

具有拥塞缓解策略的动态虚拟通道研究 及其 VLSI 实现

赖明澈 王志英 郭建军 戴 葵

(国防科学技术大学计算机学院 长沙 410073)

摘 要 虚拟通道技术改善了片上网络性能,却带来了巨大的面积与功耗开销.通过分析静态虚拟通道的不足,提出了基于拥塞缓解策略的动态虚拟通道结构.它采用链表方式组织缓冲,可以自动调整通道结构来适应各种流量负载:在较低流量下,该结构扩展通道队列深度,减小了报文传输延迟;在较高流量下,它增加虚拟通道数量,消除队列头阻塞与通道不足阻塞,并缓解拥塞现象发生,减少流反馈次数,提高了网络吞吐率.在 90nm CMOS 工艺下完成了 DVC 路由器的 VLSI 设计,与传统路由器相比,不仅改善了报文传输延迟与吞吐率,而且有效降低了面积与功耗开销.

关键词 片上网络;虚拟通道;延迟;吞吐率;VLSI 实现

中图法分类号 TP302 **DOI 号:** 10.3724/SP.J.1016.2008.02026

Research and VLSI Implementation of a Dynamic Virtual Channel Structure with Congestion Awareness Scheme

LAI Ming-Che WANG Zhi-Ying GUO Jian-Jun DAI Kui

(School of Computer, National University of Defense Technology, Changsha 410073)

Abstract The virtual channel flow control approach provides an efficient way for the high throughput of on-chip routers. However, allocating the virtual channels statically results in a waste of area and energy consumption. Through the analysis towards shortcomings of statically-allocated virtual channels, a novel dynamic virtual channel structure with congestion awareness scheme is proposed. The buffer resources are organized by linked lists and their structures regulated according to the traffic conditions. In low traffic, it produces few deep channels to reduce the packet latency. In high traffic, it dispenses many VCs and avoids congestion situations to improve the throughput. The VLSI implementation of DVC router is completed under 90nm CMOS process. The experiment results show that the DVC router which suits for the various inject ratios and traffic patterns can provide throughput increase and latency decrease, with the obvious savings of silicon area and power consumption when compared to traditional routers.

Keywords network-on-chip; virtual channel; delay; throughput; VLSI implementation

收稿日期:2008-05-31;最终修改稿收到日期:2008-09-11. 本课题得到国家自然科学基金(60773024)、国家“九七三”重点基础研究发展规划项目基金(2007CB310901)和国家“八六三”高技术研究发展计划项目基金(2007AA01Z101)资助. 赖明澈,男,1982 年生,博士研究生,研究方向为高性能处理器体系结构、片上网络、ASIP 设计. E-mail: lmc82@163.com. 王志英,男,1956 年生,教授,博士生导师,研究领域为高性能计算机体系结构、VLSI 设计、信息安全等. 郭建军,男,1981 年生,博士研究生,研究方向为高性能处理器存储体系结构、片上网络、VLSI 设计. 戴 葵,男,1968 年生,副教授,研究方向为高性能计算机体系结构、计算机可靠性.

1 引言

随着半导体工艺迅猛发展,单芯片集成能力越来越强,系统设计者借鉴网络领域相关概念提出了新的通信技术——片上网络(network on chip)^[1],用于支持多核间的互连通信.片上网络采用点对点基本构架,能够适应超深亚微米工艺下较大传输线延迟,具有较高的通信带宽与良好的可扩展性.

目前,大多数片上网络均采用虫孔交换技术,其特点为:每个报文被分割成多个切片单元,头切片选择传输路径,体切片沿着路径呈流水方式传输,每个节点仅需配置较少缓冲,不需缓存完整报文内容.虫孔交换技术提供了较低传输延迟,但是无法解决队列头阻塞问题,即缓冲队列头部被阻塞的报文会阻碍其它报文的路由转发.文献[2]研究表明,队列头阻塞使得整个网络至多获取 58% 左右的吞吐率.为了解决该问题,大量片上网络研究采用了虚拟通道技术^[3-5],通过多个报文共享链路进行传输,减少了队列头阻塞次数,改善了网络性能.但虚拟通道技术也给芯片设计带来了不可忽视的面积与功耗负担^[6],文献[7]研究表明:在 70nm CMOS 工艺下,缓冲占据了片上路由器(包括链路)约 50% 的面积,同时耗费了 64% 左右的漏电流功耗.

本文从缓冲结构组织角度出发,讨论如何挖掘片上路由器潜在性能,同时降低其代价开销.针对该问题,近几年学术界曾展开过相关研究. Hu 等人利用排队论模型提出了基于程序通信特征的局部缓冲分析算法^[8],该算法与均匀分配策略相比节省了 85% 左右的缓冲资源,并且没有任何性能损失.随后, Huang 等人采纳虚拟通道技术,通过分析各节点处的报文冲突与期望带宽来进行虚拟通道分配,节省了 40% 的缓冲资源^[9].这些工作效果明显,但均采用了硬件资源定制的做法,无法适应网络流量与传输模型的变化.与上述工作不同,本文主要研究一种动态虚拟通道结构,它能依据网络负载的变化动态组织缓冲资源,因此具有更高实用价值.在多处处理器互连通信领域,动态通道队列早有研究. Tamir 等人最早提出了动态多队列结构(DAMQ)^[10],每个输入端口分别为各输出方向设置一个通道队列,通过共享缓冲来提高缓冲利用率. DAMQ 结构通道队列数量有限,队列头阻塞依然明显.另外, DAMQ 结构的控制硬件复杂,访问缓冲耗时 3 个周期,不易被片上网络采纳. Ni 等人提出了一种全相连环形缓冲

结构(FC-CB)^[11],它实现了单周期切片访问,具有较低的传输延迟与较高的吞吐率,但其通道数量仍然有限,无法满足较高流量负载下的报文传输需求.同时, FC-CB 结构访问切片时还需要执行缓冲环形旋转操作,消耗了大量动态能耗,违反了芯片低功耗设计原则.随后, Nicopoulos 等人还提出了一种虚拟通道动态分配结构(ViChar)^[12],其主要设计原则是在较高负载下通过增加通道数量来减少队列头阻塞,提高链路吞吐率.然而,这种结构硬件设计比较复杂,无法控制虚拟通道、控制表格体积,反而增加了路由器硬件开销.其次, ViChar 结构难以扩展通道深度,传输延迟较大,并且对各种流量与传输模型的适应性较差,流控制反馈频繁发生,网络性能有待进一步提高.

上述工作存在着一定的局限性,通过分析静态虚拟通道不足,本文提出了基于拥塞缓解的动态虚拟通道结构,它的特点主要体现在:首先,采用链表方式组织大量虚拟通道共享缓冲资源,具有较小代价开销,提高了缓冲利用率;其次,该通道结构可以自动调整来适应不同流量负载.在低流量情况下,扩展通道深度,减少报文分布的节点数目,降低传输延迟;在高流量情况下,通过增加通道数量来消除队列头阻塞与通道不足阻塞,并且缓解拥塞发生,减小流反馈次数,提高网络吞吐率.本文对路由器中的通道控制部件、通道分配部件加以改造,并且设计了拥塞缓解电路,完成了 DVC 路由器的 VLSI 实现.实验结果表明: DVC 路由器较好地适应了各种网络流量与传输模型,与典型路由器比较,降低了约 19.6% 的传输延迟,提高了约 7.4% 的吞吐率,同时节省了约 27.4% 的面积与 17.5% 的功耗.

