文章编号:1001-2486(2011)06-0042-06

片上网络中一种单周期2GHz无缓冲路由器*

冯超超[1],鲁中海[2],张民选[1],李晋文[1]
(1. 国防科技大学 计算机学院,湖南 长沙 410073; 2. 瑞典皇家工学院 电子系统系,瑞典 斯德哥尔摩 16440)

摘 要:近年来,无缓冲路由器由于不需要缓冲器而成为片上网络低开销的解决方案。为了提高无缓冲路由器的性能,提出一种单周期高性能无缓冲片上网络路由器。该路由器使用一个简单的置换网络替换串行化的交换分配器与交叉开关以实现高性能。虚通道路由器与基准无缓冲路由器相比,该路由器在TSMC 65nm工艺下可以以较小的面积开销达到2GHz的时钟频率。在合成通信负载与真实应用负载下的模拟结果表明,该路由器的包平均延迟远小于虚通道路由器和其他无缓冲路由器。

关键词:片上网络; 无缓冲路由器; 偏转路由; 置换网络
中图分类号:TP302.8 **文献标识码**:A

A 1-cycle 2GHz Bufferless Router for Network-on-chip

FENG Chao-chao[1], *LU Zhong-hai*[2], *ZHANG Min-xuan*[1], *LI Jin-wen*[1]
(1. College of Computer, National Univ. of Defense Technology, Changsha 410073, China;
2. Department of Electronic Systems, Royal Institute of Technology, Stockholm 16440, Sweden)

Abstract: Recently, bufferless router, which does not need buffers, has become a low-cost solution for Network-on-Chip. To improve the performance of the bufferless router, a 1-cycle high-performance bufferless router was proposed for Network-on-Chip. The router used a simple permutation network instead of the serialized switch allocator and the crossbar to achieve high performance. Compared with the virtual channel router and the baseline bufferless router, the proposed bufferless router can achieve the frequency of 2GHz with small area cost under TSMC 65nm technology. Simulation results under both synthetic and application workloads demonstrate that the proposed bufferless router achieves much less average packet latency than the virtual channel router and other bufferless routers.

Key words: network-on-chip; bufferless router; deflection routing; permutation network

近年来,随着集成电路工艺尺寸的降低、集成度的提高,单芯片上已可以集成越来越多的处理单元。片上网络(Network-on-Chip, NoC)技术已逐渐成为单芯片多处理器(CMP)片上互连的设计范例[1]。在之前的片上网络研究中,大部分均采用虫孔或虚通道路由器,其特点是路由器每一个输入/出端口均包含缓冲器缓存网络中传输的包。虽然缓冲器可以有效改善网络的带宽利用率,减少丢包以及绕道路由的情况,但是也存在以下缺点:缓冲器消耗大量的能耗;需要复杂的流控策略来管理缓冲,增加了网络设计的复杂度;并且需要占据较大的芯片面积,在TRIPS原型芯片中,路由器输入缓冲占整个芯片面积的75%[2]。

为此,无缓冲路由器应运而生,为片上网络提供了一个低开销的解决方案[3-6]。在无缓冲路由器中,除了流水线寄存器外,路由器中没有额外的缓冲器,可以在很大程度上降低路由器的能耗和面积。无缓冲路由器可以分为基于丢包(drop-based)的路由器[3-4]和基于偏转(deflection-based)的路由器[5-6]两种。基于丢包的无缓冲路由器需要额外的机制来处理丢包的情况,在一定程度上增加了实现开销。在基于偏转路由的无缓冲路由器中,路由器接收到包后立即将其转发到下一个路由器,在出现竞争的情况下,一些包将偏离最短路径路由。由于偏转路由的非最短路径自适应路由特性,需要采用相应的技术避免包被无限制地偏转以产生活锁(livelock)。目前的偏转路由算法通常按包优先级从高到低的顺序进行输出端口分配来避免活锁[7],但是,串行化的端口分配也导致了无缓冲路由器中的关键路径过长,

