ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.2, February 2009, pp.415−424
doi: 10.3724/SP.J.1001.2009.03168

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

H-Torus拓扑结构等分带宽的计算*

乐祖晖+, 赵有健, 吴建平, 张小平

(清华大学 计算机科学与技术系,北京 100084)

Calculation on the Bisection Width of H-Torus Topology

YUE Zu-Hui+, ZHAO You-Jian, WU Jian-Ping, ZHANG Xiao-Ping

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

+ Corresponding author: E-mail: yuezuhui@chinamobile.com

Yue ZH, Zhao YJ, Wu JP, Zhang XP. Calculation on the bisection width of H-Torus topology. *Journal of Software*, 2009,20(2):415−424. http://www.jos.org.cn/1000-9825/3168.htm

Abstract: Two methods are presented to calculate the lower bound and upper bound on the bisection width of H-Torus topology. These two methods can also be applied to the 2D Torus topology. A method is presented to calculate the exact bisection width for H-Torus too. But this method has unacceptable complexity and can only be accepted with small scale. It is shown that H-Torus topology has larger bisection width. Regarding precision, the lower bound and upper bound introduced in this paper are greatly improved. This result strongly supports the design of scalable routers.

Key words: H-Torus; 2D Torus; bisection width; direct network; scalable router

摘 要: 针对 H-Torus 拓扑结构,给出两种确定该拓扑结构等分带宽上、下界的方法.这些方法同样适用于 2D Torus拓扑结构.还提出了 H-Torus结构等分带宽的精确求解方案,但是该算法的复杂度过大,只适用于网络规模较小的情况.实验表明,H-Torus 拓扑结构的等分带宽大于同等规模的 2D Torus 结构,更有利于提高路由器的吞吐率.与现有的研究结果相比,所提出的等分带宽上、下界在精度上有了较大的提高,这为可扩展路由器的性能评估提供了有力的支持.

关键词: H-Torus;2D Torus;等分带宽;直连网络;可扩展路由器

中图法分类号: TP393 **文献标识码**: A

路由器和链路将 Internet 中原本孤立的端用户互连起来,分组在由源端用户传输到目的端用户的过程中,可能需要经过多台路由器.一个典型路由器主要由线卡和交换网络两部分组成.分组由输入端口到达线卡,线卡根据分组中包含的目的地址信息在路由表中进行查找,得到相应的输出线卡和输出端口编号,同时对分组内容作出修改.分组在进入交换网络之前通常需要进行分片处理,即切割成长度相等的数据单元(通常称为信元,本文中如果没有特别说明,提及的分组均指信元);交换网络将分组由输入线卡转发到输出线卡;在离开线卡前,分组通常还需要进行重组处理.

* Supported by the National Natural Science Foundation of China under Grant No.90604029 (国家自然科学基金); the National Basic Research Program of China under Grant No.2003CB314801 (国家重点基础研究发展计划(973))

Received 2006-12-25; Accepted 2007-09-06

Internet发展迅速,网络用户对网络容量的需求(通过对用户流量进行测量)按照每年大约提升1倍的速度增长[1],而商用路由器的容量增长速度只是接近摩尔定律,即每18个月增加1倍[2].为了满足不断增长的用户需求,路由器需要提供更大的容量,也就意味着路由器需要支持更多的端口数目或更高的端口速率.但是,很多因素制约了端口速率的增加[3].因此,路由器的研究重点正由单纯提高端口速率逐渐转移到交换网络如何支持更多的端口数目,即设计可扩展的交换网络上.目前,绝大多数的商用路由器都采用集中式交换网络.这类交换网络很难在扩展性、扩展粒度和调度算法实现复杂度等方面寻找到一个平衡点.例如,共享缓存和共享总线结构的扩展性受到带宽的限制;MIN类交换网络的扩展性较好,但是扩展粒度大,调度算法复杂;负载均衡结构[4]虽然具有调度算法简单和扩展粒度低的优点,但是全连接的要求却限制了其扩展性;而对于crossbar结构,则存在扩展性差和调度算法复杂的缺点.

直连网络最初主要应用于处理器和存储器之间的互连,或是I/O端口间的互连;后来,这一技术又成功应用于可扩展路由器的设计(传统的分组交换问题转换为直连网络内部的分组路由问题).其中,基于3D Torus(又称为k-ary 3-cube)结构[5]的交换网络在Avici公司的TSR路由器中得到应用[6].TSR中的每个线卡作为该拓扑结构的一个节点,在源节点和目的节点之间存在多条可选路径.这一设计具有下述优点:扩展粒度小,规模大;有利于实现负载均衡;具有较高的容错性.虽然3D Torus拓扑结构本身具有较好的扩展性,但是在TSR具体实现时,当线卡数目达到320[6]后,整个交换网络的等分带宽不再增加,为了提供所需的加速比,再增加线卡就会降低整个路由器的性能,因此限制了TSR的扩展规模.H-Torus[7]结构表现出更好的拓扑属性,并且在具体部署时能够保障等分带宽随着节点数目的增加而增加.

