



FALB: 一种 FC 协议链路聚合算法

王兆辉, 沈剑良, 张霞, 陈艇

(国家数字交换系统工程技术研究中心, 河南 郑州 450003)

摘要: 链路聚合是一种通过将多条物理链路捆绑成一条逻辑链路来增加带宽的常用技术, 但是在基于光纤通道 (fibre channel, FC) 协议的交换网络中, 随着转发流量区域的不断集中, 传统静态链路聚合算法存在负载均衡性差、可靠性低的问题。因此提出了一种基于 FC 协议的动态负载均衡 (flow adaptive load balancing, FALB) 链路聚合算法, 可以实现对流量的动态自适应匹配以及断链可重连保护机制。结果表明, 该算法机制相较于静态负载均衡的算法设计, 可以将有效带宽比最终收敛到 95% 以上, 有效提升聚合链路的负载均衡特性和可靠性。

关键词: 光纤通道; 链路聚合; 负载均衡; 可靠性; 带宽

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2021215

FALB: a FC protocol link aggregation algorithm

WANG Zhaohui, SHEN Jianliang, ZHANG Xia, CHEN Ting

China National Digital Switching System Engineering & Technological R&D Center, Zhengzhou 450003, China

Abstract: Link aggregation is a common technology to increase the bandwidth by bundling multiple physical links into a logical link. However, in a switching network based on the fibre channel (FC) protocol, as the forwarding traffic continues to concentrate, traditional static link aggregation algorithms have problems of poor load balance and low reliability. A dynamic load balancing link aggregation algorithm flow adaptive load balancing (FALB) based on FC protocol messages was proposed, which could realize dynamic adaptive matching of traffic and a reconnection protection mechanism for broken links. The results show that compared with the static load balancing algorithm design, the algorithm mechanism could maintain the effective bandwidth ratio above 95%, effectively improving the load balancing characteristics and reliability of the aggregated link.

Key words: FC, link aggregation, load balancing, reliability, bandwidth

1 引言

光纤通道 (fibre channel, FC)^[1] 协议是美国

国家标准委员会定义的高速串行通信协议, 由于其传输距离远、抗干扰能力强等特点而广泛应用在航空航天环境下的电子指挥、控制、仪器仪表、

收稿日期: 2021-03-22; 修回日期: 2021-07-05

基金项目: 国家科技重大专项资助项目 (No.2016ZX01012101)

Foundation Item: The National Science and Technology Major Project (No.2016ZX01012101)



仿真、信号处理、传感器以及视频数据分发等^[2]领域,具有高可靠、高容错和高带宽的性能指标需求,而链路聚合技术是在不改变原设计架构基础上增加通道有效带宽的一种常用方案。

链路聚合是将离散数量的多条物理链路聚合为一条逻辑链路的技术,这是在 Internet 服务提供商 (Internet service provider, ISP) 网络中实现高带宽连接的一种广泛使用的方法^[3]。当不存在更高带宽的技术时,通过捆绑几条带宽较小的链路实现带宽增加是一种广泛用于高速交换芯片的数据链路技术应用^[4],该技术可以提升数据包传输的可伸缩性和可靠性,大幅降低额外开销。同时为了实现成本和公平性的平衡,必须在可用链路上平均分配数据包,否则数据包过度集中在少量链路上,进而造成链路有效通道利用效率低下,导致传输时延增加,同时拥塞链路端点处的缓冲区溢出也可能导致数据包丢失^[5]。

本文针对基于 FC 的交换芯片设计中链路聚合负载均衡、负载不均衡以及可靠性差的问题,提出一种动态负载均衡 (flow adaptive load balancing, FALB) 链路聚合算法,该设计可以在更加复杂多变的网络环境中提升对各种聚合链路的负载均衡调度灵活性、增加实际可用带宽、提高链路聚合技术的整体功效和系统稳定性。