* **收稿日期**:2011-06-20
基金项目:国家自然科学基金资助项目(60970036、60873212、61003301);国家"核高基"重大专项"超高性能CPU新型架构研究"资助项目(2011ZX01028-001-001)
作者简介:冯超超(1982—),男,博士生。

限制了其实现的时钟频率。为了消除路由器中输出端口分配器的复杂逻辑,文献[6]提出一种简化的无缓冲路由器(CHIPPER),采用一个部分置换网络替换输出端口分配器和交叉开关。但是,CHIPPER 路由器仍然采用 2 级流水线实现,并且需要采用一种称为 golden Packet 的复杂规则来避免活锁。

本文提出一种单周期高性能片上网络无缓冲路由器(BLESS_PERM)。该路由器使用一个简单的置换网络(permutation network)替换原路由器中串行化的交换分配器与交叉开关以缩短路由器中关键路径的长度,并且不需要采用流水线,单周期即可完成包交换。在 TSMC 65nm 工艺下对路由器进行综合,可以达到 2GHz 的时钟频率,并且具有较小的面积开销。在合成通信负载与真实应用负载下的模拟结果表明,BLESS_PERM 路由器的包平均延迟远小于虚通道路由器、基准无缓冲路由器和 3 级流水线无缓冲路由器。

1　基准无缓冲路由器

本文采用瑞典皇家工学院开发的 Nostrum 无缓冲路由器[8]作为基准路由器。该路由器是基于偏转路由的单周期路由器,除了输入寄存器外,路由器中不包含额外的缓冲,其结构如图 1 所示。包中的每一个包片(flit)均包含一个包头,在网络中独立路由,包片可以乱序到达目标节点。因此,需要在目标节点的网络接口中包含一个大小为 $n \cdot k$(其中 n 为项数,k 为每一个包中的包片数)的重排序缓冲(reassembly buffer)对包进行重组。路由器采用偏转路由算法,即包片不在路由器中缓存,路由器接收到包片后立即将其转发到下一个路由器。每一个包片长度 128 位,包片格式如图 2 所示,其中包头 32 位,数据 96 位。包头分为 4 个域:一位有效位(V)表示该包片是否有效(0:无效包片;1:有效包片);目的地址域(Dst_addr)与源地址域(Src_addr)均采用相对寻址,即:当前节点分别到目标节点与源节点的相对地址。其中,高 6 位为行地址,低 6 位为列地址,采用补码编码。跳步数域(Hop_counter,7 位)记录该包片已在网络中经历的跳步数,用于表示包片的优先级以避免活锁。包片每经过一个路由器,地址域与跳步数域均由图 1 所示的包头更新模块进行更新。

