# *K*-ary *N*-cube 网络中的维度气泡流控与无死锁完全自适应路由

肖灿文 张民选 过 锋

(国防科学技术大学计算机学院 长沙 410073)

**摘 要** 利用虚跨步切换技术中消息的依存关系只与相邻缓冲区队列相关的特点,设计了一种称为维度气泡流控(DBFC)的新型流控策略.该流控策略建立在虚跨步(VCT)切换和信约流控机制之上,通过分析端口信约值和路由信息实现点点间的流控.在无边带 $k$-ary $n$-cube 网络中,如果采用 DBFC 流控策略,即使网络中存在环相关,设计的自适应维度气泡路由(ADBR)算法仍可实现无死锁的最短距离的路由.对于以上结论,文中提供了详细的证明.最后,通过修改模拟工具 RSIM 的网络模拟器——NETSIM 的代码,实现了 DBFC 流控策略和 ADBR 算法.模拟结果显示,ADBR 算法在性能上比常用的维序路由优越,在报文延迟上有近 17.5% 的降低.

**关键词** 基于信约的流控;死锁;无边带 $k$-ary $n$-cube 网络;虚跨步切换

**中图法分类号** TP302

# DBFC: Dimensional Bubble Flow Control with Deadlock-Free and Fully Adaptive Routing in the *K*-ary *N*-cube Network

XIAO Can-Wen ZHANG Min-Xuan GUO Feng

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

**Abstract** Focusing on the particular characteristics of virtual cut-through switching network, the authors find that message dependencies are localized between adjacent queues. Using this characteristic, the authors design the novel flow control strategy called dimensional bubble flow control (DBFC). The flow control strategy of DBFC builds on virtual cut-through switching and credit-based flow control mechanism and analyzes the credit value of port and the routing information of the packets to realize the point-point flow control. In the $k$-ary $n$-cube network without wraparound connections, when the flow control strategy of DBFC is accepted, the adaptive dimensional bubble routing (ADBR) algorithm designed in this paper can get the goals including deadlock-free and minimal distance even if the cyclic dependencies exist. In this paper, the detail proof is provided for these conclusions. Lastly, the authors adapt the source code of NETSIM that is a simulator of network and realize the flow control of DBFC and ADBR algorithm in NETSIM. The authors adopt the actual application programs to test the performance of ADBR on RSIM. The simulation performance shows the preposed scheme is superior to the approach of usual dimension-order routing, with nearly 17.5% reduction in the packets latency.

**Keywords** credit-based flow control; deadlock; $k$-ary $n$-cube network without wraparound connections; virtual cut-through switching

**收稿日期**:2005-12-13;**修改稿收到日期**:2006-01-19.**本课题得到国家自然科学基金**(90307001)**资助**.**肖灿文**,男,1970 年生,副教授,主要研究方向包括计算机系统结构、计算机互连网络系统研究以及 ASIC 芯片设计. E-mail: c_w_xiao@sohu.com. **张民选**,男,1954 年生,教授,博士生导师,主要研究领域包括高性能计算机系统结构和微电子技术. **过 锋**,男,1977 年生,博士研究生,主要研究方向包括计算机系统结构和计算机互连网络系统.

## 1 引 言

目前,大规模并行处理(MPP)系统仍然是构造高端超级计算机的主要形式. 当处理器个数达到上万个之后,MPP 系统的性能极大地依赖于互连网络子系统的性能. 而互连网络子系统的设计包括网络拓扑、流控策略、路由算法和路由器芯片等方面.

对于网络拓扑,一些研究[1,2]显示了低维网络的优越性,如 $k$-ary $n$-cube 网络,它具有拓扑结构规整、结点度低和在单芯片上容易实现等优点. Ring, Mesh, Torus 和 Hypercube 就是这种典型的 $k$-ary $n$-cube 网络. 许多并行计算机系统采用此种类型的拓扑结构,像 Intel Option-Red(3D Mesh), SGI/Cray T3D/T3E(3D Torus)和 BlueGene/L(3D Torus). 图 1(a)给出了一个 16 个结点的 Hypercube. 图 1(b)则是一个 3-ary 2-cube 或 2D Torus. 把 $k$-ary $n$-cube 中环绕通道去除可以实现无边带 $k$-ary $n$-cube 网络. 图 1(c)显示了一个由 3-ary 3-cube 去除环绕通道后构成的 3-ary 3-cube(3D) Mesh.

