

Paleyfly: 一种可扩展的高速互连网络拓扑结构

雷 斐 董德尊 庞征斌 廖湘科 杨明英

(并行与分布处理国家重点实验室(国防科学技术大学) 长沙 410073)

(leifei@nudt.edu.cn)

Paleyfly: A Scalable Topology in High Performance Interconnection Network

Lei Fei, Dong Dezun, Pang Zhengbin, Liao Xiangke, and Yang Mingying

(National Laboratory for Parallel and Distributed Processing (National University of Defense Technology), Changsha 410073)

Abstract High performance interconnection network is one of the most important parts in high performance computing system. How to design the topology of interconnection networks is the key point for the development of larger scale networks. Therefore, we contribute a new hierarchical topology structure Paleyfly (PF), which not only utilizes the property of strong regular graph with Paley graph but also supports the continued scale like Random Regular (RR) graph. Compared with other new high performance interconnection networks, Paleyfly can solve the problems of the scalability of Dragonfly (DF), the physical cost of Fat tree (Ft), the wiring complexity and the storage for routing table of Random Regular and so on. Meanwhile, according to the property of strong regular graph for load-balanced routing algorithm, we propose four routing algorithms to deal with congestion. Finally, through the simulation we briefly analyze the performance of Paleyfly comparing with other kinds of topologies and different routing algorithms. Experimental results show that our topology can achieve better effect compared with Random Regular under the various scales of network and different traffic patterns.

Key words topology structure; scalability; Random Regular (RR); Dragonfly (DF); Paley graph

摘 要 高速互连网络是高性能计算系统的重要组成部分. 随着网络规模需求的扩大, 如何搭建更大规模的网络是高速互连网络拓扑结构设计的关键. 因此, 提出一种新型层次化的拓扑结构 Paleyfly(PF), 其结合了 Paley 图强正则的特性和 Random Regular(RR)图支持任意规模大小的特点. 相比其他新型高速互连网络拓扑结构, Paleyfly 能够有效解决在路由芯片端口数受限的背景下, Dragonfly(DF)可扩展性受限、Fat tree(Ft)物理成本高、RR 结构物理布局难、路由表规模大等问题. 同时, 根据强正则属性在路由策略上负载均衡的优势, 提出了 4 种路由策略来解决网络的拥塞问题. 最后, 通过模拟器实验比较分析 PF 结构与其他拓扑结构及 PF 结构不同路由策略的性能, 验证了 PF 结构在不同规模以及不同通信模式配置下网络延迟优于 RR 结构.

关键词 拓扑结构; 可扩展性; 随机正则图; Dragonfly; Paley 图

中图法分类号 TP302.1

收稿日期: 2015-02-16; 修回日期: 2015-05-04

基金项目: 国家“八六三”高技术研究发展计划基金项目(2013AA014301, 2013AA01A208); 全国优秀博士学位论文作者专项资金项目(201450); 国家自然科学基金项目(61272482)

随着串行传输技术的发展,路由芯片引脚数不断增加,使用高阶路由芯片搭建低延迟、低能耗、高吞吐率的高性能互连网络拓扑结构是今后高性能计算系统的主流,如Cray BlackWidow^[1]系统是第1个采用64个端口的高阶胖树互连网络系统,Cray Casscade系统则采用48个端口的Aries ASIC路由芯片搭建Dragonfly高阶互连网络系统^[2],IBM的PERCS系统也采用的是48个端口的POWER7路由芯片搭建Dragonfly高阶互连网络系统^[3]以及连续4次TOP500排名第1的天河-2超级计算机采用的是高阶胖树互连网络结构.高阶路由芯片不仅能够提供丰富的链路使得拓扑结构设计灵活多样而且通过增加节点间的互连密度提高网络通信带宽和减少网络延迟.但是,路由芯片端口的增长速率远远跟不上网络规模的增长速率,国际半导体技术蓝图ITRS在2011年就已指出未来10年内,受限于芯片的边界带宽和芯片能耗,引脚数量和每个引脚带宽的增长速率将会十分缓慢,这直接影响了路由芯片的端口数^[4].尽管有研究者提出不需要转接口,光信号直接用芯片转发以避免芯片边界工艺的限制.但是,这种光互连技术利用额外的部件仍然不能克服边界工艺的限制,而且受到电互连及成本的约束.

在当前的高阶互连网络的拓扑结构中,Dragonfly(DF)^[5]拓扑结构是未来高速互连网络拓扑结构的热门选择^[6].DF拓扑结构的分层特性可以有效利用电互连和光互连的特点在保证低延迟的网络性能的同时还降低了成本开销.同时,DF拓扑结构相比其他高阶互连网络Flattened Butterfly(FB)^[7],Fat tree(Ft)^[8]等拓扑结构,如图1所示,在使用同样端口数的路由芯片搭建网络时,DF结构可以支持更大规模的网络.其中,FB结构采用的是三维网络,Ft则采用的是3层标准胖树结构.但是,一旦路由芯片端口数确定,DF所能搭建的最大规模也随之确定,如图2所示,路由芯片端口数为64,

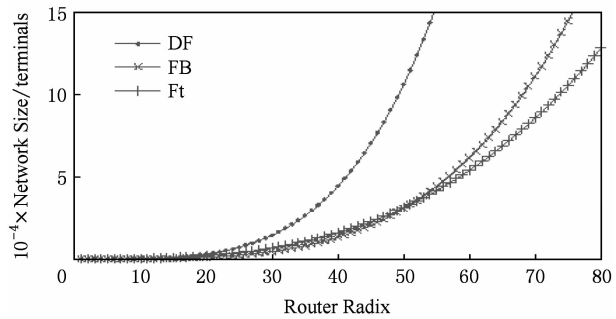


Fig. 1 Network size and radix for DF, FB and Ft.

图1 DF,FB和Ft规模与端口数的关系

每个路由芯片连接的终端数 p 与全局链路数 h 相等且终端数 p 等于局部组内路由芯片的个数 a 的一半时,网络规模达到最大.

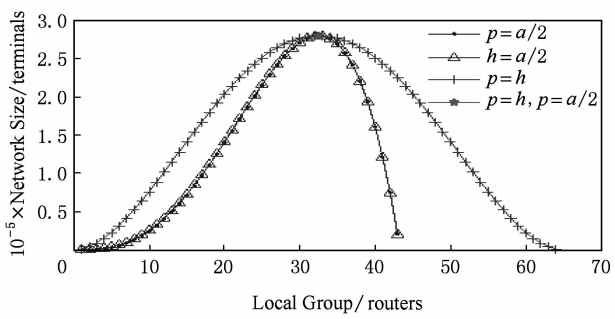


Fig. 2 Network size and local group for DF with different ratios among parameters.

图2 DF的规模和局部组大小在不同参数比值下的关系

有研究者提出利用随机拓扑结构取代规则拓扑结构来搭建高速互连网络使得网络的可扩展性更加灵活,如数据中心网络中的Jellyfish结构^[9]、小世界(small world)结构^[10],高性能计算系统中Random Shortcut Topology^[11],Layout-conscious Random Topology^[12],Skywalk^[13]等.但是,随机拓扑结构的不确定性和链路布局的复杂性以及路由表可扩展的局限性都给随机拓扑的实际应用带来困难.有研究者提出利用异构网络或者近似随机的规则图来替代随机拓扑结构实现低延迟、低能耗、高吞吐率的高速互连网络,如Penaranda等人提出的新型混合网络^[14],Singla等人提出的High Throughput Data Center Topology^[15]以及Besta等人提出的Slim Fly结构^[16].