网络直径和等分带宽是评价直连网络交换性能的两个重要指标.如果将直连网络N的节点集合V划分为两个子集合V_1和V_2,并且满足下述条件,则需要去除一些边,称为对网络N进行等分:1) $|V_1 \cap V_2|=0$;2) $V_1 \cup V_2 = V$;3) $||V_1|-|V_2||\le 1$.等分带宽表示所有等分中所需切除边数的最小值,记为B_N.等分带宽直接关系到拓扑结构的吞吐性能,因此对于路由器这类强调吞吐率的设备就显得更为重要.路由器针对注入流量不提供反压机制,因此路由器必须支持任意类型的注入流量.以等分类型流量为例,由V_1中节点注入的流量目的节点均属于V_2,由V_2中节点注入的流量目的节点均属于V_1,此时所有的流量均经过等分截面.显然,等分带宽是该类流量下的交换瓶颈.

但是,求解任意拓扑结构的等分带宽属于NP-complete问题[8],即便是针对d-regular图(图中所有节点的度均为d,且$d>0$)[9].可以将等分带宽问题与矩阵的特征值问题联系起来,定义直连网络N的拉普拉斯变换矩阵$L(N)=\{l_{v,w}\}$[9]为

$$l_{v,w}=\begin{cases}\deg(v), & \text{if } v=w\\ -1, & \text{if } v\ne w \text{ and } (v,w)\in E\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

矩阵$L(N)$的$|V|$个特征值记为$\lambda_1,\lambda_2,\ldots,\lambda_{|V|}$,且满足$\lambda_1\le\lambda_2\le\ldots\le\lambda_{|V|}$.由此得出的等分带宽下界为$\lambda_2|V|/4$[9].文献[9]指出,在特定条件下,可以得到一个更精确的下界$\lambda_2^{\beta}|V|$,$1/2\le\beta<1$.

对于任意给定的拓扑结构,如果选择一个等分切割方式,则显然根据定义该切割得到的结果就是等分带宽的一个上界.文献[10]针对循环图给出了在特定等分下等分带宽上界,但是在H-Torus结构中,这个上界的误差太大.因此,在求解特定拓扑结构等分带宽的上界时,需要充分结合拓扑结构本身的特点,选择特定的切割方式,以降低上界与精确解之间的差距.

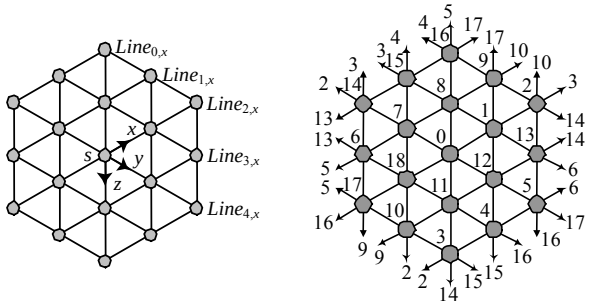
本文第1节在讨论H-Torus拓扑结构的基础上,介绍一种称为LNF的最短路径路由算法.第2节给出精确求解H-Torus结构等分带宽的方法,同时给出H-Torus结构等分带宽上、下界的求解方法.第3节将本文的结论同现有研究成果进行比较,用实验来评测上、下界在精度上的改进,并在较小规模下与精确解进行比较.在同等规模下,H-Torus结构比2D Torus结构具有更高的等分带宽.第4节对全文进行总结,指出下一步的研究工作.

1 H-Torus结构简介

H-Torus 结构是 H-Mesh[11]结构的变体,本节将详细讨论该结构的一些拓扑特性.

1.1 H-Mesh和H-Torus结构

H-Mesh 结构如图 1(a)所示,该结构的外围节点位于同一个正六边形上.在 H-Mesh 结构中,外围节点的度小于 6,所以该拓扑结构属于非规则图.文献[11]针对 H-Mesh 结构,对外围节点进行了统一处理,称为连续类型的连接(continuous type wrapping).如图 1(b)所示,在 H-Mesh 结构中引入了 x,y 和 z 坐标轴,在 3 个方向上分别定义第 0 行,第 1 行,…,第 $2t-2$ 行(t 表示 H-Mesh 结构的圈数).每个方向上,第 i 行的尾节点和第 $j(j=(i+t-1) \bmod (2t-1))$ 行的首节点相连.由文献[11]可知,经过这样的处理,所得拓扑结构为自同构图,该拓扑结构记为 H-Torus.为了方便讨论,定义只包含 1 个节点的 H-Mesh 结构的圈数为 1,简记为 HM_1,相应的 H-Torus 结构简记为 HT_1,…;圈数为 t 的 H-Mesh 结构简记为 HM_t,相应的 H-Torus 结构简记为 HT_t.HT_t 中节点数目$|V|=3t^2-3t+1$[7],边数目$|E|=3|V|$[7].



(a) H-Mesh (b) H-Torus

Fig.1 H-Mesh and H-Torus topologies

图 1 H-Mesh 和 H-Torus 拓扑结构

在 HT_t 中建立如图 1(b)所示的 3 轴坐标系,分别沿 x,y 和 z 轴方向对节点 v_1 和 v_2 进行编号,所得坐标记为(x_1,y_1,z_1)和(x_2,y_2,z_2).记符号$[x]_y \equiv x \bmod y$,其中 x 为整数,y 为正整数.由文献[11]可知,HT_t 中满足下述关系式:$[x_2-x_1]_{|V|}=[(3t^2-6t+3)(y_2-y_1)]_{|V|}$,$[x_2-x_1]_{|V|}=[(3t^2-6t+2)(z_2-z_1)]_{|V|}$.