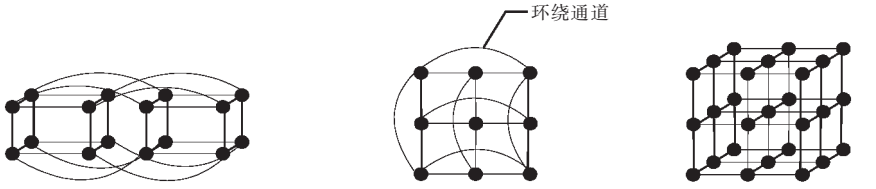


图 1 不同的网络拓扑

路由芯片是互连网络子系统的核心部分,其通用结构如图 2 所示. 每个路由器通过直接的链路与其相邻的路由结点相连. 计算结点和路由器之间有两条链路,一条为注入链路,用于计算结点发送消息;另一条为消耗链路,用于接收来自路由器的消息.

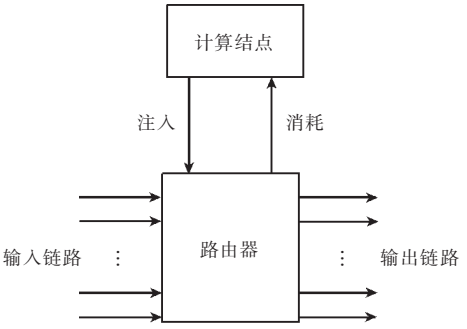


图 2 一种通用的结点结构

大规模互连网络设计的一个关键问题是如何消除死锁. 文献[3]提出的维序路由和转弯模型(turn model)算法严格规定路由规则,防止阻塞的报文形成环相关,然而它们要容忍更长的等待延迟. 在文献[4,5]中,当检测到死锁时,通过从环形等待队列中移除一个或多个消息来消除死锁. 文献[6]介绍了一种避免死锁的方法,该方法实现了一个完全自适应通道和一个逃逸虚通道,Cray T3E 中的自适应路由采用了这种方法. 以上几种避免死锁的方法的缺点是较高的硬件复杂性和缓冲空间的不均衡使用.

在文献[7,8]中,Puente 首先提出了一种"气泡"流控策略,用来消除 VCT 网络中由于环绕链路导致的死锁. 不需要额外的虚通道,在任何 $k$-ary $n$-cube 网络中,气泡流控机制支持在同一维中无死锁的自适应路由. 如果报文从一个维流向另一个维,则它们不得不遵循维序路由,而不是自适应路由. 气泡流控未能解决 $k$-ary $n$-cube 网络中消息在不同维间路由时的死锁问题.

在本篇文章中,我们提出了一种新型的流控策略和自适应路由算法来避免消息在 $k$-ary $n$-cube 网络的不同维间路由时形成死锁.

本文的主要贡献如下:

我们提出了一种称为 DBFC(Dimensional Bubble Flow Control)的流控方式,该方式在 $k$-ary $n$-cube 网络中可以避免自适应路由算法在不同维间形成的死锁;并介绍了一种基于 DBFC 的 $n$ 维的自适应路由算法 ADBR(Adaptive Dimensional Bubble Routing). 在一个被称为 RSIM 的多处理器模拟器中,通过两个基准测试程序 LU 和 FFT 模拟了该自适应路由算法. 模拟结果显示,我们提出的路由算法在性能上比常用的维序路由优越,在报文延迟上有近 17.5%的降低.

本文第 2 节介绍 DBFC 流控策略以及一种被称为 ADBR 的完全自适应路由算法,同时在 2D 网格中证明了 DBFC 方式的无死锁特性;第 3 节则是定量的性能评价;最后,第 4 节对本文作了总结.

