

基于 Fat-tree 的高性能互联网络性能优化与分析

张 毅,何卫锋

(上海交通大学 微电子学院,上海 200240)

摘 要: 高性能互联网络对于高性能计算机的性能具有重要的影响,而高性能网络中各个模块的设计都会影响到网络的最终性能,因而前期的网络仿真显得尤为重要. HPCSim 是基于 OMNeT++ 平台自主开发的高性能网络仿真器,包括了网络拓扑和仿真参数自动生成工具 PySim 以及仿真结果分析和可视化工具 PyVisual. 仿真器中对全缓存高阶路由器结构进行了改进,提出了一种 5 级流水线的高阶路由器,能够减少路由信息缓存开销同时将网络吞吐率从基准路由器下的 0.75 提升至 0.95. 此外,还对硅光互联链路进行了逻辑抽象,并在 Fat-tree 拓扑上对链路性能和网络吞吐率的关系进行了分析.

关键词: 高性能互联网络;仿真器;全缓存路由器;硅光互联

中图分类号: TP391.9

文献标识码: A

文章编号: 1000-7180(2018)08-0036-06

Optimization and Evaluation of Fat-tree Based High Performance Interconnect Network

ZHANG Yi, HE Wei-feng

(School of Microelectronics, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: High performance interconnect network has a significant influence on the performance of high performance computer. As the difference in network model design will result in different network performance, the early network simulation is needed. HPCSim is an independently developed high performance interconnect network simulator based on OMNeT++. It includes topology and parameter file generator named PySim, and simulation data analyzer and visualizer named PyVisual. An improved fully-buffered router structure is proposed, which achieves 5 stage pipeline, and reduces routing computation states buffer as well as promotes throughput from 0.75(baseline router) to 0.95. Moreover, a link model is inferred from photonic network, and the link performance simulation is run based on Fat-tree network to analyze the relation between link bandwidth and network throughput.

Key words: high performance interconnect network; simulator; fully-buffered router; silicon photonic

1 引言

随着集成电路工艺与并行计算网络技术的不断发展,高性能计算机的性能在不断的提升. 由国防科技大学设计的高性能计算机天河 2 号在 2013 年 6 月被 41 期 TOP500 名单评为性能最高的计算机,其拥有 3 120 000 个处理器核心,理论峰值性能为 54.9 PFlop/s^[1]. 之后在 2016 年 6 月,由国家并行计算机工程技术研究中心研发的神威太湖之光获得了 47 届 TOP500 名单中超级计算机排名的第一,它拥有

10 649 600 个处理器核心,理论性能峰值为 125 PFlop/s^[2],而功耗仅为 15.3 MW,小于天河 2 号的 17.8 MW.

高性能计算机的性能提升主要通过提高网络的并行性,随着计算节点数的不断上升,给并行网络的设计带来了巨大的挑战. 高性能互联网络拓扑的选择在高性能计算机的设计过程中显得至关重要,网络拓扑决定了高性能计算机处理节点的规模以及系统最后能达到的性能上限. 国内外的高性能计算机

设计机构选用了不同的互连网络来实现其高性能计算机,如 Intel 的 Paragon^[3]采用的是 2D Mesh 网络,而 Cray 的 T3D^[4]采用的是 3D Torus 网络,天河 2 号以及神威太湖之光采用的则是 Fat-tree 网络^[5]. 本文采用 Fat-tree 网络作为高性能互连网络仿真器的基础网络,下文的网络性能仿真都在此基础上进行. 在高性能计算机设计中,除了网络拓扑之外,路由器的设计对网络性能的也有非常重要的影响. 目前学术界和工业界主要采用高阶(High-radix)路由器来设计高性能计算机,高阶路由器由 Kim 和 Dally^[6]提出,在网络总节点数相同的情况下,高阶路由器能够降低网络的平均延时并提高网络的吞吐率. 本文在 Dally 提出的全缓存路由器结构的基础下,提出了改进的路由器结构,使其对比基准(Baseline)路由器有 20%吞吐率的提升. 此外,在高性能计算机中,由于芯片大小和散热的限制,传统的电互联方案遇到了性能的瓶颈,在下一代高性能计算机中,硅光互联技术以更高的带宽和更低的功耗将逐渐取代原有的电互联方案. 本文对硅光链路进行了分析,并提出了硅光链路的建模方式,分析了硅光链路性能提升对网络性能的影响.

