三维片上网络故障及拥塞感知的容错路由器设计

欧阳一鸣[1],张一栋[1],梁华国[2],黄正峰[2]
(1.合肥工业大学计算机与信息学院,安徽合肥 230009;2.合肥工业大学电子科学与应用物理学院,安徽合肥 230009)

摘 要: 三维片上网络中路由器的输入端口和交叉开关出现故障,将严重影响整个网络的性能,因此文章提出了一种故障及拥塞感知的容错路由器.通过增加一个冗余的输入端口和旁路总线,不仅能实现对输入端口和交叉开关容错的目的,而且还能在没有端口故障的情况下使用冗余端口有效地解决拥塞问题.实验表明此容错机制能够使得网络在故障路由器多、拥塞严重的情况下,仍然保持良好的性能.

关键词: 三维片上网络;故障;拥塞;容错路由器;旁路机制
中图分类号: TP302 **文献标识码:** A **文章编号:** 0372-2112 (2013)05-0912-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.05.013

A Fault-Tolerant Design of Faults and Congestion-Aware Router in Three-Dimensional Network-on-Chip

OUYANG Yi-ming[1], ZHANG Yi-dong[1], LIANG Hua-guo[2], HUANG Zheng-feng[2]
(1. *School of Computer and Information, Hefei University of Technology, Hefei, Anhui* 230009, *China*;
2. *School of Electronic Science and Applied Physics, Hefei University of Technology, Hefei, Anhui* 230009, *China*)

Abstract: The faults occurring in the input ports and crossbar of the router will seriously affect the performance of the entire network in Three-dimensional Network-on-Chip. This article proposes a fault and congestion-aware fault-tolerant router. By adding a redundancy of the input port and the bypass bus, our scheme can achieve fault tolerance of input ports and crossbar faults, and can effectively solve the congestion problem in the case of no fault port using the redundant port. The fault tolerance mechanism proposed can tolerate more fault-routers and still maintain good performance in a serious case of congestion.

Key words: three-dimensional network-on-chip; fault; congestion; fault-tolerant router; bypass mechanism

1 引言

三维片上网络[1~3](Three-Dimension Network-on-Chip,3D NoC)将多个晶片(die)在垂直方向堆叠,层间通过高速且高密度的硅通孔(Through Silicon Via,TSV)相连.使用垂直方向的短互连代替水平方向的长互连,缩短了内连线长度,减少了延时、降低了功耗、提高了性能.目前常见的架构有:3D Mesh、三维纤毛 Mesh、三维堆叠 Mesh 和 3D torus 等.其中基于 TSV 层间互连的 3D Mesh 结构被广泛研究,结构如图1所示.

但是由于硅特征尺寸接近原子量级时,VLSI 的偏差性和易于老化的脆弱性变得更加突出[4],因此其在可靠性方面也受到了严峻的挑战.例如,由制造缺陷、电路老化、工艺不稳定性和电子迁移等引起的路由器或 TSV 的永久性故障,并且这种故障是无法恢复的[5,6].这就意味着,为了提高系统的可靠性,在 3D NoC 设计时必须考虑潜在的硬件故障.

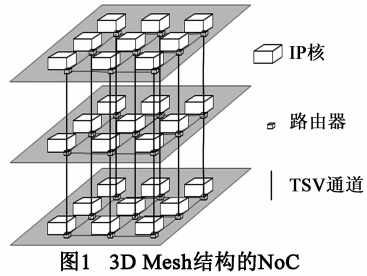

图1 3D Mesh结构的NoC

路由器作为 NoC 中重要的通信资源,负责数据包的存储转发,其一旦发生故障必然会导致 NoC 的通信效率降低,甚至整个芯片的报废.鉴于此,路由器的故障

收稿日期:2012-11-13;修回日期:2013-01-17
基金项目:国家自然科学基金(No.61274036;No.61106038);安徽高校省级自然科学研究重点项目(No.KJ2010A269);安徽省科技攻关项目(No.11010202190)