## 2　维度气泡流控策略(DBFC)

### 2.1　基本概念

在并行计算机的互连网络子系统中，流控策略决定了报文的流动和停止. 流控通常基于信约，信约是下一级路由器中空余缓冲的大小，相邻路由器使用信约通知彼此的空余缓冲大小. 当报文停止时，它们缓存在路由器的缓冲中，消耗信约. 按照流控单位的不同，有两种常用的切换技术：VCT 切换和虫孔切换. 这两种切换技术都以流水方式工作，其中虫孔切换可以获得最低的消息延迟，而 VCT 切换则可以达到最大的吞吐率. 然而，在虫孔切换中，路由器缓冲中的缓存单位是 1 个微片(flit)，当一个报文被阻塞时，它的微片占据了多个路由器的缓冲器. 而在 VCT 切换中，缓存的单位是一个报文，大小是 10 个以上微片.

路由算法决定报文的行走路径，为其建立了一条从源端到目的端的通路. 路由算法有两种类型，确定性路由和自适应路由. 确定性路由在消息发送出去之前就为其规定了行走路径，故算法比较简单，但是带宽利用率较低. 自适应路由则意味着消息在运行期间的行走路径是不确定的，可以根据路由器的动态工作负载调整. 相比于确定性路由，自适应路由有较高的通道资源利用率和较低的网络阻塞可能性. 然而，自适应路由算法必须解决网络死锁问题. 当一个消息的拥有者没有获得下一级的空余缓冲空间，而拥有者本身又不释放所占据的资源时，死锁就形成了. 环相关是死锁形成的必要条件.

下面我们介绍一种新型的流控策略和自适应路由算法，可以解决消息在 VCT 切换的 $k$-ary $n$-cube 网络不同维间路由时引起的死锁问题.

在 $k$-ary $n$-cube 网络中，根据 VCT 切换机制，报文从一个路由器的缓冲移动到相邻路由器的缓冲. 所有报文在释放它们占有的资源之前，必须获得要求的下一级缓冲空间. 因此，如何正确地调度缓冲空间，避免死锁十分关键. 出于对这种问题的考虑，我们采用基于信约的流控[9]，并利用报文的路由信息实现了一种称为 DBFC 的新型流控，DBFC 可以解决消息在 $k$-ary $n$-cube 网络的不同维间路由时引起的死锁问题. 文献[7,8]已经介绍了在 $k$-ary $n$-cube 网络的同一维中避免死锁的方法，我们工作的重点主要放在 $k$-ary $n$-cube 网络的不同维间存在的死锁问题上. 而对于无边带连接的 $k$-ary $n$-cube 网络(称为 NW $k$-ary $n$-cube 网络)，死锁只会出现在不同维间，因此我们的研究对象是 NW $k$-ary $n$-cube网络.

下面给出 DBFC 流控策略的定义：

NW $k$-ary $n$-cube网络中，当一个报文在$N(N\leq n)$个方向上剩有路由步时，报文进入下一级缓冲的条件是，下一级缓冲必须有不少于 $N$ 个报文大小的空余空间；否则，报文等待.

按照 DBFC 流控策略，在两个维上拥有路由步的报文能够进入下一级缓冲，仅当下一级缓冲至少有 2 个空余的报文空间. 根据报文信约机制，每个路由器包含有相邻路由器缓冲空间大小的信息，故 DBFC 流控策略用在分布式路由器中，而不采用集中式调度.

基于 DBFC，我们设计了一个称为 ADBR 的完全自适应路由算法. 首先，根据源路由可知，一个报文在每一维上行走所必须的最少跨步数是确定的. 而最少的跨步数保证了通路上没有活锁存在. 其次，报文在不同方向上请求目标缓冲，这些方向上的跨步数不为零. 请求成功之后，选择多个成功的请求中的一个，报文沿着该请求的方向流出，同时该方向上的跨步数减 1. 报文继续向前移动直至所有方向上的跨步数为 0. 最终，报文被目的结点吸收.