2 高性能计算机

天河 2 号超级计算机具有 18 304 个计算节点,每个节点由 2 个 Intel Xeon E5-2600 和 3 个 Intel Xeon Phi 组成. 整个网络是一个三层 Fat-tree 结构,其路由器(NRC, Network Routing Chip)具有 24 个 8 路双向端口,每条通路的单向带宽为 14 Gbps,路由器总带宽为 5. 376 Tbps^[1]. 其路由器设计参照了 WJ Dally 的高阶路由器^[6],其内部分为 24 个子模块(Tile),这种结构能够解决仲裁器的布线复杂度并提高整体的吞吐率. 同时,神威太湖之光采用国产 SW26010 多核处理器,单个 CPU 包含 260 个计算单元, CPU 的工作频率为 1. 45 GHz,整个系统包含了 40 960 个 CPU,由一个 2 层的 Fat-tree 网络连接,整个系统可以达到 125 PFlop/s 的性能^[2]. 本文以天河 2 号的拓扑以及路由器结构作为原型,实现了可自定义参数的 Fat-tree 网络以及高阶路由器.

3 仿真器设计

高性能计算机仿真器 HPCSim 由 Shell 脚本,

Python 组件 PySim, 离散事件仿真器 OMNeT++^[7]和 Python 可视化组件 PyVisual 组成. Shell 脚本调用 PySim 组件和 OMNeT++ 组件进行仿真. PySim 模块根据 Shell 脚本传入的仿真配置参数如注入率,链路带宽等参数来自动生成 OMNeT++ 所需的 ned,ini 和 h 文件. 其中 ned 文件是网络拓扑文件,描述了网络中计算节点和路由器之间的连接方式以及链路的带宽和延时参数,ini 文件为 OMNeT++ 配置文件,提供仿真拓扑选择和仿真时间设置等功能. 此外,PySim 还会生成 h 头文件,主要宏定义相关的网络参数,如节点数量,缓存大小,流量模型等参数. OMNeT++ 是开源的离散事件仿真器,它利用 PySim 生成的 ned 和 ini 文件,以及路由器和计算节点对应的 cc 和 h 文件以及 msg 消息文件编译为可执行文件,Shell 脚本再调用可执行文件运行仿真,仿真结果保存在 sca 和 vec 文件中. PyVisual 为可视化组件,它通过解析 sca 和 vec 文件中的数据,统计并分析仿真数据,最后绘制相应图表,如图 1 所示.

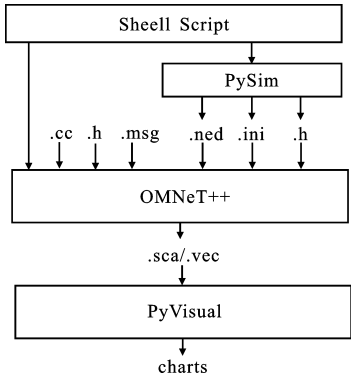


图 1 HPCSim 框图

4 Fat-tree 网络结构

Fat-Tree^[5]是由 Leiserson 提出的一种网络拓扑,它是一种层次化的网络拓扑结构,它具有高二等分带宽(bisection bandwidth),低网络直径(diameter)以及路由路径的多样性被广泛用于高性能计算机以及商业数据中心如天河 2 号和神威太湖之光等.

图 2 是一个 3 层的 Fat-Tree 网络,由上到下的三层是由交换机(switch)连接而成的 Fat-Tree 拓扑,最下面一层的计算单元(Processing Unit, PU)与第三层交换机直接相连. 其中交换机(switch)部分实现了路由策略,流控机制和虚通道仲裁等;计算

