

基于控制中心的新型 SAN 架构的设计与实现

陆一飞¹ 张震伟² 陶 军³ 唐 玲¹

¹(南京理工大学计算机科学与工程学院 南京 210094)

²(华为技术有限公司南京华为研究所 南京 210012)

³(东南大学计算机科学与工程学院 南京 210096)

(luyifei@njust.edu.cn)

Design and Implementation of New SAN Architecture Based on Controller

Lu Yifei¹, Zhang Zhenwei², Tao Jun³, and Tang Ling¹

¹(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094)

²(Huawei Nanjing Research Institute, Huawei Technologies Co. Ltd, Nanjing 210012)

³(School of Computer Science and Engineering, Southeast University, Nanjing 210096)

Abstract With the rapid development of Internet technology and the explosive growth of data, the storage system designed by tightly coupled hardware and software is limiting the development of storage technology severely, and unable to meet the fast changing needs in the mobile Internet and big data era increasingly. Software defined network (SDN) as new network architecture is more suitable for the development of the next-generation data centers. This paper proposes a new controller-based storage area network (SAN)—CSN architecture using the idea of SDN technology. CSN decoupling protocol control plane of FC switch from data plane deploys protocol control plane and distributed functions in the controller. At first, we introduce the architecture of CSN and discuss the specific design of controller for CSN in this paper. Then, the implement of communication and overall mechanism are described in detail. After that, CSN-based on-demand bandwidth-available first routing protocol is introduced and discussed. Finally, we verify the feasibility through the actual development environment. As a result of several experiments, server can establish a connection with the storage more quickly in CSN, and furthermore, CSN has better throughput and faster convergence, as well as more reliability and scalability. In addition, stress testing is conducted in the central controller for CPU, memory and bandwidth.

Key words FC protocol; storage area network (SAN); SAN architecture; software defined storage network (SDSN); controller

摘 要 随着互联网技术的快速发展,数据的爆炸式增长,存储系统的软硬件紧耦合设计严重地限制了存储技术的发展,也越来越无法满足移动互联网和大数据时代下对存储系统快速、多变的需求.软件定义网络(storage defined network, SDN)作为一种新的网络架构,更适合下一代数据中心的发展.在 SDN 思想基础上,提出了一种基于控制中心的新型 SAN 架构(controller-based SAN, CSN),该架构分离 FC 交换机的协议控制面和数据面,将 FC 协议的控制功能和分布式服务功能部署在控制器中,随后

收稿日期:2015-01-12;修回日期:2015-06-23

基金项目:国家自然科学基金项目(61370209);江苏省青年自然科学基金项目(BK20140797)

This work was supported by the National Natural Science Foundation of China (61370209) and the Natural Science Foundation for Youth of Jiangsu Province of China (BK20140797).

详细阐述了该架构的控制器设计、通信机制和整体机制的设计与实现. 在 CSN 基础上, 又提出了按需可用带宽优先路由协议. 最后, 通过实际开发环境验证该架构的可行性. 通过实验测试, 得出 CSN 下服务器能够更快地与存储服务器建立连接, 并且存储网络具有更高的吞吐量、更快的收敛性、更好的可靠性和可扩展性, 此外, 还对中心控制器的 CPU、内存和带宽进行压力测试.

关键词 FC 协议; 存储区域网络; 存储区域网络架构; 软件定义存储网络; 控制器

中图法分类号 TP393

随着 Internet 和网络技术的飞速发展, 特别是数据中心网络的兴起, 信息系统的数据呈爆炸式增长, 数据的高速访问和平滑简单的扩容要求日益迫切. 目前, 业界使用专门的存储区域网络(storage area network, SAN)^[1]支持服务器与存储设备之间的数据高速传输. SAN 一般使用光纤通道(fibre channel, FC)^[2]技术, 其拥有自动寻址、设备发现、状态变更通知等先进的存储机制, 为发起端设备(服务器系统)和目标端设备(存储器系统)之间的数据交换提供了便利. 一般使用 FC 的 SAN(FC-SAN)传输速度可以达到 4 Gbps, 8 Gbps 或者更高的 16 Gbps. 近年来基于以太网 FC 的 SAN(FCoE-SAN)^[3]也在快速地发展和应用. 本文只讨论 FC-SAN 下的相关研究.

文献[4]的相关研究指出: 有别于 TCP/IP 网络架构中的端节点协议复杂(各种控制协议)、网络操作简单(基于 IP 的转发), FC-SAN 架构的特点是端节点协议简单、网络操作复杂(自动寻址、设备发现、安全控制和分布式服务等). 因此, FC-SAN 的维护成本更高、配置更加复杂. 同时, SAN 中的关键技术 FC 协议实现复杂, 目前只有少数几家公司能够完全地实现 FC 协议, 如 IBM、思科和博科等. 由于这种技术壁垒的存在, FC-SAN 在发展的过程中存在一个较大问题: 尽管各个厂商的 FC 协议是兼容标准的, 但是上层的功能却无法完成很好的互操作, 不同硬件厂商的 FC 交换机不能互通. 而且, FC-SAN 的软硬件紧耦合设计严重限制了存储技术的发展, 也越来越无法满足移动互联网和大数据时代下对存储系统快速、多变的需求.

软件定义网络(software defined network, SDN)^[5]技术通过分离网络设备的控制面与数据面, 将控制功能从网络设备中分离出来, 形成集中化的控制中心(控制器). 网络设备维护流表结构, 数据分组按照流表进行转发, 而流表的生成、维护、配置则由中心控制器管理. 当前, 控制中心对交换机的控制主要通过 OpenFlow 协议^[6]实现.

SDN 技术为研发网络新应用和未来互联网技

术提供了一种新的解决思路. FC-SAN 架构下网络侧的功能较为复杂, 如果利用 SDN 技术把 FC-SAN 的控制面功能放到控制中心, 就能极大地减轻网络复杂度, 提高数据面的转发能力. 这种基于 SDN 的存储网络解决方案称为软件定义存储网络(software defined storage network, SDSN)^[7], 这个概念首先由 Jeda 公司在 2013 年的 SNIA^[8]开发者大会上提出, 随后 SDSN 思想迅速成为存储界关注的焦点.

本文在 SDN 思想基础上, 创新性地提出了将 FC 协议与控制器相结合的新型 SAN 架构: 基于控制中心的 SAN 架构(controller-based SAN, CSN). CSN 分离 FC 交换机的 FC 协议控制和数据转发功能, 将 FC 协议的控制机制和分布式服务部署在 FC 交换网控制器(FC fabric controller, FFC), 使存储网络具有灵活控制机制的同时, 还拥有高可靠性、扩展性以及更高的网络性能.

本文的主要贡献有 2 点:

1) 有别于传统 FC-SAN, 本文将 FC 协议的控制面协议和分布式服务在 FFC 中实现, 设计了控制器和 FC 交换机间的通信机制, 使存储网络具有更加的灵活性、可靠性和扩展性.

2) 利用 FPGA 开发板实现了 FC 协议和 CSN 架构, 并通过实验验证了 CSN 架构的优势.

1 相关研究

1.1 FC-SAN 协议

FC 协议是 SAN 的事实标准. FC 交换机通过 E 端口的分布式服务形成存储交换网, 而交换网又为整个网络提供名字服务(name server, NS)和分区服务(zone)等分布式服务, 如图 1 所示. 服务器和存储通过注册过程登录到交换网.