2 相关工作

链路聚合技术由于是对子链路的“捆绑重组”,在流量传输时如何对负载合理分配是研究的核心难点。在配置好一组链路聚合组之后,每条子链路的流量负载不均衡可能会导致部分链路拥塞,同时也会造成整体聚合链路的实际可用带宽过低。为了避免这种情况,系统需要在多条链路上执行负载的均衡分配^[6],提升负载均衡算法的均衡性和可靠性是优化聚合链路设计的主要研究方向^[7]。

(1) 均衡性

IEEE 802.3ax 标准的链路聚合控制协议 (link

aggregation control protocol, LACP) 是关于动态链路聚合配置的协议^[8],它通过协议报文链路聚合控制协议数据单元 (link aggregation control protocol data unit, LACPDU) 和相连的设备交互信息。当端口启用 LACP 后,端口通过发送 LACPDU 给 CPU,上报优先级、系统 MAC 地址、端口的优先级、端口号和操作指令等信息^[9-10],然后 CPU 再根据这些端口和链路信息对数据包路由传输动态调度,但是该协议仅支持以太网协议传输的设备,如何在基于 FC 协议的交换设备中通过 LACP 进行动态链路聚合还少有相关机构研究。而 TFRC (TCP-friendly rate control) 算法是一种针对视频流数据传输的调度算法^[11],该算法使用传输速率作为链路拥塞控制的基础,并以此来反映链路质量的动态变化。通过系统实时获得的链路传输速率,可以做到对各链路的动态调度。可以有效地提高链路带宽并监控链路状态。但是由于该算法仅针对视频流数据,在拥塞控制的基本单元选择上不适用于 FC 交换网络。

(2) 可靠性

Akbari 等^[12]提出的 STreLo 算法是用于在聚合链路存在的情况下定位链路故障,该算法使用分段路由,通过发送周期探测包来定位链路的故障位置,保证链路的可靠性,具有毫秒级的定位速度和可扩展性,但同时也存在功耗大、设计复杂的不足,在对于一些拓扑简单的聚合链路组中的实用性较低。Wenzel 等^[13]提出的单独的分发策略方案,可以在聚合链路部分失败的情况下通过引用备份链路来保障整个链路组的正常运作,具有高度的灵活性,但同时增加备用链路也意味着资源开销的增加。在目前实际工程设计中,链路聚合的主流设计还是采用哈希算法对子链路进行静态权重配置,随着技术的不断发展,对于聚合链路的实际可用带宽比要求不断提高,需要对传统静态负载均衡算法进行平衡性和可靠性优化。

本文根据 LACP 的动态交互和备份链路设计思想, 结合 FC 协议特点, 提出了动态负载均衡算法 FALB, 可靠性方面提出断链重配置机制, 保证在个别子链路出现掉链的情况时不影响整个链路组的数据传输, 均衡性方面支持流量自适应的负载均衡调度, 实验表明可以在多种流量转发模式下均收敛到 95% 以上的有效带宽比, 并通过了多种可靠性功能验证。

3 动态链路聚合设计

本节首先介绍基于 FC 协议的动态链路聚合整体设计思路, 然后介绍 FALB 负载均衡调度算法, 本文针对 FC 协议报文特点, 通过流量监控加哈希条目重分配的方式实现基于流量的动态负载均衡。

3.1 整体设计框架

由于需要对子链路的负载分配权重进行动态调整, 所以如何设计流量监控和负载均衡算法是核心要点。为了实现基于动态的负载均衡, 本设计采用流量监控反馈和链路权重计算的方式实时负载调整, 整体设计框架如图 1 所示。首先配置好整个链路聚合组的子链路通道端口号, 在未有流量通过时平均分配哈希条目权重, 在有流量通过链路组子链路进行路由传输时, 对每个子链路的通道入端口进行流量统计和状态监控, 并将每个端口的状态信息数据实时上传控制平面。