单元部分(PU)主要发送和接收网络数据包。

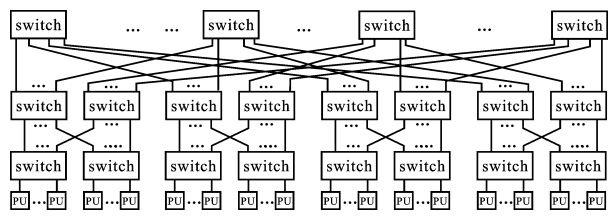


图 2 Fat-tree 拓扑结构

5 路由器设计

5.1 基准路由器

本文以 Kim 和 Dally 提出的基准路由器 (baseline router) 和全缓存 (fully buffered) 路由器^[6]为原型,在仿真器中实现了 4 级流水线的基准路由器,并提出了改进的 5 级流水线的全缓存路由器. 基准路由器的结构框图如图 3 所示,一个 k 个端口 v 个虚通道的基准路由器的主要模块分为虚通道数量为 v 的输入缓存 (input buffer) 模块,路由计算 (routing computation) 模块,虚通道仲裁 (virtual channel allocation) 模块,交叉开关仲裁 (switch arbitration) 模块,交叉开关传输 (switch traversal) 模块以及交叉开关 (crossbar) 模块. 其中各部分的功能如下.

输入缓存模块 (Input buffer): 路由器某个输入端口接收到数据包后,根据微片 (flit) 中的 VCID (virtual channel id) 字段将该微片放入对应的虚通道缓存中.

路由计算模块 (Routing computation, RC): 根据收到数据包的头微片 (head flit) 中所包含的目标节点地址,计算出该数据包在当前路由器中的输出端口以及分配输出虚通道 VCID (output virtual channel id) 号,该阶段只对每个数据包 (packet) 计算一次.

虚通道仲裁模块 (Virtual channel allocation, VCA): 在虚通道仲裁阶段,每个输出虚通道对所有请求它的输入虚通道进行仲裁,只有一个输入虚通道胜出. 仲裁胜利后,该输出虚通道一直被仲裁胜利的输入虚通道占用,直到数据包的尾微片 (Tail flit) 从该路由器转发出去.

交叉开关仲裁模块 (Switch allocation, SA): 交叉开关仲裁分为两个阶段,第一个阶段 (SA1) 为输入端口仲裁,同一输入端口的不同虚通道竞争与交叉开关 (crossbar) 相连的物理通道;第二阶段 (SA2)

为输出端口仲裁,每个输出端口对请求它的所有输入端口进行仲裁,只有一个输入端口仲裁胜利.

交叉开关传输模块 (Switch traversal, ST): 根据交叉开关仲裁阶段的结果,将仲裁胜利的微片从输入缓存通过交叉开关传送到输出端口.

上述几个阶段中,路由计算和虚通道仲裁过程以数据包为单位,只对头微片进行,交叉开关仲裁和交叉开关传输对每个微片都会进行,基准路由器的流水线如图 3 所示.

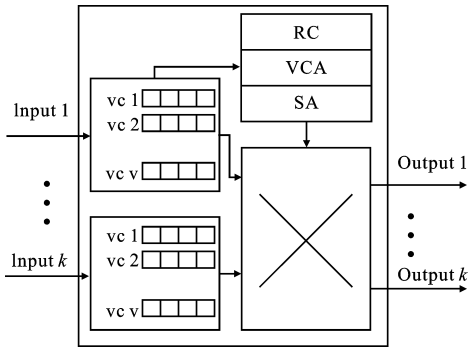
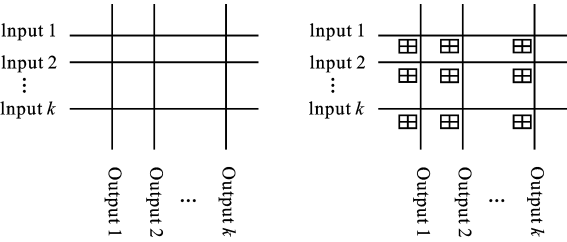


图 3 基准路由器

5.2 全缓存路由器

全缓存路由器与基准路由器的区别在于交叉开关是否带有缓存,图 4(a) 是基准路由器的交叉开关,不带有缓存,因此微片通过交叉开关时,需要同时获得交叉开关输入仲裁和输出仲裁的胜利,如果输入仲裁胜利但输出仲裁失败,则输入仲裁无效,这个阶段交叉开关输入通道无法被其他虚通道所使用,造成了资源的浪费. 图 4(b) 是全缓存路由器的交叉开关,在原有交叉开关的基础上在每个交叉点 (Crosspoint) 加入了缓存,其中每个交叉点的缓存大小为 $d * v$, 其中 d 为缓存的深度, v 为虚通道的数量,因此这种结构引入的总缓存数量为 $k * 2 * v * d$, k 为路由器端口数. 全缓存路由器相较于基准路由器,可以将输入仲裁和输出仲裁解耦,当输入仲裁胜利后,先将微片存入交叉点缓存中,然后再进行输出端口仲裁,即使输出端口仲裁失败,也不需要重新进行输入端口仲裁,这样可以提高交叉开关输入端的利用率,从而提高路由器的吞吐率.