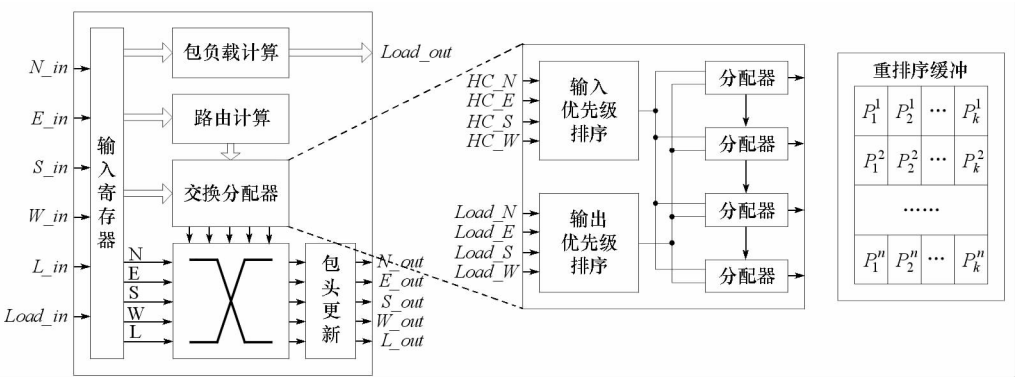


图 1　基准无缓冲路由器结构

Fig. 1　Architecture of baseline bufferless router

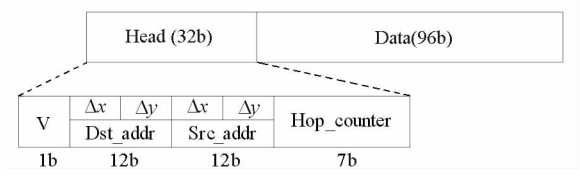


图 2　包片格式

Fig. 2　Flit format

偏转路由是一种非最短路径完全自适应路由算法,由于包片不需要在路由器中缓存,因此该路由算法不存在死锁(deadlock)。在偏转路由中,当属于最短路径的输出端口都被占用时,包片就被偏转,使其距目标节点更远。如果一个包片被连续偏转,就可能产生活锁,即包片永远不能到达目标节点。因此,需要通过限制偏转次数来避免活锁的发生。基准无缓冲路由器根据包片在网络中经历的跳步数对输入包片进行优先级排序,按优先级从高到低的顺序进行输出端口分配以避免活锁。在网络中经历时间越久的包片优先级越高。当两个或多个包片竞争同一个属于最短路径的输出端口时,优先级最高的包片通过该端口路由,而其他包片被偏转到非最短路径输出端口。

交换分配器按输入包片优先级从高到低的顺序将输出端口分配给每一个输入包片。分配策略采用一种负载感知的策略[9],每一个路由器均将 4 个周期内处理的包片数通过专用信号线(如图 1 所示 Load_in 信号)作为负载信息传递给相邻路

由器,当前路由器根据 4 个输出方向的负载信息进行输出端口分配可以有效避免网络拥塞,实现负载均衡。输出端口根据与其相连的路由器负载信息进行优先级排序,负载最小的端口优先级最高。交换分配器根据输入及输出优先级排序的结果进行输出端口分配。如图 3 所示,对于同一个包片,如果有 2 个最短路径输出端口可用,则从中选择一个负载较小的端口分配给该包片;如果只有一个最短路径输出端口可用,则选择该端口;如果最短路径输出端口均被高优先级包片占用,则从剩余空闲端口中选择一个负载较小的端口对该包片进行偏转。所有来自相邻路由器的包片输出端口分配完毕后,若有空闲端口才允许本地节点向网络中注入包片并进行输出端口分配。交换分配器按优先级逐个分配的串行结构使其成为基准无缓冲路由器的关键路径,降低了路由器实现的时钟频率,因此有必要对其进行优化设计以提高无缓冲路由器的性能。

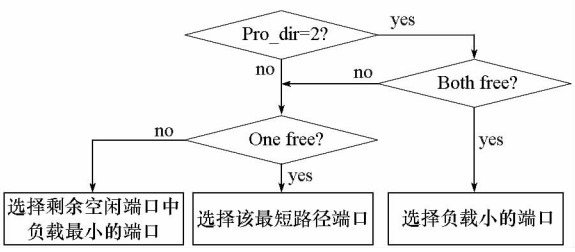


图 3 基准无缓冲路由器交换分配策略

Fig. 3 Switch allocation scheme for baseline bufferless router

2 流水线无缓冲路由器

交换分配器的复杂逻辑降低了基准无缓冲路由器的性能。对无缓冲路由器采用流水线优化设计缩短关键路径长度可以有效提高其时钟频率。为此,我们将基准无缓冲路由器分为 3 级流水线实现,其结构如图 4 所示。输入与输出优先级选择可以并行执行,放在第一级流水站中实现。输入包片经过路由计算后,如果发现有包片到达目标节点,则可以通过第一级流水站中的 4 选 1 多路选择器输出到局部端口。如果同时有多个包片到达目标节点,则选择优先级最高的包片排出到本地节点。此外,第一级流水站中还包含负载计算模块,用于计算当前路由器在最近 4 个周期内处理的包数。交换分配器被进一步分为两个部分,分别放置于第二、三级流水站中。在第二级流水站中的交换分配器,根据第一级流水站产生的输入、输出优先级选择结果以及路由计算结果,对输入包片进行输出端口分配。如果输入包片数少于 4,则允许本地节点向网络注入包片,同时进行路由计算。此外,为了进一步减少第三级流水站的逻辑级数,我们将包头更新放在第二级流水站中执行。由于此时最终的输出端口分配还未完成,因此,需要对输入包片的包头进行前瞻更新,即同时产生 4 个输出方向的包头更新结果。由于到达包片排出在第一级流水站中进行,因此,基准无缓冲路由器中的 5 ×5 交叉开关中的一个 4 选 1 多路选择器被移至第一级流水站中,第三级流水站仅包含一个 5 ×4 交叉开关。第二级交换分配器根据第一级的分配结果以及本地输入包片的路由计算结果产生 5 ×4 交叉开关的选择信号,选择包片从对应的输出端口输出。