另外,国内一些研究组针对数据中心网络中服务器网络端口数受限的情况,平衡了可扩展性和网络性能等因素,设计出“雪花结构”^[17]、Dcent^[18]以及CH^[19]等递归化层次拓扑结构,有效解决可扩展性和吞吐率的问题.

本文在层次化拓扑结构的基础上,提出一种新型可扩展的高速互连网络拓扑结构Paleyfly(PF),该结构在不改变路由芯片端口数的前提下,可以有效构建相比DF结构更大规模的网络,而且在路由策略上根据Paley图的强正则特性可以有效避免拥塞的路由策略,并在结构中引入随机的概念,使得PF降低了随机结构链路布局的复杂性以及路由表可扩展的局限性.该结构可以满足高性能计算系统在实际物理布局的需求,以层次化的方法划分机柜内的电互连链接方式以及机柜间光纤通信的链接方式.实验结果表明,与当前具有代表性的高速

互连网络拓扑结构相比, 该结构有效支持多种通信模式下更大规模的网络同时具有良好的物理布局性.

1 相关工作

1.1 DF 拓扑结构

2008 年 Kim 等人提出 DF 结构^[5], 其不仅突破了传统 k 元 n 立方体中维的概念, 用层次化全互连的方法进行拓扑互连, 而且有效利用了高阶路由芯片的特性和光缆与电缆在成本与距离上的差别, 使得性能指标提升. DF 拓扑结构具体构造是将多个高阶路由器全互连成一个超级节点, 然后在超级节点的层次上再进行一次全互连, 如图 3 所示, 每个超级节点内部有 5 个路由节点, 每个路由节点引出 2 条全局链路链接其余超级节点并连接 2 个终端. DF 结构实际上是一个 3 级的层次结构, 第 1 级是单个路由节点, 第 2 级是多个路由节点全互连而成的超级节点, 第 3 级则是多个超级节点全互连而成的 DF (p, a, h) . 每个路由节点端口链接类型分成 3 类: 1) 链接 p 个终端的端口; 2) 链接在同一个超级节点内其余 $a-1$ 个路由器的端口; 3) 链接 h 个超级节点的端口. 因此, 每一个路由节点的度数为 $k = p + a + h - 1$.

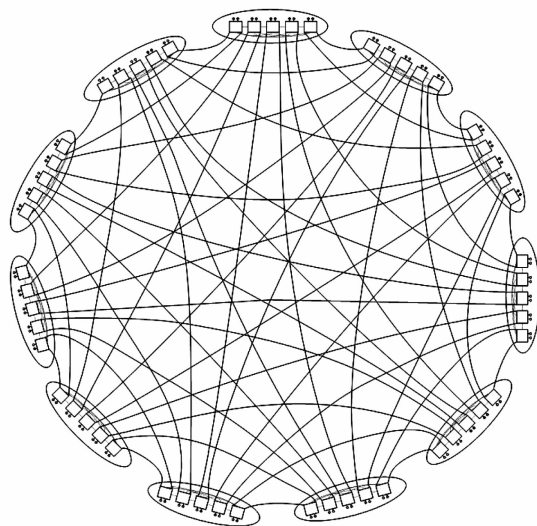


Fig. 3 DF(2,5,2).

图 3 DF(2,5,2)

DF 网络最大可支持 $N = ap(ah + 1)$ 个终端, 网络中有 $g = ah + 1$ 个超级节点, 任意 2 点之间都有一条全局链路相连, 每个报文转发最多只要经过 2 条本地链路和一条全局链路, 则任意 2 个终端通信最多只要经过 3 跳即可. 若要做到网络负载均衡, 参

数配比最好满足 $a = 2h = 2p$, 因为全局链路较长, 采用成本较高的光纤, 本地链路的数量要大于全局链路的 2 倍才能保证全局链路的利用率. 如图 2 所示, 引言也分析了 DF 结构使用端口数为 64 的路由芯片在不同的参数配比关系下规模的走势. 当 $a = 2h = 2p$ 时, 网络规模达到最大值. 因此, 当路由芯片端口数一定时, DF 结构所能搭建的最大规模也随之确定. 但是, 随着 E 级计算的到来, 路由芯片引脚技术的工艺限制, DF 结构已不能满足网络规模的飞速增长, 如何应对这个挑战才是下一代高性能互连网络拓扑结构设计的关键. Cray XC30 为了解决 DF 可扩展性问题, 在局部组内采用的是 16×6 的 FB 结构, 每个路由芯片引出 10 条全局链路, 局部组内每 4 条全局链路捆成一束链接一个超级节点, 系统中一共有 241 个超级节点, 网络规模达 92 544 个终端. 另外, DF 结构全局网络中任意 2 个超级节点有且只有一条全局链路相连, 容易造成全局链路拥塞, Jiang 等人提出用广播的方式在超级节点内部通知全局链路的拥塞信息来判断是否走最短路径路由^[20]. 而 Marina 等人发现一些通信模式下除了易造成全局链路的拥塞, 还容易造成超级节点内部本地链路的拥塞, Marina 等人提出逃逸子网^[21-22]、偏转模型^[23]的方法解决 DF 链路拥塞问题. 虽然通过广播、逃逸子网等方式可以延缓甚至控制拥塞, 但是真正的解决办法还是在于拓扑结构的性质和路由策略的选择.

1.2 随机拓扑结构

自 Shin 等人在文献[10]中提出使用小世界网络搭建数据中心网络后, Singla 等人^[9]也相继提出将随机拓扑 Jellyfish 应用在数据中心上, 保证每个路由芯片度数一致的前提下路由芯片之间随机互连, 即随机规则 Random Regular(RR)图. 如图 4 所示, RR(13, 6)图中有 13 个节点, 每个节点的度数最多为 6, 节点之间随机互连. 当新的节点加入网络中, Jellyfish 去掉已有链路, 分别与新加入的节点建立新连接. 不仅可以不受路由芯片端口数的影响连续扩展网络规模, 而且 Jellyfish 与 Ft 结构相比, 在实现同样二分带宽情况下 Jellyfish 结构的物理开销可以减少至 Ft 结构的一半左右. Koibuchi 等人在文献[11]提出的 Random Shortcut Topology 将随机拓扑应用在超级计算机系统传统互连网络拓扑结构上, 分析比较在 DLN(distributed loop networks)^[24], Mesh, Torus, Hypercube 以及 FB 结构上添加随机链路的性能, 添加随机链路后的 DLN 结构相比其他结

构在多种通信模式下性能更优. 随机结构的随机特性支持任意规模并有效减少了网络平均路径长度和网络直径,从而有效降低了网络延迟. 但是,随机链路的长度是不可控制的,这对采用随机结构的实际系统的物理布局是一个挑战. Koibuchi 等人在文献[12]提出的 Layout-conscious Random Topology 以及 Fujiwara 等人提出的 Skywalk 结构^[13]都针对随机拓扑结构的物理布局作进一步优化. Layout-conscious Random Topology 通过控制链路的长度重新分布机柜间的连线; Skywalk 结构则在建立网络初始化时就将随机链路分布在三维的机柜布局中,既保证了随机拓扑的随机特性又控制了缆线长度的大小. 但是,随机拓扑的路由策略仍然是网络规模可扩展性的瓶颈. 因此,如何更好地结合随机图和规则图的优点来设计拓扑结构、解决路由策略和物理布局以及可扩展性的问题,是未来拓扑设计的一个重要课题. 引言也提到有研究者提出使用异构网络或者近似随机图的规则图构建网络, Penaranda 等人提出的新型混合网络^[14]就是把 Ft 结构嵌入到 Mesh 结构中; Singla 等人提出的 High Throughput Data Center Topology^[15]指出未来的网络只有异构网络结构才能满足高吞吐率的网络需求; Besta 等人提出的 Slim Fly 结构^[16]近似节点度相等; 网络直径为 2 或 3 的 Moore bound 图^[25]即用有限的端口搭建更大规模网络直径为 2 或 3 的网络.