NPort 登录^[9]是指 NPort 端口注册到交换机和注册到 NS 过程. 一般来说, 该过程有 2 步: ① NPort 发送 FLOGI 到 FC 交换机, FC 交换机返回的 ACC(accept)报文, 该 ACC 报文中向 NPort 分配了一个

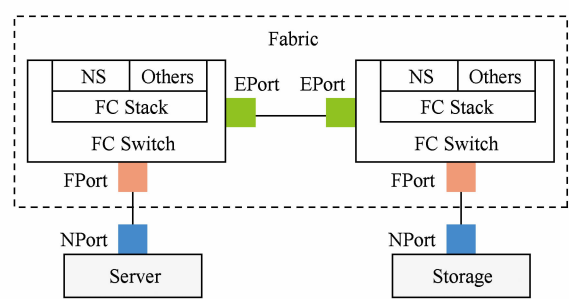


Fig. 1 The architecture of FC-SAN.
图 1 FC-SAN 架构

FC ID(FC identity),并规定链路间的服务类型等参数;②NPort 发送 PLOGI 报文到 FC 交换机,向 NS 注册登记.具体过程如图 2 所示:

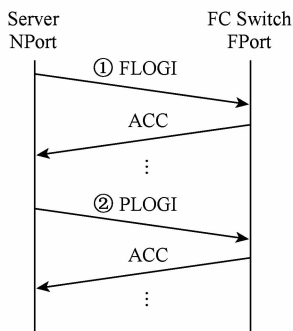


Fig. 2 Process of NPort login.
图 2 NPort 注册过程

交换网的配置^[10]是指 FC 交换机间的 EPort 连接时建立连接和生成路由的过程.交换网配置过程一般分为 5 步:EPort 端口间交换链路参数、协商交换机能力、主交换机选择、域 ID 分配以及 FSPF (fabric shortest path first)路由算法,如图 3 所示.另外还有些其他高级功能也会在交换网配置过程中实现,譬如 zone 的合并、分布式名字服务等.总体来看,交换网配置是一个漫长和复杂的过程,最终生成交换网的路由.

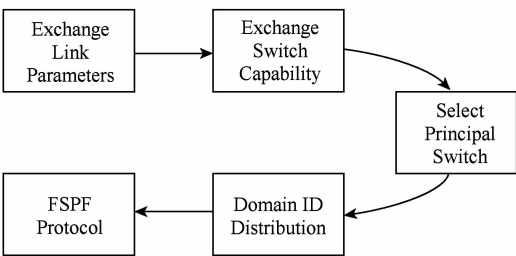


Fig. 3 Process of EPort protocol.
图 3 EPort 协议过程

在 FC 协议中,NPort 登录时服务器信息记录在 NS 中,在服务器间通信时需要通过分布式的查询服务获取对端的信息,所需的时间较长.同时,交

换网配置过程不仅所需的收敛时间较长,而且其中 FC 交换机一旦变化就需要重新进行主交换机选择等步骤,对网络性能影响较大.本文提出的 CSN 架构主要是优化这部分协议.

1.2 相关研究

文献[11]中分析了 FCoE-SAN 下的数据流都要通过 FCoE 转发者(FCoE forwarder, FCF),限制了 FCoE-SAN 的网络规模.因此在文献中提出了一种在网络边缘交换机分离 FCoE 数据帧的方法(AFCoE),数据帧分为用户面帧和控制面帧,数据面帧通过以太网交换机转发,而用户面帧被传输到控制服务器中处理.这种方法可以在一定程度上缓解 FCF 的压力,从而增加 FCoE-SAN 的规模,但是这种方法只把 FCoE 的控制面帧转移到控制器,用户面帧还是要经过 FCF,而控制帧只占有数据帧中的很小部分,因此该方法对于缓解 FCF 的压力比较有限.

Jeda 公司在 2013 年的 SNIA 开发者大会上提出基于 SDN 思想的 SDSN 方案^[12],该方案在 FCoE-SAN 网络中分离 FCoE 交换机的控制层和数据层,并使用软件交换网控制器(fabric network controller, FNC)替代物理 FCoE-SAN 中的网络控制层,建立更加灵活、可扩展性更强的动态存储网络.FNC 以 VMware 虚拟机的形式提供服务,与服务器、存储服务器和交换机交互以获取所需信息.另外,FNC 对边缘交换机添加 ACL 规则,以此优化数据流的转发,提高存储网络性能.

文献[13]在 FCoE-SAN 环境下,利用 SDN 思想对 FCoE 协议的部分控制进行分离,将其中的 FIP 报文转发到控制器,而其余的报文转发到 FCF.控制器通过分析 FIP 报文记录服务器的位置,然后对边缘以太网交换机下发新的转发表,优化数据流路径,数据流不再通过 FCF 处理.该文献也讨论了 OpenFlow 协议在此的作用.

以上的研究都聚焦于 SDSN 研究,但都只通过改变边缘交换机的转发表优化数据流,而且,只分离了 FCoE 协议中的部分控制报文(如 FIP 报文或部分控制帧).另外,目前的研究都聚焦于融合网络,即 FCoE-SAN,它的特点在于架构简单但是规模很大,这与 FC-SAN 有一定的区别.最后,以上的研究都没有对控制器的处理能力进行分析.

本文聚焦于基于 SDN 思想的 FC-SAN 研究,分离 FC 协议本身的控制面流量和数据流量,并且将 FC 协议和分布式服务在控制器中处理,以此简化 FC 协议的实现,建立更加灵活、可扩展性更强的存储架构.

2 基于控制中心的 SAN 架构

传统的 FC-SAN 架构下存在交换网配置复杂、网络收敛速度慢和各硬件厂商间的设备不兼容等问题. 同时,FC 交换机需要实现各种功能,如分布式名字服务、分布式分区功能和路由功能等,因此交换网中 FC 交换机的变更(功能增加、节点的增删)会对整个网络带来较大的变动,网络控制流量大量增加,

影响网络传输性能,网络管理异常复杂. 为了改善这些问题,本文基于 SDN 思想提出 CSN 架构.

2.1 CSN 整体架构

CSN 架构在传统的 FC-SAN 基础上增加了控制器 FFC,FFC 与 FC 交换机通过专门的管理网络(以太网)连接,且通过 TCP 协议进行通信,服务器和 FC 交换机间、FC 交换机间的链路不变,仍然使用 FC 链路.

CSN 架构如图 4 所示:

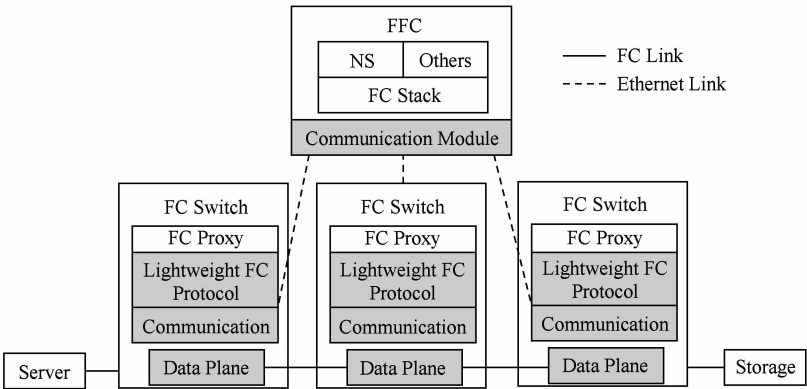


Fig. 4 The architecture of CSN.
图 4 CSN 架构

在 CSN 架构下,FC 交换机保留了原有的数据面处理功能、增加了 FC 代理模块,该模块包括了简化 FC 协议处理功能和通信功能. 简化 FC 协议功能负责解析 FC 交换机收到的 FC 控制报文,并根据需要配置交换机或者将报文转交给通信模块处理. 通信模块负责建立和维护与 FFC 的 TCP 链路,同时对 FC 控制报文进行封装和解封装.

答,下发到 FC 交换机,FC 交换机根据情况保留该应答报文或者转发给服务器.