如果沿 x 轴方向对所有节点进行编号(如图 1(b)所示),则该编号可以唯一标识 H-Torus 结构中各个节点,由此不难得出节点 0 的 6 个邻居分别为 $1,3t-2,3t-1,3t^2-6t+2,3t^2-6t+3$ 和 $3t^2-3t$.当 $t=2$ 时,满足 $3t^2-3t>3t-2>3t^2-6t+3$;当 $t>2$ 时,$3t-1<3t^2-6t+2,3t^2-6t+3<3t^2-3t$.所以,当 $t=2$ 时,定义邻接矢量:$L=(1,3t^2-6t+2,3t^2-6t+3,3t-2,3t-1,3t^2-3t)$;当 $t>2$ 时,定义邻接矢量:$L=(1,3t-2,3t-1,3t^2-6t+2,3t^2-6t+3,3t^2-3t)$.因此,对于 $HT_t(t>2)$中任意给定节点 $i(0\le i<|V|)$,其 6 个邻居分别为$[i+1]_{|V|}$,$[i+3t-2]_{|V|}$,$[i+3t-1]_{|V|}$,$[i+3t^2-6t+2]_{|V|}$,$[i+3t^2-6t+3]_{|V|}$,$[i+3t^2-3t]_{|V|}$.所以,HT_t 属于自同构图和循环图,同时还是 6-regular 图;HT_t 还满足节点对称和边对称.图 2 是 HT_3 的另一种组织方式.

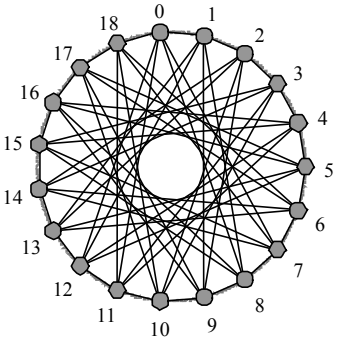


Fig.2 A redrawn HT_3

图 2 HT_3 的另一种组织方式

显然,HT_t 的邻接矩阵属于循环矩阵,该矩阵第 1 行中邻接矢量对应的列,元素值为 1,其余各项均为 0;第 $i(1<i\le|V|)$行元素的值可以由第 $i-1$ 行元素循环右移 1 列得到.

1.2 H-Torus结构中的LNF算法

1.2.1 相对坐标的建立

因为 H-Torus 结构为自同构图,所以其中任意节点都可视为对应 H-Mesh 结构的中心节点.如图 3 所示,以节点 s 为中心,定义 6 条轴线:$axis_{s,0}$,$axis_{s,1}$,…,$axis_{s,5}$;另外,定义 6 个区域:$region_{s,0}$,$region_{s,1}$,…,$region_{s,5}$.显然,s 的 6 个

邻居节点 $n_0,n_1,\ldots,n_5$ 分别位于 $axis_{s,0},axis_{s,1},\ldots,axis_{s,5}$ 上.除 s 外的任意节点 d 必然位于某条轴线 $axis_{s,i}(i\in\{0,1,\ldots,5\})$上或某个区域 $region_{s,j}(j\in\{0,1,\ldots,5\})$中.例如,图 3 中的 d_1 位于 $region_{s,4}$ 中,d_2 位于 $axis_{s,3}$ 上.

实际系统中,任意给定节点 s 除了记录自己的编号以外,还需要记录其余各个节点 d 相对于 s 的位置信息:$(D_{d,s},pos_{d,s})$,其中 $D_{d,s}$ 表示 d 到 s 的距离,$pos_{d,s}$ 表示 d 相对节点 s 所处的轴线或区域.例如,图 3 中,节点 s 记录的 d_1 的位置信息为$(2,region_4)$,d_2 的位置信息为$(2,axis_3)$.

对于 $HT_t(t>2)$中任意给定节点 s,其余的节点可以分为两类:s 的邻接节点和非邻接节点.邻接节点的 ID 可以根据邻接矢量得到,$n_0=[s+1]_{|V|}$,$n_1=[s+3t^2-6t+3]_{|V|}$,$n_2=[s+3t^2-6t+2]_{|V|}$,$n_3=[s+3t^2-3t]_{|V|}$,$n_4=[s+3t-2]_{|V|}$,$n_5=[s+3t-1]_{|V|}$.如图 4 所示,按箭头所示方向,可以得到沿途各个节点的 ID,而且可以得到各个节点相对节点 s 的位置信息和距离信息.

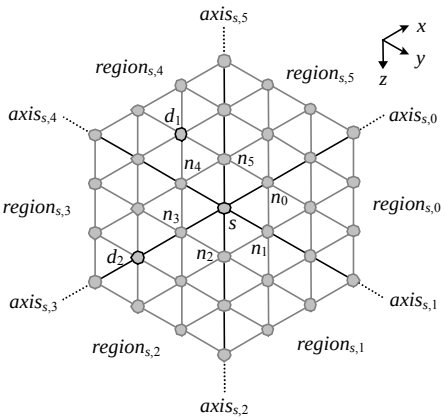


Fig.3 Division of topology based on node s

图 3 基于节点 s 的拓扑分割

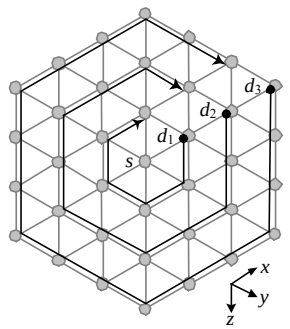


Fig.4 Computing the information of all nodes based on node s

图 4 基于节点 s 计算所有节点的信息

1.2.2 LNF 路由算法

假设分组由节点 s 注入,目的节点为 d,例如,图 3 中的节点对(s,d_1).节点 s 上记录的关于节点 d_1 的信息为$(d_1,2,region_{s,4})$,记节点 n_4 为节点对(s,d_1)的左邻居(left neighbor,简称 LN),节点 n_5 为节点对(s,d_1)的右邻居(right neighbor ,简称 RN).节点对(s,d_2)的左邻居与右邻居重合,即为节点 n_3.如果在分组的路由过程中始终选择当前节点和目的节点的左邻居,则记为左邻居优先最短路径路由算法,简记为 LNF 算法.