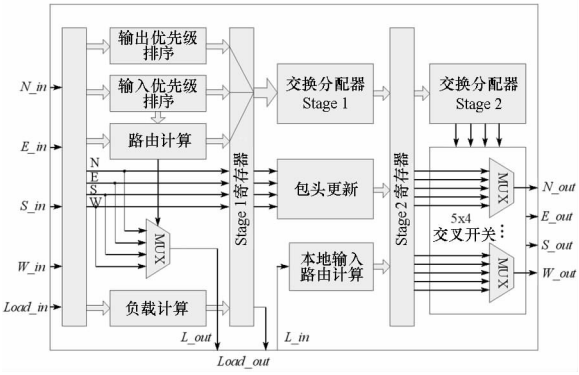


图 4 三级流水线无缓冲路由器

Fig. 4 Bufferless router with 3 pipeline stages

3 采用置换网络的单周期高性能无缓冲路由器

虽然流水线无缓冲路由器可以有效提高基准无缓冲路由器的时钟频率,但是,流水线也增加了包片的单跳步延时。用于间接网络(indirect network)中的置换网络[10]经过修改后可以有效应用于无缓冲路由器中以替代原有的串行化交换分配器与交叉开关。为此,采用一个两级置换网络替换基准无缓冲路由器中的交换分配器与交叉开关对其进行优化。与同样采用置换网络的 CHIPPER 无缓冲路由器[6]的不同点在于,CHIPPER 路由器仍然采用 2 级流水线实现,并且使用一种复杂的 Golden packet 规则避免活锁,而本文提出的路由器采用单周期实现,并对置换单元(permutation cell)进行了修改,在包片置换的同时可以有效避免活锁。此外,输入与输出优先级排序模块从原路由器中移除以进一步减少逻辑级数。

图 5 所示为采用置换网络的单周期无缓冲路由器结构。包片排出器(flit ejector)首先判断输入包片是否到达目标,如果有包片到达目标,则产

生 4 选 1 多路选择器的选择信号。包片排出器同时还确定本地节点是否可以向网络注入包片,如果输入包片数小于 4,则本地节点可以通过置换网络的一个空闲输入端口注入包片。置换网络主要用于间接网络中,本文采用的置换网络包含 2 级,每一级包含 2 个 2×2 的置换单元,如图 5 所示。传统的置换单元通常采用阻塞的方式处理竞争,本文对置换单元进行修改,如果两个包片同时竞争一个最短路径端口,优先级高的包片通过最短路径端口输出,而另一个包片则被偏转。置换网络中输出端口顺序(S, N, W, E)与输入端口顺序(N, E, S, W)的对应关系可以有效避免输入包片在发生偏转时从输入端口返回。第一级 2×2 置换单元的实现算法如图 6 所示。置换单元首先根据包片中的目标地址(d_addr)计算出两个输入包片所需的最短路径端口,然后根据包片的跳步数进行置换选择。具有较大跳步数的包片(高优先级)可以选择最短路径端口,而另一个包片则从另一个端口输出。两级置换网络保证了最高优先级的包片总是可以从最短路径端口输出,因此可以有效避免活锁。