2.2 控制器设计

FFC 作为 CSN 架构的核心,由软件实现并部署在专门的主机设备上. FC 协议栈、路由协议和分布式服务都要在 FFC 上实现,也即 FFC 是整个网络的控制中心.

FC 交换机将收到或自身产生的 FC 控制协议消息转发给控制器,控制器收到 FC 报文后进行应

答,下发到 FC 交换机,FC 交换机根据情况保留该应答报文或者转发给服务器.

FFC 的功能模块如图 5 所示,它的功能模块包括 TCP 通信模块、控制报文解析器、FFC 处理单元

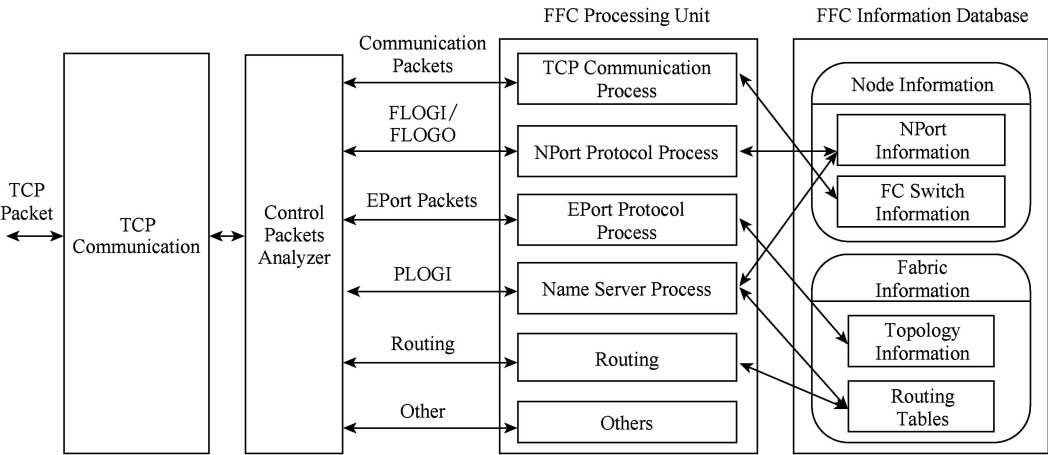


Fig. 5 The function modules of FFC.
图 5 FFC 功能模块

和 FFC 信息库. TCP 通信模块负责与 FC 交换机通信,它们之间通信的 TCP 报文中封装了控制信息和 FC 控制报文,因此,需要报文解析器对收到的 TCP 报文进行解析,并将解析后的报文递交给 FFC 处理单元. FFC 处理单元和 FFC 信息库是控制器的核心模块. FFC 处理单元包含多个子单元,分别处理解析器递交来的报文,如 TCP 通信处理子单元会处理 TCP 链路的建立、保活和关闭;NPort 协议处理子单元维护 FC 协议中的 NPort 登录和登出等控制信息;EPort 协议处理子单元维护 EPort 启动和关闭等控制信息;名字服务处理子单元为 FC 交换网中的名字服务器,记录了交换网中的所有 NPort 信息;路由处理子单元负责生成和维护交换网的路由. FFC 处理单元处理后的相关信息会保存至 FFC 信息库中,生成节点信息库和交换网信息库,其中的全局拓扑信息库为路由计算提供依据,最终形成路由表并下发给 FC 交换机.

2.3 CSN 通信机制

CSN 架构下,FFC 与 FC 交换机间通过以太网链路进行通信,它们间的通信根据功能不同,可以分为 2 类:FC 交换机和 FFC 间建立和维护 TCP 连接的通信;转发 FC 协议报文的通信.

在 FC 交换机启动时,与 FFC 间通过 3 次握手建立 TCP 连接,其中 FC 交换机端为客户端,FFC 为服务器端. 同时,为了维护 TCP 连接,FFC 会周期性向 FC 交换机发送保活报文. FC 交换机收到后,同样返回保活报文,当控制器或者 FC 交换机连续 3 次没有收到该报文时,表示对端失效,释放相关资源.

由于 FC 交换机与 FFC 间使用 TCP 协议,当 FC 交换机与 FFC 间传输 FC 控制报文时,需将把 FC 报文封装成 TCP 报文,当然,FC 交换机和 FFC 收到 TCP 报文时,也需要解析报文得到 FC 报文或其他控制报文. FC 交换机与 FFC 间的 TCP 报文封装格式如图 6 所示:

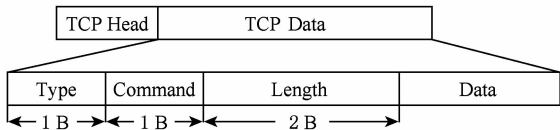


Fig. 6 TCP encapsulation format of FC packet.

图 6 TCP 封装 FC 报文格式

其中,Type 字段大小为 1 B,表示 TCP 报文中数据的类型. 目前,我们把 FC 交换机与 FFC 通信过程的报文分为 4 类:1)FC 交换机与 FFC 间的控制报文;2)NPort 端口登录时产生的控制报文;3)

EPort 开启关闭时产生的控制报文;4)控制器向 FC 交换机下发的流表报文.

Command 字段大小为 1 B,表示每种类型下的操作符,具体的含义如表 1 所示. Length 字段表示包括 Type 和 Command 字段在内的整个数据字段的长度,是以 4 B 为单位. Data 字段在不同 Type 和 Command 数值下有不同的格式,在此不再详述.

Table 1 Meaning of Type and Command Field

表 1 Type 字段和 Command 字段含义

Type	Command	Meaning
1	1	Keep-alive Packet from FC Switch to FFC
	2	Keep-alive Packet from FFC to FC Switch
	3	Login Request Packet from FC Switch to FFC
	4	Login Request ACK Packet from FFC to FC Switch
	5	Logout Request Packet
	6	Logout Request ACK Packet
2	1	NPort FLOGI Request
	2	NPort FLOGI Request ACK
	3	NPort FLOGO Request
	4	NPort FLOGO Request ACK
	5	NPort Control Packet
3	1	EPort UP Request Packet
	2	EPort UP Request ACK Packet
	3	EPort Down Request Packet
	4	EPort Down Request ACK Packet
4	1	Distribute Routing Table
	2	Distribute Routing Table ACK

目前,控制器只向 FC 交换机下发整个路由转发表操作,并没有区分增加流表、删除流表等操作,这部分内容会在以后的研究中改进.

从上面的通信机制描述可以看出,FC 交换机和 FFC 间的通信协议与 SDN 中的 OpenFlow 协议有些类似,都是在交换机和控制器间建立连接、维护链路,并且可以下发转发表(路由表),但是本文并没有使用成熟的 OpenFlow 协议,原因是:1)OpenFlow 协议中目前还没有控制存储数据流的方法;2)FC 协议中虽然有类似 IP 和 MAC 地址的概念,但这些字段的大小跟 IP 和 MAC 有差异,OpenFlow 的流表结构不适用于 FC-SAN;3)CSN 中的通信机制是基于 TCP/IP 的一个独立模块,这种方式更利于在传统的 FC 交换机上进行补丁升级,而不需要另外重新设计支持 OpenFlow 协议的设备.

