述集中器是按照权利要求 16 所述的互连结构。

26. 按照权利要求 2 所述的互连结构，其中，

数据分组独立于包含在该数据分组中的目标输出信息在所述互连结构中移动。

27. 按照权利要求 1 所述的互连结构，其中，

所述数据路径 Q 与所述数据路径 P 是分离的且是不同的。

28. 按照权利要求 1 所述的互连结构，其中，

所述数据路径 Q 和所述数据路径 P 是相同的数据路径。

可伸缩蠕虫洞路由选择集中器

背景技术

通信或计算网络由几个或许多物理上通过例如金属或光纤电缆这样的通信媒介相互连接的装置组成。一类可以包括在网络的装置是集中器。例如，一个大规模时分开关网络可以包括一个中央开关网络和一系列在该开关网络中与其它装置的输入和输出端连接的集中器。

集中器典型地用于支持网络的多端口连接性。集中器是一个连接到多个将信息集中到较少的几条线的共享通信线的装置。集中器内在地通过增加阻塞和数据丢失的发生或者通过调用缓存器中的信息存储而使互连路径的容量降低。

当将数据移动到处理器和用户时，会出现在大型并行计算和通信中发生的持续问题。由于集中器固有的容量的降低，这一问题使在包括集中器的系统中的问题变得更糟。

所需要的是一种集中器结构，该结构通过避免阻塞快速地为数据选择路由并改善信息流，该结构是不受限制地可伸缩虚拟的，并支持低延迟和高流量。

发明内容

一种互连结构通过在使用控制信号的控制单元之间的单比特（single-bit）路由选择大大改进了信息集中器的操作。术语“单元”或“控制单元”指简单的开关元件。术语“节点”指作为一个单位操作的一个单元或一组单元。该互连结构和操作技术支持蠕虫洞路由选择和消息流。进入该结构的消息分组永远不会被丢弃，从而保证了任何进入该结构的分组被送出。

按照本发明的一个方面，互连结构包括一连接不相交路径中多个节点的互连带状线。该互连带状线从源级到目的级绕过多级。缠绕的转数从源级到目的级在减少。该互连结构还包括由耦合与经过缠绕的各级的带状线上的节点的互连线而形成的多个列。一种在互连结构上交换数据的方法结

合了一种用于为数据分组向下多层级选择路由的高速最小逻辑方法。

附图说明

所描述实施方式的被认为新颖的特征由所附权利要求进一步说明。但是，本发明的关于结构和操作方法的实施方式，可以通过参考下面的描述并结合附图理解。

图 1 示出了一个以多级 32:8 集中器形式的数据互连结构例子的方框图。

图 2 示出了以数据分组的形式在互连结构中传送的数据格式的数据结构图。

图 3A 和图 3B 示出了可以用于图 1 所示的包括不同输入和输出端口配置的互连结构中的单元的例子。

图 4A、图 4B 和图 4C 示出了适用于如图 1 所示结构的互连结构中单元间互连的多个例子的示意框图。

图 5 是说明互连结构中相对于各级不同的单元的优先权的示意图。

图 6A 和图 6B 示出了将互连的单元分组成节点的两个例子的示意图。

图 7 是说明使用多个集中器的系统的示意框图。

具体实施方式

参照图 1，一个表示以多级 32:8 集中器 100 形式的数据互连结构例子的框图，该互连结构包括三个级和 K 个列。各级是分层的、并从底部向上编号为 0、1 和 2，而列则从左至右编号为 0 到 K-1。集中器 100 从 32 条轻负载数据传送线接收输入数据，并将数据传送到 8 条负载较重的线。在所示例子中，数据分组在八路宽“条带”行中从 32 个端口输入端 120 被传送到 8 端口输出端 122。条带 112 包括一个位于每一列的控制阵列 110 的一组八路宽的控制单元 102，和位于列 K 右边的移位寄存器（FIFO）118。底部输出级不包括 FIFO。条带 112 的拓扑结构可以描述成如图 1 所示理发店招牌柱子样式中的螺旋。数据格式如图 2 中的分组 200 所示。条带 112 绕级 2 盘绕四次，绕级 1 两次，而经过级 0 一次。因此，级 2 有 32 行单元和 32 个 FIFO，级 1 有 16 行单元和 16 个 FIFIO，而级 0 有 8 行单元没有 FIFO。条带的宽度典型地等于输出端口的数目，尽管其它配置也是实用的。

图 1 所示的互连结构具有一八路宽的数据传输线条带。每条线被划分

成 7 段，每段的长度足够包含一条消息。在该八路宽的条带中，每条数据传输线与一系列节点相连接。该八路宽条带被绕成典型的螺旋形状并在不同的绕匝上有一些节点之间的相互连接。数据可以沿数据传输线在节点之间按先入先出操作的方式前进。数据从系统输入端口通过该互连结构移动到系统输出端口。除了互连结构上部的 8 条线在级 2 列 0 上的节点以外，数据传输线 L 上的每个节点 B 都有一个在数据传输线 L 上的直接前任节点 A。数据传输线 L 上的每个节点 B 都有一个在数据传输线 L 上的直接后续节点 C 或一个输出端口。节点 A 总是可以向节点 B 发送数据。节点 B 总是可以向节点 C 发送数据。将数据从节点 A 送至节点 B 以及从节点 B 送至节点 C 总是不会阻塞的。

一些节点位于条带的 FIFO 区域。在 FIFO 区域的节点仅有一个输入端口和一个输出端口。

一些节点具有允许数据从传输线条带的外部进入的附加输入端口。

一些节点具有能够使数据从一个传输线条带的上游节点进入到该节点的次级输入端口。该上游节点典型地位于相对于当前节点的前一个条带绕匝上。

一些节点具有一个或多个能够使数据传输进一步沿条带向下接近条带的系统输出端口的次级输出端口。具有多个输出端口的节点具有关联的逻辑，该逻辑总是试图将消息尽可能地向前朝系统输出端口传输，而不是沿传输线条带将消息传输到直接的后继节点。

具有多个输入端口的节点具有分配给输入端口的优先权。来自直接后继者的消息总是具有比来自非直接后继者的节点更高的优先权。与非直接前任节点连接的具有多个输入端口的节点对接收数据也具有相应的优先权关系。

总之，节点除了与每个节点 B 相关外，总是试图将数据尽可能地沿条带向下传送，其中，定义了优先权，以从不同的、可以向节点 B 发送数据的节点接收消息。最高优先权被赋予直接的前任节点。一个集合 S 包括可以向节点 B 发送消息的节点。在集合 S 中的节点之间定义了一种向节点 B 发送消息的优先权关系。管理信息行进的规则如下：

1. 如果节点 N 是集合 S 的成员且消息 M 抵达节点 N，则节点 N 向节点 B 传送消息 M 不会因集合 S 中其它优先权低于节点 N 的节点向节点 B

发送消息而被阻塞。

2. 对应于集合 S 中的每个节点 N 存在一个节点集合 TN，节点 N 可以向该集合中的节点传送消息。每个节点 N 的相关逻辑将 TN 的成员从最希望接收消息 M 的节点 NM 到最不希望接收消息 M 的节点 NL 进行分级。与节点 N 相关的逻辑将消息 M 发送给集合 TN 中最希望的未阻塞的成员。

3. 节点 B 的直接前任者 A 具有向节点 B 发送数据的最高优先权。

在图 1 中，具有向/从传输线条带的前一绕匝发送和接收消息能力的节点、即跳过互连结构段的节点，位于第 K 控制单元列。只有一个输入端口的节点位于互连结构的 FIFO 区域。

一消息包括一以单比特标头开始的有效负载，该标头是一个总是设为 1 的一定时比特。每个段对应于一级上的一行。一行包括跨越该行的 K 个控制单元。由于该消息必须适合于 K 个控制单元和一行上的 FIFO，因此，消息的长度不能超过长度 FIFO+CK，其中 C 是一个控制单元的比特数。因此，FIFO 的长度必须至少是最大消息长度减 CK。如果一个系统将大量消息集中成 R 个信号，则条带宽度是 R。如果互连结构包括 L+1 级而每级有 K 列，则该系统包括 $R\cdot(2^{L+1}-1)$ 行，每行有 K 个控制单元和一个长度至少为最大消息长度减 CK 的 FIFO。

在一个实施方式中，消息被送入到列 0 上的条带段中之一。当该消息移动至列 1 时，该消息可以进一步沿条带段继续前行，或者该消息的第一比特可以向前移动到更接近系统输出端口的另一段。当消息的第一比特向传输线条带的一个新的段移动时，该消息可以移动到条带的一条不同的传输线或者留在条带的同一传输线上。当消息头到达一新的列时，该消息可以再次向前跳到一个新的段上。以这种方式，一条消息可以跨越若干段，并且因为底部级没有 FIFO，因此消息的第一比特可以在整条消息进入集中器之前退出集中器。

在下面的描述中，术语分组指数据单位，典型的是以串行的方式。数据分组的例子包括互连网协议（IP）分组、以太网帧、ATM 单元、例如较大的帧的部分或数据分组的交换结构（switch-fabric）段、并行计算机处理器间的消息，或者其它具有长度上限的数据消息。传过一个级的分组经过 K 列控制单元 102。该传过一级的分组可以直接从一个单元传到在同级中下一列的单元。在一个例子中，对于在同级单元之间传输的分组，一个分组的