2 NoC 路由器结构

典型片上路由器结构如图 1 所示,包括输入缓冲、通道控制(Channel Ctrl.)、路由计算(RC)、虚拟通道分配(VA)、传输仲裁(SA)与交叉开关(Crossbar)等部件.输入端口缓冲资源组织成 N 条对称通道,彼此共享交叉开关的输入链路,每次切片单元输入均依据通道标识缓存到正确的通道队列.在该结构中,路由计算部件负责报文传输路径解析,明确报文传输方向.通道分配部件接收各报文通道申请,结合下游节点通道状态实施通道分配,把空闲通道的使用权赋予唯一获胜报文.然后,由获取通道的报文提出传输请求,传输仲裁部件负责对所有传输请求执行两级仲裁,第 1 级在每个输入端口选择唯一获胜

报文,第2级继续对各输入端口的获胜报文加以仲裁.依据传输仲裁结果,各输入端口读取获胜通道头部的切片内容,穿过交叉开关到达目标端口输出.图1中的路由器通常采用 Mesh、Torus 等规整拓扑结构.本文将要讨论的动态通道将以具有5个端口的路由器为例,分别对应 E、S、W、N 与本地处理单元.

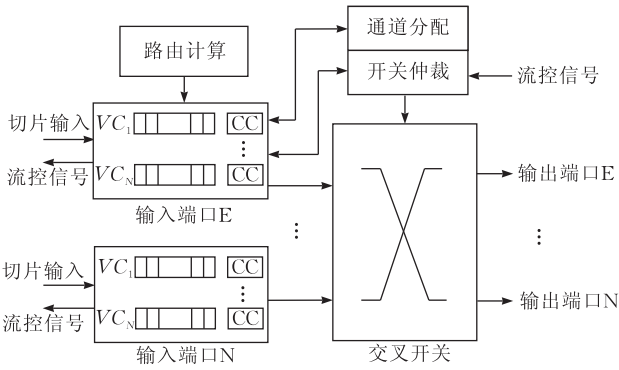


图1 片上路由器结构框图

3 静态通道结构分析

典型片上路由器采用虚拟通道来减少队列头阻

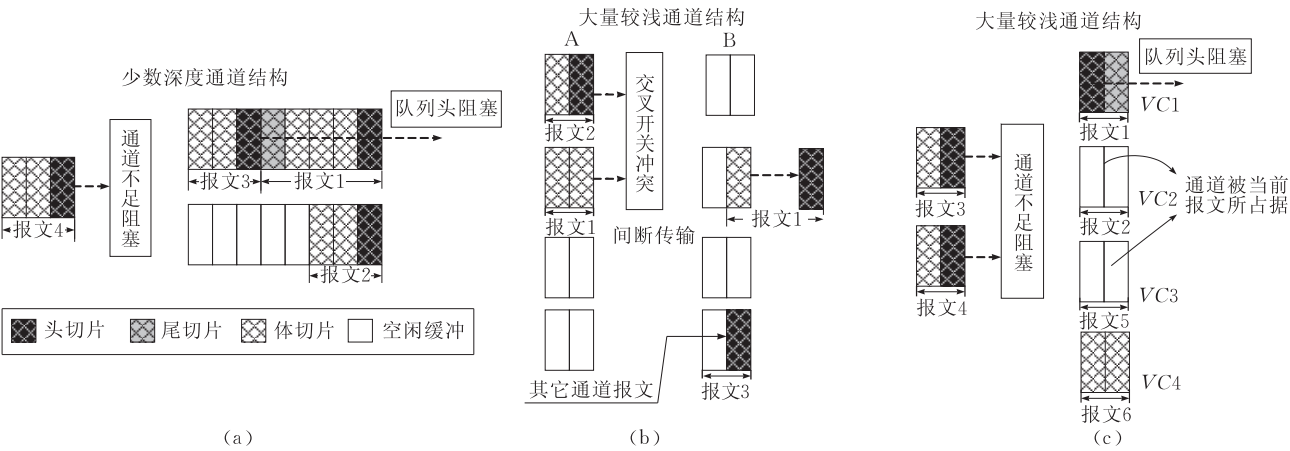


图2 静态虚拟通道结构的不足

其次,可以通过增加通道数量来减少阻塞现象,但它也暴露出了许多性能上的缺陷.(1)更多通道数量意味着通道深度较浅,在低负载情况下它会迫使单个报文分布在若干个节点之中,图2(b)显示了它的不良影响.当报文1头部切片离开节点B时,其后续切片在节点A处与其它报文发生了冲突,延长了报文1通过节点B的时间.Rezazad^[13]也曾研究表明报文连续性传输有助于减少传输延迟.(2)即使采用了更多虚拟通道仍无法彻底消除队列头阻塞与通道不足阻塞现象.在静态通道结构中,每个通道在服务周期内仅被唯一报文独占使用,如果

塞,当部分报文被阻塞时,允许其它通道报文继续传输.从理论上讲,设置更多数量虚拟通道有利于提高网络性能.但受到芯片缓冲约束,增加通道数量必然会减少通道深度,容易使得单个报文分布在大量节点中,这样无疑增加了每个节点上的路由冲突,延长了报文传输时间.针对有限的缓冲资源,Rezazad^[13]等人通过实验得出结论:在低流量负载下,较深通道对应较低传输延迟;在高流量负载下,需要继续区分不同传输模型.在随机分布模型中,增加通道数量有助于提高吞吐率;在“热点”模型中,扩展通道深度可以降低报文传输延迟.

然而,无论采用上述哪种通道结构,都无法兼顾各种流量与传输模型,事实上造成了网络性能瓶颈,暴露了缓冲利用率低、路由器代价较大等不足.首先,少量深度通道配置仅适合于低负载情况,随着流量增加,该结构极易发生各种阻塞.如图2(a)所示,由于队列头部的报文1、2无法传输,报文3与报文4会分别发生队列头阻塞与通道不足阻塞.这些现象均由通道数量不足所引起,导致链路空闲,影响了网络性能.

当前报文数量大于通道数量,那么容易发生阻塞情况.在该结构中,虚拟通道与缓冲资源相互绑定会致使缓冲利用率下降.如图2(c)所示,当前通道2、3上的切片已流出,但余下切片被阻塞在上游节点,以至于通道资源无法及时释放.这种情况会造成大量空闲缓冲无法被其它报文使用,报文3、4由于无法获得通道而发生了通道不足阻塞.

4 基于拥塞缓解的动态虚拟通道技术

根据上述分析,报文间断传输增加了低负载下

的传输延迟,队列头与通道不足阻塞限制了高负载下的网路吞吐率.虽然 ViChar 结构能依据网络流量产生大量通道,有效避免了队列头阻塞与通道不足阻塞,但其控制逻辑太复杂,特别是通道控制表格体积为 $O(vr)$,反而增加了硬件开销^[16].在低负载下,这种结构难以扩展通道深度来维持报文连续传输;在高负载下,即使分配更多通道,仍无法减少流反馈次数,报文阻塞明显.为此,提出了一种基于拥塞缓解的动态虚拟通道技术,分别采取了如下措施(如图 3 所示):利用链表动态组织虚拟通道,链表体积为 $O(v)$.在低负载情况下,缓冲动态组织成较深队列,减少报文分布的节点数目,保证报文连续传输.在高负载情况下,每个报文使用单独通道,避免队列头阻塞;同时,通过增加通道数量来解决通道不足阻塞.在高负载情况下,还专门采取了拥塞缓解机制,优先报文向低负载区域传输,减少流控制反馈出现次数,继续提高网络性能.