2.4 CSN 整体机制的设计与实现

本节将详细描述 CSN 架构下 FC 交换机启动/关闭、NPort 注册和交换网配置的设计与实现。

1) FC 交换机启动/关闭过程

FC 交换机启动时需要跟 FFC 建立 TCP 连接. 当建立 TCP 连接后,FC 交换机会向 FFC 发出登录请求报文,请求报文中包含了自身的信息和向 FFC 请求域标识(identity, ID)信息. FFC 收到请求后,返回登录请求应答报文,向 FC 交换机分配域 ID. 同样,当 FC 交换机或者 FFC 关闭时,会发出登出请求报文、对端回复登出应答报文,然后释放自身相关的信息,如内存空间、域 ID 等. 在此期间,FFC 和 FC 交换机间会周期性地发送保活报文交互信息. 具体过程如图 7 所示:

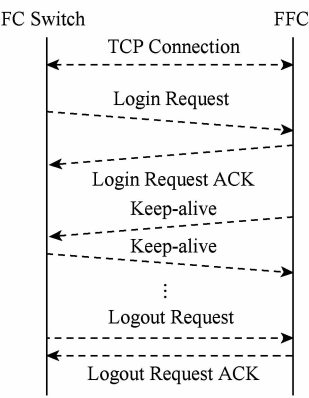


Fig. 7 Interactive process between FC switch and controller.
图 7 FC 交换机和控制器交互过程

登录请求报文和登录应答报文如表 2 和表 3 所示. 登录请求报文中 Type, Command 和 Length 字段前面已经做了描述. 会话 ID 用来唯一表示此次请求过程,FFC 的请求应答报文的会话 ID 跟此值一致. 交换机名字指交换机的全球范围节点名字(world

Table 2 Login Request Packet		
表 2 登录请求报文		B
Field	Packet Size	
Type	1	
Command	1	
Length	2	
Session ID	4	
Switch Name	8	
Domain ID	2	
Port Numbers	2	
Port Information	8	
⋮	⋮	
Port Information	8	

Table 3 Login Request ACK Packet		
表 3 登录请求应答报文		B
Field	Packet Size	
Type	1	
Command	1	
Length	2	
Session ID	4	
Domain ID	2	
ACK Code	2	

wide node name, WWNN)^[14],用来唯一标识 FC 交换机. 如果 FC 交换机申请某个具体域 ID,则域 ID 字段设为想申请的域 ID 值,否则域 ID 字段置为 0,表示向 FFC 请求分配域 ID. 端口数量字段表示该 FC 交换机拥有的端口数量. 端口信息字段又包含了端口索引、端口类型、端口速度等字段.

登录请求应答报文中会话 ID 跟登录请求报文中的会话 ID 一样. 应答码为 200 时,表示接受登录请求(ACC),此时,域标识字段值表示 FFC 为 FC 交换机分配的域 ID;应答码为 201 时,表示拒绝登录请求(RJT),域 ID 字段无效.

FC 交换机的登出过程与上面描述类似,登出请求报文和登出请求应答报文在此不再详述.

FC 交换机登录和登出之间,FFC 会周期性地(本文实现中设为 2 s)向 FC 交换机发送保活报文,FC 交换机收到保活报文后同样返回保活报文,如果控制器或者 FC 交换机连续 3 次没有收到保活报文,则认为对端失效,会发出登出请求报文,并释放本端的相关资源. 表 4 所示为 FC 交换机发送给控制器的保活报文.

Table 4 Keep-alive Packet		
表 4 保活报文		B
Field	Packet Size	
Type	1	
Command	1	
Length	2	
Switch Name	8	
Sequence ID	2	
Other Information	N	

保活报文中的 Type, Command 和 Length 字段不再详述,交换机名字唯一表示了保活通信中的 FC 交换机,序列号表示保活报文发送的顺序,该值越大表示保活报文越新. 其他信息字段作为预留,在后面的研究中使用.

2) 服务器端 NPort 注册和注销

在 NPort 的注册过程中,NPort 发出 FLOGI 报文,FC 交换机收到后解析该报文,收集相关信息生成 NPort FLOGI 请求报文,封装后转发给 FFC, FFC 收到后解封装得到 NPort 的注册信息,为 NPort 分配 FC ID 和参数配置等信息,并返回 NPort FLOGI 应答报文. FC 交换机收到 NPort FLOGI 应答报文,如果返回的是 ACC 报文,解析出其中的参数信息,对相应的 FPort 进行配置,并转发 ACC 报文到 NPort. 同样,NPort 注销过程也类似,FC 交换机需要对 NPort LOGO 报文进行重新封装,并在最后释放 FPort 和控制器中的相关资源配置.

NPort 的注册和注销过程如图 8 和图 9 所示.

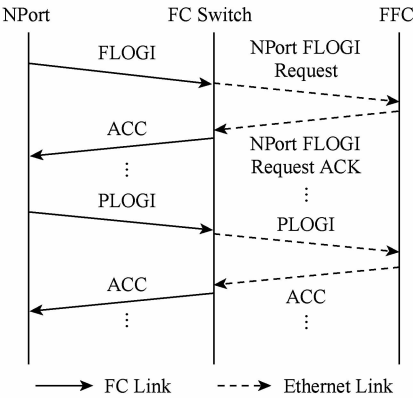


Fig. 8 Login process of NPort.
图 8 NPort 注册过程

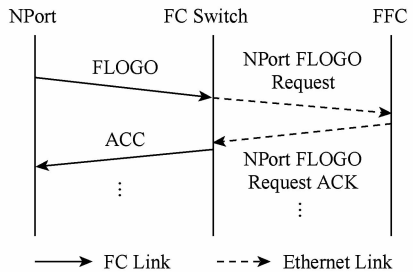


Fig. 9 Logout process of NPort.
图 9 NPort 注销过程

NPort 注册中,NPort 发出的其他 FC 控制报文(如 PLOGI 报文等),FC 交换机收到后不做处理,直接封装成 TCP 报文转发给 FFC,由 FFC 进行应答.

NPort FLOGI 请求和应答报文格式如表 5 和表 6 所示. 在请求报文中需要携带 FC 交换机的名字和与 NPort 对应的 FPort 索引信息,另外,在请求报文中还需要 NPort 的全球范围端口名字(world wide port name, WWPN)^[14]、FC ID(如果需要申请某个 FC ID,则 FC ID 字段填写该值,否则 FC ID 字

段置为 0)和该 NPort 的参数信息. 在应答报文中会携带应答码、分配给 NPort 的 FC ID 和相关配置参数.

Table 5 NPort FLOGI Request Packet

表 5 NPort FLOGI 请求报文 B

Field	Packet Size
Type	1
Command	1
Length	2
Session ID	4
Switch Name	8
Port Index	4
FC ID	4
NPort WWPN	8
Parameters	116

Table 6 NPort FLOGI Request ACK Packet

表 6 NPort FLOGI 应答报文 B

Field	Packet Size
Type	1
Command	1
Length	2
Session ID	4
Port Index	4
ACK Code	4
FC ID	4
Parameters	116

3) 交换网的配置和维护

在 CSN 架构下,FC 交换机间不再使用 FC 协议的 EPort 机制^[10],而是在 FFC 中实现交换网的配置和路由功能. FC 交换机在开机初始化时会向 FFC 注册,这样 FFC 就获得了全局的 FC 交换机信息. 随后,通过 EPort 启动时的注册过程,FFC 就获得了 FC 交换机间的连接关系,至此,FFC 获得了交换网的全局拓扑结构,最终通过路由协议计算出交换网的路由.