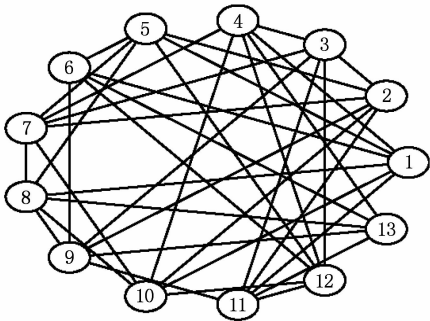


Fig. 4 RR(13,6).
图 4 RR(13,6)

2 Paleyfly (PF) 拓扑结构

下一代高性能互连网络拓扑结构不仅要求网络直径小、能够支持更大规模的拓扑结构,而且要求物理开销低、网络性能高. 但是,在路由芯片端口数固定的情况下,难以满足上述 2 个要求. 例如使用端口数为 64 的路由芯片搭建网络,标准 3 层 Ft 结构

使用 5120 个路由芯片搭建 65 536 个终端,DF 结构使用 16 416 个路由芯片支持最大规模 262 656 个终端,而三维 FB 则使用 4 096 个路由芯片支持 65 536 个终端. 虽然,相比 Ft 结构和 FB 结构,DF 结构在扩展性上更突出其优势;但是,如果需要使用 64 个端口的路由芯片搭建更大规模的网络或者支持连续扩展,DF 结构就不能满足需求. 因此,借鉴 DF 结构层次化的构造思想同时考虑路由芯片端口数的限制以及光电通信的差别,提出一种新型可扩展的高速互连网络拓扑结构 PF. PF 结构结合了 Paley 图的强正则、DF 结构的层次全互连以及随机图连续可扩展特点,有效满足网络直径小、网络规模大、连续扩展以及负载均衡的路由策略等需求.

2.1 Paley 图

Paley(p,n)是一类特殊的图,其中 p 是一个素数, n 是一个正整数,且满足 $p^n \equiv 1 \pmod 4$. 用 $P(V,E)$ 表示 Paley(p,n),点集 $V(P) = F_{p^n}$,边集 $E(P) = \{\{x,y\} : x,y \in F_{p^n}, x-y \in (F_{p^n}^*)^2\}$ ^[26]. F_{p^n} 是 $q = p^n$ 元有限域,由 $q = p^n$ 个元素组成. 构造 F_{p^n} (即 F_q) 需要先找到 F_q 的生成元 ξ . F_q 的所有非零元素都可以用 $\xi^i (i \in \mathbb{N})$ 表示^[27]. 计算生成元 ξ 的算法目前只能通过遍历的方式获得,如 $q=13$,则 $\xi=2, F_{13} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$,得出 $(F_{13}^*)^2 = \{1, 3, 4, 9, 10, 12\}$. Paley(13,1) 的点 x 与 $V(P)$ 中 $x+1, x+3, x+4, x+9, x+10$ 以及 $x+12$ 相连, $E(P) = \{\{x, x+1\}, \{x, x+3\}, \{x, x+4\}, \{x, x+9\}, \{x, x+10\}, \{x, x+12\}, \forall x \in F_{13}\}$, 如图 5 所示. 若 $q=9$,则 $p=3, n=2, F_9 = \{0, 1, 2, a, 2a, 1+a, 1+2a, 2+a, 2+2a\}$,得出 $(F_9^*)^2 = \{1, 2, a, 2a\}$, 如图 6 所示. 除了用乘法的构造方法构造 q 元有限域之外,还可以通过加法的构造方法构造 q 元有限域. Paley(p,n) 图不仅是对称图而且是自补图,即 Paley(p,n) 图与其补图同构. 除此之外, Paley(p,n) 图还具备一个很重要的特性——强正则特性,图中 $q = p^n$ 个节点,每个节点的节点度数一致,度为 $(q-1)/2$,任意 2 个相邻节点共享 $(q-5)/4$ 个公共邻居节点,任意 2 个不相邻的节点共享 $(q-1)/4$ 个公共邻居节点,因此,该图的网络直径为 2. Paley(p,n) 图可以表示为 $(q, (q-1)/2, (q-5)/4, (q-1)/4)$, 则 Paley(13,1) 可表示为 (13, 6, 2, 3). 图 5 中点 1 与点 8 虽然不相邻,但是可以通过点 4、点 5 以及点 11 中转到达,点 1 和点 2 虽然相邻,但是也可以通过点 5 和点 11 中转到达. 相比 Paley(p,n) 图,完全图 K_q 的

节点度为 $q-1$, 任意 2 个节点共享 $q-2$ 个公共邻居节点, 则表示为 $(q, q-1, q-2)$.

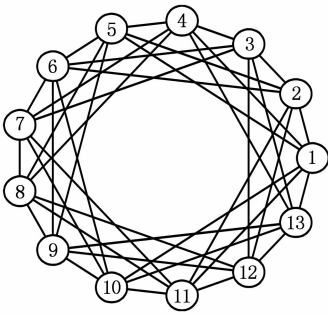


Fig. 5 Paley(13,1).
图 5 Paley(13,1)

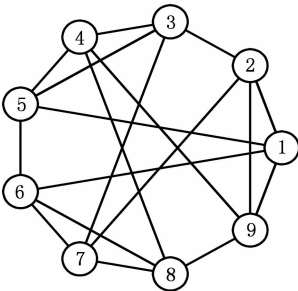


Fig. 6 Paley(3,2).
图 6 Paley(3,2)

2.2 基本结构

PF 结构是一个层次化的结构, 分成路由节点层和超级节点层, 具体参数见表 1. 路由节点层, 每个路由节点除了链接 p 个终端外, 全互连属于同一个超级节点内的其余 $a-1$ 个路由节点, 并引出 h 条全局链路链接其他超级节点内的路由节点. 因此, 每个路由节点的端口数 $k=a+h+p-1$. 超级节点层, 每个超级节点就像是一个虚拟的超高阶路由节点, 端口数为 $a(h+p)$, 引出 ah 条全局链路链接 ap 个终端. 为了充分利用电互连距离短成本低的特点, 每个

超级节点内的所有路由节点采用电信号传输全互连相连, 而超级节点构成的全局网络则采用距离长成本高的光纤传输.