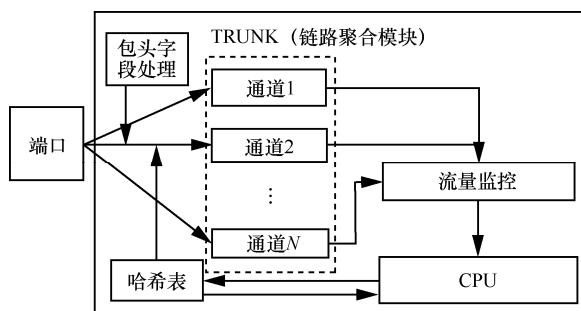


图 1 聚合链路组设计整体框架

在数据包进入聚合链路组后首先进行包头字段处理。FC 协议包头字段格式如图 2 所示,

FC-FS-4 协议规定最大可支持 2 148 bit (不含扩展帧头) 的 FC 协议报文。为了提取数据包路由指示信息, 保证数据包的传输序列不变, 本文中选取包头域段中的 D_ID 和 S_ID 作为哈希函数的计算输入值, 域段 D_ID 表示数据包的目的地址, S_ID 则表示数据包的源地址, 根据两组共 48 bit 位宽的域段信息可以将每个数据包的次序定位, 以保证不会出现数据包传输乱序的情况。在完成包头字段提取操作后将两组位宽 24 bit 的域段值进行拼接后作为每个数据包的哈希函数输入值, 哈希表共设置 256 个条目值作为每条子链路的映射地址, 通过对这 256 个条目的分配来决定每条子通道的负载分配权重。在流量监控方面采用对每条子链路端口的进包数进行统计并上传, 根据各子链路的负载情况进行调度重配置后实现对哈希输出条目映射域的动态控制。

Bits	Word	31—24	23—16	15—08	07—00
0	R_CTL	D_ID			
1	CS_CTL/ Priority	S_ID			
2	TYPE	F_CTL			
3	SEQ_ID	DF_CTL	SEQ_CTL		
4	OX_ID			RX_ID	
5	parameter				

图 2 FC 协议包头字段格式

3.2 调度算法 FALB

由于是哈希查表结果决定具体每条子链路的负载分配方案, 所以动态调度算法是整个聚合链路的核心设计。FALB 动态调度算法的设计是通过计算并更新每条子链路对应的哈希条目地址权重值来实现负载的动态均衡配置。这里首先计算每条子链路的端口负载速率, 记每条子链路的端口速率为 X_t , X_t 的计算式如式 (1) 所示:

$$X_t = \frac{S_t}{T} \quad (1)$$

其中, X_t 表示第 t 个子物理链路的端口负载速率, S_t 表示第 t 个子物理链路的端口负载长度, T 表示时钟周期。根据从控制平面中得到的端口速



率数值反馈, 权重调度算法可以得到每个子链路的哈希地址条目有效值 P_t , 其计算式如式(2)所示:

$$P_t = \frac{X_t \cdot R \cdot c}{\sum_{t=0}^n X_t} (0 \leq c \leq 1) \quad (2)$$

其中, R 代表总哈希条目值, c 代表敏感系数, 流量自适应的负载均衡效果受敏感系数的影响, 取值为 1 时为完全动态负载均衡(默认为 1), 根据每时刻的 P_t 值对哈希查表的结果条目值映射域变化实现动态的负载均衡控制。具体流程如图 3 所示, 首先提取 S_ID 和 D_ID 的输入值, 然后根据式(2)计算每条子链路的哈希有效条目值(P_t), 当有流量通过入端口时就会触发流量监控模块, 此时会将各子通道的流量数据上传 CPU 并重新计算 P_t 值, 如果不变则不改变子链路通道的哈希条目值映射范围, 如果 P_t 值改变则重新计算和分配哈希条目的地址映射。