FC 交换机间 EPort 连接时,会通过端口的交互得知端口是 EPort,并且也可得知对端 FC 交换机名字. 随后,FC 交换机向 FFC 发送 EPort 启动请求报文,该报文中携带了该 EPort 的参数(FC 交换机 ID、端口索引、端口带宽、缓存大小等)和对端 FC 交换机名字. FFC 收到 EPort 启动请求报文后,返回 EPort 启动请求应答报文,报文内容包括协商后的带宽、缓存大小等. 最后,FFC 通过全局的路由计算,

向所有 FC 交换机下发新的路由表报文. 该过程如图 10 所示:

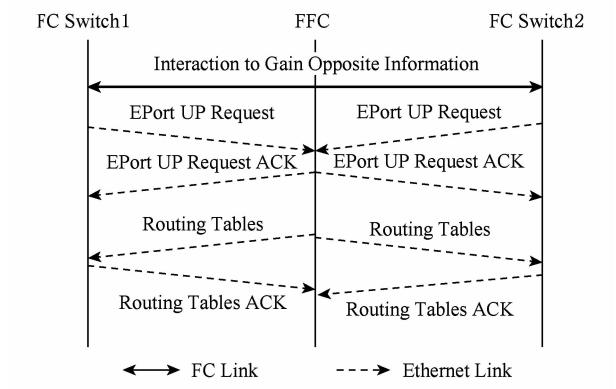


Fig. 10 Fabric configuration process.

图 10 交换网配置过程

EPort 启动请求报文如表 7 所示. EPort 启动请求报文包括了自身交换机名字(交换机 Name_1)和对应 EPort 的端口索引、对端 FC 交换机名字(交换机 Name_2)和对应 EPort 的端口索引以及该 EPort 中的参数(包括链路速率、缓存等). EPort 启动请求应答报文在此不再赘述,如表 8 所示.

Table 7 EPort UP Request Packet

表 7 EPort 启动请求报文

B

Field	Packet Size
Type	1
Command	1
Length	2
Session ID	4
Switch Name_1	8
Port Index	4
Switch Name_2	8
Port Index	4
Parameters	N

Table 8 EPort UP Request ACK Packet

表 8 EPort 启动请求应答报文

B

Field	Packet Size
Type	1
Command	1
Length	2
Session ID	4
Switch Name	8
Port Index	4
ACK Code	4
Parameters	N

在 CSN 架构下,交换网的路由算法使用 FC 协议中基于最短路径的 FSPF 算法,在此不再介绍该算法. 表 9 是路由表报文格式,其中最重要的是路由表项,路由表项包括 FC ID、前缀和接口索引,该表项类似于传统网络中的路由表.

Table 9 Routing Table Packet

表 9 路由表报文

B

Field	Packet Size
Type	1
Command	1
Length	2
Session ID	4
Switch Name	8
Item Numbers	4
Routing Item	12
⋮	⋮
Routing Item	12

4) FFC 性能优化

FFC 作为 CSN 架构的控制中心,一旦出现问题会对整个网络造成不可估量的危害,因此有必要对影响 FFC 性能的因素进行评估优化. 由 FFC 的功能模块可以知道 FFC 有 2 个服务对象:FC 交换机和服务器/存储服务器. 因此,主要从 FC 交换机和服务器/存储服务器数量 2 个方面考虑对 FFC 的性能影响.

① 我们知道,FC-SAN 网络的建设成本很高,维护的难度和费用较大,因此,在企业的存储网络中只有特别重要的应用(如重要的审批流程、财务应用等)才会使用高速稳定的 FC-SAN,其他应用会使用相对便宜的 IP-SAN. 这就决定了 FC-SAN 的规模不大,FC 交换机的数量不会很多,在此我们不考虑 FC 交换机同步登录对 FFC 的影响,而主要优化 FFC 与 FC 交换机维护连接的开销影响,即定期发送 keep-alive 报文的影响. 在此,我们在设计 keep-alive 机制时,该报文是由 FFC 主动发出,FC 交换机收到后回复,因此,可以通过建立 keep-alive 发送规则避免同时收发 keep-alive 报文影响 FFC 的性能.

keep-alive 发送规则:假定发送 keep-alive 报文的周期为 T ,FC 交换机的数量为 N ,则 FFC 每隔 T/N 的时间发送 keep-alive 报文给其中一台 FC 交换机.

② FC-SAN 的规模虽然不大,但是其中的服务器数量肯定要远大于 FC 交换机的数量,因此我们必须针对网络中的服务器数量进行压力测试. 这部分工作将在第 4 节的实验 6 中做详细的分析.

3 COBF 路由

SAN 是一种专注于信息存储、访问和管理的高速网络,它采用高速通道连接服务器和存储系统以此获得较高的传输速率,但是在 SAN 中使用的路由协议为最短路径协议,这种路由协议仅考虑使用最小跳数作为路由判据,并没有考虑利用链路的可用带宽来提高网络整体传输效率,因此,本节我们在 CSN 架构的基础上提出了按需可用带宽优先(CSN-based on-demand bandwidth-available first, COBF)的路由协议.

COBF 路由协议在 CSN 的 FFC 端实现,并且以路径上的可用带宽作为选择路由的判据,其 COBF 路由判据 M_{COBF} 定义如下:

定义 1. 对于某条路径,COBF 路由判据的数学定义为

$$M_{\text{COBF}}(p) = \sum_{\text{link } l \in p} \alpha \frac{1}{l_a},$$

其中, p 表示网络上的一条路径, l 是路径 p 上的一条链路, l_a 为链路 l 上的可用带宽, α 为常量,在此 $\alpha=1.0$.

在此链路的含义是 FC 交换机端口间的连线,因此,链路的可用带宽是指链路两端交换机端口的可用转发速率的最小值,即 $l_a = \min\{r_m, r_n\}$, 其中 r_m, r_n 分别是链路两端 FC 交换机端口的可用转发速率. FC 交换机端口的可用转发速率在 CSN 架构下可以通过 FC 交换机定期上报获取,在此,我们扩展了 CSN 中的保活报文,在保活报文中增加了交换机端口的可用转发速率,扩展后的保活报文如表 10 所示:

Table 10 Keep-alive Packet of COBF
表 10 COBF 下的保活报文 B

Field	Pactet Size
Type	1
Command	1
Length	2
Switch Name	4
Sequence ID	2
Interface Numbers	2
Port Index	4
Surplus Bandwidth	4
⋮	⋮
Port Index	4
Surplus Bandwidth	4

至此,CSN 架构中的 FFC 端就具有了计算路由的所有信息.对于每个路由请求,FFC 可以计算出服务器和存储间每条路径的路由判据,选出路由判据最小值的路径为最终的路由,并向 FC 交换机下发路由转发表.

4 实验和测试

4.1 验证环境

CSN 架构的验证环境包括了 4 块 FPGA 开发板、4 台 PC 机、1 台 FFC、1 台服务器和 1 台存储设备,组成了如图 11 的验证环境,其中验证环境中所用的设备清单如表 11 所示. FPGA 开发板与 PC 机通过 PCI-E 扩展槽连接,实现了 FC 交换机功能,其中 FPGA 开发板主要实现交换机的端口和数据转发功能,PC 机实现交换机的 CPU 功能.同时,PC 机通过自身的以太网卡连接到 FFC. FPGA 板卡上收到的 FC 控制报文转交给 PC 机中 FC 代理模块,FC 代理模块对 FC 报文简单处理并重新封装后通过 TCP 通信模块转发给 FFC.

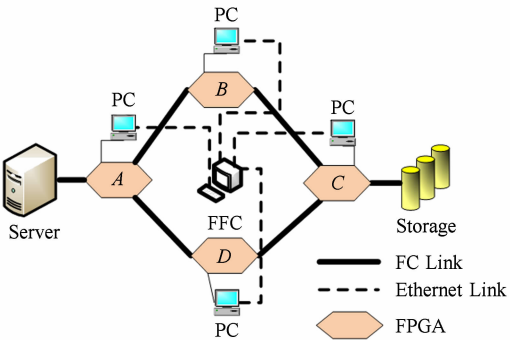


Fig. 11 CSN test environment.
图 11 CSN 验证环境

Table 11 List of Soft and Hardware in Test Environment
表 11 验证环境所用软硬件清单

Device/Software	Manufacturer and Model	OS	Number
Server	Huawei Tecal RH2285 V2	VMware ESXi 5.0	4
Storage	Huawei OceanStor S2200		1
PC	Lenovo Think Center Series	Ubuntu12.04/Windows XP	5
FPGA Board	Stratix IV EP4SGX230		4
Iometer	Iometer 1.1.0		1

4.2 实验分析

在 CSN 架构下,FC 交换机上数据面的处理跟传统的 FC-SAN 是一样的,都是通过路由表转发数据,因此这部分实验主要是针对 FC 协议的控制层面.这部分测试 CSN 架构下交换网络的延迟、收敛速度、数据转发吞吐量和针对 FFC 的压力测试.实验结果都是重复 10 次实验的平均值.