测试和容错问题引起了学术界和工业界的广泛关注.文献[7]提出了片上网络的测试方法.一般对于 NoC 中的硬件故障,解决方法分为以下两类:(1)通过提出容错路由算法绕过故障节点达到容错目的.如文献[8]提出了偏转路由算法,当路由器某端口出现故障时,数据包随机的向其它端口偏转.此方法增加了其它端口的竞争,容易产生拥塞,增加网络的延时.文献[9]中提出的容错路由算法,虽然能够保证正常通信,但是没有充分利用可用资源,而且算法复杂,使路由器设计变得复杂,可靠性降低;(2)通过改变路由器结构容错.文献[10]提出了一种可靠的路由器架构,通过对故障缓冲区的重构容错,此方法实现较为复杂并且只能对输入缓冲区容错,如果输入端口其它部件出现故障,该方法无法解决.文献[11]提出了共享虚通道的路由器,所有端口虚通道进行共享,若某一端口虚通道故障时,可以利用其他端口虚通道存储数据.该方法也能有效解决虚通道故障问题,不足之处在于,通过共享增加了其它输入端口的竞争,而且控制逻辑也变得复杂.

本文提出了故障及拥塞感知的容错路由器,通过增加一个冗余的输入端口和旁路总线实现容错.使得某一输入端口出现故障时,可以启用冗余的输入端口传输数据;在没有端口故障的情况下,若某个输入端口竞争激烈,出现拥塞,可以通过冗余端口平衡负载;在交叉开关出现故障情况下,通过旁路机制传输数据.通过本文方法,有效的解决了路由器故障问题,提高了系统的可靠性.

2　3D NoC 中路由器结构

在 3D NoC 中,路由器作为重要的通信资源.以 3D Mesh 中路由器为例,其结构如图 2 所示.此路由器有 7 个输入、输出端口,采用了虚通道技术.每个输入端口由若干个输入缓冲队列组成,每个输入缓冲队列分别对应着一个虚通道.路由器主要由输入缓冲模块(input buffer,IB)、路由计算模块(routing computing,RC)、交叉开关模块(crossbar)、虚通道分配器(virtual channel allocator,VA)和交叉开关分配器(switch allocator,SA)组成.

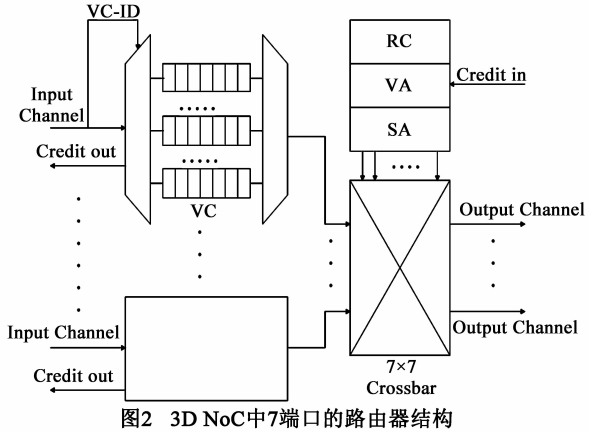

图2　3D NoC中7端口的路由器结构

各功能部件协同工作,可是一旦某个部件出现了永久性故障无法恢复时,路由器将丧失部分功能甚至全部功能.本文将分析这些潜在故障.首先给出以下两个定义:

定义 1　端口故障是指输入端口中数据分配器(DMUX)、FIFO 和多路选择器(MUX)故障.

定义 2　交叉开关故障是指输出端口的多路选择器出现故障.

2.1　端口故障模型

图 3 为 3D NoC 中路由器端口故障模型.图中右边路由器的西输入端口发生故障时,与之相邻的上游路由器的数据便不能向本级路由器转发,只能存储在上游路由器中或者丢弃,这样会造成上级路由器的拥塞或者重传丢弃的数据包,增大了网络延时,降低了吞吐量.更为重要的是路由器输入端口一旦发生硬故障,便无法恢复,如果没有相应的容错措施,只会使网络状况越来越差.因此在路由器设计时必须考虑输入端口故障问题.