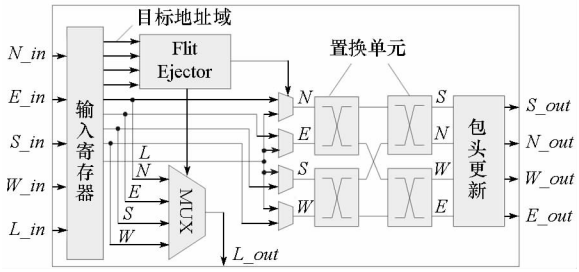


图 5　采用置换网络的单周期无缓冲路由器

Fig. 5　1-cycle bufferless router with a permutation network

```
Permutation_cell
Input: in0, in1(Flit type)
Output: out0,out1 (Flit type)
1: {d_productive0} ←get_productive_dir(int0.d_addr)
2: {d_productive1} ←get_productive_dir(int1.d_addr)
3: if in0.HC >= in1.HC then
4:    if d_productive0 = North or d_productive0 = South then
5:       out0←in0; out1←in1;
6:    else
7:       out0←in1; out1←in0;
8:    end if
9: else
10:   if d_productive1 = North or d_productive 1 = South then
11:      out0←in1; out1←in0;
12:   else
13:      out0←in0; out1←in1;
14:   end if
15: end if
```

图 6　置换单元实现算法

Fig. 6　Algorithm for permutation cell

4　实验结果与分析

本文采用 VHDL 语言设计实现了基准无缓冲路由器(BLESS_noPL)、三级流水线无缓冲路由器(BLESS_PL)以及采用置换网络的无缓冲路由器(BLESS_PERM)。为了与有缓冲路由器进行比较,我们也实现了一个四级流水线虚通道路由器(VC)[10]。在本节中,我们分别采用合成通信负载以及真实应用负载对无缓冲与有缓冲路由器的性能进行评估,并在 TSMC 65nm 工艺下对路由器进行综合以对比各个路由器的硬件实现开销。

4.1　实验配置

模拟实验采用一个周期精确的 NoC 模拟器在一个 8×8 的二维 Mesh 上执行。每一个包包含 4 个包片,每一个包片的长度为 128 位。在每一个资源节点的网络接口中均包含一个 FIFO 队列用于缓存由于无缓冲路由器没有空闲端口或虚通道路由器没有空闲虚通道而暂时不能注入网络的包片。无缓冲路由器在网络接口中还包含一个重排序缓冲对乱序到达的包片进行重排序。在模拟中,我们对合成通信负载与真实应用负载下所需的重排序缓冲项数进行了评估。

对于合成通信负载,本文使用了均匀随机(uniform random)、置换(transpose)与位补(bit complement)三种通信模式。对于均匀随机通信,每一个资源节点以相同的概率向其他节点发送包。在置换通信中,位于(x, y)(其中$x \neq y$)的源节点向位于(y, x)的目标节点发送包。对于位补通信模式,源节点号为$\{s_i \mid i \in [0,5]\}$的源节点向节点号为$\{\neg s_i \mid i \in [0,5]\}$的目标节点发送包。为了模拟真实应用,我们采用了 Splash-2 基准测试程序[11]的运行踪迹。该运行踪迹通过一个全系统模拟器 Simics[12]运行产生。全系统的配置如表 1 所示,该多核系统包含 64 个处理器,通过一个 8×8 二维 Mesh 网络互连。

表 1　真实应用运行踪迹产生的全系统配置

Tab. 1　Full system configuration for trace generation

系统特征	配置
处理器数	64
指令集	SPARC
一级 Cache	32K - I/D, 4 路组相联, 64B/line
二级 Cache	共享 S - NUCA 结构, 512KB/bank, 64 bank, 64B/line, 8 路组相联
一致性协议	MOESI_CMP_directory
存储控制器	8 个片上存储控制器
Splash-2 应用程序	barnes, cholesky, fft, fmm, lu, radix, raytrace, water