实验 1. NPort 注册时间

该实验测试 FC-SAN 和 CSN 架构下服务器 NPort 注册时间的差异. NPort 的注册时间是指 NPort 发出 FLOGI 报文到发送向存储设备登录的 PLOGI 报文的间隔. FC-SAN 架构下的 NS 是分布式服务,服务器和存储位置的不同对 NPort 注册时间有较大的影响,因此,该实验针对图 12 中的 3 种场景进行测试.实验结果如图 13 所示.

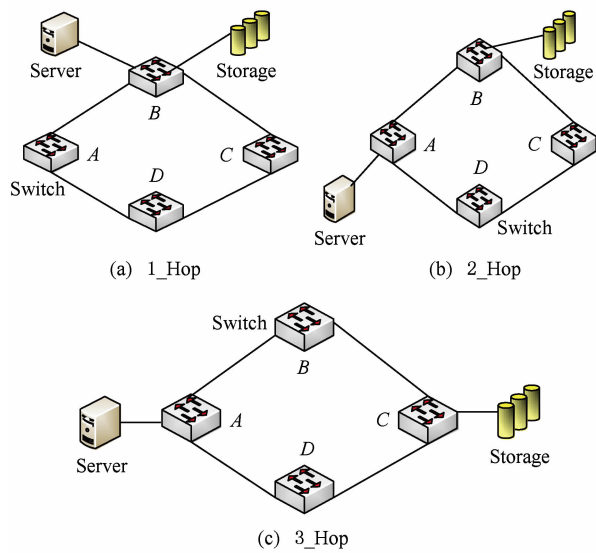


Fig. 12 NPort login scenes.

图 12 NPort 注册场景

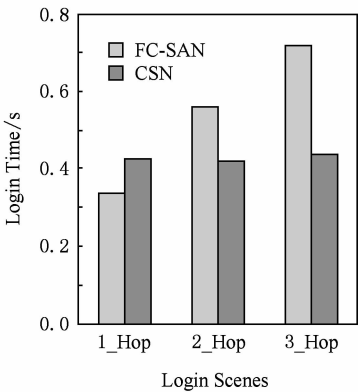


Fig. 13 NPort login time for three scenes.

图 13 3 种场景下的 NPort 注册时间

从图 13 可以看出,单跳场景下 CSN 架构的 NPort 注册时间要高于 FC-SAN 架构,这是由于 NPort 注册时 CSN 架构下 NPort 需要与控制器进行交互,这比 NPort 直接与 FC 交换机交互的时间要长.在其他 2 种场景下,传统 FC-SAN 架构下的注册时间要明显高于 CSN 架构下的注册时间,这是由于服务器跟存储通信时需要获得全网的 NS 功能,而此功能在 FC-SAN 架构下是分布式的,类似于 TCP/IP 中的域名服务,当 FC 交换机本地没有存储的信息时,需要通过查询的方式获得,服务器与存储间的距离(跳数)越远,查询时间也越久.而 CSN 架构下,NS 功能部署在控制器中能直接获取,因此,在 3 种场景下,CSN 架构的 NPort 注册时间基本不变.

实验 2. 交换网初始化

这部分实验测试 FC-SAN 和 CSN 架构下交换网初始化所需的时间.交换网初始化时间是指 FC 交换机上电到整个交换网路由收敛完成所需的时间,即 FC 交换机启动完成并获得了全局路由表的时间.为了把其他影响因素去除,这部分实验并没有把服务器和存储设备连接到网络中.

本次实验分 4 种情况:只开启 1 台 FC 交换机、同时开启 2 台 FC 交换机、同时开启 3 台 FC 交换机和同时开启 4 台 FC 交换机,分别记录下最终的交换网初始化时间,结果如图 14 所示.从图 14 看出,单台 FC 交换机时,CSN 架构所需时间要略多于 FC-SAN 所需时间,这是因为 CSN 下 FC 交换机需要向 FFC 注册并通过 FFC 下发路由表,而传统的 FC-SAN 不要执行路由协议等操作;在多于 2 台 FC 交换机下,CSN 架构的网络初始化时间要比 FC-SAN 架构小,原因是 FC-SAN 架构下交换网初始化需要全局交换信息,选举主交换机和执行 FSPF 路由分布式算法等步骤,而 CSN 下的路由只在 FFC

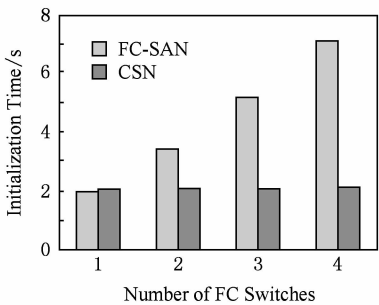


Fig. 14 Network initialization time.

图 14 网络初始化时间

上计算完成,因此 CSN 下的网络初始化时间要比 FC-SAN 架构小,而且网络规模越大,CSN 架构的优势更加明显.

实验 3. FC 交换机增加

这部分实验测试随着 FC 交换机的增加,交换网路由收敛时间的变化.该实验同样没有把服务器和存储连接到交换网中.实验中,FC 交换机从 1 台增加到 4 台,交换网路由的收敛时间变化如图 15 所示:

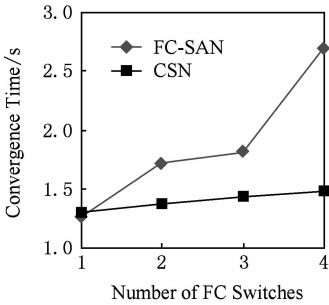


Fig. 15 Routing convergence time when increasing the number of FC switches.

图 15 FC 交换机增加时路由收敛时间

从图 15 可以看出,在只有 1 台 FC 交换机情况下,FC-SAN 下的收敛时间要略小于 CSN 架构下的时间,原因在实验 2 中已经阐述.但是随着 FC 交换机的增加,FC-SAN 架构下的路由收敛时间要明显多于 CSN 架构下所需的时间,而且网络规模的越大,两者的收敛时间差别越大,这是由于 FC-SAN 架构下需要执行复杂的分布式交互,且需要通过全局泛洪的方式计算路由,所用时间较多.而 CSN 架构下,FC 交换机只需要向控制器进行注册,使控制器拥有全网的拓扑结构,使用集中式路由计算就可以得出全局路由,所用时间较少,CSN 下的路由计算时间主要受 FFC 本身计算性能等因素的影响.

实验 4. 删除 FC 交换机

该部分实验测试在 FC-SAN 和 CSN 架构下,4 个 FC 交换机删除其中 1 个 FC 交换机时对交换网路由的影响.由于在 FC-SAN 架构下,有主 FC 交换机的概念,所以该部分实验分别删除主 FC 交换机和非主 FC 交换机得到交换网路由的收敛时间,测试结果如图 16 所示.