图 3 是 ADBR 路由算法描述. 输入参数包括报文的路由场和目标缓冲剩余空间的大小. 报文路由场的表达式是$(\Delta Z_1, \Delta Z_2, \cdots, \Delta Z_n)$，其中 $\Delta Z_i$ 是报文在第 $i$ 维剩余的跨步数. 目标缓冲剩余空间大小的表达式为$(N_1, N_2, \cdots, N_n)$，其中 $N_i$ 表示第 $i$ 维的缓冲还可存放 $N_i$ 个报文. Internal 表示报文将被计算结点吸收. $E$ 是报文可以选择进入的一组维度通道的集合. $i+$ 表示第$i$维的正方向，$i-$ 表示第$i$维

```
ADBR 路由算法.
输入：报文路由场(ΔZ1，ΔZ2，…，ΔZn)和目标
      缓冲剩余空间(N1，N2，…，Nn)
输出：被选中的输出通道
过程：
    k := 0；E := { }；
    For i := 1 to n do
      if (ΔZi ≠ 0) then
        k := k+1；
      endif
    end
    if (k == 0) Channel := Internal；
    else
      For i := 1 to n do
        if (ΔZi ≠ 0 and Ni ≥ k) then
          if (ΔZi > 0) E := E+{i+}；
          else E := E+{i-}；
        endif
      end
    Channel := Select(E)；
```

图 3　ADBR 路由算法

的负方向. Select 函数从可选通道集 $E$ 中选择任一通道. ADBR 算法使报文总是沿着最短距离的路径行走,不存在自身的环,不会出现活锁. 如果报文总是向前流动,它们将最终到达目的结点. 这种情况下,网络无死锁.

### 2.2 ADBR 算法无死锁性的证明

下面,证明 ADBR 算法是无死锁的. 不失一般性,分析 1 维和 2 维网络的情况.

证明.

(1)当维度为 1 时,NW $k$-ary $n$-cube 网络成为如图 4 所示的线性结构.


图 4 无边带连接的线性结构

从图 4 可以看出,由于网络拓扑是线性结构,不存在环,报文路由也是单向的,因此结论显然成立,ADBR 算法无死锁.

(2)当维度为 2 时,NW $k$-ary $n$-cube 网络的拓扑结构如图 5 所示.

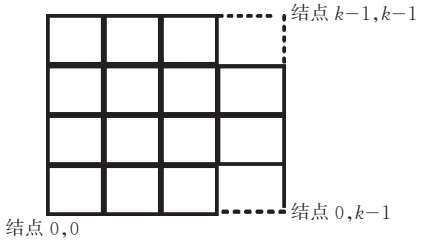


图 5 2D Mesh 结构

路由器的结构如图 6 所示. 路由器使用输入缓冲,报文被缓存在路由器的输入端口. 路由器包括 4 个用于链路连接的双向端口,1 个注入口和 1 个消耗口.

路由器中,沿 $X+$, $X-$, $Y+$, $Y-$和注入方向有5个缓冲,每个缓冲中的所有报文按照它们的路

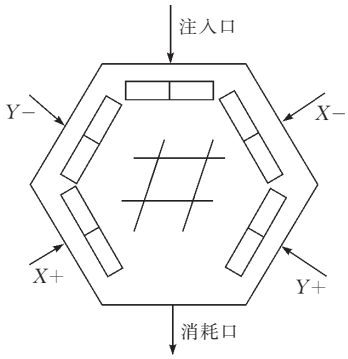


图 6 路由器结构图

由信息预备参加仲裁. 为了支持自适应路由算法和避免出现 FIFO 结构中,由于队列第一个报文无法流出,导致所有报文阻塞的情况出现,缓冲按以下模式组织:当一个报文进入缓冲,一个新的指针被用来指向报文头,使得缓冲中的每个报文都可以根据各自的路由信息申请仲裁. 缓冲的组织类似于 DAMQ[9].

通过分析报文在缓冲中的所有可能情况,证明在任何情况下报文都能到达目的结点,从而可得出网络中无死锁的结论.