针对全缓存路由器,本文提出了 2 种不同的路由器结构. 图 5(a) 是 6 级流水线结构,其流水线如图 6(b) 所示,图 5(b) 是 5 级流水线结构,其流水线如图 6(c) 所示. 下文假设每个数据包只包括一个头



(a)基准路由器交叉开关 (b)全缓存交叉开关

图 4 交叉开关

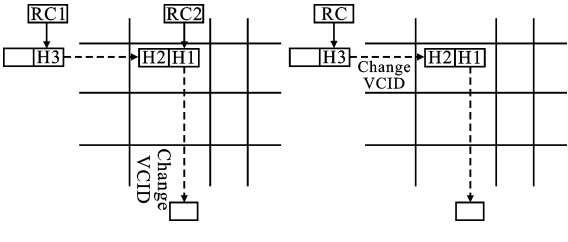
微片(head flit),图 5 中 H1,H2,H3 分别代表三个数据包的头微片,以此为前提来讨论三种结构下资源的占用率.

两种全缓存路由器都将交叉开关仲裁(SA)和交叉开关传输(ST)分别分成了两个阶段,包括输入仲裁(SA1),输出仲裁(SA2),输入传输(ST1),输出传输(ST2).图 5(a)的路由器结构将路由计算(RC)分成了两个阶段,分别为 RC1 和 RC2.其中 RC1 阶段根据头微片的目标地址信息计算输出端口,并把结果保存在 RC1 状态寄存器中,RC2 阶段则计算头微片的输出虚通道,并将结果保存在 RC2 状态寄存器中.这样,在 H3 到达输入缓存的头部,并进入交叉开关前,通过 RC1 判断将其存放在哪一列的交叉点缓存中.在 H1 到达交叉点缓存的头部后,通过 RC2 判断它的输出虚通道,并进行虚通道仲裁(VA),在 H1 通过交叉开关后,将其中的 VCID 字段进行更新为 RC2 的计算结果,即下一跳路由器的输入 VCID.分析图 5(a)的 6 级流水线路由器,其输入缓存部分需要 $k \times v$ 个 RC1 状态寄存器,交叉开关部分需要 $k \times 2 \times v$ 个 RC2 状态寄存器.

图 5(b)中的结构在微片进入交叉开关前(即 ST1 阶段),根据 RC 阶段得到的输出端口 VCID 号,立刻更新微片的 VCID 字段,同时将它传入交叉开关中与输入虚通道 VCID(即更新前的 VCID)对应的缓存中.因此图 5(b)中的结构相比于图 5(a)的结构能够节省 $k \times 2 \times v$ 个 RC2 状态寄存器,同时由于省略了 RC2 阶段,可以将流水线级数缩减到 5 级.

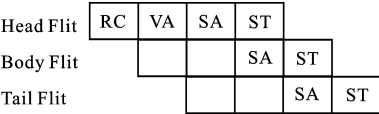
6 硅光链路模型

硅光互联技术相较于电互联技术具有更低的功耗以及更高的带宽,同时由于可以将激光器和调制器等硅光器件集成在硅基板上,因而相比于传统的光互连技术,具有更高的集成度和通信密度.图 7 是



(a)6 级流水线路 (b)5 级流水线路

图 5 全缓存交叉开关



(a)基准路由器流水线路

(b)6 级全缓存路由器流水线路

(c)5 级全缓存路由器流水线路

图 6 路由器流水线路

硅光链路的模型,发射端包括高速串并转换(SerDes)模块,驱动模块(Driver),马赫曾德尔调制器(MZM)和激光器(Laser),接收端包括光学探测器(PD),均衡器(CTLE,DFE),线性放大器(LA)等.