2 计算H-Torus结构的等分带宽

2.1 H-Torus结构等分带宽的精确求解

根据等分带宽的定义,必须遍历所有等分方案,从中选择最小值.在 HT_t 中,节点数目$|V|=3t^2-3t+1=3t(t-1)+1$,所以$|V|$为奇数.从$|V|$个节点中选择$(|V|-1)/2$ 个节点作为节点集合 V_1,剩下的$(|V|+1)/2$ 个节点作为节点集合 V_2,所以总的等分方案数为 $C_{|V|}^{(|V|-1)/2}$.将 V_1 中的节点按序号从小到大的次序排列,得到 $v_1v_2\ldots v_{(|V|-1)/2}(v_1<v_2<\ldots<v_{(|V|-1)/2})$,由于 HT_t 结构属于同构图,所以这里只需考虑 $v_1=0$ 的情况,其余情况都可以由该情况求解,所以总的方案数目降为 $C_{|V|-1}^{(|V|-3)/2}$.

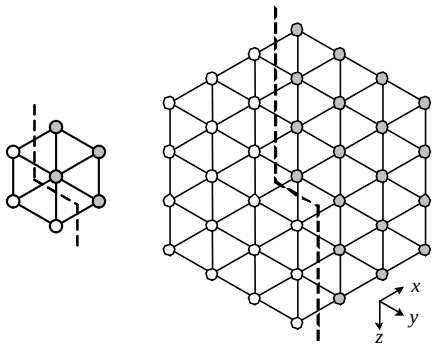
得到子集 V_1 和 V_2 后,源节点属于 V_1 的边共计 $6\cdot(|V|-1)/2=3\cdot(|V|-1)$条,设目的节点属于 V_1 的边的数目为 M,则目的节点属于 V_2 的边的数目为 $3\cdot(|V|-1)-M$,根据等分带宽的定义可知,该值恰好对应该等分所切除的边数.求解 M 的时间复杂度为 $\frac{(|V|-1)}{2}\cdot\frac{(|V|-3)}{2}=\frac{(|V|-1)(|V|-3)}{4}$.所以,总的算法复杂度为

$$\frac{(|V|-1)(|V|-3)}{4}C_{|V|-1}^{(|V|-3)/2}=\frac{(|V|-1)(|V|-3)}{4}\cdot\frac{(|V|-1)!}{\left(\frac{|V|-3}{2}\right)!\left(\frac{|V|+1}{2}\right)!} \tag{2}$$

取 t=2,得$|V|$=7,对应的时间复杂度为 90;取 t=3,得$|V|$=19,对应的时间复杂度为 3 150 576;取 t=4,得$|V|$=37,对应的时间复杂度为 2 630 833 959 600.不难看出,当 t 较大时,精确求解 HT_t 的等分带宽是不可行的.

2.2 利用等分法求解H-Torus结构等分带宽的上界

文献[7]给出了 H-Torus 的一种等分方案,如图 5 所示(忽略了外围节点的连接).该等分方案所切除的边数记为 HT_t 结构等分带宽的上界:B_u=10t–10(t>3).这种求解等分带宽上界的方法称为等分法.

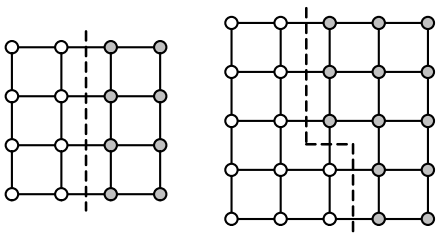


(a) t=2 (b) t=4

Fig.5 A method to bisect HT_t

图 5 HT_t 的一种等分方法

2D Mesh 和 2D Torus 结构是目前可扩展路由器[6]设计中常选的拓扑结构,这里选取规模均为 n·n 的 2D Mesh 和 2D Torus 进行分析.若 n 为奇数,利用等分法,等分方案如图 6(a)所示;若 n 为偶数,等分方案如图 6(b)所示.不难得出,当 n 为奇数时 2D Mesh 和 2D Torus 的等分带宽上界分别为 n+1 和 2·(n+1);当 n 为偶数时,2D Mesh 和 2D Torus 的等分带宽上界分别为 n 和 2n.类似地,由图 6 容易得到 3D Mesh 和 3D Torus 结构按照这种切割方式得到的等分带宽上界.选取规模为 n·n·n 的 3D Mesh 和 3D Torus 结构:若 n 为奇数,则 3D Mesh 和 3D Torus 的等分带宽上界分别为 n·(n+1)和 2n·(n+1);若为偶数,则 3D Mesh 和 3D Torus 的等分带宽上界分别为 n^2 和 2n^2.