在 2D Mesh 中,缓冲内的报文有以下 4 种情况:

(1)缓冲中的报文是正等待被结点消耗的报文. 此类报文不可能永远阻塞其它类型的报文,因为它们不久就会被消耗掉.

(2)缓冲中的报文仅有一个方向的路由,并且路由方向与所在缓冲方向一致. 我们使用图解方法来分析.

假设:

①报文路由方向为 $X+$;

②一个缓冲器可以容纳 2 个报文.

报文在路由器中的情况如图 7 所示.

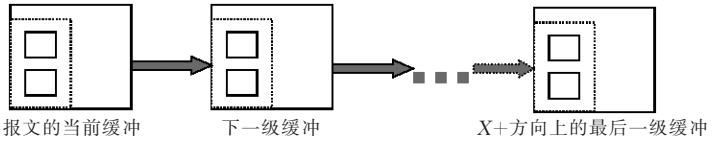


图 7 $X+$方向的缓冲

对于下一级缓冲,存在两种可能的情况:① 如果下一级缓冲有一个报文的剩余空间,按照 DBFC 流控策略,报文可以进入下一级缓冲. ② 如果下一级缓冲没有剩余空间,则可以得出这样的结论:下一级缓冲中至少有一个报文的路由方向仅剩 $X+$或为第 1 种情况的报文. 假设下一级缓冲的两个报文剩余的跨步含两个方向或有一个不同于 $X+$方向,则根据 DBFC 流控策略,一个拥有两个跨步方向的报文进入下一级缓冲器的必要条件是:目标缓冲必须至少含有两个报文空间的大小. 因此,以上假设不成立. 故可以认为,下一级缓冲至少有一个报文的路由仅剩 $X+$方向或为第 1 种情况的报文. 依此类推,可以知道后面缓冲区中的报文情况都类似.

由于 $X$ 方向和 $Y$ 方向都不存在环形连接,故在 $X$ 方向或 $Y$ 方向上不可能形成环相关. $X+$方向最后一级缓冲中的报文不可能有 $X+$方向的跨步,因

此这个缓冲区中肯定有报文已完成路由等待吸收，即属于第 1 类报文. 所以，这些只在 $X+$ 方向上有路由步的报文总可以流动，终将到达目标结点. 总之，第 2 类报文将进入下一级缓冲存储区，最终将到达目标结点.

(3)缓冲中的报文仅有一个路由方向，且该方向与所在缓冲方向不一致. 当目标缓冲没有剩余缓冲空间时，分析目标缓冲中报文的可能情况. 假设占有目标缓冲的报文是第 1 种情况或第 2 种情况的报文，由于这两种情况的报文总能向前流动，它们不会永久阻塞第 3 种情况的报文. 假设目标缓冲中的报文有 $X$ 和 $Y$ 两个方向的路由或者属于第 3 种情况，则根据 DBFC，此种报文不可能占据所有的缓冲空间，它们进入目标缓冲后，应该剩余一个报文空间. 因此，第 3 种情况的报文可以进入目标缓冲. 所以，第 3 种情况的报文总能到达目的结点.

(4)缓冲中的报文还有 $X$ 和 $Y$ 两个方向的路由. 根据 ADBR 算法，报文会同时向 $X$ 和 $Y$ 两个方向申请仲裁. 我们考虑那些与所在缓冲方向相一致的报文的情况. 假设当前缓冲方向为 $X+$，如图 8 所示.

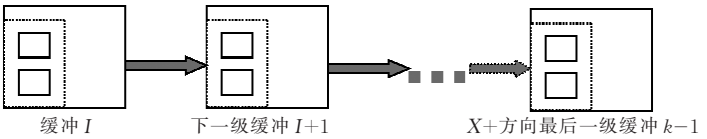


图 8　$X+$ 方向缓冲

根据前面的分析，如果占据下一级缓冲的报文都是前面所述的 3 种情况的报文，它们不会导致其它情况的报文永久阻塞，故死锁仅可能在最后一种情况的报文之间产生. 根据 ADBR 算法，可以确定的是，$X+$ 方向上最后一级缓冲中的报文至多有一个方向的路由，因为它们在 $X+$ 方向不可能再有路由了，这种报文就是前面 3 种情况的报文. 所以，$X+$ 方向缓冲存储区中的第 4 类报文总是可以流动的. 最终，第 4 类报文将走完 $X+$ 方向的路由，成为第 3 种类型的报文. 所以，第 4 类报文也不会永久阻塞，最终将到达目标结点.