两个比特被定位在每个单元中。尽管在本发明中单元的设计被简化，在这里所公开的系统中仍可使用相同的时序。FIFO 包含分组的适当的比特数，使得当分组在列 0 进入一个阵列时，消息头比特 202 与传过同一级的分组的其它消息头比特对齐。在该例中，一个分组被设置成单行，使得分组进入列 0 的时序与从位于上层单元到达的分组同步。

定时和控制通过保证最长分组的比特长度不超过 FIFO 中的比特数加上列中的比特数之和得以实现。例如，对于具有上限长度 400 比特的分组，在具有 12 列控制单元且每个控制单元保持 2 比特、FIFO 的长度为 376 比特的结构中，需 400 个时钟周期到达。分组的第一比特在两次时钟滴答或节拍中，从一特定级上的单元移动到同一级上的下一列的单元。分组的第一比特在一次时钟滴答中，从一特定级上的单元移动到低一级的下一列的单元。因此，连接级之间的 FIFO 的长度比考虑级之间的时序差别要少一比特。连接级之间的 FIFO 的长度比考虑级之间的时序差别要少一比特。级时序将在下面详细讨论。

分组被从顶级的 32 个数据输入端口之一插入到输入阵列 120。一个输入端口服务于顶级 32 条线的每一个。一个分组以串行的方式插入到每条线中。从阵列 120 中的条带 112 进入到一个单元中的分组其优先权高于试图从互连结构外部进入到该单元的分组的优先权。没有内部分组可以阻塞一个分组进入到结构的顶端 8 行，但是，在一定的阻塞条件下，穿过顶级顶行的分组可以绕行“理发店柱”并重新进入阵列 120 中顶级的较低的 8 行。进入结构的分组决不会被丢弃，使得任何进入结构的分组都保证能退出，由此，对常规的集中器作出了实质的改善。

该分组传输到开关阵列的下一列的优先权高于试图从结构外部进入到互连结构的分组的优先权。在这种情况下，试图进入结构的分组被禁止进入。尽管操作的多方面影响穿越互连结构，但较早的分组在概率上有较高在的优先权在更新的分组之前退出。

在另一实施方式中，分组可以在顶级以多个角度进入条带。对于顶级的一特定单元，从结构内部进入到该单元的分组的第一比特与从结构外进入到同一列中的单元的分组的第一比特相符合。

换言之，在顶级的 32 条输入线从结构外接收分组。在这 32 条输入线中，顶端 8 个输入端口连接在条带的开始处，并不会被已在结构内的分组

所阻塞。而其余 24 个输入端口可以被阻塞。在对附图 4A、4B 和 4C 的描述中将详细讨论阻塞。

作为另一种绕行螺旋携载数据条带的传输，分组可以从分层结构中一个较高级的单元跳跃到较低级的单元。这种跳跃处理使分组在条带中集中，使得在底级上的条带段具有优点地、平均比在顶级上条带段携带更多的分组。

分组在一个给定级上进入控制单元阵列 110 中的控制单元 102。参照附图 4A，当分组 P_A 410 进入控制单元 B，与控制单元 B 相关的逻辑可以将分组通过单元 B 路由到同一级上的另一个单元 C。作为另一种选择，与控制单元 B 相关的逻辑还可以将分组 P_A 路由到较低级上的单元 X，这是使用集中器结构和操作方法所希望的结果。

当分组从上部线 106 或 404、例如从单元 B 或 E 进入单元 W 时，则该分组在线 104 上没有延时地被送至下一列上的单元 X。因此，当分组 P_A 从单元 B 被路由到单元 X 时，分组 P_A 先被传送到单元 W，然后直接进入单元 X。结合附图 4A 参照附图 4C，当单元 B 将分组 P_A 向下路由到线 406 时，分组 P_A 也类似地直接进入单元 X。附图 4A 和 4B 中所示的是在单元之间路由分组的功能上等效的例子。实现的考虑可能会对影响到路由技术的选择。

分组向较低级的直接“移下”或“跳跃”提高了集中器互连结构的有效操作。基本上，如果在一给定的较低级的单元没有被分组占据，即可以接收数据，而且该单元与更高级上的另一个单元连接，则在该更高级上单元的分组将移下以填充该更低级单元的空位。有几种移下或跳跃处理是适用的。由控制单元结构确定移下或跳跃技术。下面将详细讨论控制单元结构和整个数据流以及定时。

附图 1 示出了一个 32:8 集中器的实施方式，该集中器具有 8 路宽条带和三个级，级之间的比率是 2:1，由此得到整体集中比率为 4:1 的集中器。当希望提供多种输入和输出端口数时，采用其它条带宽度，则其它的集中器比率也是实用的。集中器比率可以通过增加附加级、通过增加级之间的集中器比率或者通过两者来增加。一种具有较高的 4:1 整体比率的实施方式有多于三个级。其它实施方式可以使用不是 2:1 的级之间的比率。在图 1 所示的控制单元中，仅示出了部分数据携载线和控制信号携载线。下面将参照图 3A 和 3B 所示结构对附加线进行讨论。

图 2 示出了一个分组的布局。分组 200 的头 202 包括一个总是设为 1 并表示存在一个分组的单一比特。该分组的其余部分被称为有效负载 204。图 2 中举例示出了不同的有效负载。分组可以从任何输出端口退出互连结构 100，因此不需包含路由消息的附加头。

如果分组在离开集中器 100 之后进入一个网络路由设备，供该路由设备使用的路由信息可能会在有效负载前加上数据比特。集中器 100 不使用这些路由比特。集中器 100 总是忽略有效负载的内容 204。

另一集中器的实施方式（没有示出）可能使用表示服务质量的附加的标头比特。

图 3A 和 3B 示出了一个单元的输出和输入端口的例子。在所示出的端口配置中，单元 102 在两条数据输入线 106 上从较高级的单元接收数据，并在数据输出线 106 上将数据发送到一个较低级的单元。垂直连接的 2 比 1 的比率支持集中器的这种数据集中操作。单元 102 还包括一条来自一同级单元的数据输入线 104 和一条至一同级单元的数据输出线 104。除了单元之间的数据输入和输出线，附加的互连线 108 用于传送单元间的控制信息。接收单元逻辑用该控制信息进行判断，如何从接收单元为数据分组进行路由选择。这些控制线用于增强在互连结构中管理数据流的基于位置的优先权规则。

如这里所述，两个试图向第三节点发送数据的节点使用控制信号进行通信以解决上游争端问题。在其它网络中，进入节点 N 的分组争夺输出端口。在这里所描述的系统中，分组争夺节点输入端口，而这种争端使用控制信号在上游解决。在本发明的网络中，唯一的拓扑结构允许上游数据流，即允许两个分组竞争一个特定的下游节点的输入端口。这种争端的解决至少部分基于节点在网络中的位置。如果节点 E 具有比节点 B 高的、向第三节点 V 发送数据的优先权，则直接或间接向 B 发送一个控制信号以执行该优先权。图 3A 描述了一个接收一控制信号并发送一控制信号的节点。图 3B 描述了一个接收一控制信号并发送两个控制信号的节点。

在图 4A 所示的对互连结构的讨论中，对这两种节点的使用进行了更详细地描述。

图 3B 示出了一个其优先权高于一个较高级单元和一个同级单元的优先权的单元的例子。

图 4A 是图 1 所示集中器 100 的一部分的放大描述，其中，未示出所有数据和控制线。单元 E 和 B 在同一级上。单元 E 和 B 中的每个都可以被分别视为单独的节点，或者可以将两个单元 E 和 B 一起视为一个单一节点。

集中器 100 按以下方式工作。对于在同一级上的单元 V、W 和 X，如果分组 P_V 由单元 V 送至单元 W，单元 W 将分组 P_V 的第一比特送至单元 X，然后单元 W 向较高级的单元 E 发送一个控制信号。单元 W 经线 104 将分组 P_V 送至单元 X。单元 W 经线 108 向单元 E 发送控制信号，通知单元 E 不要沿线 106 向下发送数据。单元 E 又通过线 402 发送控制信号，通知单元 B 不要沿线 404 向下发送数据。在单元 W 向单元 X 发送分组 P_V 的情况下，任何从单元 D 进入单元 E 的分组 P_D 412 将由单元 E 的逻辑路由到下一列上的单元 F。此外，任何从单元 A 进入单元 B 的分组 P_A 将被从单元 B 送至单元 C。

如果在给定的分组到达时刻，单元 W 不向单元 X 发送分组，则单元 W 将单元 W 没有在向 X 发送分组 M 的情况通知单元 E。

再次参考图 4A，控制线 108 携载的控制信号中包含了控制信息，该控制信号由单元 W 送至单元 E，以通知单元 E 没有分组从单元 W 送至单元 X。从单元 W 至单元 X 的线 104 没有被分组占用并可以接收数据。如果在这种条件下单元 D 向 E 发送分组 P_D，则单元 E 将通过读分组 P_D 的一比特头 202 来检测分组 P_D 的存在。与单元 E 关联的逻辑使用来自单元 W 的控制信号和分组 P_D 的消息头比特，来确定是否将分组 P_D 通过单元 W 路由到单元 X。因为单元 W 当前没有使用数据线 104 向单元 X 发送其它分组，因此数据线 104 对分组 P_D 是空闲的，其可经数据线 104 从单元 W 至单元 X。