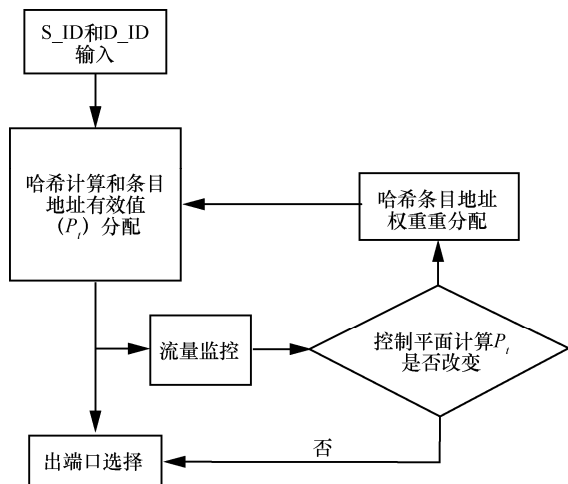


图3 流量监控调度流程

3.3 断链重配置机制

可靠性优化方面, 为了保证链路的稳定性, 设计了断链重分配机制, 如图 4 所示。每条子链路对应一段哈希地址的映射域。由于该调度算法可以实时监控每条子链路的流量情况, 所以当其中一条或几条链路出现故障时, 该机制可以使故

障链路的映射赋值使能关闭, 并将属于该链路的哈希映射地址平均分配给该链路组的其他成员, 同时根据目前已建链的子链路数重新分配哈希表的条目地址映射, 从而不影响整个链路组的数据传输。由于断链后仅仅是将问题链路的映射使能位关闭而其物理端口号仍属于原聚合链路组, 所以当故障排除完毕后, 可以立即恢复该链路的正常工作状态, 节省了二次配置的时间, 同时也提升了整体链路组的可靠性。

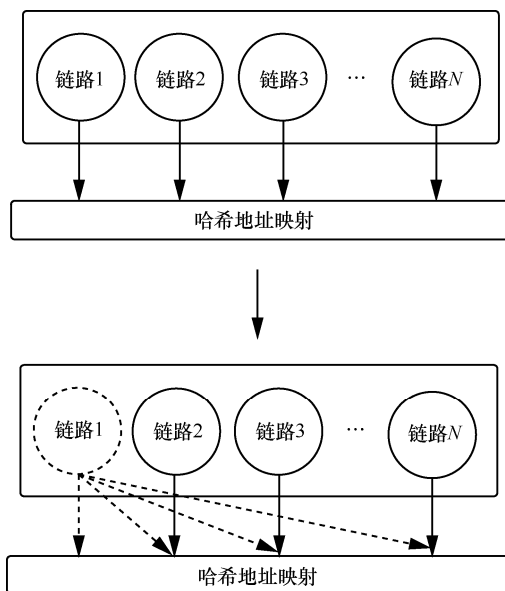


图4 断链重分配机制

4 实验结果分析

本文实验使用数据仿真加 FPGA 功能验证作为实验设计方案, 仿真使用 Vivado 验证平台对基础功能模块进行时序分析验证, FPGA 验证使用 Xilinx Virtex UltraScale440 的 FPGA 板卡作为本实验的环境开发板卡和 Xgig 光纤通道发包仪作为 FC 数据包发送设备来构建实验环境, 并对实验结果进行了分析。