各种不利因素	采取相关措施
队列头阻塞现象	每个报文占用独立通道队列
通道数量不足阻塞现象	配置大量虚拟通道,通道与缓冲解除绑定,提高缓冲利用率
报文间断性传输(低负载)	动态形成较深的通道队列
大量流控制反馈(高负载)	拥塞缓解,优先向低负载区域的报文传输

图 3 基于拥塞缓解动态虚拟通道采取的策略

4.1 动态虚拟通道结构

在本节中,动态虚拟通道采用链表方式组织,它能够自动调整通道结构来适应流量负载的变化.其内部结构如图 4 所示,包括 N 条通道队列与一个缓冲 Tracker 单元.每条通道队列由各自通道控制单元负责控制,其长度在 0 与 L 间变化,记录的主要信息包括:Valid(有效域,标识报文是否已经获取下级节点中的通道使用权)、 H_P (头指针域,表示通道队列头索引)、 T_P (尾指针域,表示通道队列尾索引)、 ID (下级通道标识,表示申请的通道标识)、 DIR (方向域,表示报文流出方向)与 N_DIR (下级方向域,表示下级节点中的流出方向).每个链表项信息则包括 $Data$ (数据域,表示切片内容)与 N_P (链表域,表示后续切片索引).另外,图 4 还包括一个缓冲 Tracker 单元,它利用位图方式记录输入缓冲状态,在每个周期提供一个空闲缓冲索引来缓存新切片内容,同时回收流出切片使用过的缓冲项.

动态虚拟通道在切片输入或流出过程中通过执行少量操作来组织缓冲资源.其中,每次切片输入执

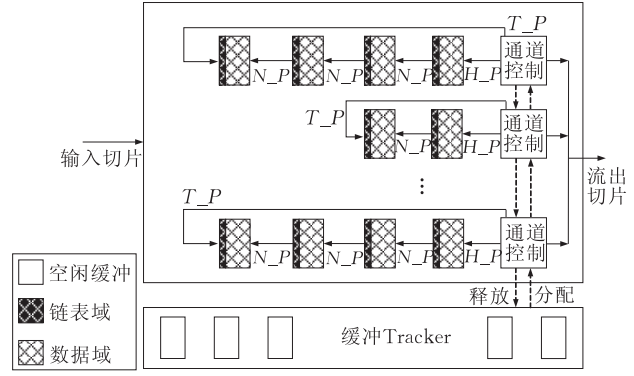


图 4 动态虚拟通道链表结构

行 3 个操作.首先,利用缓冲 Tracker 单元提供的空闲缓冲索引来缓存切片单元内容,如式(1)所示;其次,式(2)与式(3)表示利用相同索引更新通道队列尾指针与通道队列尾切片的 N_P 域.而每次切片流出也需要执行 3 个操作.式(4)表示依据 SA 部件返回的通道标识 i 读取获胜通道队列头部的切片内容;式(5)表示更新通道队列的头指针域,让其指向后续切片;然后,式(6)表示缓冲 Tracker 单元更新位图,回收通道队列 i 头部的缓冲项.

$$data[tracker()] \leq flit_incoming \quad (1)$$

$$VC[i].T_P \leq tracker() \quad (2)$$

$$N_P[VC[i].T_P] \leq tracker() \quad (3)$$

$$flit_departure \leq data[VC[i].H_P] \quad (4)$$

$$VC[i].H_P \leq N_P[VC[i].H_P] \quad (5)$$

$$tracker[VC[i].H_P] \leq 1 \quad (6)$$

基于上述操作,每个虚拟通道可以动态申请或释放缓冲资源,并且依据流量负载来自动调整通道结构,具备了一定自适应能力.在较低负载情况下,由于报文数量较少,每个节点仅使用少量虚拟通道.输入端口拥有足够空闲缓冲来分配给每个输入切片,从而形成少量较深的通道队列,流反馈控制不会产生,此时保证了报文连续性传输.在较高负载情况下,报文数量越来越多,该结构给每个报文分配单独的通道队列,使得通道数量迅速增加.另外,在较高负载下,频繁流控制反馈还会产生更多通道队列.例如,本地节点接收下游节点的流控制反馈后取消当前报文传输,将优先级赋予沿其它方向传输的报文.等待该方向的流控制反馈信号解除,所有报文又重新竞争传输优先权.此时,如果其它报文竞争获胜,会产生多个报文交叉传输现象,显然会增加下游节点中的通道队列数量.总而言之,在较高负载情况下,动态通道结构分配了更多虚拟通道,消除了队列头阻塞与通道不足阻塞,让更多报文复用传输链路,这样有利于提高网络吞吐率.

4.2 拥塞缓解原理

在较高负载情况下,动态通道结构通过产生大量通道消除了队列头阻塞与通道不足阻塞,然而,网络中还存在着大量流控制反馈阻塞现象.如果下游节点输入端口没有空闲缓冲,那么它会立即产生流控制反馈信号阻止本地报文继续传输,直至空闲缓冲再次出现,流控制反馈才被解除.假如下游节点产生流控制反馈信号,即使当前路由器分配再多的虚拟通道,报文传输也会被抑制,传输链路空闲将会导致网络吞吐率下降.尤其是,在下游节点输入端口不剩空闲缓冲的情况下,如果它的报文再被阻塞,那么将长时间持续产生流控制反馈信号,我们称这种现象为拥塞状态.以图 5(a)为例,由于 C 节点反馈的流控制信号或 B 节点处的路由冲突,报文 1、2 在向 C 节点传输过程中被阻塞在 B 节点.报文 2 的后续切片继续涌现 B 节点容易造成 B 节点 w 端口进入拥塞状态.此时,不仅报文 1、2 被阻塞,而且后续报文 3、4 也会由于 B 节点发出的流反馈控制而停滞在 A 节点.很显然,链路 $A \rightarrow B$ 、 $B \rightarrow E$ 、 $B \rightarrow D$ 都将空闲.

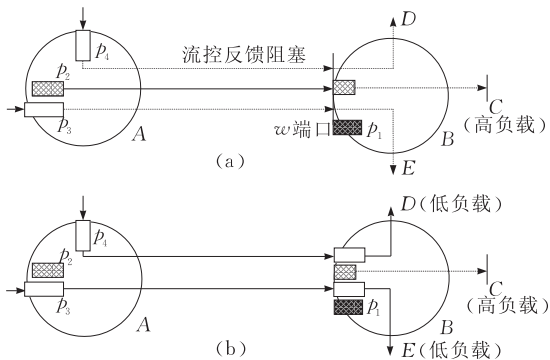


图 5 拥塞缓解策略示意图

针对上述情况,本文提出了拥塞缓解策略,其主要思想是优先向邻居节点周围低负载区域的报文进行传输.根据该策略,邻居节点周围低负载区域的空闲缓冲较多,优先报文向其传输很少受到流反馈控制,并且路由冲突概率也较小,这样降低了报文阻塞在下级节点的可能性,缓解了拥塞现象发生.图 5(b)继续举例说明拥塞缓解原理.首先,预测 B 节点 w 输入端口是否即将发生拥塞.假设 B 节点 w 输入端口在未来 k 个周期内没有任何切片输入,计算 k 个周期后 w 端口剩余的空闲缓冲数量.如果预测数量小于 k ,那么表示继续向 B 节点 w 端口传送切片可能会导致 k 个周期后的拥塞发生.然后, A 节点根据 B 节点各邻居输入端口的负载,立即取消报文 2 的传输请求,优先报文 3、4 向下游节点周围低负载区

域传输,这样能够最大程度避免 B 节点 w 端口的拥塞发生.