从图 16 看出,删除主 FC 交换机对 FC-SAN 的影响较大,因为删除主 FC 交换机后需要重新通过泛洪的方式选举主 FC 交换机,因此,所需的网络收敛时间更长.在删除非主 FC 交换机时,FC-SAN 架

构所需收敛时间也要大于 CSN 架构,因为删除 FC 交换机后需要扩散链路状态信息,并重新计算路由.而这部分时间在 CSN 下是不需要的,CSN 下删除 FC 交换机只需要执行 FC 交换机的登出请求过程,并在 FFC 重新计算全网路由,时间较短.

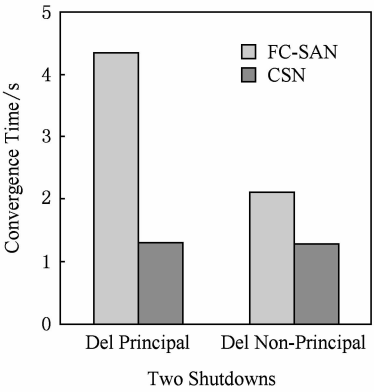


Fig. 16 Routing convergence time when shutdown FC switch.

图 16 FC 交换机删除时网络路由收敛时间

实验 5. 网络吞吐量

该部分实验针对 FC-SAN 和 CSN 两种架构,在 CSN 下针对 CSN 的 FSPF 和 COBF 路由,分别标识为 CSN-FSPF 和 CSN-COBF.在图 12(c)的环境下测试交换网整体吞吐量.该实验过程包括交换网初始化,然后连接服务器和存储,最后服务器通过 Iometer^[15]工具读写存储中数据,并且在读取过程中断开非主 FC 交换机链路和主 FC 交换机链路.整个实验过程中记录下交换网整体的吞吐量,实验结果如图 17 所示:

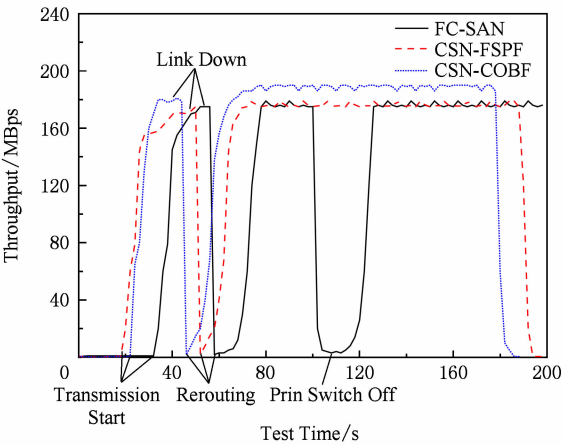


Fig. 17 Throughput of fabric.

图 17 交换网吞吐量

从图 17 可以看出,在网络初始化和读写存储中

数据时,FC-SAN 和 CSN-FSPF 的最大吞吐量基本一致,这也表明 CSN 架构对网络数据面的转发并没有太大的影响;而当使用 CSN 架构下 COBF 路由时,网络的最大吞吐量要比 FC-SAN 和 CSN-FSPF 高 10%左右.当断开非主 FC 交换机链路时,CSN-FSPF 和 CSN-COBF 架构下的吞吐量恢复速度要比 FC-SAN 架构更快.而断开主 FC 交换机链路时,FC-SAN 网络吞吐量的恢复需要更多的时间.因此我们可以看出,CSN 架构可以对存储网络进行更合理的路由规划、获得更优的吞吐量,同时该架构反应更加灵敏、控制更加高效和灵活.

实验 6. FFC 压力测试

CSN 架构中的核心为 FFC,一旦 FFC 的性能下降或者失效,将影响整个存储网络的性能,甚至崩溃.因此,这部分实验我们将对 FFC 进行压力测试.本实验中 FFC 控制器的配置如表 12 所示,存储网络中有 4 台 FC 服务器,分别接入 4 台 FC 交换机.

该实验主要通过增加服务器并发启动的数量来测试 FFC 的 CPU、内存和带宽利用率变化情况.由于实验室的服务器数量有限,因此实验中通过虚拟机来代替物理服务器,一共创建了 20 台虚拟机.实验时通过同时打开 1~20 台虚拟机,在 FFC 控制器中记录 CPU、内存和网络带宽的瞬时利用率最大值,并通过重复 10 次实验,得到瞬时利用率最大值的平均值和 10 次实验瞬时利用率的最大值,同时记录下 FFC 稳定状态下的各项指标利用率,得到实验结果如图 18 所示.

Table 12 Hardware of FFC

表 12 FFC 硬件配置

OS	Windows XP
CPU	Intel CORE i3 3.4 GHz, Cache 3 MB
Memory	4 GB DDRⅢ 1600
NIC	1 Gbps

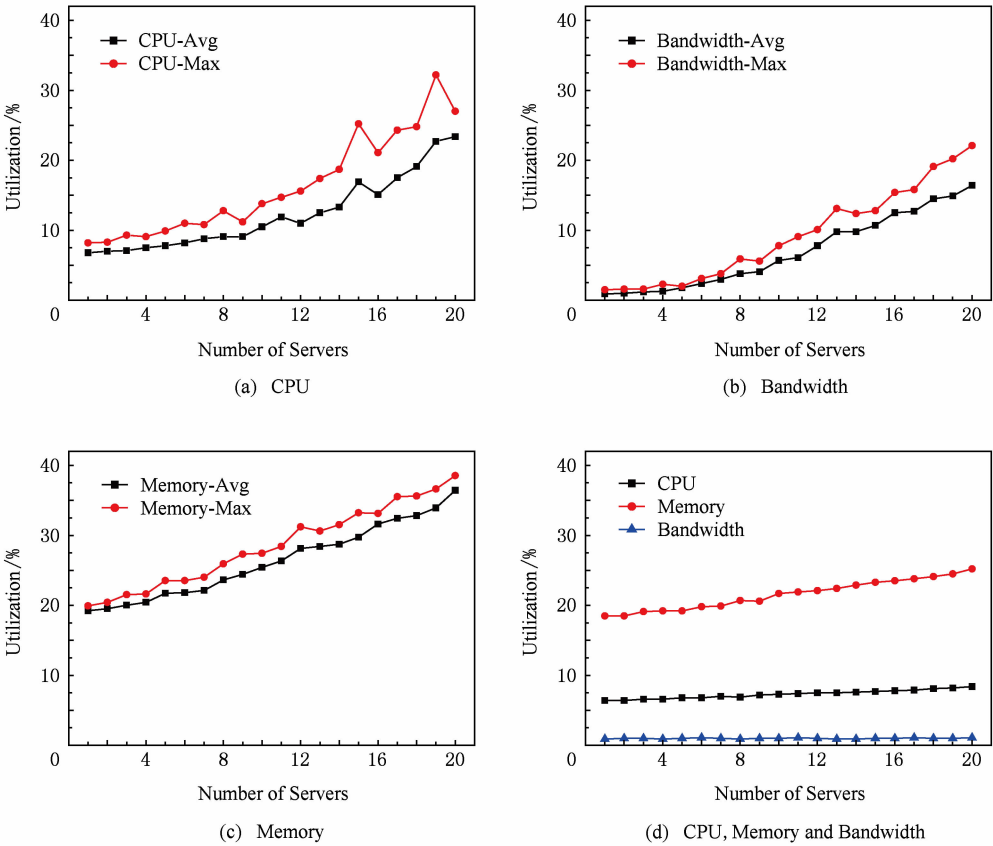


Fig. 18 Stress Testing of FFC.

图 18 FFC 压力测试

图 18(a)(b)(c)可以看出,随着服务器数量的增加,CPU、内存和网络带宽的瞬时最大值利用率也随之增加;但是在服务器数量较少时,CPU 和网

络带宽的利用率变化较少,随着服务器数量增加,利用率增加得较快,而内存利用率基本随着服务器数量的增加而增加.