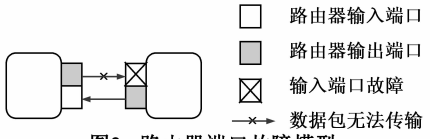

图3　路由器端口故障模型

2.2　交叉开关故障模型

路由器交叉开关的功能可以通过多路选择器来实现.3D NoC 中路由器的每个输出端口处都有一个六选一的多路选择器和仲裁器(Arbiter).图 4 所示的为北输出端口(North Ouput Port)的结构图.北输出端口接受来自东、南、西、上、下及本地输入端口的请求,根据仲裁信号选择一路数据输出.一旦输出端口的多路选择器出现故障,则请求该端口输出的数据将无法传输,只能保存在原来的输入端口中,可能会造成了输入端口拥塞.因此,交叉开关故障也严重影响了网络性能,对交叉开关的容错也很必要.

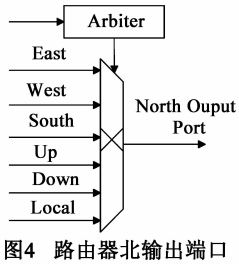

图4　路由器北输出端口

3 故障及拥塞感知的容错路由器工作原理

本文通过对 3D NoC 中路由器端口故障、交叉开关故障和端口拥塞三种情况的分析.提出了故障及拥塞感知的容错路由器.通过增加了一个冗余的输入端口,使得某一端口出现故障时,可以启用冗余端口传输数据;在没有端口故障的情况下,若出现拥塞也可启用冗余端口,起到平衡负载的作用;在交叉开关出现故障时,通过旁路机制向输出端口传输数据.

3.1 容错路由器结构及其工作原理

本文提出的容错路由器具体结构如图5所示.输入端口由东、南、西、北、上、下、本地七个基本端口和一个冗余端口(Redundant Port)组成.东、南、西、北、上、下、本地七个方向的输入通道(Input Channel)分别通过三态门与冗余输入端口相连.增加了一个 7 × 1 MUX 和一个 Arbiter,Arbiter 的输入信号为 XFPR(X Fault Port Request)和 XCPR(X Congestion Port Request),其中 $X \in$ {east、south、west、north、up、down、local}.XFPR 请求信号在 X 输入端口出现故障时有效,XCPR 请求信号在 X 输入端口无输入缓冲时有效.在通信过程中,当七个基本端口中出现故障或拥塞时,会申请使用冗余端口.当多个基本端口出现故障或拥塞同时向冗余端口申请时,采用主-辅两级优先级仲裁机制进行仲裁.优先级用 $P_{i,j}$ 表示,其中 $i \in \{F、C\}$,F 表示申请冗余端口的原因是出现故障,C 表示申请的原因是出现拥塞.$j \in$ {east、south、west、north、up、down、local},表示申请冗余端口的请求来自七个基本端口的哪一个.优先级规则如下:

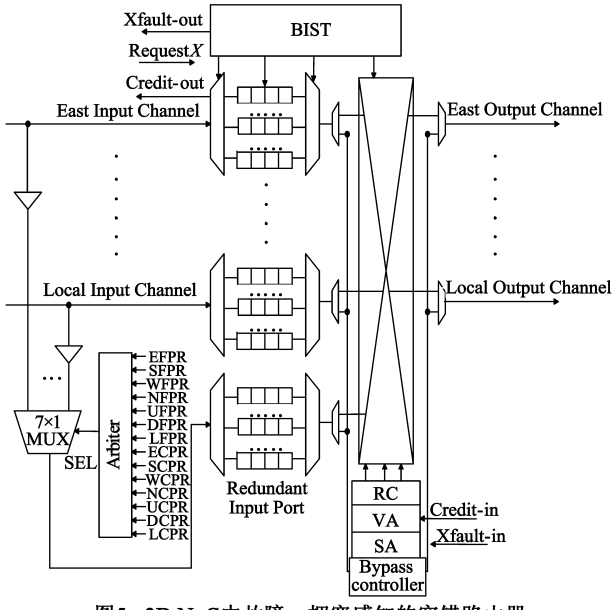

图5 3D NoC中故障、拥塞感知的容错路由器

(1) $P_{F,X1} > P_{C,X2}$(主优先级:故障优先级高于拥塞优先级)

(2) $P_{i,\text{up}} > P_{i,\text{down}} > P_{i,\text{east}} > P_{i,\text{south}} > P_{i,\text{west}} > P_{i,\text{north}} > P_{i,\text{local}}$,$i \in \{F、C\}$(辅助优先级:端口间的初始优先级)

(3)主优先级始终不变,辅助优先级可以动态变化.对于出现故障或者拥塞两种不同情况的某个基本端口的请求一旦得到响应,则该端口优先级降为最低.