任何从上面的级进入到单元 W 的分组总是被直接送至单元 X。在示出的集中器中，与单元 W 关联的逻辑能够将分组从与单元 W 同级上的其它单元路由到单元 W，但是对从更高级进入到单元 W 的分组没有任何控制。来自更高级的分组通过单元 W 到达单元 W 至单元 X 的数据互连线 104。在单元 E 将分组 P_D 向下通过单元 W 路由到单元 X 的情况下，单元 E 通过控制线 402 向单元 B 发送一个控制信号。该控制信号携带有规定阻塞单元 B 此时通过线 404 发送分组的信息。在存在来自单元 E 的阻塞控制信号时，如果单元 B 从单元 A 接收一个分组 P_A，则与单元 B 关联的逻辑将使分组 P_A 通过线 104 被从单元 B 路由到单元 C。

在一种情况下，在一特定分组到达时刻，与控制单元 W 相关的逻辑不是将分组路由到单元 X，而单元 E 没有沿线 106 向下发送分组，则单元 B 空闲，可沿线 404 向下发送分组。单元 W 通过控制线 108 发送例如单比特形式的控制信号，通知单元 E 单元 W 未被阻塞从单元 E 或单元 B 接收分组。单元 E 响应来自单元 W 的该控制信号，并在没有来自单元 D 的消息时，通过线 402 发送控制信号，通知单元 B 单元 W 未被阻塞从单元 B 接收分组。如果分组 P_A 此时到达单元 B，则单元 B 经线 404 将分组 P_A 通过单元 W 送至单元 X。该分组 P_A 首先通过线 404 然后通过线 104 传输。因为线 104 此时未被用来携带从单元 W 或单元 E 路由到单元 X 的分组，所以线 104 可以携带分组 P_A。

继续参照图 4A，节点 W 被连接成通过单个控制信号携载线 108 向节点 E 发送控制信号。除了所示互连线，在级 J 上的节点也具有能够携载来自级 J-1 节点的控制信号的控制信号携载线（没有示出），和能够从级 J 节点向级 J-1 节点携带分组的数据携载线（没有示出）。例如，节点 W 具有图 3A 所示的端口结构，其包括三个数据输入端口、两个数据输出端口、一个控制信号输入端口和一个控制信号输出端口。再次参照图 4A，节点 E 通过线 108 从节点 W 接收控制信号。节点 E 将控制信号发送给同级节点 B，并且如果节点 E 不是在顶级上，则节点 E 也向级 J+2 上的节点发送控制信号。除了所示互连线，如果不是在最高级，则在级 J+1 上的节点还具有连接到级 J+2 上节点的附加数据和控制互连线。例如，节点 E 具有如图 3B 所示的包括两条来自级 J+2 节点的数据携载线的端口结构。

在单元 B 不在顶级的情况下，单元 B 与用来向更高级单元发送控制信号的一条控制线（没有示出）连接。该控制线的功能与从单元 W 至单元 E 的控制线的功能相同。如果单元 B 向单元 C 发送分组 P_A，则单元 B 向上一级的单元发送一个阻塞信号。

如果单元 B 在列 K-1 并在顶级的上 24 行中，则单元 B 位于通过 FIFO 向从集中器以外的设备接收输入数据的单元发送消息的位置。如单元 B 这样可以向其它也能接收来自结构外的消息的单元发送信息的单元，具有特殊的控制信号线，用于控制从结构外设备进入互连结构的消息流。该特殊控制线上的控制信号通知结构外设备关于数据输入列 120 的可能的忙碌状态。

如果单元 B 是在顶级列 K-1 上但不在顶级的上 24 行中，则单元 B 不向结构外发送控制信号。

如果单元 E 不在顶级上，则单元 E 发送两种控制信号。当单元 E 向单元 W 发送分组时，单元 E 用控制信号线 402 向单元 B 发送一个阻塞信号。当单元 E 向单元 F 发送分组时，单元 E 用控制信号线 403 向更高一级发送一阻塞信号。

如果单元 E 在列 K-1 并在顶级的上 24 行中，则单元 E 发送一个控制从结构外设备进入互连结构的消息流的特殊控制信号。

总之，在分层结构中具有向下发送分组优先权的节点具有两条控制信号发送线，如图 4B 所示。没有向下发送分组优先权的节点只有一条控制信号发送线。

图 4A 和 4B 描述了逻辑上相同的另一种互连结构实施方式。其在物理上的差别在于，图 4B 所示结构先通过单元 E 再通过单元 W 将分组从单元 B 发送至单元 X。当单元 E 没有另外使用线 106 发送数据且单元 E 和单元 W 都没有使用线 104 向单元 X 发送数据时，消息可以从单元 B 跳跃至单元 X。图 4A 和 4B 所示的控制结构相同的并包括控制线。

图 4C 描述了在逻辑上和图 4A 和 4B 所示互连结构相同的互连结构的第三实施方式。在图 4C 中，单元 E 和 B 不经中间节点直接将分组发送至单元 X。图 4A、4B 和 4C 所示互连结构中的控制线结构相同。图 4A、4B 和 4C 所示的结构物理上不同但逻辑上等效。

集中器的成功运行至少部分地取决于定时。分组 P_V 的第一比特在预定分组到达时刻到达节点 W。与节点 W 关联的逻辑根据分组 P_V 的单比特头 202 和级 J-1 节点的控制信号的到来，作出路由选择判断。如果分组 P_V 出现在节点 W 且未被较低级的节点阻塞，则节点 W 将分组 P_V 发送至较低级并将控制信号发送至更高级上的节点 E。来自节点 W 的控制信号与分组 P_V 到达节点 E 的同时或接近同时到达节点 E。逻辑控制这种定时，使得分组在到达节点 E 先到达节点 W。

参照图 4C，例如一种光学实施方式，逻辑为分组到达节点 X 预先确定一个分组到达时刻。从节点 E 送至节点 X 的分组必须在与从节点 W 送至节点 X 的分组同时到达节点 X。如上所述，分组在到达节点 E 之前先到达节点 W。因此，分组从节点 E 传送到节点 X 的时间必须少于分组从节点 W 传

送到节点X的时间。在一个光学实施方式中，定时是通过选择从节点W到节点X的互连线的光纤长度短于从节点E到节点X互连线的光纤长度来调节的。按这种方式，来自更高级的分组可以赶上较低级的分组以同步到达时刻。

在电子实施方式中，在同级上的两个节点之间传送的分组经过两个一比特移位寄存器单元。向下一级移动的节点旁路（bypass）掉一个寄存器单元，因此在图4A、4B和4C所示结构中，一个分组比特从单元E至单元F在两个时钟周期中传输，而从单元E至单元X用一个时钟周期传输。

图1、3A、3B、4A、4B和4C所示的结构具有相同的通过互连结构发送消息的优先权。节点X的直接前任节点W具有向节点X发送数据的第一优先权。节点E具有向节点X发送数据的第二优先权。节点B具有向节点X发送数据的第三优先权，其中，节点X比节点B更靠近条带的输出端口。在另一实施方式中，最远离条带输出端的节点具有优先权。参照图5，优先权相对于列是可变的，因此如果节点B是节点A的直接后续者而节点A是一个没有优先权的单元，则节点B是一个有优先权的单元。

图5示出了互连结构的三列和三级。在每个级和每个列上有一个控制单元102的阵列110。图5所示结构中的单元至单元的互连与在图4B中所示的互连相同。一对单元形成一个节点502，因此节点对的两个单元均被定位于向较低一级的一个单一节点发送数据。在集中器高度适合的实施方式中，控制阵列110内在行104上的单元102被随机放置。对多个随机放置的软件模拟可以按需要选择最佳性能的安排。实框504表示节点中具有较高优先权的单元。空框表示具有低优先权的单元。每个所示节点包括一个较高优先权单元和一个较低优先权单元。沿行104，在列上有低优先权的分组优选地总是在同级的下一列上有高优先权。

如果一个节点被考虑仅包含一个单一的控制单元，则每个节点具有一条来自同级的数据输入线。不是位于顶级上的节点具有两条来自更高级的数据输入线。每个节点具有一条至同级节点的数据输出线。不是位于底部级上的节点具有另外一条至较低级的附加数据输出线。每个不是位于底部级上的节点具有一条控制输入线。不是位于顶级上的节点具有一个或者两个控制信号输出端口。只有在输入列上的节点具有至外部输入源的控制信号输出线。只有在由全局时钟信号指示的时刻且只有在顶级节点没有收到