综上所述，维数为 2 的情况下，ADBR 算法不会产生死锁. 类似地，可以证明，当维数为 $N$ 时($N$ 是大于 2 的自然数)，ADBR 算法不会产生死锁.　　证毕.

## 3　ADBR 算法性能分析

我们在 RSIM 模拟器中实现了 ADBR 算法，通过执行 SPLASH-2 套件中的应用程序来评价算法的性能，并同其它常用路由算法进行了比较.

RSIM(Rice Simulator for ILP Multiprocessors)[10]是由 Rice 大学开发的一款共享存储多处理机模拟器. RSIM 中的互连网络子系统模拟器称为 NETSIM，它模拟了一个 2D Mesh 的互连网络，该网络采用维序路由和虫孔切换，通过修改参数，可以实现 VCT 切换. NETSIM 中有两套网络，分别为请求网络和响应网络，其中请求网络中流动的是请求报文，而响应网络中流动的是响应报文.

FFT 和 LU 是由斯坦福大学开发的 SPLASH2[11] 套件中的两个并行应用程序. 我们在三种路由算法下运行这两个应用程序，这三种路由算法包括带 VCT 切换和虫孔切换的维序路由算法以及带 VCT 切换的 ADBR 算法. 当处理器数目非常大时，并行系统的模拟非常耗时，因此我们把处理器的个数限制在 64 个以内.

通过设置处理器个数为 32 和 64，缓冲分别为 10，20，30，40，50 和 60 个微片大小，可以获得不同配置下的性能值. 图 9～图 12 是这些模拟的最终结果. $Y$ 轴是以 CPU 周期为单位的运行时间，$X$ 轴是针对不同处理器数以微片为单位的缓冲大小. 在 RSIM 中，一个报文由 10 个微片组成，因此 VCT 切换的流控单位是 10 个微片. 对于 ADBR 算法，输入缓冲至少要设成 2 个报文大小，即 20 个微片. 以下模拟结果中，响应网络时间是程序执行完时消息在响应网络中延迟时间总计. 请求网络时间类似.

### 3.1　LU 模拟结果

从 LU 的模拟结果可以得出以下结论：

在相同缓冲大小的情况下，ADBR 算法的网络延迟要低于维序路由算法的网络延迟，不论后者采用 VCT 切换还是虫孔切换. 同时，当缓冲为 20 个微片大小时，网络延迟达到最低点，再增加缓冲影响不大.

### 3.2　FFT 模拟结果

从 FFT 的模拟结果图中可以看出：

虽然没有在 LU 中明显，ADBR 算法的网络延迟还是低于维序路由算法的网络延迟. 同时，当缓冲空间从 20 片到 30 个片增加时，网络延迟也相应减少，当处理器个数为 64 时尤为明显. 再增加更多的缓冲空间收益就甚微了.