4.1 仿真验证

数据仿真方面 Vivado 仿真验证平台是 FPGA 厂商赛灵思公司 2012 年发布的集成设计环境, 采用层次化器件编辑器和布局规划器设计可以用于

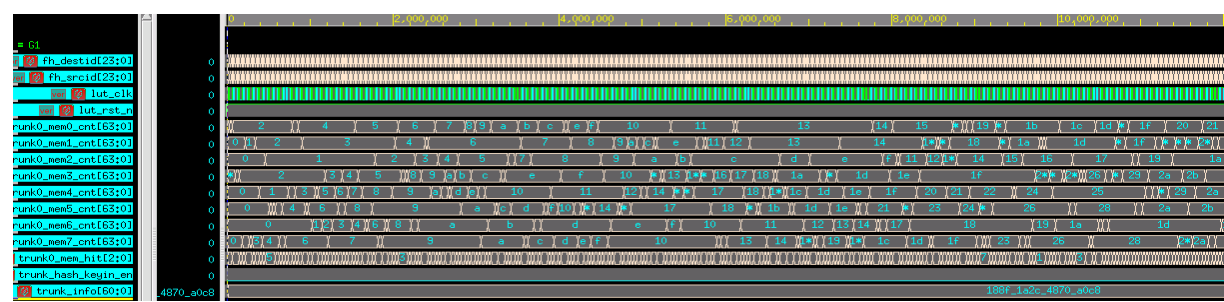


图 5 Vivado 仿真结果

表 1 流量测试模型

序号	FC 数据包参数设置	流量集中度	应用场景
模式 1	S_ID 和 D_ID 值在 0~255 随机	0%	空闲状态
模式 2	S_ID 和 D_ID 值在 0~191 随机	25%	部分节点拥挤
模式 3	S_ID 和 D_ID 值在 0~127 随机	50%	部分节点处于高峰期

快速综合和 verilog 模块级代码的逻辑功能仿真和时序分析,本实验以 Virtex 7vx980tffg1930-2L 为核心芯片,采用 312.5 MHz 的时钟频率进行时序仿真验证。

Vivado 仿真结果如图 5 所示,8 条子物理链路的聚合链路组配置后,设置随机类型域特征的 S_ID 和 D_ID 值的包头,共配置 8 条子通道计数器(trunk0_num0_cnt0-7),每个计数器代表每条通路的实际流量情况。结果显示正常发送不同的 S_ID 和 D_ID 数值的 FC 数据包后,每条子通路计数正常且符合预期,符合时序功能要求。

4.2 均衡性验证

由于链路聚合技术是将多条子链路整合为一条逻辑电路的技术,所以理论上整合后的链路组带宽应该为所有子链路带宽之和,但是实际工程应用中,链路组的实际带宽取决于其负载均衡特性,所以本实验分别设计 3 种流量发送模型,具体参数见表 1。通过缩小数据包的 S_ID 和 D_ID 取值范围,可以有效地增加流量的集中度(范围越小,代表流量转发端口区域越集中,均衡难度越大)。

建立如图 6 所示实验连接拓扑模型,将两块

440FPGA 开发板的两个 bank 对接后,设置 port0~port7 端口为一条聚合链路,并采用 8 GB/s(满负载)的速率发送数据包,计算每组 3 种模式的有效带宽比(实际有效带宽与理论带宽之比)。实验结果如图 7 所示,其中横轴为时间点,纵轴为聚合链路组的有效带宽比。

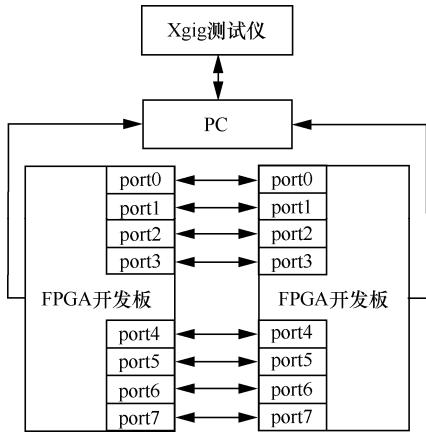


图 6 实验连接拓扑

根据实验数据结果可以看出,在发送测试数据流时两种算法的流量负载均衡效果区别明显,随着流量集中度的增加(模式 1~模式 3),传统静态负载均衡算法的均衡性不断下降,在模式 3 流量集中度到达 50%时,聚合链路组的有效带宽