阻塞信号时，输入源才被允许将分组送至集中器。输入源使用相同的时序和将分组向下发送到集中器中的路由规则，内部单元遵循这些规则来将分组向下传送至较低级单元。

本领域的普通技术人员容易实现对在这里所描述的基本集中器的许多变形、改动、增加和改进。

可选拓扑实施方式

在上面披露的集中器中，一条离开行 J 的 FIFO 的线与在行 J-8 上进入列 0 的线相连。在第一可选的例子中，该离开行 J 中 FIFO 的线与在同一行 J 上进入列 0 的线相连。拓扑结构被这样改变，以使顶级有 32 个环，在层次结构中的下一级有 16 个环，而在底级有 8 个环。

在第二可选的例子中，在列 K-1 和 FIFO 列之间作了改变。一条离开行 J 中 FIFO 的线与在同一行 J 上进入列 0 的线相连。对于某些变化，产生的拓扑结构在每级上有一个环。在一些情况下，数据总是被允许从在级 0 列 k-1 的输出端口离开集中器。在这种情况下，图 1 所示新颖的“理发师柱”结构优选地运行，使得进入结构顶部的消息总是被保证在由一个固定常数设定的时间量之内退出该结构。

在其它情况下，在一些条件下，数据可能被阻塞离开集中器输出端口，每级有一个环的结构可能更适合。

服务质量（QoS）实施方式

在网络和集中器中实现服务质量（QOS）优先权的一种简单技术是包括一个或多个消息头比特来指明服务级的质量。参照图 4A，实现了一种 QOS 优先权技术。如果分组 P_D 比分组 P_A 具有相同或更高的 QOS 优先权，且单元 W 未向单元 X 发送分组 P_V，则将分组 P_D 发送至单元 X。然而，如果分组 P_D 具有比分组 P_A 较低的 QOS 优先权且单元 W 未向单元 X 发送分组 P_V，则将分组 P_A 送至单元 X。为了实现 QOS 优先权，单元 E 和 B 能够读取 QOS 消息头比特，而一条从单元 B 至单元 E 的控制线携载 QOS 信息。单元 B 能够通过现有的线 108 或通过一条附加的控制线将 QOS 信息发送至单元 E。

附加级实施方式

所示例子排除了节点的缓存器。但是，集中器确实按缓存器的方式工作，因此可以处理突发业务。例如，如果进入集中器的消息的平均数小于8，但偶尔多于8条消息进入集中器，则可能没有要进入集中器的消息被阻塞。这种处理突发业务的能力可以通过为集中器增加一个附加级，在例子中为级3而增强。级3可以有64行，数据进入上部32行。加入的附加级增加了整个集中器的有效缓冲规模。在其它实施方式中，可以向结构中增加若干附加级，以进一步增强处理突发业务的能力。

多输入列实施方式

实现64:8集中器的一种技术包括增加一个具有使用8绕条带64行的附加级3。在另一技术中，为级2增加一个输入端口端列，可能使在级2上32线上接收的数据量加倍。可以为级2增加一个附加移位寄存器FIFO列，来处理该32条线上增加的业务。附加移位寄存器FIFO列是否得到保证取决于从输入设备加到集中器消息的时序。与消息从集中器内部节点进入到输入列一样，消息从集中器外进入一个输入列。该可选技术在每个输入通道的平均数据率较低且突发业务少的应用中是有用的。

级之间不同互连的实施方式

在分层结构中，每个节点包括一个单一控制单元，而每个在较低级上的节点与唯一一个在更高级上的节点连接。可以这样改变集中器，使得在底部级上一个特定级只有一半的线在一个特定的时间节拍将消息向下发送至下一较低级。

在一个实施方式中，四个控制单元被组合成一个可以向四个单元发送数据的单一节点。本领域的普通技术人员可以改变该结构，使得在特定级上的四个单元形成一个可以向较低级上的两个单元发送数据的节点。

再次参照图5，在分层结构中，在一级上安排两个控制单元将数据发送至下一较低级上一个单一控制单元。节点502示出一个2:1节点集中器的结构，在该集中器中，节点中的两个单元能够将数据发送至较低级上的一个节点。在一个运行的例子中，消息M_1从级1的行2列0的控制单元被发送至级1的行2列1上的单元。在同一运行时间周期中，一消息M_2到达级2

的行 8 列 0 的控制单元。消息 M_1 被用于阻塞消息 M_2，而消息 M_2 会保留在同一级。因此，在图 5 所示的拓扑结构中，一级上的单一消息可以阻塞一更高一级的消息。

参照图 6A 和 6B，单元被分组成例如节点 602，使得一个级上的四个单元被设置为用互连线 604 向较低级上的两个单元发送数据。每级上的节点包含四个控制单元。级 2 上最左边节点的四个控制单元中的每个都能够发送数据至级 1 上两个单元中的一个。在图 6A 所示的例子中，节点 N 中的四个单元能够发送数据至节点 P 的单元 Q 和 R。类似地，节点 M 中的四个单元能够发送数据至节点 P 的单元 S 和 T。在一个级上没有消息可以阻塞更高级上的消息，因此可以增加从一级向下一较低级的流量。图 6A 示出了一个 4:2 节点集中器的结构，在该集中器中节点中的四个单元能够将数据发送至较低级上的两个节点之一。在其它实施方式中，例如如图 6B 所示，可以添加逻辑以增加节点 610 中单元的数目。例如，在节点 T 和 U 中的所有 8 个单元可以联合被控制，来将数据发送至节点 V 和 W 之一或两者中所有四个单元，而形成一个 8:4 节点集中器。更复杂的节点用于以每个节点更多逻辑的成本增加流量的设计。

参照图 7，示意框图示出了一个使用多个集中器和多个开关构成有数个芯片的大网络的系统。在另一实施方式中，将较小的网络安排成双扭立方体。图 7 所示系统改进了双扭立方体结构。

在实践中，超大网络可以使用所示多个芯片模块构成。例如，可以结合 128 个开关芯片和 128 个集中器芯片。例如，128 个芯片中的每个可以包含 64 个单线输入端口和 64 个三线输出端口。这种组合形成了一个具有 64^2 输入线和 64^2 输出线的单一芯片。图 7 所示例子是一个为了说明的目的、有用的很小的系统。在实际中可以构造大得多的系统。

图 7 所示网络 702 和 704 具有三个级和多个列。每级有四行。在级 0 上的三列包含输出端口。在所示系统中，每个芯片 702 具有三条送至地址 0 的输出线、三条送至地址 1 的输出线、三条送至地址 2 的输出线和三条送至地址 3 的输出线。底部集中器与地址 0 连接的所有 12 条输出线。其它三个集中器每个接收与地址 1、2 和 3 适当连接的 12 条输入线。

其它网络在不同的时间从不同的列发送数据。在图 7 所示系统中，来自不同列的消息传过适当的延时线 FIFO（没有示出），使得来自所有列的消

息同时到达集中器 700。

集中器 700 有在三个级上 4 路宽条带的行，使得级 0 有 4 行、级 1 有 8 行，而级 2 有 16 行。在级 2 上，16 行的低 12 行设置成在集中器 700 的输入端口（没有示出）从开关 702 接收数据。

对于数据预期特别突发的应用，可以为集中器 700 增加附加的层。集中器 700 将数据从在第一列每个芯片的 12 条通道输出线集中到在第二列的四条输入线。该集中器还将数据在时间上分散或散布以减小可能的热点。来自集中器 700 的数据加到开关芯片 704 的第二列。

消息同步地退出集中器 700，使得在集中器模块 700 和开关芯片 704 之间不需要 FIFO。消息从开关 704 出现并传过 FIFO（没有示出）以便在时间上与进入到集中器模块 706 第二列的消息对准。每个集中器芯片 706 包含四个集中器。在集中器芯片 706 的这四个集中器中的每个具有一宽度为一行的条带。

第二列集中器可以被设计成三个级，级 0 有 1 行、级 1 有 2 行而级 2 有 4 行。顶级 4 行中的 3 行能够接收输入数据。同样，对于突发业务可以在集中器芯片 706 中为集中器增加附加的行。

控制线（没有示出）从下游芯片向上游芯片提供控制信号，以通知上游芯片下游芯片中的的数据阻塞情况。例如，如果开关 704 不能从集中器 700 接收数据，则该数据被重新导向集中器 700 的一顶端行。由于只有集中器 700 较低的 12 行从上游开关 702 接收消息，顶端 4 行总是可以接收数据。

返回到顶端开关的控制线可以来自若干地方。在一个实施方式中，可以在集中器 800 和开关 804 之间设置缓存器。当缓存器充至容量水平之上时，可以有选择地将控制信号送至馈给充满的缓存器的开关 802 的输出端口。可供选择的是，控制信号起源于集中器内位于较高级中在通道拥挤时接收消息而在通道清闲时不接收消息的列的左边列的节点。该控制信号根据集中器中的业务阻塞开关 800 的特定输出端口。

对于所有规模的集中器，将阻塞的消息反馈到集中器的顶行总是成功的，因为反馈消息的最大数目等于条带的宽度，而条带宽度在集中器的顶端行总是开放的。

在本发明参照各种实施方式被描述的同时，可以理解，这些实施方式是示意性的，本发明的范围并不局限于此。所述实施方式的许多变形、修

改、增加和改进是可能的。例如，本领域的普通技术人员容易实现必要的步骤来提供这里披露的结构和方法，并理解处理参数、材料和尺寸只是举例地给出且可以被改变以实现希望的结构以及在本发明范围内的改动。这里披露的实施方式的变形和改动，可以在这里展开的描述的基础上作出，而没有超出所附权利要求设定的本发明的范围和精神。例如，本领域的普通技术人员可以对这里描述的其它互连结构使用第一和第二服务质量技术。