在实验中,我们评测了网络吞吐率以及包平均延迟。网络吞吐率(*TP*)定义为网络达到饱和所能接收的最大通信量,单位为:flits/cycle/node,由式(1)计算得出。其中,$N_{total_received_flits}$为网络中所有节点所接收到的包片总数;N_{nodes}为网络中的节点数;$T_{measure_time}$为测量时间。吞吐率为 1 表示每一个节点在每一个周期可以接收一个包片。

$$TP = \frac{N_{total_received_flits}}{N_{nodes} \cdot T_{measure_time}} \quad (1)$$

包延迟由式(2)表示,单位为:ns。其中,$T_{pkt_gen_time}$表示包产生的时钟周期时间,$T_{lastflit_arrival_time}$表示该包的最后一个包片到达目标节点的时钟周期时间。该延迟包含了包在源队列中的等待时间以及网络发送时间。为了进行公平比较,我们在评估中考虑了各个路由器在 TSMC 65nm 工艺下的时钟周期(T_{clk_cycle})。

$$T = (T_{lastflit_arrival_time} - T_{pkt_gen_time}) \cdot T_{clk_cycle} \quad (2)$$

4.2 合成通信负载实验结果

图 7 所示为 3 种合成通信负载下,8 × 8 二维 Mesh 网络配置 4 种路由器的吞吐率。在均匀随机通信模式下,使用 VC 与 BLESS_PL 路由器的网络吞吐率最大,比使用 BLESS_noPL 与 BLESS_PERM 路由器的网络分别高 9% 和 24%。在置换与位补通信模式中,使用无缓冲路由器的网络吞吐率要高于使用虚通道路由器,这是由于偏转路由的完全自适应特性带来的路径多样化的优势,而 VC 路由器采用确定性 XY 路由算法在这两种通信模式下容易造成网络拥塞,使吞吐率降低。此外,在 BLESS_noPL 与 BLESS_PL 路由器中,使用了负载感知的输出端口分配策略,因此,网络吞吐率要高于使用 BLESS_PERM 路由器的网络。

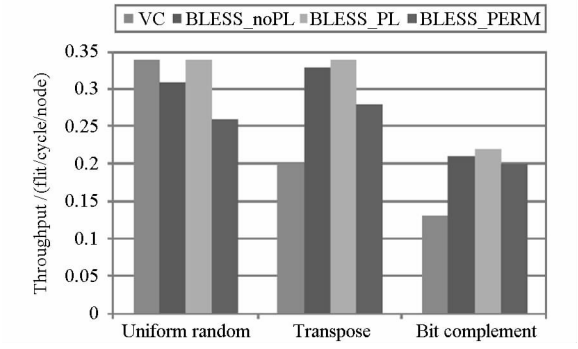


图 7 合成通信负载下的吞吐率

Fig. 7 Throughput with synthetic workloads

图 8(a) ~ (c)所示为 4 个路由器在 3 种合成通信负载下的包平均延迟。在网络未达到饱和前,BLESS_PERM 路由器具有最小的包平均延迟。在均匀随机通信模式下,BLESS_PERM 路由器的包平均延迟比 VC、BLESS_noPL 以及 BLESS_PL 路由器分别少 66%,64% 和 51%。在置换通信模式下,BLESS_PERM 路由器的包平均延迟比 VC、BLESS_noPL 以及 BLESS_PL 路由器分别少 73%,69% 和 61%。在位补通信模式下,BLESS_PERM 路由器的包平均延迟比 VC、BLESS_noPL 以及 BLESS_PL 路由器分别少 73%,67% 和 62%。在模拟中,我们发现三种合成通信负载下无缓冲路由器所需的重排序缓冲项数不超过 10,这相对于虚通道路由器中的大量缓冲器而言,只需要较小的面积开销。