本文对硅光链路模型进行抽象,来提取仿真所需的相关参数.通常在高性能计算机仿真中,用带宽、延时、功耗这几个指标来描述信道,因此对于图 7 的硅光链路,可以获得如下的仿真参数:

$$\text{Latency} = \text{Overhead}_m + \text{Latency}_f + \text{Overhead}_r$$
$$\text{Bandwidth} = \text{Min}(\text{Bandwidth}_m, \text{Bandwidth}_f, \text{Bandwidth}_r)$$
$$\text{Power} = \text{Sum}(\text{Driver}, \text{Laser}, \text{MZM}, \text{PD}, \text{CTLE}, \text{LA}, \text{DFE})$$

其中,延时主要集中在光纤上,为 5 ns/m,目前硅光器件和 SerDes 的单通道带宽为 25 Gbps 左右.

7 仿真条件

利用 HPCSim 对高性能计算机进行仿真时,需要设定路由器端口数量,Fat-tree 层数,链路带宽等相关参数,具体的仿真参数设置为端口数量为 16,网络层

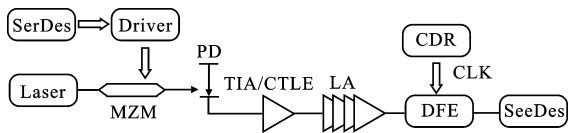
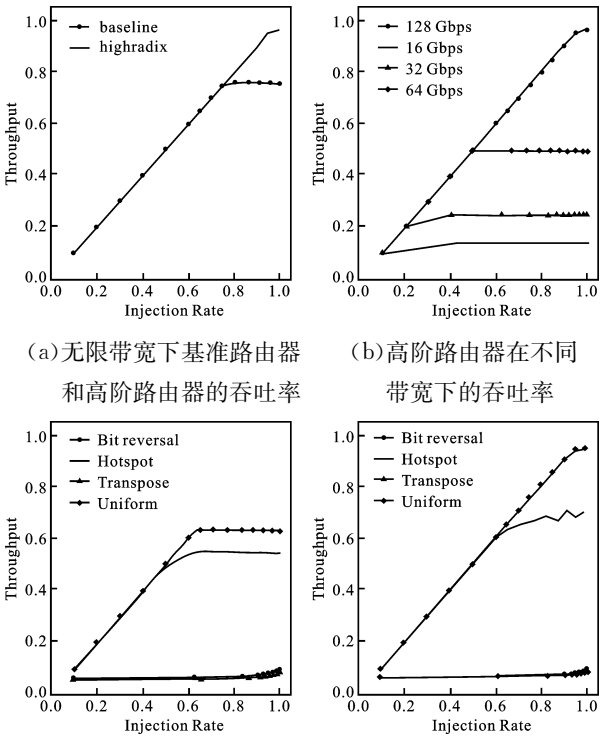


图 7 硅光链路模型

数为 3, 微片大小为 4 Bytes, 包大小为 16 Bytes, 输入缓存大小为 32 Packets, 交叉点缓存深度为 8 Flits, 链路延时为 5 ns, 链路带宽为 128 Gbps, 虚通道数量为 3, 计算节点为 1 024。为了匹配链路带宽, 计算节点发送数据的频率为 $\text{Freq} = 128 \text{ Gbps} / 32 \text{ bis} = 4 \text{ GHz}$ 。本文采用了计数制 (Credit-based) 流控机制^[8], 和均匀模型 (Uniform), 热点模型^[9] (Hotspot), 转置模型 (Transpose), 位翻转模型 (Bit reversal) 这几种流量模型对网络吞吐率进行分析。

8 仿真结果

图 8 是利用 HPCSim 得到的仿真结果, 图 8(a) 表示基准路由器和高阶路由器在带宽无限大的情况下, 网络的吞吐率曲线, 基准路由器的最大吞吐率为 0.75, 而高阶路由器能够达到最大 0.95 的吞吐率, 验证了全缓存路由器设计的合理性。