在权利要求中，除非另外指明，冠词“一个”是指“一个或多个”。

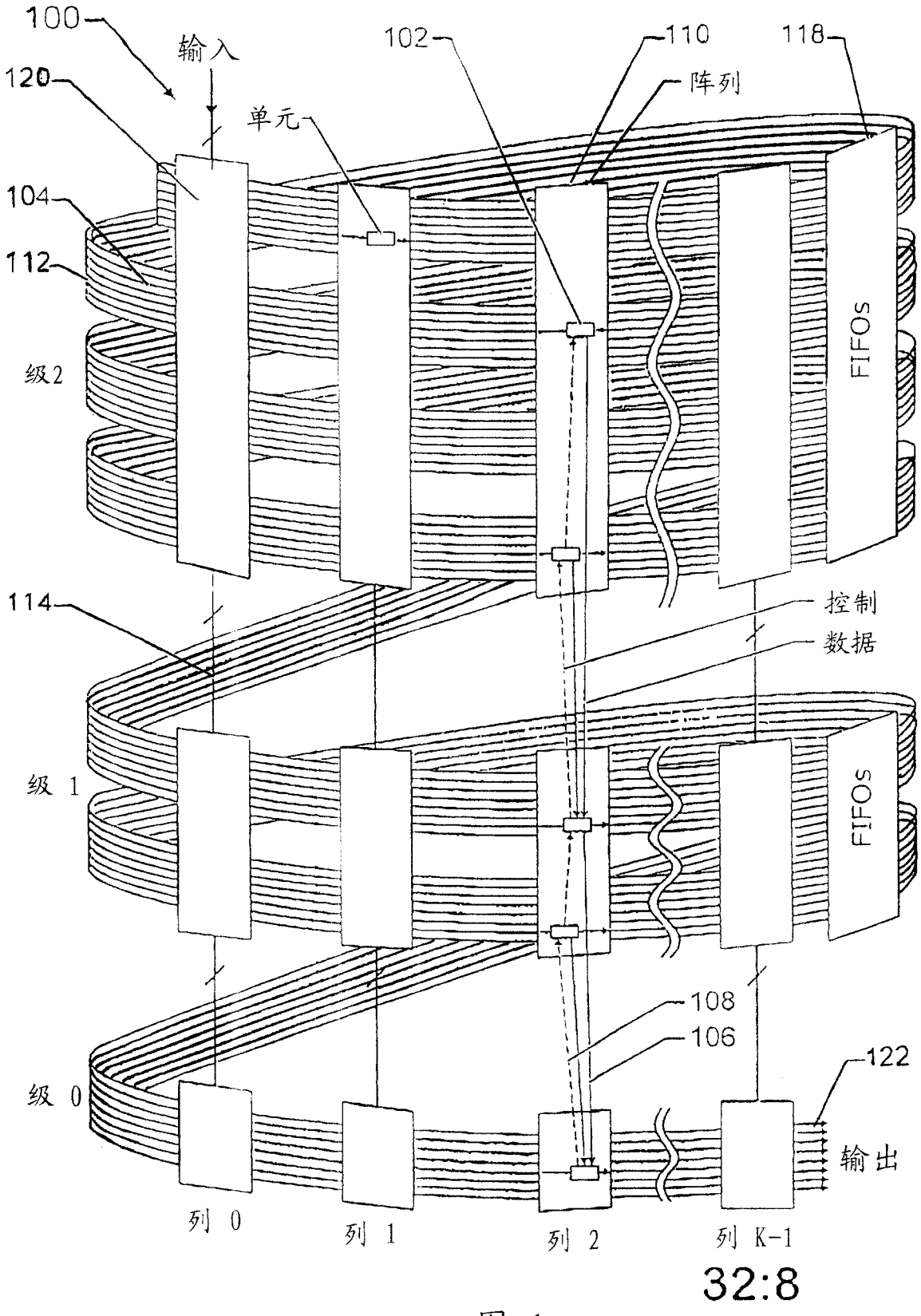

100
输入
120
102
110
118
阵列
单元
104
112
级2
FIFOs
114
控制
数据
级 1
FIFOs
108
106
122
级 0
输出
列 0
列 1
列 2
列 K-1
32:8