主-辅优先级的设计,使得各基本端口都能公平的使用冗余端口解决问题.

图中采用了 BIST 电路,用于对输入端口中的 DMUX、FIFO 和 MUX 及路由器中的 Crossbar 进行故障检测.图 6 为测试路由器输入端口及交叉开关模块的 BIST 电路原理图.TPG(Test Pattern Generator)为测试向量生成器;DMUX、FIFO、MUX 和 Crossbar 为被测电路;MISR(Multiple Input Signature Register)为多输入特征寄存器,对被测电路的响应进行压缩,产生响应压缩特征值;ROM 主要用于存储正确特征值;输出响应分析器 ORA(Output Response Analyzer)的功能是把测试后所得到的响应特征值与 ROM 中的正确特征值比较,最终得出测试的结果,是故障还是无故障;BIST 控制器功能为控制测试的开始、结束和协调整个测试过程.其工作原理为:当系统进入测试模式时,在 BIST 控制器控制下 TPG 产生测试向量,并施加到被测电路.MISR 将被测电路的响应输出压缩得到特征值.ORA 对响应压缩特征值与 ROM 中的正确特征值进行比较分析,最终得出测试结果是故障还是无故障.

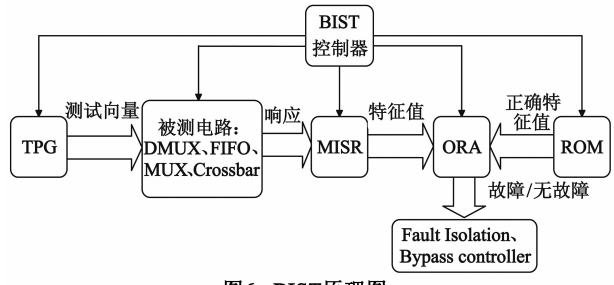

图6 BIST原理图

本文提出的路由器工作在两种不同的模式下,分别为测试模式和正常模式.路由器在进入正常模式工作前,首先工作在测试模式下.系统会自动给出启动 BIST 信号,启动测试模块对路由器输入端口及交叉开关进行测试.测试结束后 BIST 模块输出测试结果作为路由器重配置的控制信号.具体如下:

(1)端口故障及拥塞

端口故障包括输入端口 DMUX、FIFO 和 MUX 的故障.本文提出的路由器,增加了一个输入端口,不仅能够容端口故障,同时也能起到缓解输入端口拥塞的作用.

数据包申请 X 输入端口(Request$X$ 信号有效),如果 BISI 检测结果为端口故障,则使 XFault 信号有效,返

回故障信息给上级路由器.上游路由器得知故障情况,便请求冗余端口,使 *X*FPR 有效.如果此时 *X* 输入端口无故障,但是 *X* 输入端口可用的 Input Buffer 为 0,则 credit-out 信号为 0,上游路由器得知拥塞情况,便请求冗余端口,使 *X*CPR 信号有效.最终根据优先级,Arbiter 会授权给某个请求,并选通相应的三态门,数据存入冗余端口,随之进行 RC、VA、SA、ST.某时刻拥塞状况得到缓解,则返回 credit-out 信号为 1,并且使拥塞端口的请求信号 *X*CPR 无效.通过以上的办法,能够有效的解决路由器端口故障问题,提高系统的可靠性;在路由器无端口故障情况下还能利用冗余端口解决其它端口拥塞问题,降低了延时,提高了吞吐量.

(2)交叉开关故障

本文提出的路由器的交叉开关的功能是通过多路选择器实现的.当某输出端口多路选择器出现故障,本方案通过添加了一条旁路总线(Bypass Bus)容错.总线控制器(Bypass controller)根据 BIST 测试的结果,动态的选择数据包是通过交叉开关还是通过 Bypass Bus 传输.如果有多个 flit 同时申请 Bypass Bus 时,采用分时复用的方法,一个时钟周期传输一个 flit.图 7 为北输出端口多路选择器故障容错原理图.此时 Bypass controller 控制输入端口数据通过 Bypass Bus 绕过故障多路选择器从输出端口输出.