(c)(d) 基准路由器和高阶路由器在不同流量模型下的网络吞吐率

图 8 仿真结果

图 8(b) 在其他仿真参数不变的前提下改变了链路的带宽, 观察图表可以发现在链路带宽低于计算节点数据发送频率时, 带宽是限制网络吞吐率的主要原因。随着带宽的提升, 网络的吞吐率也不断提升, 直到链路带宽与计算节点发送频率相匹配后, 网络吞吐率达到性能上限, 此时继续提升链路带宽已经无法提高网络吞吐率。其主要原因是, 当链路带宽小于计算节点发送数据频率时, 节点向网络中注入新的数据, 信道还在传输上一个数据包, 从而导致冲突的发生, 最终使吞吐率下降。

图 8(c) 和图 8(d) 分别是基准路由器和全缓存高阶路由器在不同的注入率下的吞吐率, 当网络在均匀流量模型下时, 各路由器负载相对均衡, 能够使网络达到较高的吞吐率, 而热点流量模式下, 由于网络中流量的不均匀性, 会导致一些热点路由器中缓存阻塞以及仲裁冲突增加, 从而导致系统吞吐率的下降。

9 结束语

高性能计算机仿真工具 HPCSim 能够支持仿真前的网络拓扑和仿真参数自动化生成, 以及仿真后仿真数据的自动化分析并生成图表的功能。本文对全缓存高阶路由器提出了改进逻辑结构, 将流水线级数从图 5(a) 的 6 级缩减为图 5(b) 的 5 级, 并将 VCID 更新逻辑从输出开关仲裁 (ST2) 阶段提前至输入开关仲裁 (ST1) 阶段, 从而使 RC2 路由状态缓存从 $k \cdot 2 \cdot v$ 削减为 0。同时, 通过仿真结果验证了采用 5 级流水的全缓存高阶路由器可以提升网络的吞吐率, 可以将原先采用基准路由器的网络从 0.75 的吞吐率提升到 0.95。此外, 本文对硅光链路进行建模, 并通过仿真分析链路性能提升对网络性能的影响, 得出在链路带宽低于计算节点数据发送频率时, 链路性能提升可以提高网络的吞吐率, 而在链路带宽高于计算节点数据发送率后, 链路性能提升并不会带来网络吞吐率的提升, 并分析了现象产生的原因。文章最后分析了不同流量模型下, 网络吞吐率的差别。

参考文献:

[1] Liao X K, Pang Z B, Wang K F, et al. High performance interconnect network for Tianhe system[J]. Journal of Computer Science and Technology, 2015, 30

- (2): 259.
- [2] Fu H, Liao J, Yang J, et al. The sunway taihu light supercomputer: system and applications[J]. Science China Information Sciences, 2016, 59(7): 072001.
- [3] Hwang K, Wang C, Wang C L. Evaluating MPI collective communication on the SP2, T3D, and Paragon multicomputers[C]//High-Performance Computer Architecture, Third International Symposium on. San Antonio, Texas, IEEE, 1997: 106-115.
- [4] Scott S, Thorson G. Optimized routing in the Cray T3D[J]. Parallel Computer Routing and Communication, 1994(853): 281-294.
- [5] Leiserson C E. Fat-trees: universal networks for hardware-efficient supercomputing[J]. IEEE transactions on Computers, 1985, 100(10): 892-901.
- [6] Kim J, Dally W J, Towles B, et al. Microarchitecture of a high radix router[C]//Computer Architecture, 2005. ISCA'05. Proceedings. 32nd International Symposium on. Dresden, Germany, IEEE, 2005: 420-431.
- [7] Pongor G. Omnet: Objective modular network testbed[C]//Proceedings of the International Workshop on Modeling, Analysis, and Simulation On Computer and Telecommunication Systems. Erice, Italy, Society for Computer Simulation International, 1993: 323-326.
- [8] Kung N T, Morris R. Credit-based flow control for ATM networks[J]. IEEE network, 1995, 9(2): 40-48.
- [9] Fulgham M L, Snyder L. Performance of chaos and oblivious routers under non-uniform traffic[R]. University of Washington, Seattle WA, Technical Report UW-CSE-93-06-01, 1993.

作者简介:

张毅 男, (1992-), 硕士研究生. 研究方向为高性能计算机仿真器设计. E-mail: vcheungyi@163.com.

何卫锋 男, (1976-), 副研究员. 研究方向为低功耗可重构多核处理器结构、计算机系统结构与微体系结构.