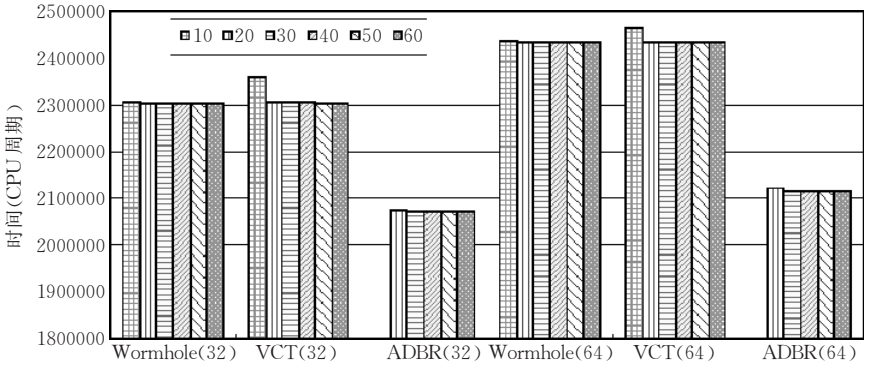


**图 9 LU 响应网络时间**

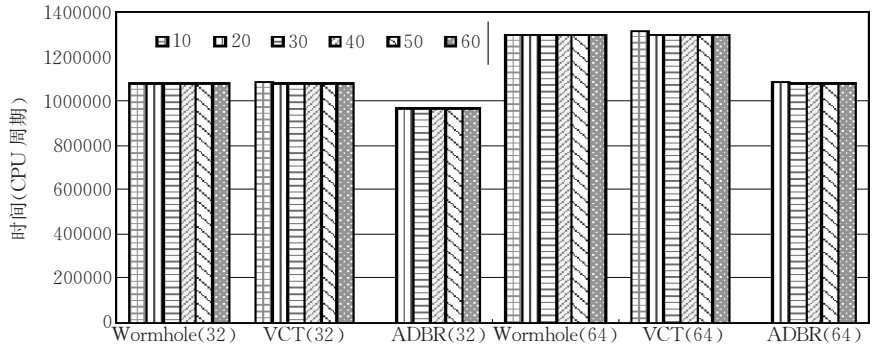


**图 10 LU 请求网络时间**

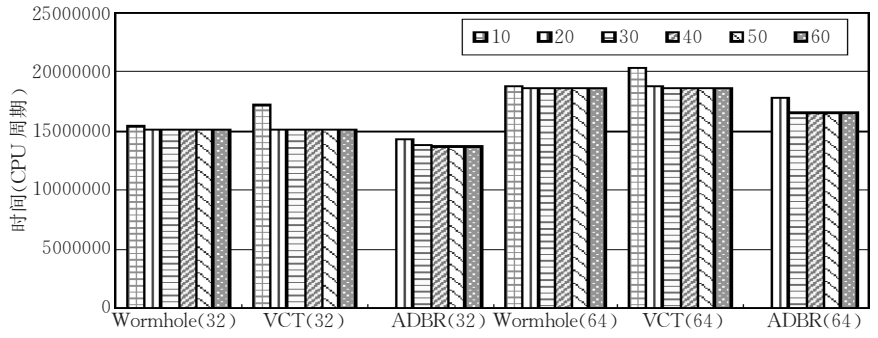


**图 11 FFT 响应网络时间**

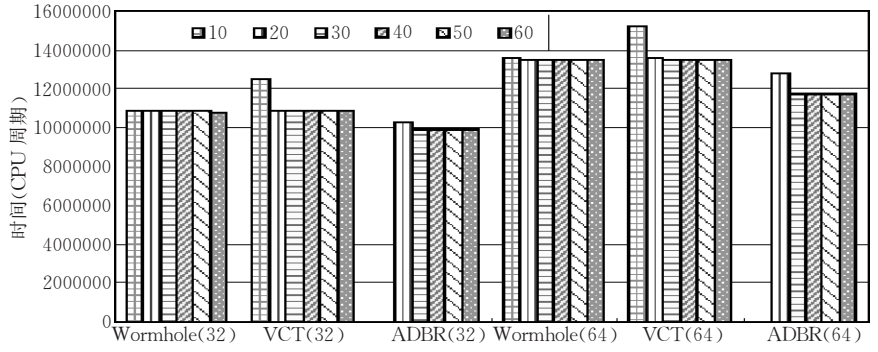


**图 12 FFT 请求网络时间**

## 4 结 论

本文中,针对无边带 $k$-ary $n$-cube 网络,我们设计了一个简单的 DBFC 流控策略和 ADBR 路由算法,不用考虑虚通道就可以实现无死锁的完全自适应路由. ADBR 算法需要一个较小的缓冲空间,所有类型的报文均可共享此缓冲空间. ADBR 算法充

分利用了结点度，降低了网络系统的阻塞时间. 模拟结果显示，ADBR 算法拥有出色的性能.