拥塞缓解的符号说明如图 6 所示.在 t 时刻, $x_d(t)$ ($d = E, S, W, N, L$) 表示 d 方向输入端口中的空闲缓冲数量; $nx_d(t)$ 表示 d 方向邻居输入端口中的空闲缓冲数量; $c_d(t+1, k)$ 预测 d 方向输入端口在后续 k 个周期内即将流出的切片数量; $nc_d(t+1, k)$ 预测 d 方向邻居输入端口在后续 k 个周期内即将流出的切片数量; $b_d(t+1, k)$ 负责计算 d 方向邻居输入端口在未来 k 个周期后所剩空闲缓冲数量.另外, f_d 表示 d 方向输入端口当前报文传输剩余的切片数量; $\Delta_d = \{\sigma_{d,E}, \sigma_{d,S}, \sigma_{d,W}, \sigma_{d,N}, \sigma_{d,L}\}$ 表示来自 d 方向邻居的负载向量,而 $\sigma_{d,E}, \sigma_{d,S}, \sigma_{d,W}, \sigma_{d,N}, \sigma_{d,L}$ 分别表示邻居周围各方向上的负载因子.

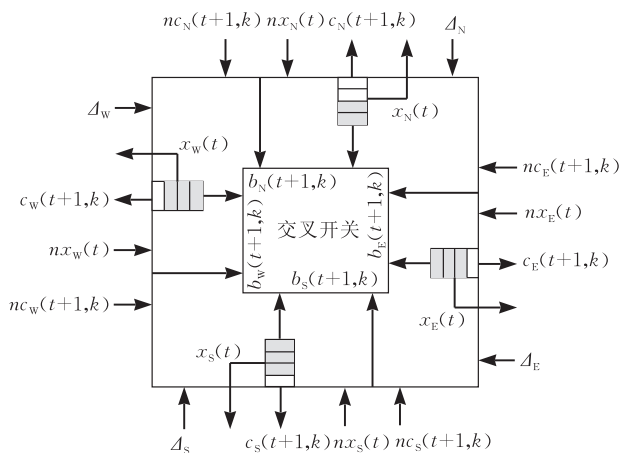


图 6 拥塞缓解符号示意图

拥塞缓解步骤如下:

首先,路由节点利用输入参数 $nx_d(t)$ 与 $nc_d(t+1, k)$ 预测 d 方向邻居输入端口在未来 k 个周期后所剩空闲缓冲数量,如式(7)所示.在初始时刻, $nx_d(0) = C$, $nc_d(1, k) = 0$, C 表示输入端口缓冲数量.如果 $b_d(t+1, k)$ 小于等于预测步幅 k ,那么表示邻居节点输入端口未来可能发生拥塞现象.

$$b_d(t+1, k) = \max\{C, nx_d(t) + nc_d(t+1, k)\} \quad (7)$$

同时,路由节点预测从各输入端口即将流出的切片数量 $c_d(t+1, k)$,发送给邻居节点.由于采用虫孔交换技术,体切片呈流水方式传输,所以根据第 1 级 SA 仲裁结果,按照式(8)预测出每个输入端口在后续 k 个周期内至少流出的切片数量,等于 $b_q(t, k)$, k , f_d 三者的最小值.

$$c_d(t+1, k) = \min\{b_q(t, k), k, f_d\},$$

$$d, q \in \{E, S, W, N, L\} \quad (8)$$

其次,在预测 d 方向邻居输入端口可能发生拥

塞的前提下,还要根据负载向量 Δ_d ,优先那些向邻居周围低负载区域的报文进行传输.采取措施是:如果通道 i 输出方向上预测到拥塞状态,立刻判断式(9)、(10)是否成立.如果两者均成立,取消通道 i 上的传输请求.式(9)表示通道 i 上报文流向邻居周围的高负载区域,而式(10)表示路由节点还存在其它报文沿相同方向流出,但却流向邻居周围的低负载区域.

$$\sigma_{VC[i].DIR, VC[j].N_DIR} = 1 \quad (9)$$

$$\exists j (VC[i].DIR = VC[j].DIR \text{ and}$$

$$\sigma_{VC[j].DIR, VC[j].N_DIR} = 0) \quad (10)$$

其中,每个路由节点依据式(11)计算自己周围的负载因子,然后发送给邻居节点. $\sigma_{d,i}$ 为‘0’表示 i 方向属于低负载区域.

$$\sigma_{d,i} = 0 \text{ iff } b_i(t+1, k) < C/2 \quad (11)$$

5 DVC 路由器 VLSI 结构

基于典型 NoC 路由器结构^[14],本节进行了 DVC 路由器详细设计.接下来,将对通道控制部件、通道分配部件加以改造,并且重新设计拥塞缓解电路.

5.1 通道控制部件改造

典型 NoC 路由器输入端口配置了 N 条对称虚拟通道,每条通道固定分配 L 项缓冲.如图 7(a)所示,每次切片输入或者流出,通道控制部件依据通道标识选择正确的控制单元,按照“先进先出”次序访问通道队列中的切片内容.改造后的通道控制部件如图 7(b)所示,它在每个输入端口采取了集中式控制,其任务之一是将通道标识转换为访问缓冲索引,读写存储体中切片内容.改造后该部件取消了通道

与缓冲之间的绑定关系,每个虚拟通道动态申请缓冲,明显提高了缓冲利用率.

针对通道控制部件, DAMQ^[10] 与 FC-CB^[11] 结构曾在多处理器通信领域展开过尝试,然而,这些技术应用在片上网络中却暴露了访问时间长、操作功耗大等缺陷.动态虚拟通道设计难点在于如何以较小代价开销支持单周期切片访问.下面将具体介绍集中式控制部件缓冲索引产生的电路结构,如图 8 所示,包括队列头指针寄存器组、尾指针寄存器组、缓冲链表寄存器组与缓冲 Tracker 单元.在 4.1 节中,每次切片输入被描述为 3 个操作,在图 8 中体现为:首先,Tracker 单元每个周期提供一个空闲缓冲索引,用于缓存新输入切片内容;其次,以输入切片通道标识为地址,以 Tracker 单元提供索引为数据,更新尾指针寄存器的 T_P 域内容;另外,还以从尾指针寄存器中读取的 T_P 域内容为地址,Tracker 单元提供的索引为数据,同时负责更新链表寄存器中的 N_P 域.类似,每次切片流出也描述为 3 个操作.首先,控制单元设置了一个比较单元,用于判断上周期获胜通道在本周期第 1 级 SA 仲裁中是否继续获胜.如果比较结果为真,那么意味着报文连续性传输,我们直接利用上个周期流出切片的索引在链表寄存器中获取本周期即将流出的切片地址;否则,利用获胜通道标识从头指针寄存器组中获取访问地址.在式(4)的上述操作中,比较单元、多路选择器与第 2 级 SA 仲裁并行执行,故而不会对时序关键路径长度产生影响.紧接着,利用上个周期流出切片所对应的 N_P 域内容来更新头指针寄存器中的 H_P 域,完成式(5)描述的功能.最后,Tracker 单元根据流出切片索引来清除对应缓冲状态位,释放流出切