图 1

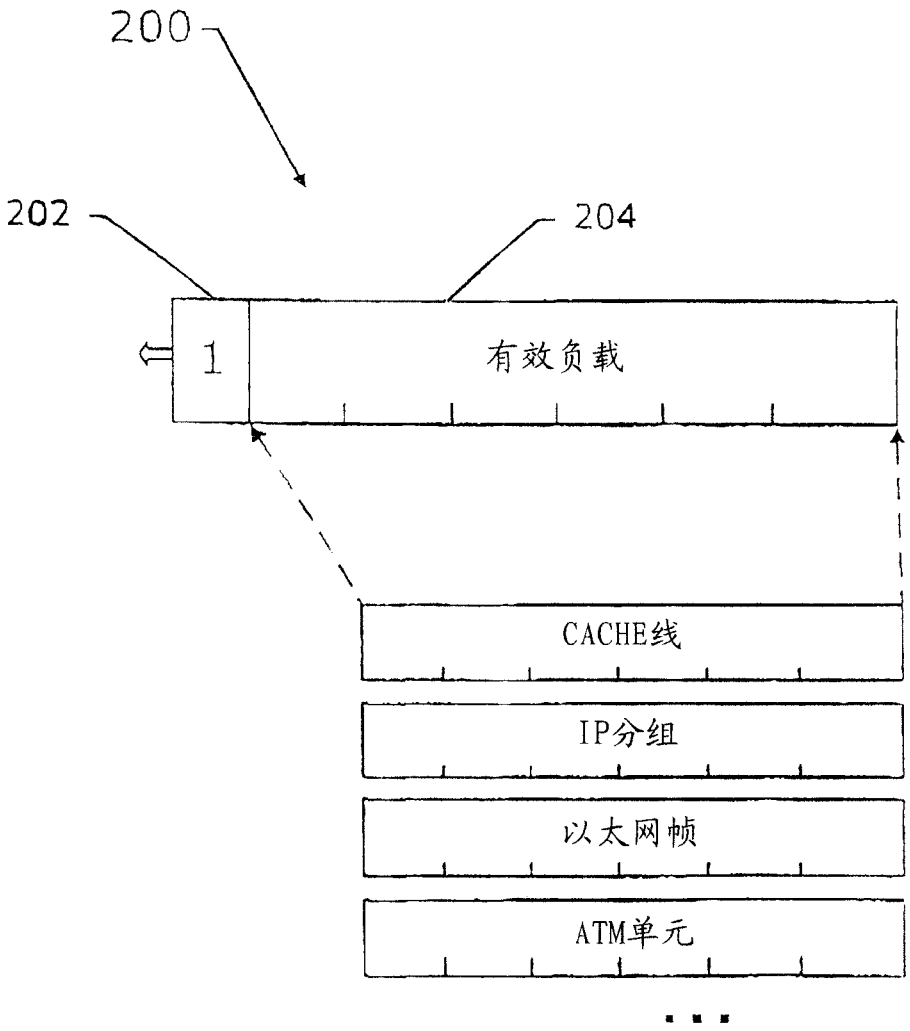


图 2

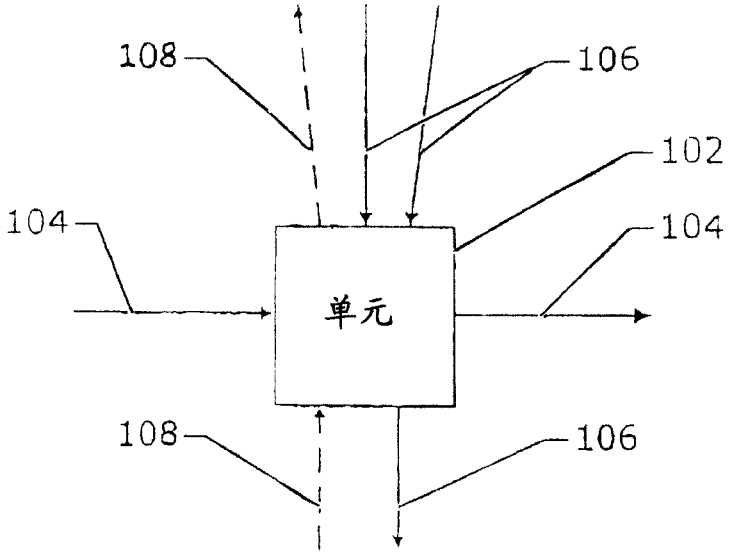


图 3A

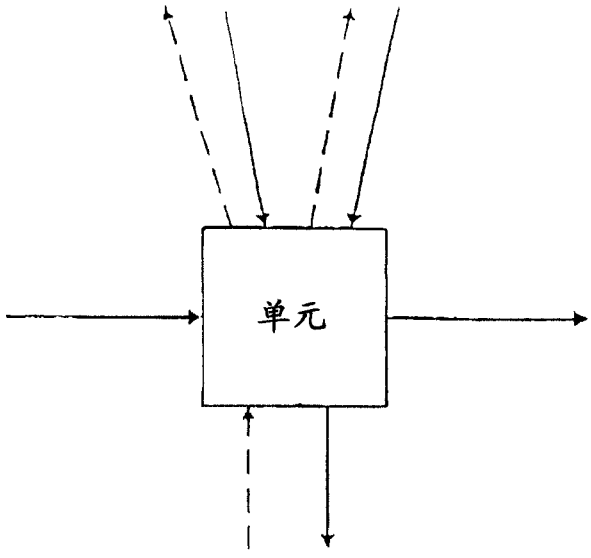


图 3B

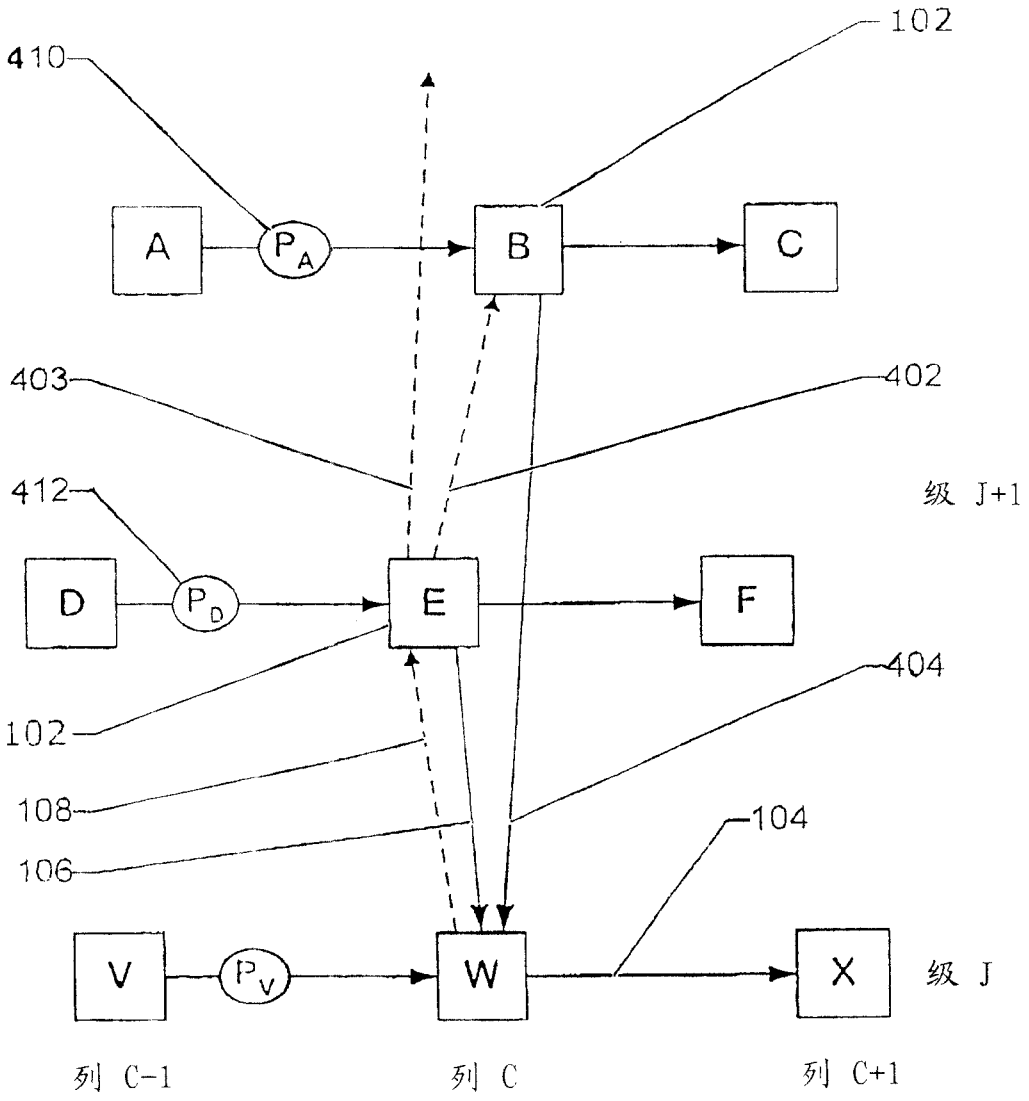

410
102
A
P_A
B
C
403
402
412
级 J+1
D
P_D
E
F
404
102
108
106
104
V
P_V
W
X
级 J
列 C-1
列 C
列 C+1