从工程实现的角度来看，该方法基于规则的 *k*-ary *n*-cube 拓扑、VCT 切换和基于信约的 DBFC 流控. DBFC 除了使用信约流控外，仅增加了路由信息部分，而在实现中并不需要特殊技术的支持. 在 DBFC 流控策略中，缓冲大小由网络维数决定. 总之，实现 DBFC 流控和 ADBR 路由的所有技术都是目前已经使用了的技术，最终的实现较为容易. 因此，ADBR 路由算法提供了一种实现高性能互连网络的新方法.

## 参 考 文 献


1 Borkar S. *et al*. Iwarp: An integrated solution to high-speed parallel computing. In: Proceedings of the Supercomputing'88, Florida, 1988, 330～339

2 Agarwal A.. Limits on interconnection network performance. IEEE Transactions on Parallel Distributed Systems, 1991, 2(4): 398～412

3 Ni L., Glass C.. The turn model for adaptive routing. Journal of the ACM, 1994, 41(5): 874～902

4 Martinez Rubio J. M., López P., Duato J.. Fc3d: Flow control-based distributed deadlock detection mechanism for true fully adaptive routing in wormhole networks. IEEE Transactions on Parallel and Distributed Systems, 2003, 14(8): 765～779

5 Song Y. H., Pinkston T. M.. A progressive approach to handling message-dependent deadlock in parallel computer systems. IEEE Transactions on Parallel Distributed System, 2003, 14(3): 259～275

6 Duato J.. A new theory of deadlock-free adaptive routing in wormhole networks. IEEE Transactions on Parallel and Distributed Systems, 1993, 12(4): 1320～1331

7 Puente V., Izu C., Beivide R.. The adaptive bubble router. Journal of Parallel and Distributed Computing, 2001, 61(9): 1180～1208

8 Puente V., Gregorio. On the design of a high-performance adaptive router for cc-numa multiprocessors. IEEE Transactions on Parallel and Distributed Systems, 2003, 14(5): 487～501

9 Laudon J., Lenoski D.. The SGI origin: A ccnuma highly scalable server. In: Proceedings of the ISCA, Colorado, 1997, 241～251

10 Pai V. S., Ranganathan P., Adve S. V.. RSIM reference manual(Version 1.0). Department of Electrical and Computer Engineering, Rice University: Technical Report: 9705, 1997

11 Woo S. C., Ohara M., Torrie E., Singh J. P., Gupta A.. The SPLASH-2 programs: Characterization and methodological considerations. In: Proceedings of the 22nd International Symposium on Computer Architecture, Italy, 1995, 24～36



**XIAO Can-Wen**, born in 1970, associate professor. His current research interests include the computer architecture, the interconnect network and the design of ASIC.

**ZHANG Min-Xuan**, born in 1954, professor and Ph. D. supervisor. His current research interests include high-performance computer architecture and micro-electronics.

**GUO Feng**, born in 1977, Ph. D. candidate. His current research interests include the computer architecture and the interconnect network.


**Background**

This work is partly supported by the National Natural Science Foundation of China under the project No. 90307001. The title is "The Research of Reconfigure-Array Architecture Based on Fixed Instruction and Multiple Data-flow".

In the research of FIMD model, the authors proposed a novel reconfigurable computing structure, called LEAP, Loop Engine on Array Processors. In the reconfigurable array, many of reconfigurable processing units(RPU) are connected by a mesh network. The network is the critical component of the array, which provides channels for data transfer between different RPUs. As the structure of the array change for different computation, the network must provide a non-blocking channel between two RPUs which have explicit data dependency. The adaptive routing algorithm for such on-chip network plays an important role in this research area. In this paper, the authors present a dimensional bubble flow control and adaptive dimensional bubble routing algorithm which provide a new way to implement high performance on-chip network.