PF 结构的全局网络是一个可以根据网络规模的实际需求重新配置的结构, 全局网络的基本结构为 Paley(p, n) 图同时复合一个 RR 图. 根据网络规模的需求, 每个超级节点引出的 ah 条全局链路中 $ah-d$ 条全局链路用于链接 Paley(p, n) 图, 剩余 d 条用于链接 RR 图. Paleyfly 结构最大规模可达 $N=(2ah+1)ap$ 个终端, 此时, 每个超级节点引出的全局链路全部使用在构造 Paley(p, n) 图上, 有 $q=(2ah+1)$ 个超级节点且满足 $q \equiv 1 \pmod{4}$, 每个超级节点引出的随机链路的个数 $d=0$, 网络直径为 2. $d=ah/2$ 表示 PF 结构的全局网络采用的是全互连结构, 此时, PF 结构等价于 DF 结构, 全局网络超级节点的个数 $q=(ah+1)$, 网络直径为 1. 当 $0 < d < ah/2$ 时, 全局网络超级节点的个数 $q=2(ah-d)+1$, 网络中的链路分成 2 部分, 一部分用于链接规则图 Paley(p, n) 图, 另一部分则用于链接 RR 图. 用 PF(q, a, h, p, d) 表示 PF 结构, 如图 7 所示, PF(17, 4, 2, 2, 0) 结构的全局网络有 17 个超级节点, 每个超级节点内有 4 个路由节点, 每个路由节点链接 2 个终端和引出 2 条全局链路.

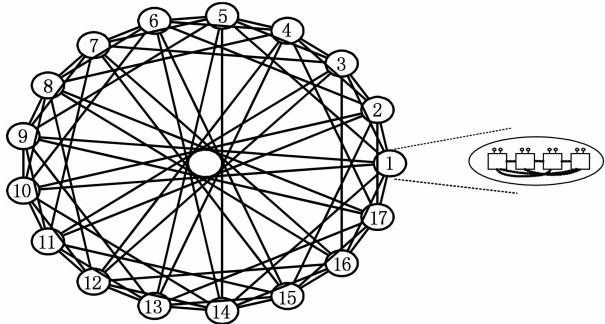


Fig. 7 PF(17,4,2,2,0).
图 7 PF(17,4,2,2,0)

2.3 参数分析

当 $0 < d < ah/2$ 时, PF 结构的规则图部分采用 Paley(p, n) 结构, 一方面是为了在有限的路由芯片端口数下网络支持更大的规模; 另一方面则是 Paley(p, n) 结构的强正则属性. Paley(p, n) 结构保证了全局网络的网络直径最大为 2, 而且, 无论是相邻的超级节点还是不相邻的超级节点都有数目只相差 1 的公共邻居个数, 这给自适应路由算法提供了多条路径选择, 有效缓解网络中的拥塞和负载均衡.

Table1 Parameters of PF
表 1 PF 的参数

Parameters	Descriptions
N	Number of Terminals in the Whole Network
q	Number of Supernodes in the Whole Network
d	Number of Random Links in a Supernode
k	Router Radix
a	Number of Routers in a Supernode
p	Number of Terminals in a Router
h	Number of Global links in a Router

图 8 所示为 PF 结构的可扩展性与每个超级节点随机链路数量的关系,相比 DF 结构,PF 结构调整随机链路的数量可以连续扩展网络规模.当网络规模不需要达到 PF 结构所能搭建的最大规模时,即 $0 < d < ah/2$,每个超级节点的剩余 d 条链路用来搭建 RR 图,不仅降低网络平均最短路径的长度,而且 RR 图和 Paley(p,n)图相结合有效解决了随机图路由策略扩展性的问题,缩小了随机图节点公共邻居的范围,达到均衡网络负载的目的.如表 2 所示,数据均为全局网络拓扑结构的参数.从表 2 可以看出,101 个超级节点,节点度为 60 的 PF 和 RR,PF 全局网络由 Paley(101,1)图和节点度为 10 的 RR 图复合而成,两者的全局链路数量和平均最短路径长度相差不大,但是,Paleyfly 的公共邻居数范围更加集中使得任意 2 个超级节点通信路径数范围收敛,在进行路由决策时各节点通信负载更加均衡;241 个超级节点,节点度为 120 的 PF 和 RR,PF 全局网络达到最大规模,拓扑结构为 Paley(241,1)图,相比 RR(241,120),在其他参数值相当的情况下 PF 相

邻节点或是不相邻节点的公共邻居数确定唯一;DF 的全局网络虽然在网络直径、平均最短路径以及公共邻居数上有明显的优势,但是对节点度数以及链路数要求过高,难以满足实际的需求.图 9 和图 10 分别是 PF(17,5,2,2,2)的全局网络和 RR(17,10),2 个结构的超级节点个数相同,节点度也相同,PF(17,5,2,2,2)的每个超级节点有 2 条随机链路.

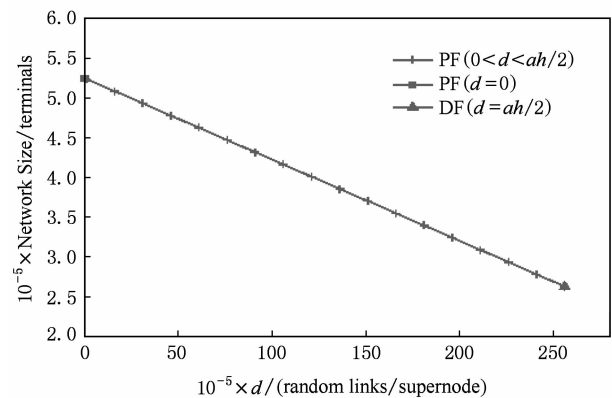


Fig. 8 Scalability of PF.
图 8 PF 的可扩展性

Table 2 Parameters of PF,RR and DF
表 2 PF,RR 和 DF 全局网络参数

Parameters	RR		PF			DF	
	RR	RR	PF	PF	PF	DF	DF
	(101,60)	(241,120)	(101,10,5, $p,0$)	(101,10,6, $p,10$)	(241,12,10, $p,0$)	($p,10,10$)	($p,20,12$)
Number of Supernodes	101	241	101	101	241	101	241
Degree of a Supernode	60	120	50	60	120	100	240
Random Degree of a Supernode	60	120	0	10	0	0	0
Number of Global Links	3 011	14 412	2 525	3 029	14 460	5 050	28 920
Mean Minimal Path Length	1.403 8	1.501 7	1.5	1.446 9	1.5	1	1
Diameter	2	2	2	2	2	1	1
Common Neighbours for Adjacent Nodes	21-44	40-76	24	24-37	59	99	239
Common Neighbours for Non-Adjacent Nodes	21-44	39-78	25	25-37	60		

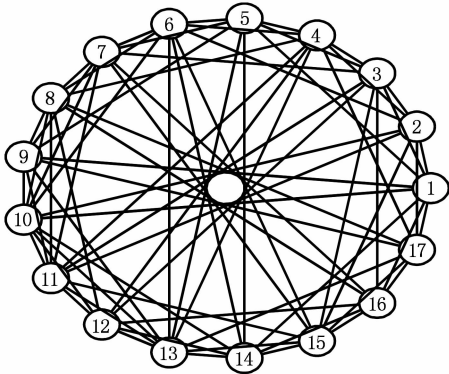


Fig. 9 The global network of PF(17,5,2,2,2).
图 9 PF(17,5,2,2,2)的全局网络

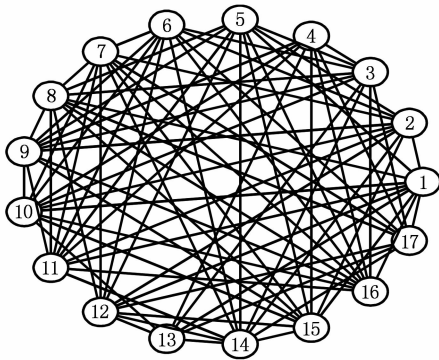


Fig. 10 RR(17,10).
图 10 RR(17,10)

图 11 对 PF,DF,FB 和 Ft 四种结构的可扩展性和二分带宽 (bisection bandwidth, BB) (假设每条链路传输速率为 10 Gbps) 进行了比较,FB 结构和 Ft 结构分别是三维和 3 层的结构. PF 结构相比其他结构,其可扩展性是 4 个结构中最好的,BB 则跟三维的 FB 结构和 DF 结构的性能一样, $BB \approx N/4$; 而 Ft 结构的 $BB \approx N/2$, 如图 11(b) 所示.