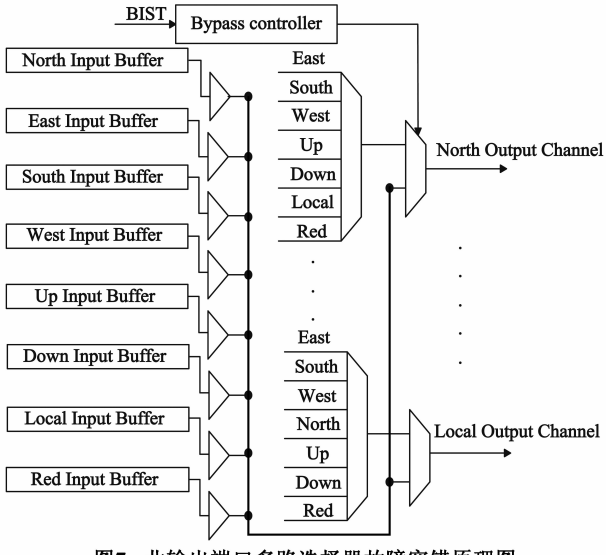

图7　北输出端口多路选择器故障容错原理图

3.2　Fault Isolation 模块

如果测试结果为某输入端口或者某输出端口的 MUX 出现故障,BIST 的输出结果作为 Fault Isolation 模块的控制信号,对故障输入端口或输出端口故障 MUX 隔离.北输出端口的 Fault Isolation 模块如图 8,主要是由 RII(Request-In Isolation),RRI(Redundant Port Request-In)和 ROI(Request-Out Isolation)组成.

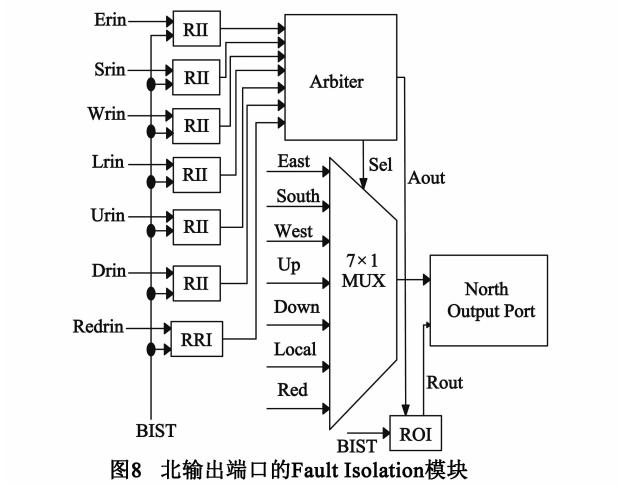

图8　北输出端口的Fault Isolation模块

(1)RII 和 RRI 的功能

RII、RRI 和 ROI 都位于输出端口处.Arbiter 输出的 Sel 作为 MUX 的控制信号选择一个 Input Buffer 中的数据输出.RII 的输入信号为 Erin(East Request-in)、Srin(South Request-in)、Wrin(West Request-in)、Urin(Up Request-in)、Drin(Down Request-in)和 Lrin(Local Request-in).它们是输入端口中数据经过路由计算后对北输出端口的请求信号.RRI 的输入为冗余输入端口的请求信号 Redrin(Redundant Port Request-In).如果来自 BIST 的诊断信号鉴别出故障端口,则 RII 使相应的输入请求信号无效,RRI 使冗余端口的请求信号有效.

(2)ROI 的功能

ROI 功能是隔离输出端口的故障 MUX.每个 ROI 都与 Arbiter 的输出信号 Aout 及输出端口的请求输出信号 Rout(Request output signal)相连.如果 BIST 诊断信号指出输出端口 MUX 故障,则 ROI 使请求输出信号 Rout 无效,从而隔离故障 MUX.

本文通过合理的设计不仅成功实现对输入端口和交叉开关故障容错,而且还能有效的解决端口拥塞问题.系统的可靠性和整体性能因此得到很大提升.