(a) n=4 (b) n=5

Fig.6 A method to bisect 2D Mesh and 2D Torus topologies

图 6 2D Mesh 和 2D Torus 拓扑结构的一种等分方法

2.3 利用边负载法求解H-Torus结构等分带宽的下界

在包含$|V|$个节点的直连网络中,如果在任意给定源节点和目的节点之间只选择一条路由,则共有$|V|$·($|V|$–1)条路由.假设经过边 e 的路由数目为 R_e,令 $R_{\max}=\max_{e\in E}(R_e)$. 假设在某个等分切割方式下得到节点集合 V_1 和 V_2,且切断的边数恰好等于等分带宽 B_N,则至少有 2·$|V_1|$·$|V_2|$条路由被切断(可能存在源节点和目的节点同在 V_1 或 V_2 中的路由被切断).因此,可以得到:B_N·$R_{\max}$≥2·$|V_1|$·$|V_2|$,即 B_N≥(2·$|V_1|$·$|V_2|$)/$R_{\max}$.这种求解等分带宽下界的方法称为边负载法.

在 $HT_t(t>1)$中,任意一点 $v_i(0\le v_i\le|V|-1)$都可以视为对应 H-Mesh 结构的中心点.取 LNF 算法作为路由算法,则由节点 v_i 出发,去往其余各个节点 $v_j(v_i\ne v_j)$的情况可以分为 $t-1$ 组:v_j 位于第 2 圈,…,v_j 位于第 t 圈;每组节点的数目分别为 6,…,$6\cdot(t-1)$;由 v_i 去往每组节点的路由长度分别为 1,…,$(t-1)$.因此,由节点 v_i 出发去往其余各个节点的路由长度之和为

$$\sum_{j\ne i,0\le i,j\le|V|-1} r_{ij}=6\cdot1\cdot1+6\cdot2\cdot2+\ldots+6\cdot(t-1)\cdot(t-1)=6\cdot(1^2+2^2+\ldots+(t-1)^2)=t(t-1)(2t-1) \tag{3}$$

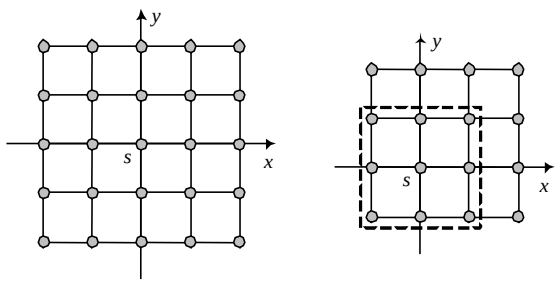
HT_t 结构满足节点对称和边对称,因此,$|V|(|V|-1)$条路由的长度之和为$|V|*t(t-1)(2t-1)$.总的链路数目为 $3|V|$,因此每条链路上平均经过的路由数目均为$|V|\cdot t(t-1)(2t-1)/(3|V|)=t(t-1)(2t-1)/3$.将节点集合等分为 V_1 和 V_2 时,满足 $|V_1|\cdot|V_2|=\frac{|V|-1}{2}\cdot\frac{|V|+1}{2}$.因此,$HT_t(t>1)$的等分带宽下界为

$$\left(2\cdot\frac{|V|-1}{2}\cdot\frac{|V|+1}{2}\right)\Big/\frac{t(t-1)(2t-1)}{3}=\frac{9(3t^2-3t+2)}{2(2t-1)} \tag{4}$$

根据拓扑结构的特点,很容易证明同等规模的 2D Torus 结构比 2D Mesh 结构具有更高的等分带宽.所以,这里只分析 2D Torus 结构的等分带宽下界,同样采用边负载法进行分析.

当 n 为奇数时,如图 7(a)所示,取相应 2D Mesh 的中心点 s,令节点 s 的坐标为(0,0),同时定义以 s 为原点的直角坐标系,则在该坐标系下,节点 $d=(x,y)$和节点 s 之间的距离为$|x|+|y|$.因此,由 s 出发去往其余各个节点的路由长度之和(采用 DOR 路由算法[12])为

$$\sum_{|x|\le(n-1)/2,|y|\le(n-1)/2} r_{sd}=n\cdot\sum_{|x|\le(n-1)/2}|x|+n\cdot\sum_{|y|\le(n-1)/2}|y|=n\cdot(n^2-1)/2=(n^3-n)/2 \tag{5}$$



(a) n=5 (b) n=4

Fig.7 Computation of distances between node s and other nodes

图 7 计算节点 s 到其余各个节点的距离

2D Torus 结构同样满足节点对称和边对称,因此,$n^2(n^2-1)$条路由的长度之和为 $n^3(n^2-1)/2$.总的链路数目为 $2n^2$,则每条链路上经过的路由数目均为 $n(n^2-1)/4$.另有$|V_1|\cdot|V_2|=(n^2-1)\cdot(n^2+1)/4$,因此,所求的等分带宽下界为

$$\left(2\cdot\frac{(n^2-1)(n^2+1)}{4}\right)\Big/\left(\frac{n(n^2-1)}{4}\right)=2n+\frac{2}{n} \tag{6}$$

当 n 为偶数时,如图 7(b)所示,s 取为左下角规模为$(n-1)\cdot(n-1)$的 2D Mesh 的中心点,定义 s 的坐标为(0,0),则该 2D Mesh 中各点到 s 的距离之和为$((n-1)^3-(n-1))/2$.剩下各点到 s 的距离之和为 $n(3n-2)/2$.所以,s 到其余各点的距离之和为 $n^3/2$,则每条链路上经过的路由数目均为 $n^3/4$.另有$|V_1|\cdot|V_2|=n^4/4$,因此,所求等分带宽下界为

$$\frac{2\cdot n^4/4}{n^3/4}=2n \tag{7}$$

由式(6)、式(7)可得:当 n 为奇数时,$B_N\ge2n+1$;当 n 为偶数时,$B_N\ge2n$.根据上一节的讨论可知,当 n 为奇数时,$B_N\le2(n+1)$;当 n 为偶数时,$B_N\le2n$.所以,当 n 为奇数时,$B\approx2(n+1)$;当 n 为偶数时,$B=2n$.