图 18(d)表示服务器登录完成后,FFC 的稳态利用率变化情况.CPU 和网络带宽的利用率变化不大,而内存利用率轻微增加,这主要是因为稳态情况下 FFC 的开销主要是用于 FFC 和 FC 交换机间的通信,本实验中交换机数量没有变化,因此 CPU、带宽的使用开销变化不大.而内存利用率稍微增加则主要是因为随着服务器的增加,FFC 中记录服务器的信息也相应增加,因此内存利用率增加.

通过以上的测试我们可以得出,在大量服务器并发启动的过程中,对 FFC 的 CPU、内存和网络带宽都会有影响,但是影响较大的是内存和 CPU,这可以通过增加内存和优化 FFC 处理代码等手段缓解;另外,通过增加 FFC 的数量并采用已有的负载均衡方案,也能得到较好地解决.

5 结论和展望

本文研究存储网络中的 FC-SAN 架构和 SDN 技术,并基于 SDN 提出了一种基于中心控制器的新型存储架构,分析了 CSN 对 FC-SAN 的改进;随后详细描述了 CSN 架构下控制器功能、FC 协议改进和通信机制等内容,并设计了基于 CSN 的按需可用带宽优先路由协议;最后,通过实际的验证平台测试了 CSN 架构下的整体性能,不管是 FC 交换机的增删或者是交换网的初始化,CSN 架构下交换网路由的收敛时间要明显优于原有 FC-SAN 架构,而且网络规模越大优势也越明显,同时,在 COBF 路由协议 CSN 架构能够获得更高网络吞吐量,最后对 CSN 的核心部件 FFC 进行了压力测试.

本文下一步工作希望对 CSN 架构的功能进行改进和增加,在 FFC 下发路由时支持增删功能;在控制器中增加 FC 协议的其他一些功能,如分区功能,使得整个 CSN 架构功能更加完善.另外,我们将针对 FCoE-SAN 环境,对 CSN 架构进行改造,使其支持 FCoE 协议,并在 FCoE-SAN 环境下进一步优化 FFC 的性能,使其能在大规模网络下工作.最后,修改 OpenFlow 协议使其能够在 FC/FCoE 环境下工作.

参 考 文 献

- [1] Vladimir V, Riabov. Storage Area Networks (SANs)[G/OL] //The Internet Encyclopedia, Volume 3. Hoboken, NJ: John Wiley & Sons, 2004: 329 - 338 [2015-01-10]. <http://onlinelibrary.wiley.com/doi/10.1002/047148296X.tie164/abstract>
- [2] T11 Technical Committee. Fibre channel [S/OL]. 2014 [2015-01-10]. <http://www.t11.org/t11/als.nsf/v2guesthome766kx?OpenForm>
- [3] T11 Technical Committee. Fibre channel over Ethernet (FCoE)[S/OL]. 2012 [2015-01-10]. <http://www.t11.org/fcoe>
- [4] Wu Longfei. The research of fibre channel over Ethernet storage technology [D]. Wuhan: Huazhong University of Science and Technology, 2013 (in Chinese)
(吴龙飞. 以太网光纤通道存储技术的研究[D]. 武汉: 华中科技大学, 2013)
- [5] Open Networking Foundation. Software-defined networking: The new norm for networks [EB/OL]. (2012-04-13) [2015-01-10]. <https://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf>
- [6] Open Networking Foundation. OpenFlow switch specification: Version 1.0.0 [EB/OL]. (2009-12-31) [2015-01-10]. <https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow/openflow-spec-v1.0.0.pdf>
- [7] Jeda Networks. Software defined storage networks: An introduction [EB/OL]. (2012-12-05) [2015-01-10]. http://www.jedanetworks.com/wp-content/uploads/2012/12/Jeda_Networks_SDSN.pdf
- [8] SNIA: Storage networking industry association [EB/OL]. 1997 [2015-01-10]. <http://snia.org>
- [9] T11 Technical Committee. Fibre channel framing and signaling (FC-FS-2) [S/OL]. 2007 [2015-01-10]. http://www.t11.org/ftp/t11/admin/project_proposals/03-092v0.pdf
- [10] T11 Technical Committee. Fibre channel-switch fabric-5 (FC-SW-5) [S/OL]. 2009 [2015-01-10]. <http://www.t11.org/ftp/t11/member/fc/sw-4/05-785v1.pdf>
- [11] Kamiya S, Ichino K, Yasuda M, et al. Advanced FCoE: Extension of fiber channel over Ethernet [C] //Proc of the 3rd Workshop on Data Center—Converged and Virtual Ethernet Switching in ITC'11. New York: ACM, 2011: 1-8
- [12] Jeda Networks. Virtualizing the SAN with software defined storage networks [EB/OL]. (2014-03-14) [2015-01-10]. <http://www.jedanetworks.com/wp-content/uploads/2014/03/140226-wp-Virtualizing-SDSN-FINAL.pdf>
- [13] Shiraki O, Nakagawa Y, Hyoudou K, et al. Managing storage flows with SDN approach in I/O converged networks [C] //Proc of the 5th IEEE MENS Workshops in 2013 GLOBECOM. New York: IEEE Communications Society, 2013: 890-895
- [14] World wide name [EB/OL]. (2014-12-26) [2015-01-10]. http://en.wikipedia.org/wiki/World_Wide_Name
- [15] Iometer [EB/OL]. 2003 [2015-01-10]. <http://www.iometer.org>



Lu Yifei, born in 1982. Received his PhD degree from the School of Computer Science and Engineering, Southeast University in 2010. Lecturer. His main research interests include data center network, cloud computing, big data and SAN.



Zhang Zhenwei, born in 1982. Received his master degree from Nanjing University of Posts and Telecommunications. Senior engineer in Huawei Nanjing R&D Center. His main research interests include SAN, big data and distributed computing (zhang zhenwei@huawei.com).



Tao Jun, born in 1975. PhD and associate professor, PhD supervisor. His main research interests include network architecture, wireless network technology (WSN, VANET, WMN, WMN), information economics, network resource allocation technology, network multimedia technology and trustworthy network (juntao@seu.edu.cn).



Tang Ling, born in 1986. Received her PhD degree in electronic science and engineering from Southeast University, Nanjing, China, in 2013. Member of IEEE. Her current research interests include cooperative communications, cloud computing and networking (ling.tang@njtu.edu.cn).

《信息安全研究》期刊简介

习近平总书记指出“没有网络安全就没有国家安全,没有信息化就没有现代化”. 数字时代信息安全工具的大众化是不可阻挡的历史潮流. 大众化的信息安全已经直接影响到我们每个人的利益,信息安全已成为国家、地方区域经济结构优化提升和转型发展的新机遇. 在信息安全上升为国家战略、行业迎来崭新发展机遇形势下,《信息安全研究》期刊应时代而生.

《信息安全研究》是由国家发改委主管、国家信息中心主办的中文学术期刊,其宗旨是集中展示和报道国际、国内网络和信息安全研究领域研究成果及最新应用,传播信息安全基础理论和技术策略,服务国家信息安全形势发展需要. 所刊登的论文均经过专家严格评审.

《信息安全研究》将于 2015 年 10 月创刊发行. 刊期为月刊,每期 96 页,由《信息安全研究》杂志社出版,国内外公开发行.

《信息安全研究》将以研究致以应用,搭建信息安全领域的学术交流平台,愿意和同行业及社会各界建立联系,友好合作,共赢美好未来. 欢迎大家积极投稿、赐稿,洽谈合作.

投稿邮箱:ris@cei.gov.cn

编辑部联系人:崔先生(185 0008 6481)

马先生(158 1058 2450)