图 4A

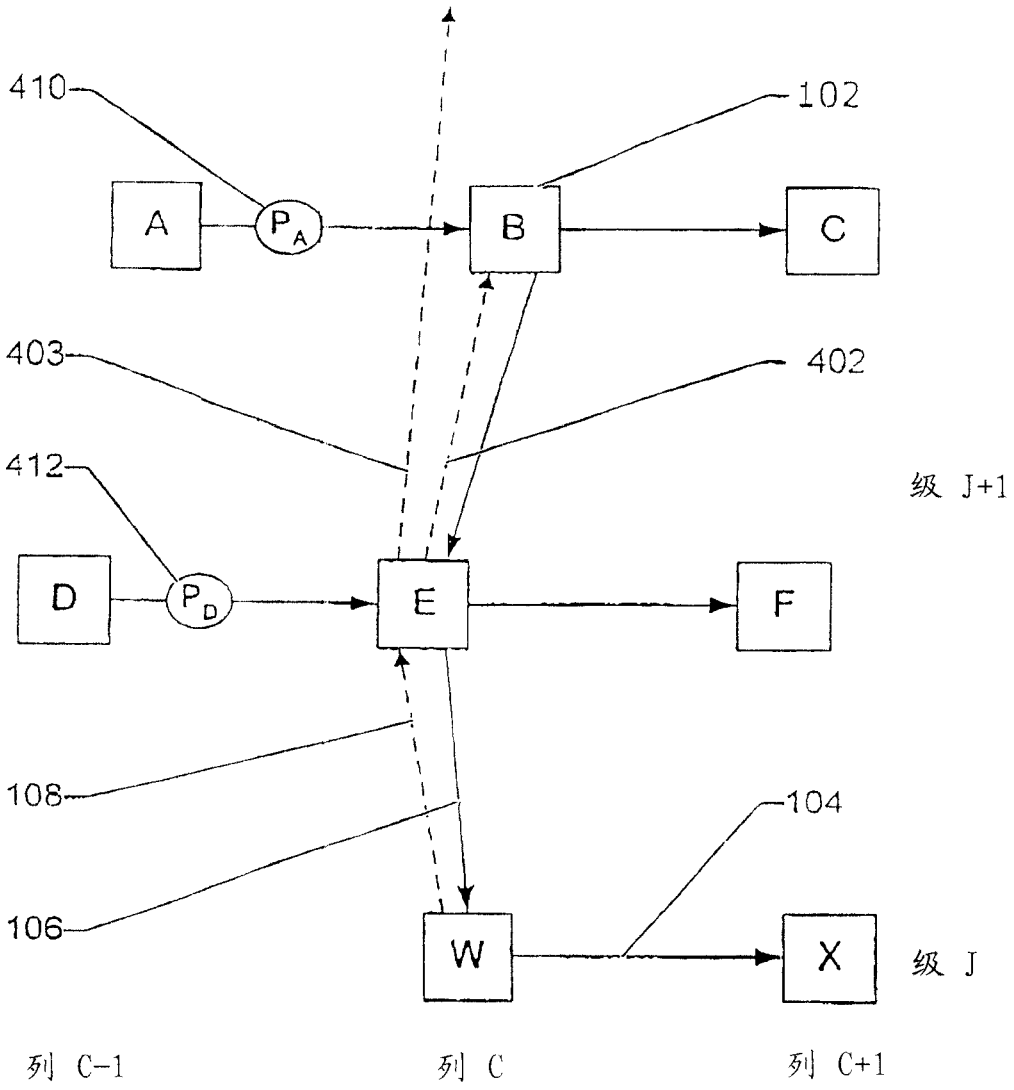


图 4B

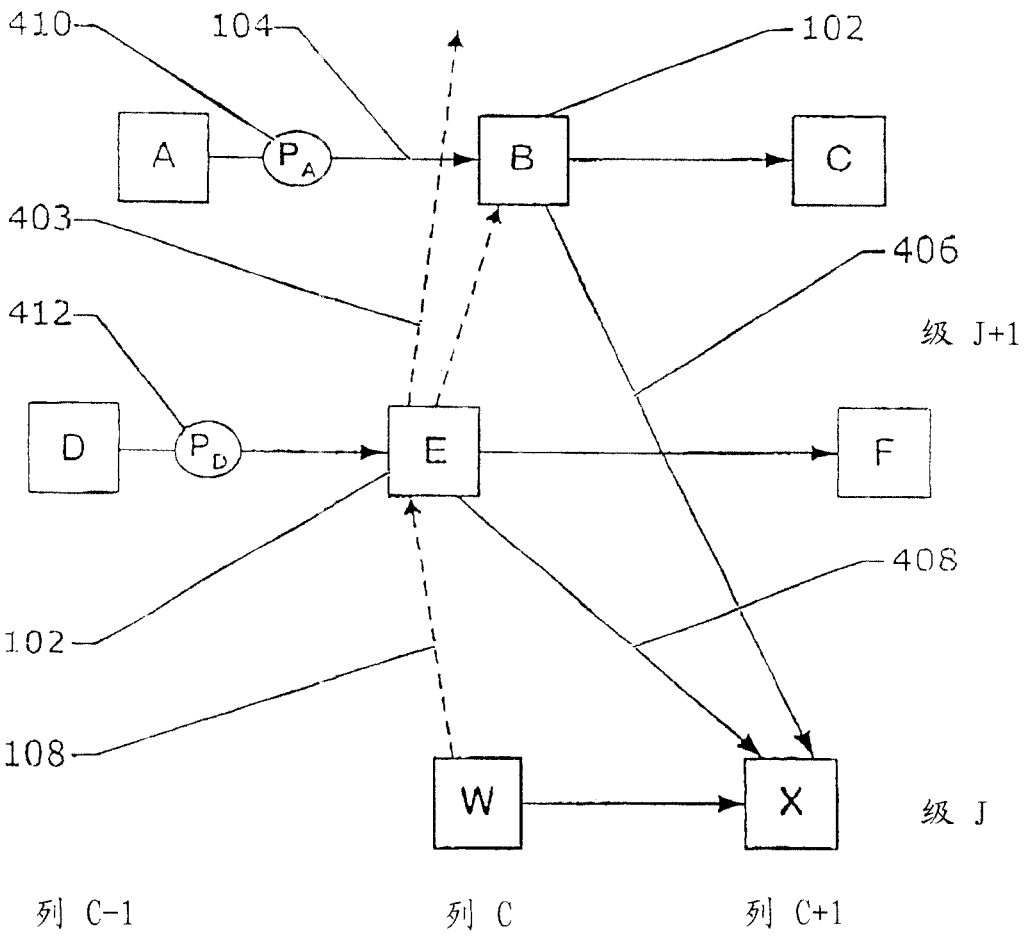

410
104
102
A
P_A
B
C
403
406
412
级 J+1
D
P_D
E
F
408
102
108
W
X
级 J
列 C-1
列 C
列 C+1

图 4C

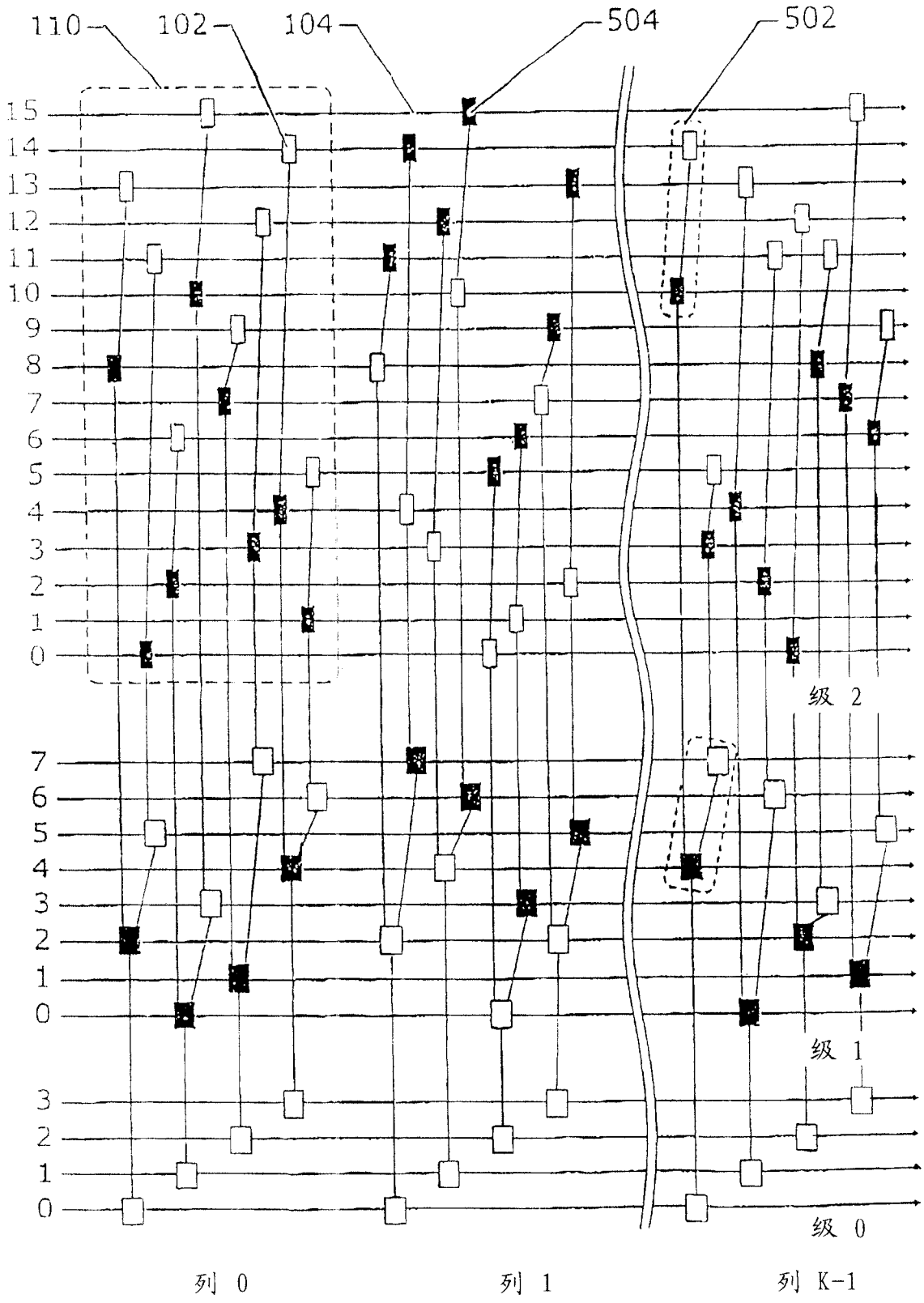

110
102
104
504
502
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
级 2
7
6
5
4
3
2
1
0
级 1
3
2
1
0
级 0
列 0
列 1
列 K-1

图 5

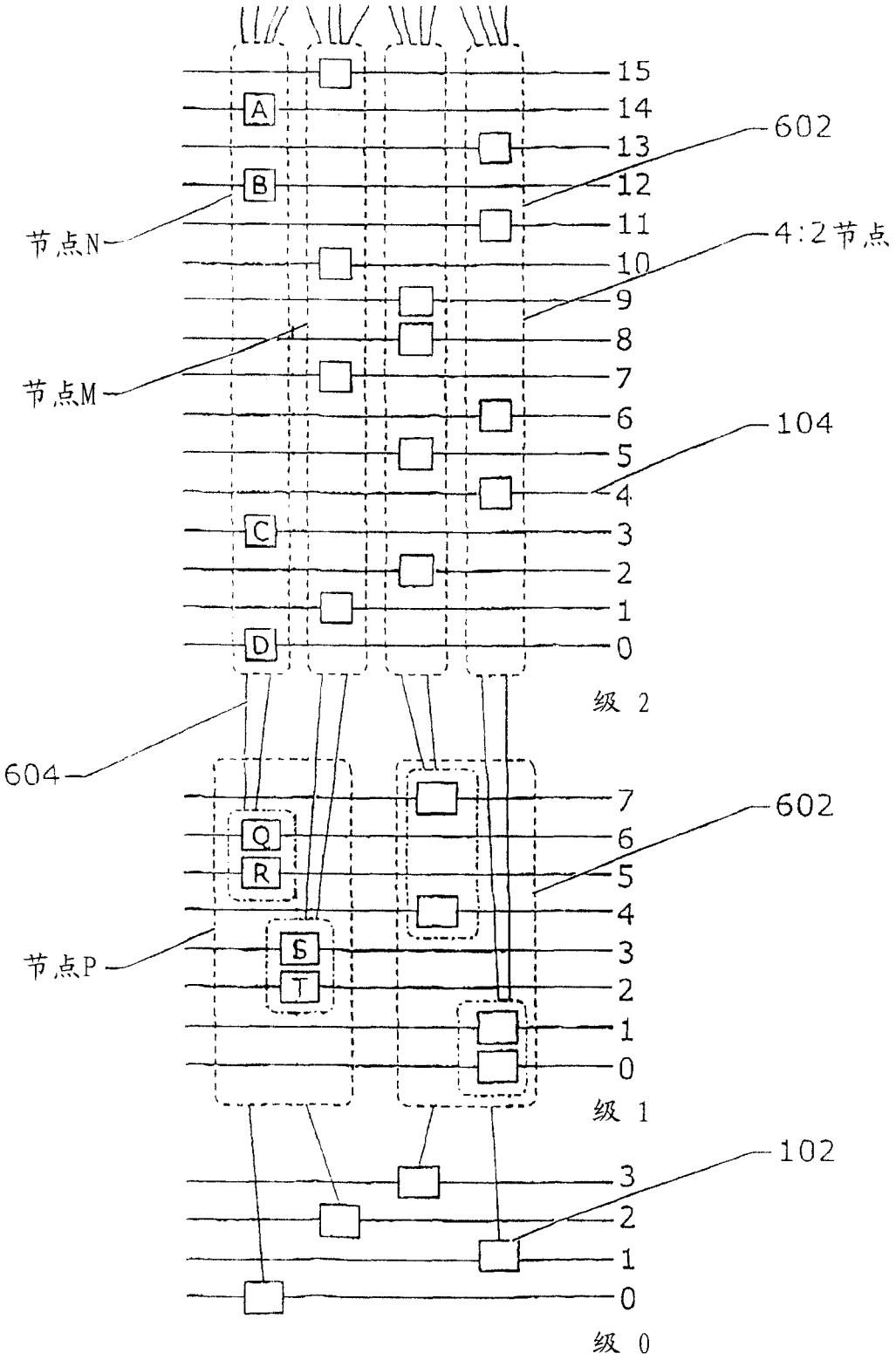

15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
A
B
C
D
602
4:2节点
104
节点N
节点M
级 2
604
7
6
5
4
3
2
1
0
Q
R
S
T
602
节点P
级 1
3
2
1
0
102
级 0