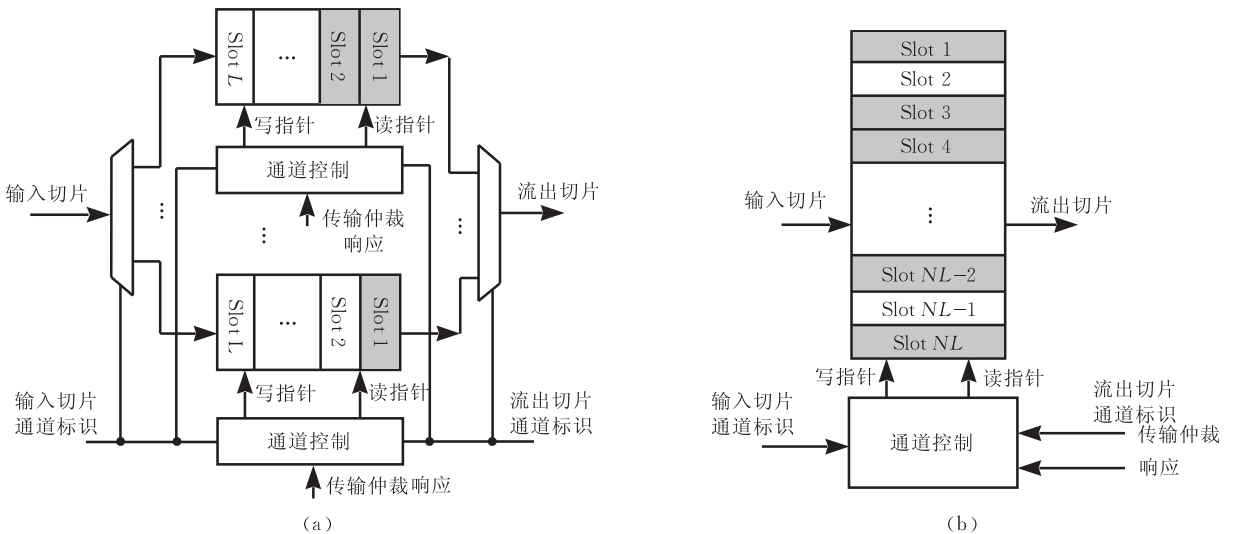


图 7 典型路由器与 DVC 路由器的通道控制部件比较

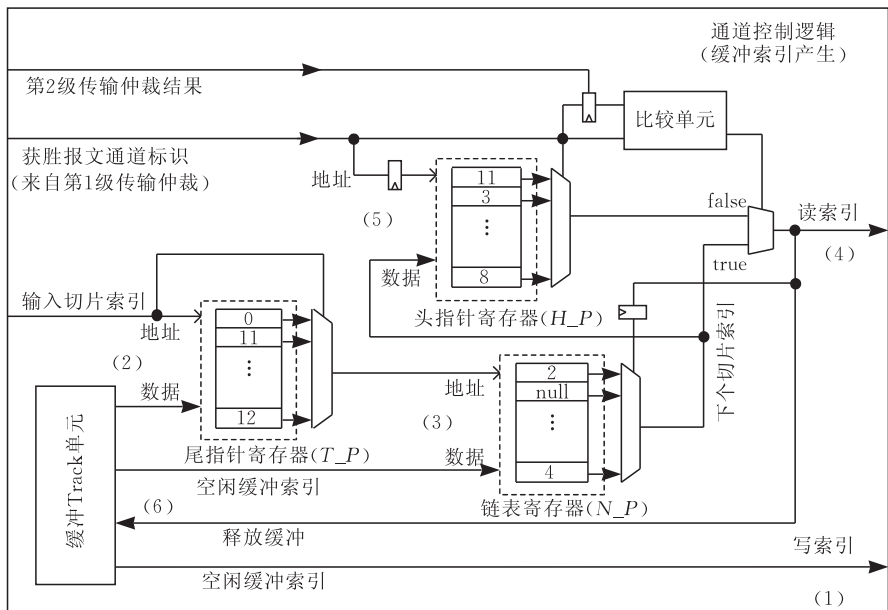


图 8 DVC 通道控制部件缓冲索引产生原理图

片占据的缓冲项.

上述通道控制部件不仅支持单周期切片访问,而且还具备较小代价开销.与典型路由器的通道控制部件比较,除了 Tracker 单元位图更新操作之外,每次切片输入仅增加一次链表寄存器写操作,而每次切片流出也仅增加一次头指针寄存器写操作与一次链表寄存器读操作.这些寄存器规模非常小,不会对能耗产生太大影响.在缓存开销方面,假设通道数量为 2^m ,缓冲数量为 2^n ,宽度为 W bits,那么上述通道控制部件额外开销如式(12)所示.如果 $m=3$, $n=4$, $W=128$,那么额外开销仅为 5.4%.

$$Overload = ((2 \times 2^m \times n) + (2^n \times n) - 2 \times 2^m \times (n - m)) / ((2 \times 2^m \times (n - m)) + 2^n \times W) \quad (12)$$

5.2 通道仲裁部件的改造

典型的通道仲裁部件采用了两级仲裁逻辑^[14]。第1级依据下级节点通道状态,为每个报文选择一个空闲通道并且提出申请;第2级对下级节点的每个输入通道进行仲裁,将通道使用权赋予唯一获胜报文。仲裁电路如图9(a)所示,第1级包括 $p_i v$ 个 $v:1$ 仲裁单元,第2级包括 $p_o v$ 个 $p_i v:1$ 仲裁单元。在DVC路由器中,随着通道数量 v 急剧增加,继续