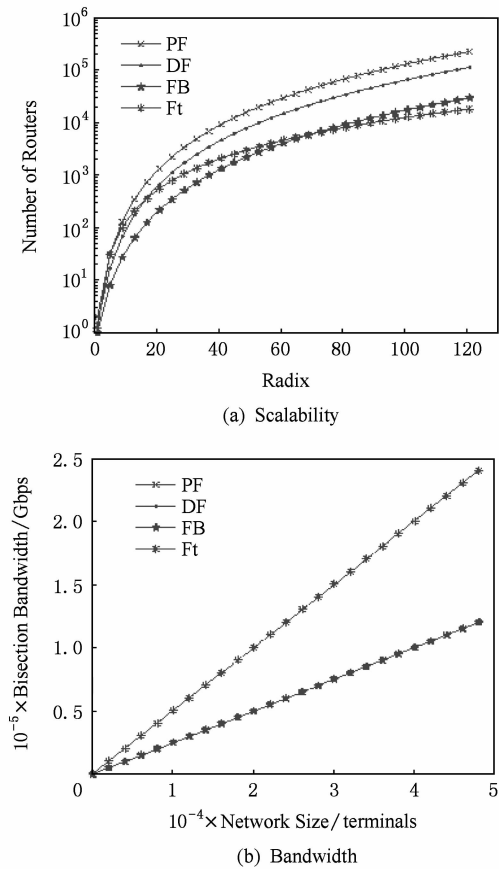


Fig. 11 Comparison of scalability and BB.
图 11 PF,DF,FB 和 Ft 结构可扩展性和二分带宽比较图

3 路由策略

本节将介绍 PF 结构的路由策略,结合 Paley (p,n) 图的特点,设计了 4 种路由策略,分别是最短路径路由策略 (min)、随机最短路径路由策略 (ranmin)、非最短路径路由策略 (nonmin) 以及自适应非最短路径路由策略 (adpnonmin).

在 PF 结构上实现的最短路径路由策略,最多需要经过 2 跳全局链路和 3 跳局部链路,源终端 T_s 在超级节点 G_s 中链接源路由节点 R_s . 到目标终端 T_d 在超级节点 G_d 中链接目标路由节点 R_d 的路由一般要经过以下 3 个步骤完成:

步骤 1. 如果 $G_s \neq G_d$ 且 R_s 乃至整个超级节点 G_s 都没有全局链路直接到达 G_d , 则顺序找 G_s 和 G_d 的公共相邻超级节点 G_i 作为中介超级节点,到达 G_i 中的 R_x , 执行步骤 2;

如果 $G_s \neq G_d$, 但是 R_s 或者 G_s 中其他路由节点有全局链路直接到达 G_d , 则通过 R_s 所连的全局链路或通过 G_s 中路由节点 R_a 的全局链路到达超级节点 G_d 中的 R_b , 直接跳转到步骤 3;

步骤 2. 如果 R_x 没有直接链接 G_d 的全局链路, 则通过本地链路到 G_i 中与 G_d 有全局链路的 R_y , R_y 通过全局链路链接 G_d 中的 R_b , 执行步骤 3;

步骤 3. 如果 $R_b \neq R_d$, 则在超级节点 G_d 内 R_b 通过本地链路到达 R_d .

图 12(a) 所示, 超级节点 3 跟超级节点 12 通信, 只需要经过 1 跳全局链路. 图 12(c) 介绍了超级节点 3 和超级节点 12 通信的整个过程, 为了避免死锁, 使用了 2 个虚通道 VC_0 和 VC_1 . 在图 12(b) 中超级节点 3 跟超级节点 10 通信, 2 个节点没有通过直接的全局链路相连, 因此需要借由超级节点 6, 7 或 13 转发, 经过 2 跳全局链路. 如图 12(d) 所示, 超级节点 3 选择超级节点 13 作为中介节点, 使用了 3 条虚通道避免死锁, VC_0 和 VC_1 用于本地链路的第 1 跳和第 2 跳以及全局链路的第 2 跳和第 1 跳, VC_2 用于本地链路的第 3 跳.

随机最短路径路由策略与最短路径路由策略的区别就在于, 2 个不相邻的超级节点通信时是随机选取公共邻节点而不是按顺序寻找公共邻节点. 这样的做法可以有效避免一些节点重复作为中介节点而影响网络性能. 如在图 12(b) 中, 随机最短路径路由策略每次选取超级节点 3 和超级节点 10 的中介节点产生的结果可能都不一样.

非最短路径路由策略和最短路径路由策略的区别在于, 即使 2 个超级节点相邻, 也要选取中介节点进行转发. 因此, 只要 2 个超级节点通信都需要经过 2 跳全局链路. 例如图 12(a) 中超级节点 3 跟超级节点 12 通信, 通过随机选取超级节点 2 或超级节点 13 转发. 无论是相邻的 2 个超级节点或是不相邻的 2 个超级节点都采用随机的方式选取中介节点.

自适应非最短路径路由策略不同于非最短路径路由策略, 前者在源路由节点的时候根据当前的网络状态判断走最短路径路由策略还是非最短路径路由策略, 网络状态根据下一级端口的缓冲占用率决定. 无论是相邻的 2 个超级节点或是不相邻的 2 个

超级节点,中介节点是根据缓冲占用率选择的.同时,为了避免缓冲区占有率最小的节点成为拥塞瓶

颈即同时多个报文选择该节点进行转发,在进行路由选择的时候引入随机因子选择转发节点.