类似地,可以由图 7 得到 3D Torus 结构的等分带宽下界,选取规模为 $n\cdot n\cdot n$ 的 3D Torus 结构,其等分带宽 $B_N=2n^2$.

3 实验仿真

本节将上一节的结论与现有的研究成果进行比较,用实验来评测上、下界在精度上的改进.通过比较可以发现,相同规模下,H-Torus 结构比 2D Torus 结构具有更高的等分带宽.

3.1 与现有结论的比较

3.1.1 H-Torus 拓扑结构的拉普拉斯矩阵

由于 H-Torus 拓扑结构的邻接矩阵为循环矩阵,所以对应的拉普拉斯变换矩阵 $L(N)=\{l_{v,w}\}$也是循环矩阵.且第 1 行(t>2)中有:$l_{0,0}=6, l_{0,1}=-1, l_{0,3t-2}=-1, l_{0,3t-1}=-1, l_{0,3t^2-6t+2}=-1, l_{0,3t^2-6t+3}=-1, l_{0,3t^2-3t}=-1$,其余元素均为 0.

HT_t(t>2)拉普拉斯变换矩阵第 1 行对应的多项式为

$$P(m)=6-m-m^{3t-2}-m^{3t-1}-m^{3t^2-6t+2}-m^{3t^2-6t+3}-m^{3t^2-3t} \tag{8}$$

1 的|V|个根为 $\cos(2k\pi/|V|)+i\cdot\sin(2k\pi/|V|)$($k$=0,1,…,|$V$|−1).代入式(8)得:

$$P(\cos(2k\pi/|V|)+i\cdot\sin(2k\pi/|V|))=6-2\cos(2k\pi/|V|)-2\cos(2k\cdot(3t-2)\pi/|V|)-2\cos(2k\cdot(3t-1)\pi/|V|) \tag{9}$$

根据文献[13]可知,式(9)即为拉普拉斯变换矩阵的特征值,由此易得λ_2,进而可以求解等分带宽的下界值$\lambda_2|V|/4$.

3.1.2 循环图的等分带宽上界

文献[10]给出了基于循环图的一个等分方案,即将序号为 0,1,…,|V|−1 的节点分为两个集合 V_1 和 V_2,并且满足:当|V|为偶数时,V_1={0,1,…,(|V|−2)/2},V_2={|V|/2,(|V|+2)/2,…,|V|−1};当|V|为奇数时,V_1={0,1,…,(|V|−1)/2},V_2={(|V|+1)/2,(|V|+3)/2,…,|V|−1}.由此得到等分带宽的上界[10].基于此,得到 H-Torus 结构的一个等分方案上界为 B_u=2·(|1|+|3t−2|+|3t−1|)=2·(1+3t−2+3t−1)=12t−4.

3.1.3 上、下界的比较

图 8 显示了利用不同方法求解得到的 H-Torus 拓扑结构等分带宽的情况,即利用拉普拉斯方程求解所得的等分带宽下界、利用循环图法求解所得的等分带宽上界,以及本文介绍的等分法求解得到的上界和边对称法求解得到的下界.随着节点数目的增加,拉普拉斯变换法的下界趋近于 20,利用循环图法求解所得的等分带宽上界误差大于等分法求解的上界,利用边对称法求解所得的等分带宽下界的误差则更加趋近于真实值.因此,利用本文介绍的方法可以得到 H-Torus 结构更为精确的等分带宽上、下界,这为路由器的设计和吞吐率分析提供了更强有力的支持.

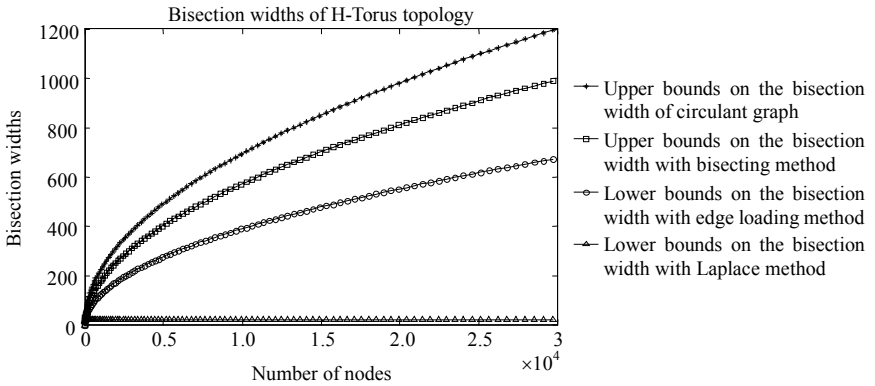


Fig.8 Upper bounds and lower bounds on the bisection width of H-Torus topology

图 8 H-Torus 拓扑结构等分带宽上、下界的比较

3.2 与精确解的比较

从前面的讨论可知,精确求解 H-Torus 结构的等分带宽是一项非常复杂的任务,只有在网络规模很小的情

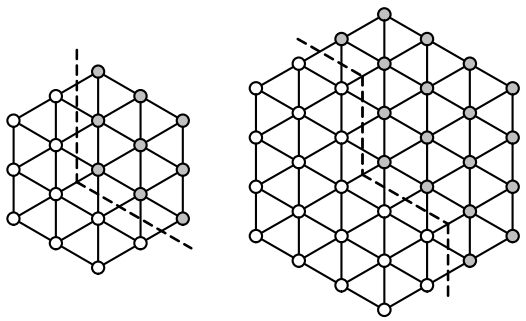
况下才可计算.在求解 H-Torus 结构等分带宽精确解的过程中,一个很重要的组成部分就是组合数的生成算法.