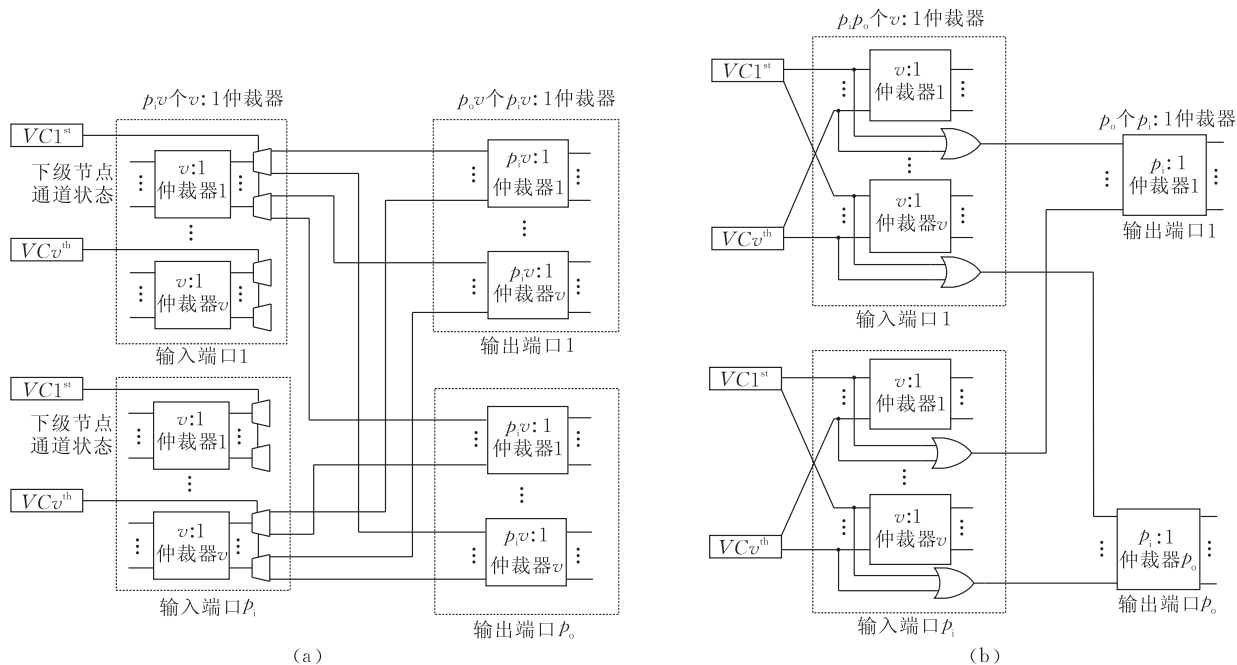


图 9 典型路由器与 DVC 路由器的通道仲裁部件

6 性能评测

6.1 实验环境

本节设计实现了片上网络的时钟精度软件仿真器,对 DVC 路由器展开性能评测.该仿真器采用 System C 语言加以描述,具有时钟模拟精度以及良好的可扩展性,可以对网络大小、拓扑结构、路由算法等灵活配置.仿真器允许扩充传输模型,我们选择了典型的随机分布与“热点”分布模型,能够反映大多数应用程序通信特征.本节采用的典型路由器配置如表 1 所示,每个输入端口设置 32 项缓冲,依据通道数量不同可分为 T-2 与 T-4 两种路由器.通过重新设计路由器关键部件,本节完成了 DVC 路由器的集成与调试工作.DVC 路由器输入端口也配置 32 项缓冲,最多支持 16 条虚拟通道.基于 3 种路由器,在各种注入激励与传输模型下对不同网络分别进行仿真,每次模拟 1×10^6 个周期.为了获取稳定网络性能,起初 2×10^5 个周期属于预热阶段,随后 8×10^5 个周期用于性能统计.

表 1 仿真器相关参数配置	
参数	取值
拓扑结构	Mesh
网络大小	8
路由算法	XY 维序路由
缓冲数量	32
路由器流水线级数	3
虚拟通道数量	2 或 4
报文大小	头切片+8 个体切片,16 字节/切片
传输模型	随机分布,热点分布

6.2 步幅参数 k

首先,分析预测步幅参数 k 对网络性能影响.当步幅 k 取值较小时,路由器虽然能预测拥塞现象,但是报文向低负载区域传输过程中还是不断受流控制反馈作用,传输频繁被中断以至链路吞吐率依然有限.相反,如果步幅参数 k 取值较大,那么又会影响低负载情况下报文的连续性传输.为此,通过调整参数 k ,图 11 显示了各种报文注入速率与传输模型下传输延迟随参数 k 的变化曲线.可以看出:参数 k 从 0 增加到 8 的过程中,高负载下的报文传输延迟明显下降.其中,步幅 k 等于 0 表示没有采用拥塞缓解机制,此时产生拥塞现象会导致较大传输延迟.随着参数 k 取值增加,优先报文向邻居周围低负载区域传输,这样可以减少来自邻居节点的流反馈次数,有助于改善链路吞吐率.但随着参数 k 从 8 增加到

16,低负载下的报文传输延迟却开始上升.设置较大参数 k 会引起邻居节点输入端口在剩余较多空闲缓冲时提前预测拥塞现象,这种做法不利于低负载情况,会强制中断连续的报文传输.从图中实验结果看来,如果缓冲数量为 32,参数 k 设定为 8 能够获得较好的片上网络性能.

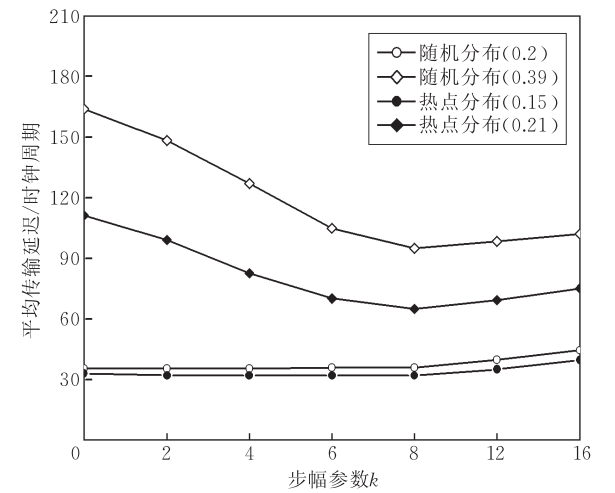


图 11 平均传输延迟随参数 k 的变化曲线

6.3 网络性能分析

接下来,对各种路由器展开性能测试,输入端口缓冲数量暂定为 32.随机分布模型下的仿真结果见图 12(a).在较低注入速率下,DVC 与 T-2 路由器具有大致相同的网络性能,但均优于 T-4 路由器.在低负载情况下,DVC 路由器提供了足够缓冲来扩展通道深度,使得单个报文分布节点数目减少,所以具有较小传输延迟.随着报文注入速率增加,T-2 路由器中较少的通道数量导致了大量队列头阻塞与通道不足阻塞,整个网络性能迅速趋于饱和.相比之下,T-4 路由器配置较多通道,减少了上述阻塞现象.但 T-4 路由器仍对应着较高的传输延迟.分析主要原因在于:一方面,T-4 路由器中虚拟通道数量有限,依然存在较多阻塞,性能损失较大;另一方面,T-4 路由器在高流量负载下出现的大量流控制反馈会导致链路频繁空闲.相比之下,DVC 路由器使用大量虚拟通道彻底消除了队列头阻塞与通道不足阻塞,同时采用了拥塞缓解机制来减少流控制反馈次数,因此对应更好的网络性能.

“热点”分布模型下的仿真结果见图 12(b)所示.在较低注入速率下,DVC 路由器与其它两种路由器具有大致相同的传输延迟.随着注入速率不断增加,T-2 与 T-4 路由器容易在“热点”附近产生大量队列头阻塞与通道不足阻塞,并频繁产生拥塞现

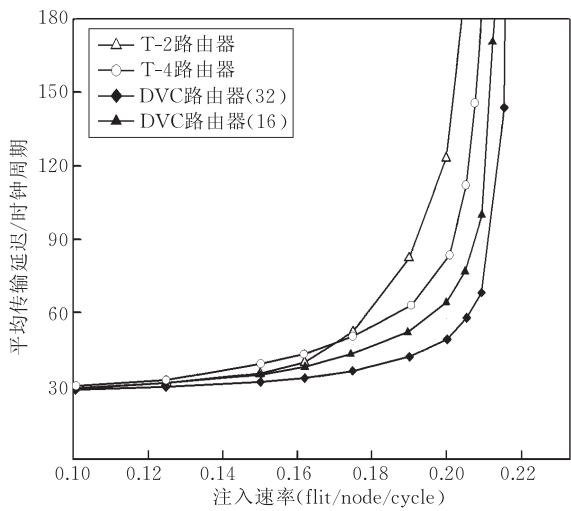
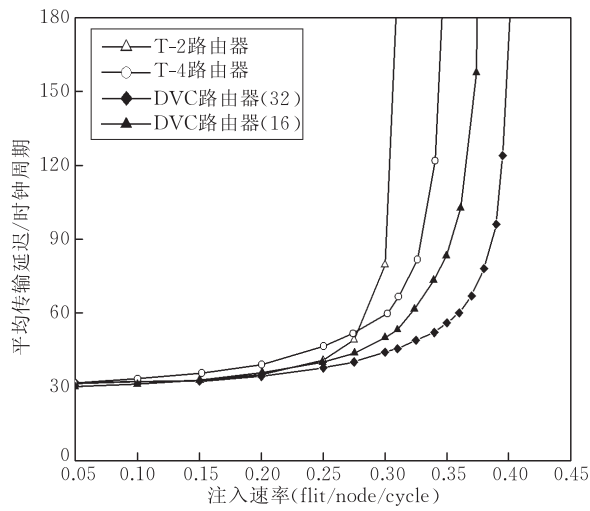