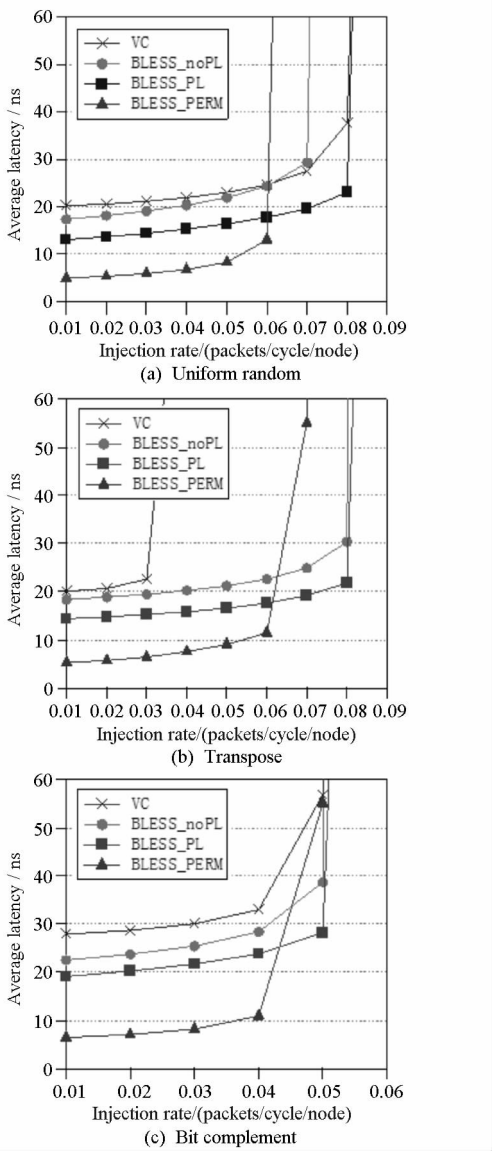


图 8 合成通信负载下的包平均延迟

Fig. 8 Average latency with synthetic workloads

4.3 真实应用负载实验结果

图 9 所示为 4 个路由器在 Splash – 2 应用程序运行踪迹下的包平均延迟。虽然在合成通信负载下,使用 BLESS_PERM 路由器网络吞吐率要低于其他路由器的网络吞吐率,但是真实应用运行

踪迹下的模拟结果表明,在真实应用中,网络通常运行于较低的包注入率模式下。因此,BLESS_PERM 路由器的高性能具有很大优势。在真实应用负载下,BLESS_PERM 路由器的包平均延迟比 VC、BLESS_noPL、BLESS_PL 路由器分别少 80%,72% 和 66%。模拟结果显示,Splash-2 应用下无缓冲路由器所需的重排序缓冲项数不超过 5。

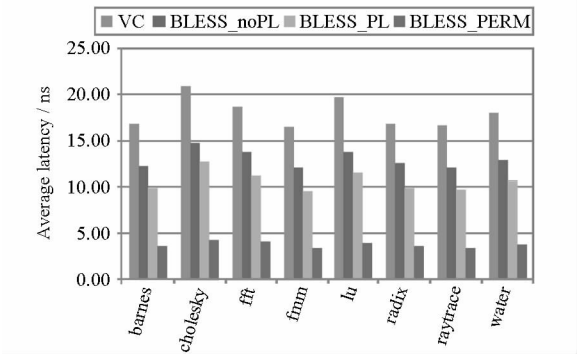


图 9　真实应用通信负载下的包平均延迟

Fig. 9　Average latency with application workloads

4.4　硬件实现开销

本文在 TSMC 65nm 工艺的典型情况(1.0v,25℃)下使用 Synopsys design compiler 对 4 个路由器进行综合。其中,虚通道路由器包含路由计算、虚通道分配、交换分配以及交叉开关传输 4 级流水站,路由器每一个输入端口包含 4 个虚通道,每一个虚通道包含 4 个缓冲。各个路由器时序、面积与功耗的综合结果如表 2 所示。表 2 同时也列出了 CHIPPER 无缓冲路由器[6]的时序和面积结果。从表中结果可以看出,本文提出无缓冲路由器的面积均远小于虚通道路由器。与基准无缓冲路由器相比,3 级流水线无缓冲路由器可以实现较高的性能而仅需要 13% 的额外面积开销,而采用置换网络的无缓冲路由器面积比基准无缓冲路由器少 49%。与同样采用置换网络的 CHIPPER 无缓冲路由器相比,本文提出的采用置换网络的无缓冲路由器可以达到 2GHz 的时钟频率,并且具有较小的面积和功耗。