根据文献[14],给定 0,1,…,|*V*|−1 共计|*V*|(这里|*V*|为奇数)个数,从中选择(|*V*|−1)/2 个数构成一个组合 $v_1v_2\ldots v_{(|V|-1)/2}$,最简单的情况是 $v_1=0,v_2=1,\ldots,v_{(|V|-1)/2}=(|V|-3)/2$,可以从该组合得到其余各个组合,具体步骤如下:

1. 求满足不等式 $v_i<(|V|+1)/2+i$ 的最大的下标 i,即 $i=\max\{j|v_j<(|V|+1)/2+j\}$;
2. $v_i\leftarrow v_i+1$;
3. 从 $i+1$ 位开始作修改: $v_j\leftarrow v_{j-1}+1$; $j=i+1,i+2,\ldots,(|V|-1)/2$.

由于 H-Torus 属于自同构图,所以只需考虑 $v_1=0$ 的情况,其余的情况等同于这种情况.

由计算结果可以得到 $t=3$ 和 $t=4$ 的最佳切割方式,如图 9 所示,相应的等分带宽分别为 20 和 28.



(a) $t=3$ (b) $t=4$

Fig.9 Bisection width of HT_t

图 9 HT_t 的等分带宽

$HT_t(t>2)$的等分带宽上界为 $B_u=10t-10$,等分带宽下界为 $B_l=\dfrac{9(3t^2-3t+2)}{2(2t-1)}$.因此,$HT_t$ 等分带宽的近似解为

$$B=(B_u+B_l)/2=10t-10+\frac{9(3t^2-3t+2)}{2(2t-1)}\Big/2=\frac{67t^2-87t+38}{4(2t-1)} \tag{10}$$

图 10 是该近似解和精确解的比较,可以看出,该近似解具有较高的精度.考虑到精确解的计算复杂度,图中只列出了规模较小的比较结果.

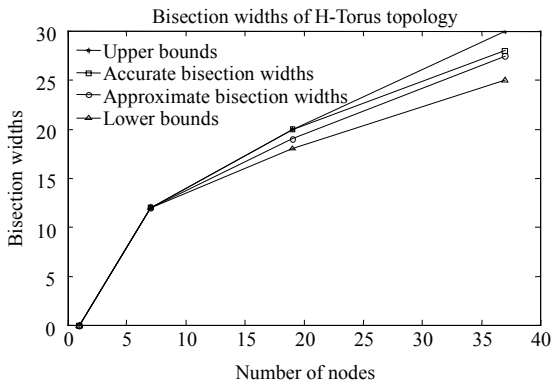


Fig.10 Approximate bisection width of H-Torus topology

图 10 H-Torus 拓扑结构等分带宽近似解

3.3 与2D Torus和3D Torus拓扑结构的比较

在前面的讨论中得出了 2D Torus 结构和 3D Torus 结构的等分带宽.图 11 给出了 2D Torus,3D Torus 和 H-Torus 等分带宽(近似解)的比较.可以看出,在节点数目相同的情况下,3D Torus 具有最高的等分带宽,H-Torus

其次,2D Torus 结构的等分带宽最低.但是,如图 12 所示,在实际部署时,3D Torus 结构受到机械结构的限制,当节点数目达到 320 时[6],再增加线卡也不会改变实际等分带宽.图 12 中的第 1 条水平线出现在线卡由 2×2×2(对应 $x\times y\times z$,下同)变到 2×2×5 时,此时等分带宽维持 8 不变;第 2 条水平线出现在线卡由 5×4×5 变到 8×4×5 时;此时等分带宽维持 40 不变;第 3 条水平线出现在线卡由 8×8×5 变到 14×8×5 时,此时等分带宽维持 80 不变.

因此,从这个意义上说,H-Torus 结构具有更好的实际等分带宽,这将有利于提高路由器中交换网络的吞吐率.