图 12 不同传输模型下的平均传输延迟

象,使得网络性能迅速趋于饱和.此时,DVC 路由器体现出了明显优势,主要原因在于 DVC 路由器中通道结构可以依据局部负载自动调整.在“热点”模型中,大量节点依然为低负载节点,仅少量节点处于高负载状态.处于低负载状态的 DVC 路由器扩展虚拟通道深度,降低了平均传输延迟.处于高负载状态的路由节点分配大量虚拟通道,消除队列头阻塞与通道不足阻塞,减少了报文停滞.尤其是,DVC 路由器在高负载节点处的拥塞缓解机制还能减少流控制反馈次数,通过优先报文向低负载区域传输,提高“热点”附近吞吐率,从而改善网络性能.

实验中还发现 T-2 与 T-4 路由器暴露出缓冲利用率低的不足.它们将缓冲绑定于特定通道,这种做法实现简单,却带来了大量阻塞现象,其中,上游报文被阻塞而本地缓冲仍有空闲的情况时有发生.在高流量负载下,统计了 T-4 路由器输入缓冲使用情况,发现有效缓冲的数量仅达 17 左右.所以,本节继续考虑裁减 DVC 路由器中缓冲数量,裁减原则为裁减后 DVC 路由器的传输延迟不应高于 T-2 与 T-4 路由器传输延迟.每次调整缓冲数量后均需要重新确定步幅参数 k .裁减后缓冲数量为 16,步幅参数为 4.图 12 还显示了裁减后路由器的传输延迟变化曲线.在随机分布模型中,DVC、T-4 与 T-2 路由器注入报文饱和速率分别为 0.375、0.335 与 0.30 flit/node/cycle.在 0.1、0.2、0.275 和 0.3 flit/node/cycle 注入样本下,DVC 路由器相对 T-2 路由器降低了 16.5% 的传输延迟;而在 0.1、0.2、0.3 和 0.325 flit/node/cycle 样本下,DVC 路由器相对 T-4 路由器降低了 21.6% 的传输延迟.在“热点”分布模型中,3 种路由器饱和注入速率分别

为 0.20、0.207 和 0.213 flit/node/cycle.本节统计了具有均匀分布特征的注入速率样本,包括 0.10、0.12 与 0.19 flit/node/cycle.其中,DVC 路由器相对 T-2 与 T-4 路由器传输延迟下降了约 20.3% 与 17.6% 左右.总体来说,T-4 路由器性能优于 T-2 路由器,而 DVC 路由器相对 T-4 路由器提高了约 7.4% 的吞吐率,并降低了约 19.6% 的传输延迟.

DVC 路由器还要考虑虚拟通道带来的硬件代价开销,冗余通道资源会给路由器设计带来较大负担.实验过程中发现在较高注入速率下路由器会使用更多数量虚拟通道,为此在两种传输模型中设定注入速率分别为 0.36 与 0.21 flit/node/cycle,观察传输延迟随通道数量变化曲线,如图 13 所示.可以看出,当虚拟通道数量为 10 与 16 时,传输延迟大致相等,这说明数量为 10 的通道配置已满足了片上网络需求.

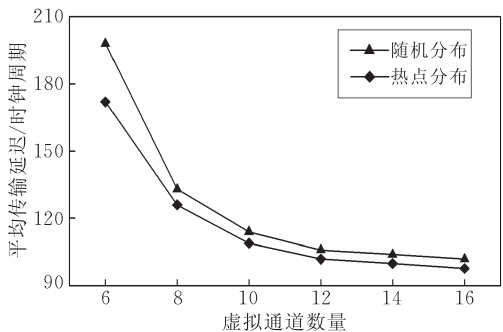


图 13 平均传输延迟随通道数量的变化曲线

紧接着,对 DVC 结构与 ViChar 结构继续加以比较. ViChar 结构的控制逻辑复杂,尤其是通道控制表格体积偏大.假设输入端口配置 16 项缓冲并支持 10 个虚拟通道, ViChar 通道控制表格及相关指

针的大小为 220B,约占输入缓冲总体积的 21.5%,而 DVC 结构的链表体积仅为 72B.由此可以看出,DVC 结构在裁减缓冲过程中具有更小额外开销.其次,DVC 结构还实现了拥塞缓解策略.在高流量负载下,ViChar 结构中频繁的流反馈现象会将报文阻塞在路由节点,不但中断当前报文传输,而且它们还会占据缓冲资源,使得后续其它报文无法传输.DVC 结构如果预测下游节点将要发生拥塞,立即优先向低负载区域的报文进行传输.这些报文在下游节点不会受到流反馈影响,并且路由冲突概率也较小,能有效减少本地产生的流反馈次数,提高网络整体性能.基于 ViChar 结构的性能测试,表 2 列出了 DVC 结构与 ViChar 结构在相同注入激励下的性能比较.在两种传输模型下,DVC 结构的传输延迟具有约 17.4%与 16.8%的优势.同时,DVC 结构在随机分布模型下还具有更高的吞吐率.

表 2 DVC 路由器与 Vichar 路由器的性能比较			
	注入速率/ (flit/node/cycle)	DVC 路由器/ cycles	ViChar 路由器/ cycles
随机 分布	0.25	37.4	39.2
	0.30	47.4	52.3
	0.325	59.5	75.9
	0.35	83.7	126.8
	0.36	103.6	∞
热点 分布	0.175	41.8	43.6
	0.19	48.2	58.6
	0.20	61.1	78.5
	0.21	98.5	128.0

6.4 DVC 路由器 VLSI 设计

本节采用 System Verilog 语言对 T-4 与 DVC 路由器(16 项缓冲,通道数量为 10)进行了 RTL 级硬件描述,并且利用 Design Compiler 工具在 90nm CMOS 工艺下完成了逻辑综合,各部件面积信息如图 14 所示.其中,传输仲裁与通道控制部件面积略微增加,前者主要是由更多数量的虚拟通道所引起,后者则源于索引产生逻辑与拥塞缓解电路.而另一方面,DVC 路由器的缓冲数量从 32 减少至 16,缓冲面积直线下降.同时,通道分配部件面积也由于仲裁器、多路选择器规模与数量的减小而下降.总体来说,与 T-4 路由器比较,DVC 路由器面积减少了约 27.4%.紧接着,还对 DVC 和 T-4 路由器进行了功耗评估.针对配置两种路由器的 4×4 网络分别展开逻辑综合并利用相同激励进行前端仿真,将获取的门级电平翻转模型与工艺库文件、网表一起输入 Prime Power 门级功耗分析工具^[15]来评估整个网络(不包括传输链路)的平均功耗.如表 3 所示,在两

种模型下 DVC 路由器平均功耗分别降低了约 16.7%与 18.3%.