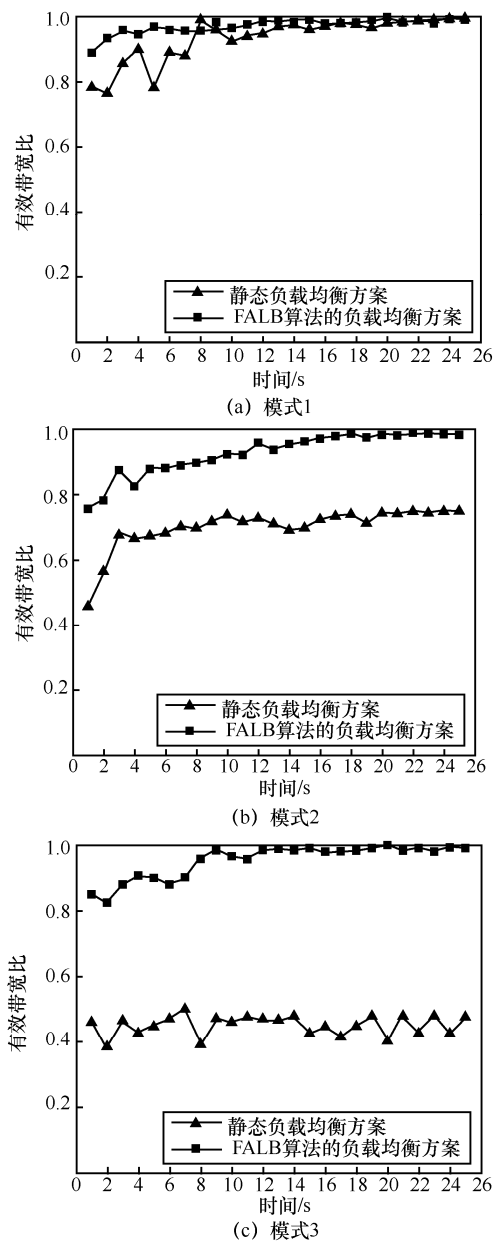


图7 FALB 负载均衡算法实验结果

比下降到 50%以下,可以看出随着流量集中度的增加,传统静态负载均衡算法的负载均衡性能在

不断下降,这是由于传统静态负载均衡算法无法根据实际流量情况动态调整各子链路的负载分配策略而导致负载分配不均,尤其当负载集中于部分子链路时由于无法及时调整分配策略从而导致有效带宽的浪费。而 FALB 在 3 种模型中均保持 80%以上且收敛于 95%的有效带宽比,且 3 种模式下的变化程度很小,说明该算法无论在高拥挤的还是分布更加发散的流量模型中均可以保持高效的负载均衡。实验结果表明 FALB 具有更好的负载均衡特性和稳定性。

4.3 可靠性验证

在实际工程中,链路的可靠性测试是检验在遇到常见故障时整体功能的完整性。本节中设置了 3 种测试方法,分别在 1 GB/s、2 GB/s、4 GB/s、8 GB/s 端口发包速率下进行热插拔(数据包正常发送过程中光纤端口的拔出和插入)、链路成员移除和链路成员增加的数据包发送可靠性测试,结果见表 2,在所有支持速率下进行的 3 种可靠性测试均可以正常工作,具有较高的可靠性。

5 结束语

本文提出一种动态负载均衡的 FALB 调度算法的链路聚合设计,并最终在 FPGA 上实现,该设计可以在保证链路可靠性的基础上实现基于流量自适应的负载均衡效果,支持断链重连,提升链路传输效率。在选择链路时,使用优化的动态平衡算法可以有效地避免子链路负载分配不均所造成的有效带宽比过低的问题。FALB 链路聚合算法在相比较于传统静态负载均衡方案具有更好

表2 可靠性测试

测试类型	1 GB/s	2 GB/s	4 GB/s	8 GB/s
热插拔测试	正常	正常	正常	正常
链路成员断链重连测试	剩余端口正常进行负载均衡传输	剩余端口正常进行负载均衡传输	剩余端口正常进行负载均衡传输	剩余端口正常进行负载均衡传输
传输过程中链路成员增加测试	新增端口与原配置端口进行负载均衡传输	新增端口与原配置端口进行负载均衡传输	新增端口与原配置端口进行负载均衡传输	新增端口与原配置端口进行负载均衡传输

的负载均衡效果, 可靠性更强, 更适用于当前流量分布不均、实际带宽要求更高的网络应用环境, 未来通过结合可编程协议解析技术, 还可以使该算法的调度选择域由用户自定义选择, 从而增加整体设计的安全性和灵活性。

参考文献:

- [1] 李攀, 田泽, 蔡叶芳, 等. FC-AE-ASM 协议优化设计[J]. 计算机工程与科学, 2017, 39(2): 280-284.
LI P, TIAN Z, CAI Y F, et al. Optimization of FC-AE-ASM protocol[J]. Computer Engineering & Science, 2017, 39(2): 280-284.
- [2] 徐士博. 基于 FC 协议的航电任务系统仿真[D]. 成都: 电子科技大学, 2018.
XU S B. Aeronautical mission system simulation based on FC protocol[D]. Chengdu: University of Electronic Science and Technology of China, 2018.
- [3] FOTIADIS P, SOLDATI P, LEGG P. Design and evaluation of LTE/WLAN traffic steering and link aggregation algorithms[C]//Proceedings of 2016 IEEE 84th Vehicular Technology Conference (VTC-Fall). Piscataway: IEEE Press, 2016: 1-6.
- [4] ARUN KUMAR S P, GOLASH M. Efficient flow-aware dynamic link load balancing[C]//Proceedings of 2009 First International Communication Systems and Networks and Workshops. Piscataway: IEEE Press, 2009: 1-6.
- [5] AL-FARES M, RADHAKRISHNAN B, RAGHAVAN N, et al. Hedera: dynamic flow scheduling for data center networks[C]//The 7th USENIX Conference on Networked Systems Design and Implementation, CA, USA: USENIX Association, 2010: 19-19.
- [6] KOROTKYI I, LYSENKO O. Hardware implementation of link aggregation in networks-on-chip[C]//Proceedings of 2011 World Congress on Information and Communication Technologies. Piscataway: IEEE Press, 2011: 1112-1117.
- [7] GÖZÜAÇIK N, NACAR B, YILMAZ E, et al. Increasing trunk capacity on central international soft-switch[C]//Proceedings of 2016 24th Signal Processing and Communication Application Conference (SIU). Piscataway: IEEE Press, 2016: 713-716.
- [8] IEEE. IEEE standard for local and metropolitan area networks-link aggregation: IEEE Std 802.1AX-2020[S]. 2020.
- [9] JUNIOR R R R, VIEIRA M A M, LOUREIRO A A F. Dynamic

link aggregation in software defined networking[C]//Proceedings of 2018 IEEE Symposium on Computers and Communications (ISCC). Piscataway: IEEE Press, 2018: 615-620.

- [10] BASKORO F, HIDAYAT R, WIBOWO S B. Comparing LACP implementation between ryu and opendaylight SDN controller[C]//Proceedings of 2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE). Piscataway: IEEE Press, 2019: 1-4.
- [11] HE Z H, HUANG J. Application of multilink aggregation and load balancing in wireless real-time video transmission system[C]//Proceedings of 2018 International Conference on Sensor Networks and Signal Processing (SNSP). Piscataway: IEEE Press, 2018: 277-282.
- [12] BASUKI A I, KUIPERS F. Localizing link failures in legacy and SDN networks[C]//Proceedings of 2018 10th International Workshop on Resilient Networks Design and Modeling (RNDM). Piscataway: IEEE Press, 2018: 1-6.
- [13] WENZEL G, HERMSMEYER C. A service aware protection overlay for Ethernet link aggregation[C]//Proceedings of 2007 ITG Symposium on Photonic Networks. [S.l.s.n.]. 2007: 1-8.

[作者简介]



王兆辉(1997-), 男, 国家数字交换系统工程技术研究中心硕士生, 主要研究方向为可编程协议解析、FC 交换芯片设计。

沈剑良(1982-), 男, 博士, 国家数字交换系统工程技术研究中心副研究员, 主要研究方向为新一代网络信息系统架构设计、大规模集成电路设计。

张霞(1980-), 女, 国家数字交换系统工程技术研究中心助理研究员, 主要研究方向为高速板卡设计、集成电路测试方法。

陈艇(1984-), 男, 博士, 国家数字交换系统工程技术研究中心助理研究员, 主要研究方向为大规模集成电路设计、高速交换电路设计、微处理器体系结构设计、高速数字信号处理。