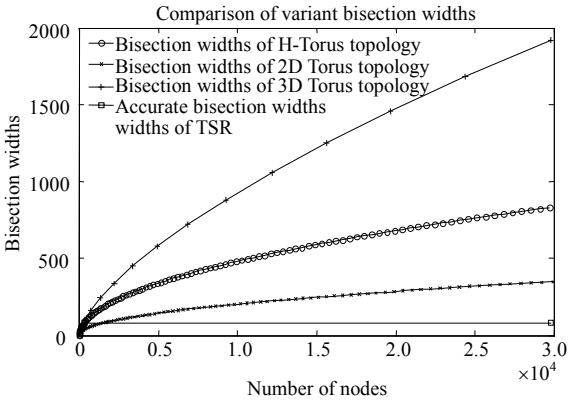


Fig.11 Comparison of variant bisection widths

图 11 各类等分带宽比较

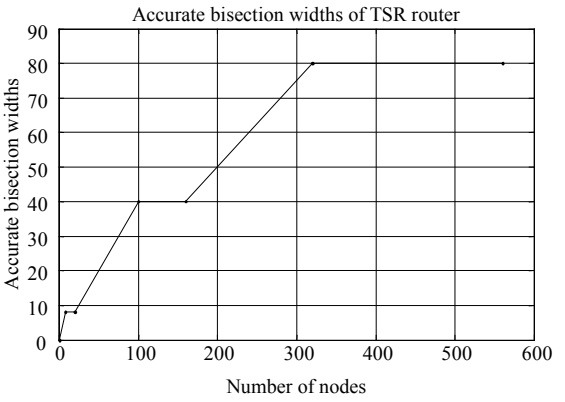


Fig.12 Bisection width of TSR router

图 12 TSR 路由器等分带宽

4 结论和未来的工作

将直连网络引入路由器中交换网络的设计可以有效提升路由器的扩展能力,拓扑结构的等分带宽在很大程度上决定了路由器的整体吞吐率.但是,针对任意拓扑结构(即便是 d-regular 图)求解等分带宽属于 NP-complete 问题.

本文以 H-Torus 结构为研究对象,提出了 LNF 路由算法.分别采用等分法和边对称法求解出 H-Torus 结构等分带宽的上、下界,该方法同样适用于 2D Torus 结构.在节点相同的情况下,H-Torus 比 2D Torus 结构提供了更高的等分带宽,更适合于可扩展路由器的设计.通过实验发现,本文求解得到的上、下界与精确解之间仍然存在一定的误差,说明在上、下界的确定上仍然存在提升空间.

拓扑结构是直连式交换网络的研究基础.在今后的研究中,研究重点将会集中在路由算法和流控方案的设计上.针对 H-Torus 结构,还可以进行组播和 QoS 交换的研究.另外,还可以针对拓扑容错性和不规则拓扑等方面展开研究.

References:

[1] Odlyzko AM. Internet traffic growth: Sources and implications. In: Dingel BB, ed. Proc. of the SPIE Optical Transmission Systems and Equipment for WDM Networking II. Orlando: SPIE, 2003. 1−15.

[2] Keslassy I, Chuang ST, Yu K, Miller D, Horowitz M, Solgaard O, McKeown N. Scaling internet routers using optics. In: Feldmann A, ed. Proc. of the Special Interest Group on Data Communication (SIGCOMM). Karlsruhe: ACM Press, 2003. 189−200.

[3] Chiussi FM, Francini A. Scalable electronic packet switches. IEEE Journal on Selected Areas in Communications, 2003,21(4): 486−500.

[4] Chang CS, Lee DS, Jou YS. Load balanced Birkhoff-von Neumann switches, part I: One-Stage buffering. Computer Communications, 2002,25(6):611−622.

[5] Dally, WJ. Performance analysis of *k*-ary *n*-cube interconnection networks. IEEE Trans. on Computers, 1990,39(6):775−785.

[6] Dally WJ. Scalable switching fabrics for Internet routers. http://www.avici.com/technology/whitepapers/TSRfabric-WhitePaper.pdf

[7] Zhao YJ, Yue ZH, Wu JP, Zhang XP. Topological properties and routing algorithms in cellular router. In: Proc. of the Int'l Conf. on Networking and Services (ICNS 2006). 2006. 101−106.

[8] Garey MR, Johnson DS. Computers and Intractability: A Guide to the Theory of NP-Completeness. San Francisco: W. H. Freeman and Company, 1979.

[9] Bezrukov S, Elsässer R, Monien B, Preis R, Tillich JP. New spectral lower bounds on the bisection width of graphs. Theoretical Computer Science, 2004,320(2-3):155−174.

[10] Mans B, Shparlinski I. Bisecting and gossiping in circulant graphs. In: Farach-Colton M, ed. Proc. of the LATIN 2004. Berlin, Heidelberg: Springer-Verlag, 2004. 589−598.

[11] Chen MS, Shin KG, Kandlur DD. Addressing, routing, and broadcasting in hexagonal mesh multiprocessors. IEEE Trans. on Computers, 1990,39(1):10−18.

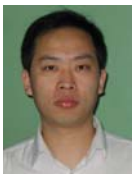
[12] Sullivan H, Bashkow TR. A large scale, homogeneous, fully distributed parallel machine, I. In: Proc. of the 4th Annual Symp. on Computer Architecture. New York: IEEE, 1977. 105−117.

[13] Kalman D, White JE. Polynomial equations and circulant matrices. American Mathematical Monthly, 2001,108(9):821−841.

[14] Lu KC. Combinatorics. 2nd ed., Beijing: Tsinghua University Press, 1991. 21−22 (in Chinese).

附中文参考文献:

[14] 卢开澄.组合数学.第 2 版,北京:清华大学出版社, 1991.21−22.



乐祖晖(1978－),男,湖北孝感人,博士,主要研究领域为可扩展路由器体系结构,交换网络,调度算法.

赵有健(1969－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为高速路由器硬件体系结构,高速大容量交换结构,IP 调度算法,混洗交换高速背板.

吴建平(1953－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算机网络体系结构,计算机网络协议测试,形式化技术.

张小平(1975－),男,博士生,助理研究员,CCF 会员,主要研究领域为网络协议体系结构,可扩展路由器体系结构,网络测量.