表 2　有缓冲与无缓冲路由器硬件实现开销对比

Tab. 2　Hardware cost comparison for buffered and bufferless routers

	关键路径延时 (ns)	面积 (μm^2)	功耗 (mW/GHz)
VC	0.7	370023	118.4
CHIPPER[6]	1.9	306165	N/A
BLESS_noPL	1.8	36660	13.4
BLESS_PL	0.7	41345	18.3
BLESS_PERM	0.5	18695	9.1

5　结束语

本文提出了一种单周期无缓冲路由器,可以在 TSMC 65nm 工艺下达到 2GHz 的高性能。该路由器使用一个简单的置换网络替代原有的串行化交换分配器以及交叉开关以缩短路由器关键路径的长度并且能够有效避免活锁。综合结果表明,VC 与 BLESS_PL 路由器分别采用三级、四级流水线达到 1.4GHz 的频率,而 BLESS_PERM 路由器单周期即可实现 2GHz 的频率,其面积远小于 VC 与 CHIPPER 路由器,比 BLESS_noPL 与 BLESS_PL 路由器分别少 49% 和 55%,并且具有较低的功耗,适用于高性能单芯片多处理器的片上互连。合成通信负载下的模拟结果表明,BLESS_PERM 路由器的包平均延迟比 VC、BLESS_noPL 及 BLESS_PL 路由器分别少 71%,67% 和 58%;在真实应用负载下,BLESS_PERM 路由器的包平均延迟比 VC、BLESS_noPL 及 BLESS_PL 路由器分别少 80%,72% 和 66%。

参 考 文 献:

[1] Benini L, Micheli G D. Networks on Chips: a New SoC Paradigm[J]. IEEE Computer, 2002, 35(1): 70-78.

[2] Gratz P, Kim C, McDonald R, et al. Implementation and Evaluation of On-chip Network Architectures[C]//Proceedings of International Conference on Computer Design, 2006: 477-484.

[3] Gomez C, Gomez M E, Lopez P, et al. Reducing Packet Dropping in a Bufferless NoC[C]//Proceedings of the 14th International Euro-Par Conference on Parallel Processing, 2008: 899-909.

[4] Hayenga M, Jerger N E, Lipasti M. SCARAB: a Single Cycle Adaptive Routing and Bufferless Network[C]//Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture, 2009: 244-254.

[5] Moscibroda T, Mutlu O. A Case for Bufferless Routing in On-chip Networks [C]//Proceedings of the 36th Annual International Symposium on Computer Architecture, 2009: 196-207.

[6] Fallin C, Craik C, Mutlu O. CHIPPER: a Low-complexity Bufferless Deflection Router [C]//Proceedings of the 17th IEEE International Symposium on High Performance Computer Architecture, 2011: 144-155.

[7] Lu Z, Zhong M, Jantsch A. Evaluation of On-chip Networks Using Deflection Routing[C]//Proceedings of the 16th ACM Great Lakes Symposium on VLSI, 2006: 363-368.

[8] Millberg M, Nilsson E, Thid R, et al. The Nostrum Backbone—a Communication Protocol Stack for Networks on Chip[C]//Proceedings of the 17th International Conference on VLSI Design, 2004: 693-696.

[9] Nilsson E, Millberg M, Oberg J, et al. Load Distribution with the Proximity Congestion Awareness in a Network on Chip [C]//Proceedings of the Conference on Design, Automation and Test in Europe, 2003: 1126-1127.

[10] Dally W J, Towles B. Principles and Practices of Interconnection Network[M]. San Francisco: Morgan Kaufmann Publishers, 2003: 305-321.

[11] Woo S C, Ohara M, Torrie E, et al. The Splash-2 Programs: Characterization and Methodological Considerations [C]//Proceedings of the 22nd Annual International Symposium on Computer Architecture, 1995: 24-36.

[12] Magnusson P S, Christensson M, Eskilson J, et al. Simics: a Full System Simulation Platform[J]. IEEE Computer, 2002, 35(2): 50-58.