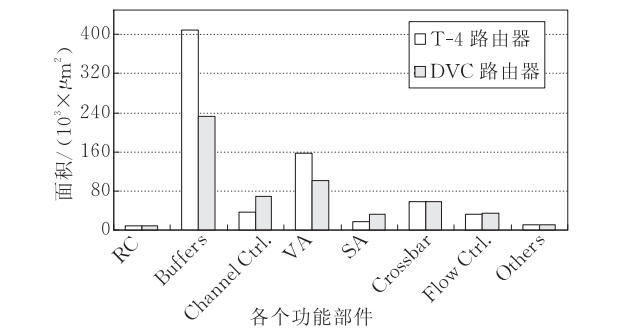


图 14 两种路由器的面积/(10³×μm²)

表 3 T-4 与 DVC 路由器的功耗比较 (单位:mW)		
	随机分布	热点分布
T-4 路由器	105.3	88.1
DVC 路由器	87.7	71.9

最后,利用 SOC Encounter 进行物理设计,利用 Formality 进行逻辑等价性检查,并利用 Prime-Time 工具进行静态时序分析.图 15(a)显示了 T-4 路由器的硬 IP 版图,其中包括 4 个 128 位 32 项的寄存器文件,分别对应于 4 个不同传输方向,总面积约 0.99mm×0.99mm.图 15(b)则显示了 DVC 路由器的硬 IP 版图,包括 4 个 16 项的寄存器文件,版图大小约 0.85mm×0.85mm,时序满足 1.6ns 约束条件.由此可以看出:基于拥塞缓解的动态虚拟通道结构不仅可以改善片上网络性能,而且有助于减小路由器的代价开销.

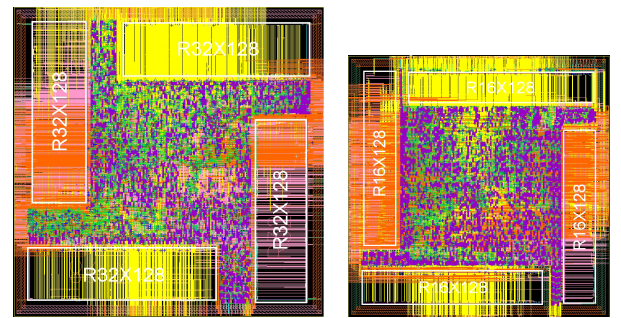


图 15 T-4 路由器与 DVC 路由器版图

7 结 论

引入虚拟通道技术来改善片上网络性能,却也同时带来了大量面积与能耗开销.本文提出了一种基于拥塞缓解的动态虚拟通道结构,并完成了 DVC 路由器硬件设计,给出了性能分析结果.它的主要特

点在于采用开销较小的链表结构来组织缓冲资源,通道结构能适应网络流量而动态变化。在较低负载下,DVC 路由器扩展通道深度,减少报文分布,降低了传输延迟;在较高负载下,通过增加通道数量消除了队列头与通道不足阻塞,并且采用缓解拥塞机制,减少流控制反馈次数,提高了网络整体性能。实验结果表明:与典型片上路由器比较,DVC 路由器降低了约 19.6% 的传输延迟,提高了约 7.4% 的吞吐量,同时节省了约 27.4% 的面积与 17.5% 的功耗。未来研究工作将主要集中在两个方面:首先,报文头切片通过每个路由节点花费 3 个周期,亟需研究一种快速通道技术;其次,利用 DVC 路由器与处理单元搭建多核处理器,利用实际应用的通信负载来评价互连网络性能。

参 考 文 献

- [1] Dally William J. Route packets, not wires: On-chip interconnection networks//Proceedings of the 38th Design Automation Conference. Las Vegas, NV, 2001: 684-689
- [2] Karol M J et al. Input versus output queueing on a space-division packet switch. IEEE Transactions on Communications, 1987, 35(12): 1347-1356
- [3] Nikolay Kavaldjiev, Smit Gerard J M. A virtual channel network-on-chip for GT and BE traffic//Proceedings of the VLSI Technologies and Architectures. Karlsruhe, Germany, 2006: 211-216
- [4] Aline Mello, Leonel Tedesco. Virtual channel in network on chip: Implementation and evaluation on hermes NoC//Proceedings of the 18th Annual Symposium on Integrated Circuits and System Design. Florianópolis, Brazil, 2005: 178-183
- [5] Mullins R, West A, Moore S. Low-latency virtual-channel routers for on-chip networks//Proceedings of the IEEE Symposium on Computer Architecture. München, Germany, 2004: 188-200
- [6] Wang H, Zhu X, Peh L, Malik S. Orion: A power-performance simulator for interconnection networks//Proceedings of

the 35th Annual IEEE/ACM International Symposium on Microarchitecture. Istanbul, Turkey, 2002: 294-305

- [7] Chen X, Peh L. Leakage power modeling and optimization in interconnection networks//Proceedings of the 2003 International Symposium on Low Power Electronics and Design. Seoul, Korea, 2002: 90-95
- [8] Hu J, Ogras U Y. System-level buffer allocation for application-specific networks-on-chip router design. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2006, 25(12): 2919-2933
- [9] Huang Ting-Chun, Ogras Umit Y. Virtual channels planning for networks-on-chip//Proceedings of the 8th International Symposium on Quality Electronic Design. San Jose, CA, 2007: 879-884
- [10] Tamir Y, Frazier G L. High-performance multiqueue buffers for VLSI communication switches//Proceedings of the Annual International Symposium on Computer Architecture. Honolulu, USA, 1988: 343-354
- [11] Ni N, Pirvu M, Bhuyan L. Circular buffered switch design with wormhole routing and virtual channels//Proceedings of the International Conference on Computer Design. California, USA, 1998: 466-473
- [12] Nicopoulos A et al. ViChaR: A dynamic virtual channel regulator for network-on-chip router//Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture. Orlando, USA, 2006: 333-344
- [13] Rezazad M et al. The effect of virtual channel organization on the performance of interconnection networks//Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium. Denver, USA, 2005: 264-272
- [14] Peh Li-Shiuan, Dally William J. A delay model and speculative architecture for pipelined router//Proceedings of the International Symposium on High-Performance Computer Architecture. Nuevo Leone, Mexico, 2001: 255-266
- [15] Data Sheet: PrimePower Full-Chip Dynamic Power Analysis for Multimillion-Gate Design. Synopsys, Inc, 2004
- [16] Avinash Kodi, Ashwini Sarathy. Design of adaptive communication channel buffers for low-power area-efficient network-on-chip architecture//Proceedings of the Symposium on Architectures for Networking and Communications Systems. Orlando, Florida, 2007: 47-56



LAI Ming-Che, born in 1982, Ph.D. candidate. His research interests include computer architecture, VLSI design, on-chip communication, and compiler optimization.

WANG Zhi-Ying, born in 1956, Ph.D., professor, Ph.D. supervisor. His main research interests include com-

puter architecture, computer security, VLSI design, reliable architecture, and asynchronous circuit.

GUO Jian-Jun, born in 1981, Ph.D. candidate. His research interests include computer memory hierarchy, on-chip communication, VLSI design.

DAI Kui, born in 1968, Ph.D., associate professor. His main research interests include high performance computer architecture, and reliable architecture.