4　实验评估

(1)性能的提高

实验使用 OPNET 模拟了一个 3D NoC 仿真平台,在此平台分别搭建两个 4×3×3 的 3D Mesh 结构的 NoC.设置了若干个路由器故障,比较不同的容错方案的性能.方案一采用文献[8]提出的容错方法,不改变路由器的结构,通过 *XYZ* + 偏转路由达到容错目的.方案二使用本文设计的故障及拥塞感知的容错路由器,路由算法为简单的 *XYZ* 路由.试验采用均匀随机模式,在相同的故障数下,比较两种方案延时、吞吐率和丢包率.

本文在使用上述不同路由器结构的 3D NoC 中,分

别设置 1~6 个路由器故障,故障发生位置相同.针对不同的故障情况比较两种方案的优越性.图 9、图 10 分别表示,在数据包注入率为 0.4 和 0.5 时,两种方法在不同故障数下延时的比较.在相同的注入率下,随着故障数增多文献[8]延时急剧增加,而本文延时增加不多,且延时远小于文献[8],如注入率为 0.4、故障数为 6 时,本文延时较文献[8]下降了 60%.在相同的故障数下,网络中数据包注入率由 0.4 增加到 0.5 时,文献[8]的延时增加较大而本文方法延时增加不大.这是因为本文提出的路由器能够有效的缓解网络拥塞.

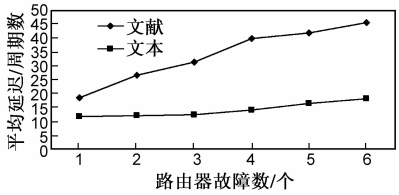

图9 注入率为0.4时平均延时比较

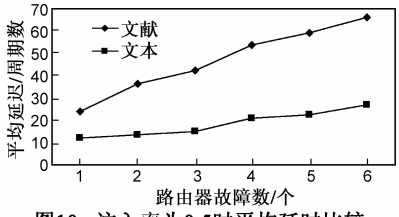

图10 注入率为0.5时平均延时比较

图 11 表明在故障路由器数为 5,注入率小于 0.2 时两种方案吞吐率基本相同.随着注入率增加本文优势明显,且文献[8]在 0.45 时网络饱和,本文在 0.55 时才饱和.注入率为 0.5 时本文较文献[8]提高 33%.

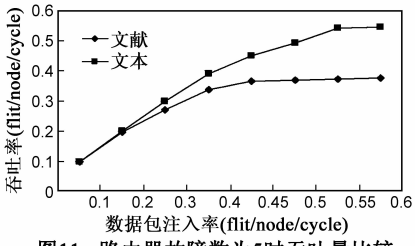

图11 路由器故障数为5时吞吐量比较

表 1 是在注入率为 0.45 时,不同故障数,两种方案丢包率的比较.当故障数小于等于 4 时,两种方案在注入率为 0.45 情况下都未达到饱和,拥塞状况良好,丢包率都很低.当故障数大于 4 以后,文献[8]网络饱和,拥塞严重,丢包率增幅较大,然而本文方案,依然能保证数据包可达.这是因为在故障数多时,本文方案能通过冗余端口保证数据正常传输及缓解拥塞,而文献[8]通过随机偏转绕过故障节点,加剧其它端口竞争,造成拥塞,因此丢包率增大.

表 1 不同故障数下丢包率比较

(数据包注入率:0.45)

故障数	2	4	5	6
文献[8]	3.68%	8.54%	18.52%	40.36%
本文方案	0.00%	0.00%	0.37%	2.37%

(2)面积开销

本文使用 Synopsis Design Compiler 在 65nm 工艺下对 3D NoC 中 7 端口路由器和本文提出的端口加固容错路由器面积仿真.两种路由器都采用虚通道,设定每个端口 4 个虚通道,每个虚通道深度为 8 个 flit[10].仿真结果如表 2,本文提出的路由器面积较传统 7 端口路由器仅增加了 21%.考虑到能够实现容错,而且性能优势较大,可以容忍增加的硬件开销.