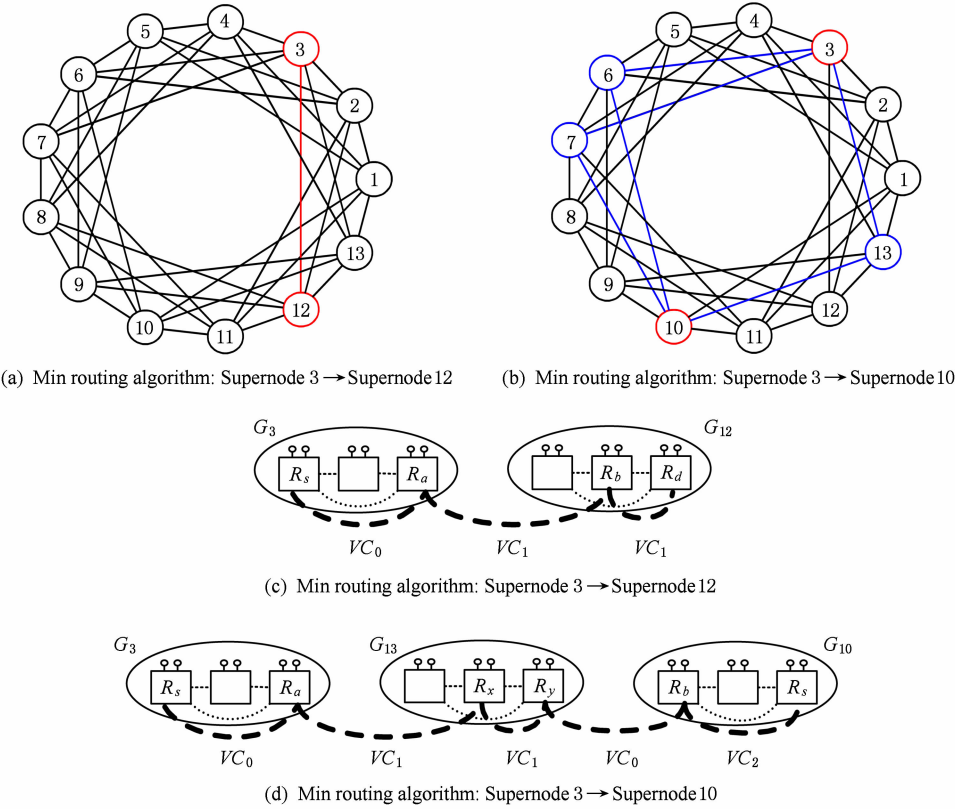


Fig. 12 Min routing algorithm of PF.

图 12 PF 结构最短路由策略

4 性能分析

本节将对 PF 拓扑结构从 2 个方面进行性能分析:1)PF 结构与 DF 结构、RR 结构在不同规模 and 不同通信模式下比较分析;2)在不同通信模式下,对第 3 节介绍的 min, ranmin, nonmin 以及 adpnonmin 路由策略进行性能比较.

性能分析采用的是一个基于 C++ 开发的微片级时钟精确的网络模拟器 Booksim. 该模拟器已被广泛地使用在各种网络环境的研究中,包括大型的高性能计算系统和多核处理器,同时,也适用于网络各方面的研究,如:网络拓扑结构、路由策略、流控机制等^[28]. 在 Booksim 中,通信报文按 Bernoulli 分布注入网络,采用 Bernoulli 分布的目的在于更贴近实际系统的报文注入情况. 服从 Bernoulli 分布的注入是一个随机注入,每个时钟周期的注入概率是所设定的注入率. 路由芯片的微体系结构采用的是输入队列结构. 模拟过程分为预热期、稳定期和排空期.

为获得精确的数据,采集的均是基于稳定期多次重复实验结果的平均值. 本节的性能比较主要从网络延迟性能指标分析不同的拓扑结构和不同的路由策略,因此,所采集的网络延迟数据要等待所测试的报文全部排空才能统计数据. 为了不受不同流控机制的影响(如虫孔路由、虚跨步路由),注入网络的报文均是单微片的报文. 每个路由芯片的缓冲区大小一致,均为 256 个微片深度. 缓冲区管理机制采用的是共享缓冲区模式. 路由结构内信用延迟设为 2 个时钟周期,虚通道分配延迟设为 1 个时钟周期,交换分配延迟设为 1 个时钟周期,链路延迟设为 1 个时钟周期,路由芯片内部链路加速比设为 2,输入/输出的加速比设为 1.

图 13~15 分别评测了在不同通信模式下 3 种拓扑结构不同规模的网络延迟. Uniform 通信模式是网络中任意节点之间随机通信;Asymmetric 通信模式是将网络终端根据编号大小分成均等的 2 部分,2 部分内的终端互相通信,也就是 2 部分节点互相通信,同一部分的节点不互相通信;Worst-case 通信

模式则是根据拓扑结构特点选择距离每个节点最远的节点进行通信。3 种拓扑结构采用的路由策略均为最短路径路由策略。参加评测的 PF 结构和 RR 结构的网络规模大小一致,分别是 1 830 个终端、1 025 个终端以及 666 个终端;DF 结构采用的则是考虑相

等规模大小的或是使用相同端口数的路由芯片以及 与 PF 网络相同参数配比的结构,规模分别为 1 056 个终端、930 个终端(约为 1 830 个终端的 PF 网络的一半且参数配比一致)以及 342 个终端(约为 666 个终端的 PF 网络的一半且参数配比一致)。

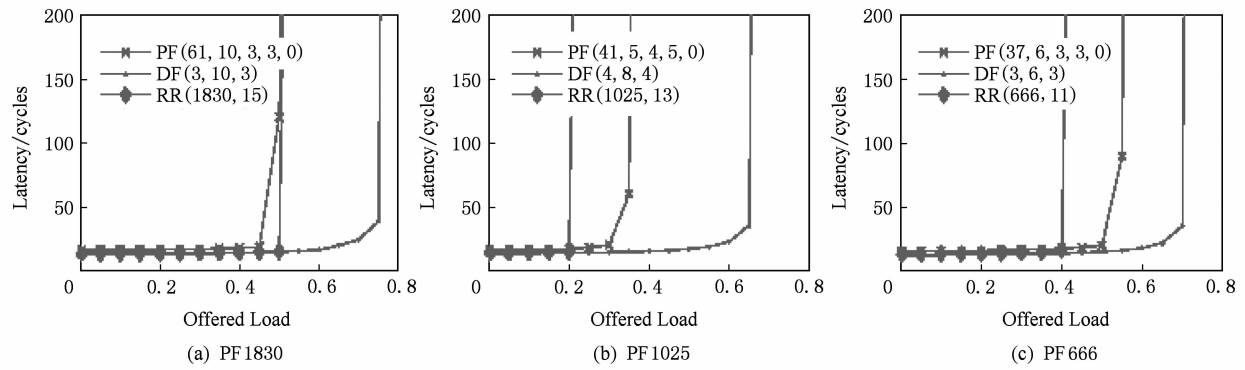


Fig. 13 Latency of PF,DF and RR under Uniform traffic patterns.
图 13 PF,DF 和 RR 网络在 Uniform 通信模式下不同网络规模的延迟

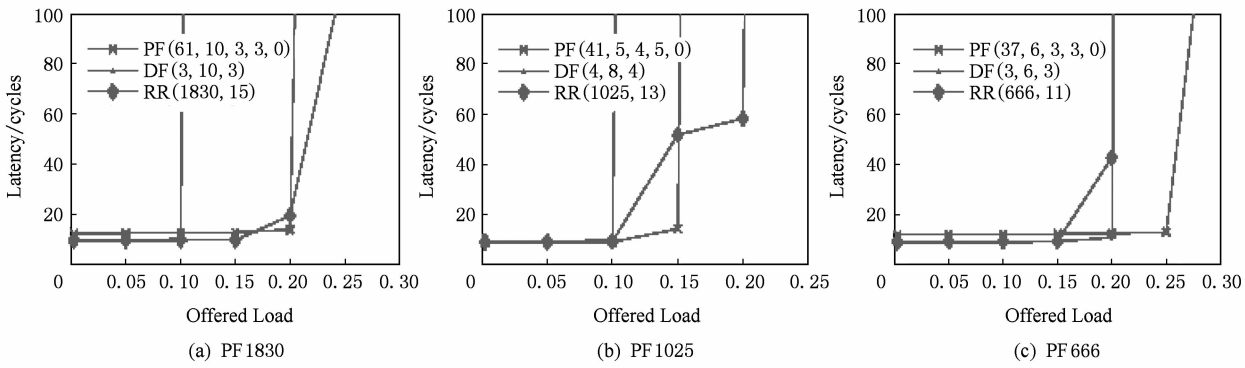


Fig. 14 Latency of PF,DF and RR under Asymmetric traffic patterns.
图 14 PF,DF 和 RR 网络在 Asymmetric 通信模式下不同网络规模的延迟