图 6A

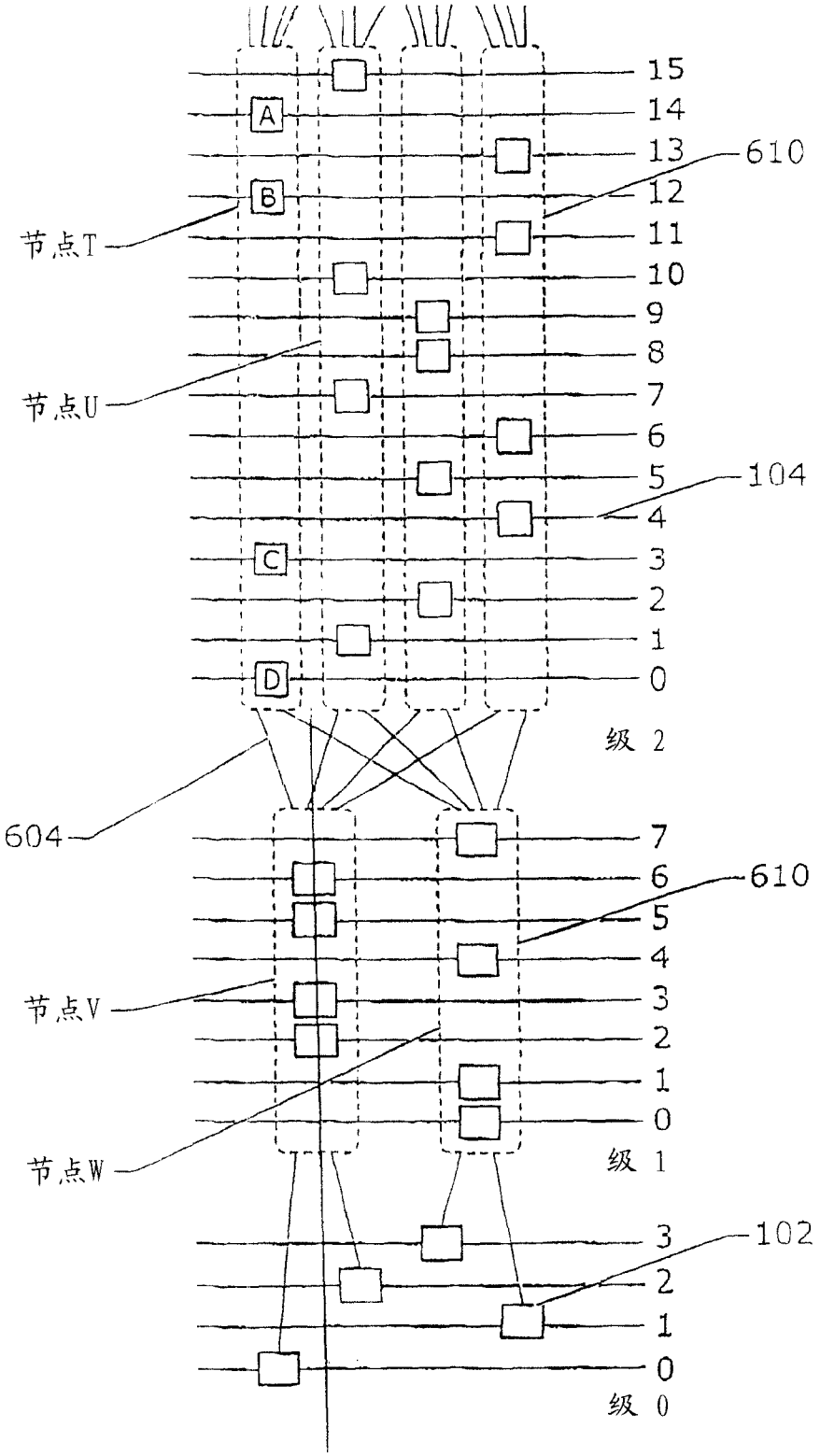

15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
级 2
7
6
5
4
3
2
1
0
级 1
3
2
1
0
级 0
A
B
C
D
节点T
节点U
节点V
节点W
604
610
610
104
102

图 6B

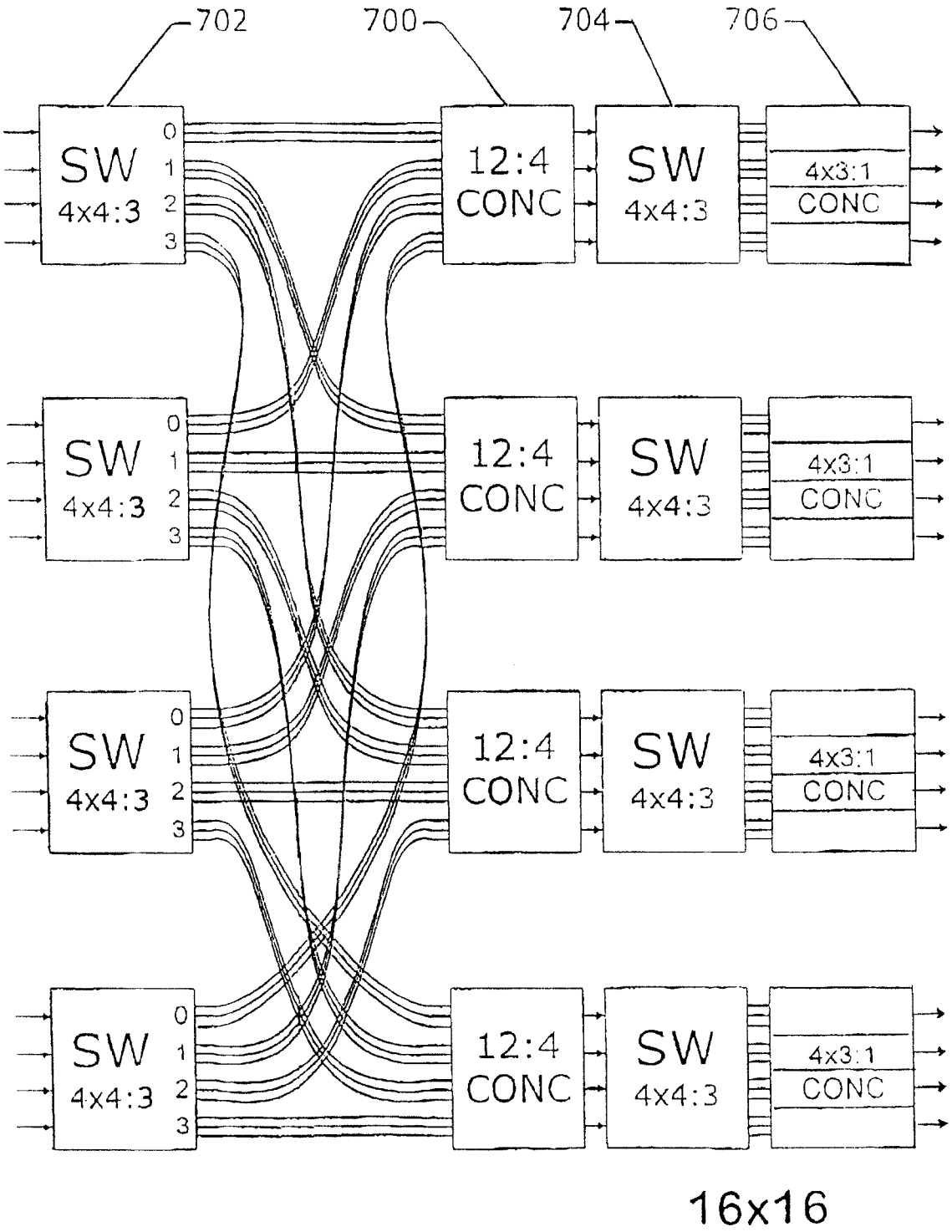


图 7