表 2 路由器面积开销

元件	7 端口路由器(μm^2)	本文路由器(μm^2)
BIST	—	14230
Bypass Bus	—	563
总面积	195154	236153
开销	21%	

5 结论

随着集成电路工艺水平的提高,可靠性成为 3D NoC 研究重点.本文通过对路由器输入端口和交叉开关故障的分析,提出了可行的容错方案.具有以下贡献:(1)通过增加一冗余输入端口和旁路总线实现对输入端口和交叉开关故障容错,提高了 3D NOC 的可靠性;(2)利用了冗余端口平衡了负载,提高了系统性能;(3)此设计与 3D NOC 拓扑无关,具有通用性.由实验可知,本文提出的容错机制优于文献[8],网络延时和丢包率相对更小、吞吐率更高,在路由器故障数多、数据包注入率大时表现尤为突出.在取得较大性能优势下,路由器面积开销仅增加 21%,在可以容忍范围内.

参考文献

[1] B S Feero, P P Pande. Networks-on-chip in a three-dimensional environment: A performance evaluation[J]. IEEE Transactions on Computers, 2009, 58(1): 32 – 45.

[2] 梁华国,李鑫,等.并行折叠计数器的 BIST 方案[J].电子学报,2012,40(5):1030 – 1033.
Lian Hua-guo, Li Xin, et al. Bist scheme of parallel folding counters[J]. Acta Electronica Sinica, 2012, 40(5): 1030-1033. (in Chinese)

[3] J D Owens. Research challenges for on-chip interconnection networks[J]. Micro, 2007, 27(5): 96 – 108.

[4] D Fick, D A Andrew, et al. Vicis: a reliable network for unreli-

able silicon[A]. Proceedings of Design Automation Conference 2009[C]. San Francisco: ACM, 2009. 812-817.

[5] 欧阳一鸣,成丽丽,梁华国.一种基于变长数据块相关性统计的测试数据压缩和解压方法[J].电子学报,2008,36(12):298 – 302.
Ouyang Yi-ming, Cheng Li-li, Lian Hua-guo. A new test data compression technique based on static relativity of variable length data block[J]. Acta Electronica Sinica, 2008, 36(12): 298 – 303. (in Chinese)

[6] 王颀,单智阳,等.串扰约束下超深亚微米顶层互连线性能的优化设计[J].电子学报,2006,34(2):214 – 219.
Wang Qi, Shan Zhi-yang, et al. The optimal design of ultra deep sub-micron global interconnect under crosstalk constraint [J]. Acta Electronica Sinica, 2006, 34(2): 214 – 219. (in Chinese)

[7] G Cristian, P Pande, et al. Methodologices and algorithms for testing switch-based NoC interconnects[A]. Proceedings of International Symposium on Defect and Fault Tolerance in VLSI System[C]. Monterey: IEEE, 2005. 238 – 246.

[8] Chao-chao Feng, Zhong-hai Lu, et al. A low-overhead fault-aware deflection routing algorithm for 3D NoC[A]. Proceedings of IEEE Computer Society Annual Symposium[C]. Chennai: IEEE, 2011. 19 – 24.

[9] N Rameshan, V Laxmi, et al. Minimal path, fault tolerant, QoS aware routing with node and link failure in 2-D mesh NoC [A]. Proceesings of IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems[C]. Kyoto: IEEE, 2010. 60 – 66.

[10] A DeOrio, D Fick, et al. A reliable routing architecture and algorithm for NoCs[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2012. 31(5): 726 – 739.

[11] K Latif, et al. A novel topology-independent router architecture to enhance reliability and performance of networks – on-Chip[A]. Proceesings of IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems[C]. Vancouver: IEEE, 2011. 454 – 462.

作者简介

欧阳一鸣　男,1963 年出生,副教授,硕士生导师,中国计算机学会高级会员,研究方向:片上网络(NoC),嵌入式系统的综合与测试,数字系统设计自动化.
E-mail: oyymbox@163.com

张一栋　男,1989 年出生,硕士研究生,研究方向:片上系统以及片上网络容错方法.
E-mail: zyd19891107@126.com

梁华国　男,1959 年出生,教授,博士生导师,中国计算机学会容错计算专业委员会委员,研究方向:嵌入式系统的综合与测试,数字系统设计自动化、ATPG 算法与分布式控制等.
E-mail: huagulg@hfut.edu.cn

黄正峰　男,1978 年出生,博士,副教授,硕士生导师,中国计算机学会容错计算专业委员会委员.研究方向:嵌入式系统的综合与测试、数字集成电路的硬件容错、星载 SoC 芯的抗辐射加固.
E-mail: hanson_hfut@sina.com