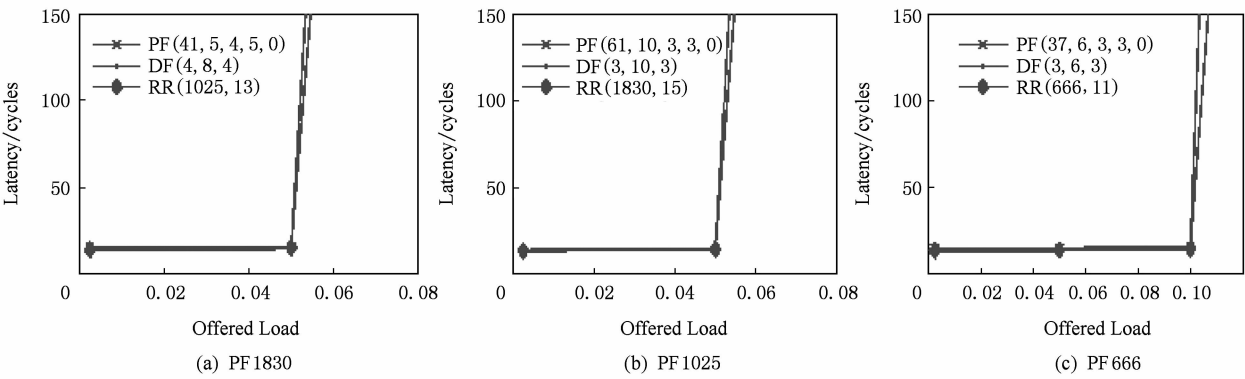


Fig. 15 Latency of PF,DF and RR under Worst-case traffic patterns.
图 15 PF,DF 和 RR 网络在 Worst-case 通信模式下不同网络规模的延迟

在 Uniform 通信模式下,DF 网络的延迟和饱和吞吐率是 3 个拓扑结构中最优的,如图 13 所示,主要原因是网络直径小而且平均最短路径小。但是,

在图 13(a)中,DF 网络使用的是端口数为 15 的路由芯片,其他 2 种结构的网络只需要使用端口数为 13 的路由芯片即可搭建相同规模的网络;图 13(b)

和(c)中,DF结构的参数配比满足负载均衡的要求,其在不同规模下的性能相差不大,而且规模只有同等参数配比PF结构的一半.同等规模下的PF结构和RR结构相比,PF结构的性能优于RR结构,尤其图13(b)(c)中PF结构的饱和吞吐率大于RR结构的饱和吞吐率.

在Asymmetric通信模式下,如图14所示,DF网络并没有展现其优势,RR结构和PF结构优于DF结构,主要原因是DF网络全局链路拥塞造成的.任意2个超级节点之间有且只有一条全局链路相连,网络平均划分两等分的终端互发消息,使得网络中链接两等分之间的全局链路集中拥塞影响网络性能.而RR结构和PF结构性能优于DF结构的原因在于,RR结构的链路是非确定性的,Asymmetric通信模式不会过于集中造成全局链路的拥塞;PF结构则是Paley图的强正则特性延缓全局链路的拥塞.

在Worst-cast通信模式下,如图15所示,3种

拓扑结构的性能都不好,主要原因是互相通信的终端过于集中,使得链路严重拥塞.

图16对PF结构上实现的4种路由策略进行了不同通信模式下的性能分析.模拟结果跟预测的一样,在3种通信模式下,ranmin路由策略通过对公共邻居采用随机顺序选择的方式明显优于顺序选择公共邻居的min路由策略.在Uniform通信模式下,ranmin路由策略的性能是最优的,adpnonmin次之,主要原因在于adpnonmin路由策略的判断条件不能及时反馈网络的实时状态造成误差.由于min策略是顺序选择公共邻居的,一些节点总是被当做中介节点导致链路拥塞,使得min路由策略比nonmin路由策略性能差.而在Asymmetric通信模式和Worst-case通信模式下,adpnonmin路由策略的性能最优,nonmin路由策略次之,主要原因是原本应该走最短路径的报文通过缓冲区的占有率或者确定性选取非最短路径的中介节点进行转发,一定程度上缓解了网络中的拥塞.

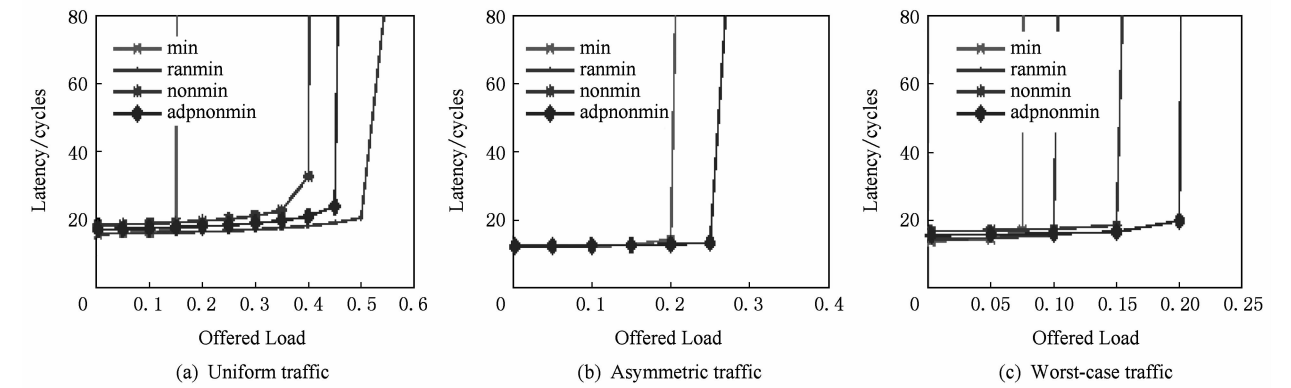


Fig. 16 Latency of PF with different routing algorithms under different traffic patterns.

图16 PF网络不同路由算法在不同通信模式下的延迟

adpnonmin路由策略适合多种通信模式,但是,如何及时获取网络的状态并作出正确的路由决策是该路由策略性能提升的瓶颈.ranmin路由策略相比min路由策略,通过随机选取非相邻的公共邻居节点有效提升网络性能,但是对于非均衡的通信模式部分链路拥塞使得网络性能难以提升;相反,采用nonmin路由策略的网络性能比ranmin路由策略的性能要好.

5 结 论

本文调研高速互连网络发展现状,结合芯片生产工艺的实际情况,对当前典型高速互连网络拓扑

结构的可扩展性进一步研究,并分析得出在路由芯片端口数无法持续增加的情况下可扩展性较优的DF结构支持的规模是最大的.本文结合Paley图强正则特性在路由策略上负载均衡的优势和RR图支持任意规模的特点,设计了一个新型层次化结构PF,有效解决了在路由芯片端口数受限的背景下DF结构可扩展性受限、Ft结构物力成本高、RR结构物理布局难、路由表规模大等问题,实现了连续可扩展网络规模、负载均衡路由策略等目标,并通过模拟器实验验证了PF结构及其路由策略的性能.但是,如何设置PF结构的参数配比和在不同通信模式下如何提高PF结构的饱和吞吐率,这也是PF结构亟需解决的问题,有待进一步的研究.

参 考 文 献

- [1] Scott S, Abts D, Kim J, et al. The blackwidow high-radix clos network [C] //Proc of the 33rd Annual Int Symp on Computer Architecture. Piscataway, NJ: IEEE, 2006: 16–28
- [2] Faanes G, Bataineh A, Roweth D, et al. Cray cascade: A scalable HPC system based on a dragonfly network [C] //Proc of the 2012 Int Conf for High Performance Computing, Networking, Storage and Analysis. Piscataway, NJ: IEEE, 2012: 1–9
- [3] Rajamony R, Arimilli L, Gildea K. PERCS: The IBM POWER7-IH high-performance computing system [J]. IBM Journal of Research and Development, 2011, 55(3): 1–12
- [4] Binkert N, Davis A, Jouppi N, et al. The role of optics in future high radix switch design [C] //Proc of the 38th Annual Int Symp on Computer Architecture. Piscataway, NJ: IEEE, 2011: 437–447
- [5] Kim J, Dally W, Scott S, et al. Technology-driven, highly-scalable dragonfly topology [C] //Proc of the 35th Annual Int Symp on Computer Architecture. Piscataway, NJ: IEEE, 2008: 77–88
- [6] John S, Sudip D, John M. Exascale computing technology challenges [C] //Proc of the 9th Annual Int High Performance Computing for Computational Science. Berlin: Springer, 2011: 1–25
- [7] Kim J, Dally W, Abts D. Flattened butterfly: A cost-efficient topology for high-radix networks [C] //Proc of the 34th Annual Int Symp on Computer Architecture. New York: ACM, 2007: 126–137
- [8] Leiserson C. Fat-trees: Universal networks for hardware efficient supercomputing [J]. IEEE Trans on Computers, 1985, 34(10): 892–901
- [9] Singla A, Hong C, Popa L, et al. Jellyfish: Networking data centers randomly [C] //Proc of the 9th Symp on Networked Systems Design and Implementation. Berkeley, CA: USENIX Association, 2012: 225–238
- [10] Shin J, Wong B, Sirer E. Small-world datacenters [C] //Proc of the 2nd Symp on Cloud Computing. New York: ACM, 2011: 2:1–2:13
- [11] Koibuchi M, Matsutani H, Amano H, et al. A case for random shortcut topologies for HPC interconnects [C] //Proc of the 39th Annual Int Symp on Computer Architecture. Piscataway, NJ: IEEE, 2012: 177–188
- [12] Koibuchi M, Fujiwara I, Matsutani H, et al. Layout-conscious random topologies for HPC off-chip interconnects [C] //Proc of the 19th Int Symp on High-Performance Computer Architecture. Piscataway, NJ: IEEE, 2013: 484–495
- [13] Fujiwara I, Koibuchi M, Matsutani H, et al. Skywalk: A topology for HPC networks with low-delay switches [C] //Proc of the 28th Int Parallel & Distributed Processing Symp. Piscataway, NJ: IEEE, 2014: 263–27
- [14] Penaranda R, Gomez C, Gomez M, et al. A new family of hybrid topologies for large-scale interconnection networks [C] //Proc of the 11th Int Symp on Network Computing and Applications. Piscataway, NJ: IEEE, 2012: 220–227
- [15] Singla A, Godfrey P, Kolla A. High throughput data center topology design [C] //Proc of the 11th Symp on Networked Systems Design and Implementation. Berkeley, CA: USENIX Association, 2014: 29–41
- [16] Besta M, Hoefler T. Slim Fly: A cost effective low-diameter network topology [C] //Proc of the 2014 Int Conf for High Performance Computing, Networking, Storage and Analysis. Piscataway, NJ: IEEE, 2014: 348–359
- [17] Liu Xiaoqian, Yang Shoubao, Guo Liangmin, et al. Snowflake: A new-type network structure of data center [J]. Chinese Journal of Computers, 2011, 34(1): 76–86 (in chinese)
(刘晓茜, 杨寿保, 郭良敏, 等. 雪花结构: 一种新型数据中心网络结构[J]. 计算机学报, 2011, 34(1): 76–86)
- [18] Zhu Guiming, Xie Xianghui, Guo Deke, et al. Dcent: A high extensible data center networking structure using dual-port servers [J]. Journal of Computer Research and Development, 2014, 51(5): 1009–1017 (in chinese)
(朱桂明, 谢向辉, 郭得科, 等. Dcent: 基于双网卡服务器的高可扩展数据中心网络结构[J]. 计算机研究与发展, 2014, 51(5): 1009–1017)
- [19] Lu Feifei, Luo Xingguo, Xie Xianghui, et al. Constant degree network for massively data center [J]. Journal of Computer Research and Development, 2014, 51(11): 2437–2447 (in chinese)
(陆菲菲, 罗兴国, 谢向辉, 等. 面向大规模数据中心的常量度数互连网络研究[J]. 计算机研究与发展, 2014, 51(11): 2437–2447)
- [20] Jiang N, Kim J, Dally W. Indirect adaptive routing on large scale interconnection networks [C] //Proc of the 36th Annual Int Symp on Computer Architecture. New York: ACM, 2009: 220–231
- [21] Marina G, Beivide R, Odriozola M, et al. On-the-fly adaptive routing in high-radix hierarchical networks [C] //Proc of the 41st Int Conf on Parallel Processing. New York: ACM, 2012: 279–288
- [22] Marina G, Beivide R, Valero M, et al. OFAR-CM: Efficient dragonfly networks with simple congestion management [C] //Proc of the 21st Annual Symp on High-Performance Interconnects. Piscataway, NJ: IEEE, 2013: 55–62

[23] Marina G, Beivide R, Odrizola M, et al. Efficient routing mechanisms for dragonfly networks [C] //Proc of the 42nd Int Conf on Parallel Processing. Piscataway, NJ: IEEE, 2013; 582-592

[24] Bermond F, Hsu D. Distributed loop computer networks: A survey [J]. Journal of Parallel and Distributed Computing, 1995, 24(1): 2-10

[25] Miller J. Moore graphs and beyond: A survey of the degree/diameter problem [J]. The Electronic Journal of Combinatorics, 2005, 14: 1-61

[26] Elsayw A. Paley graphs and their generalizations [D]. Düsseldorf, Germany: Master of Science, Heinrich Heine University, 2009

[27] Lidl R, Niederreiter H. Finite fields: Encyclopedia of mathematics and its applications [J]. Computers & Mathematics with Applications, 1997, 33(7): 136-136

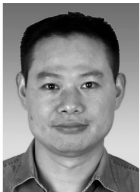
[28] Jiang N, Becker D, Michelogiannakis G, et al. A detailed and flexible cycle-accurate network-on-chip simulator [C] // Proc of the 2013 Int Symp on Performance Analysis of Systems and Software. Piscataway, NJ: IEEE, 2013; 86-96



Lei Fei, born in 1988. PhD candidate. Her current research interests include high performance interconnection network.



Dong Dezun, born in 1980. PhD, associate professor. Senior member of China Computer Federation. His current research interests include high performance interconnection network (dong @ nudt.edu.cn).



Pang Zhengbin, born in 1972. PhD, professor. Senior member of China Computer Federation. His current research interests include parallel and distributed computing, and high performance computer architecture.



Liao Xiangke, born in 1963. MS, professor. Senior member of China Computer Federation. His current research interests include computer architecture.



Yang Mingying, born in 1989. MS candidate. His current research interests include high performance interconnection network.