CSRouter：服务可扩展的路由器体系结构

吕高锋[1]，孙志刚[1]，林雨弦[2]，陈一骄[1]，李韬[1]
（1. 国防科学技术大学 计算机学院，湖南 长沙 410073; 2. 73232 部队，浙江 舟山 316200）

摘 要：互联网应用的发展促进了网络从面向端—端数据传输的“管道”向集成计算资源、存储资源和网络资源的“平台”模型的发展。以统一交换技术为基础的路由器系统区域网连接转发引擎、服务单元和存储单元等资源，建立了新型计算—存储—转发处理模型。基于 NetMagic 平台和 IEF 交换系统实现了 CSRouter 原型系统。报文缓存应用测试表明，路由器系统区域网能够在转发引擎和服务单元之间提供高带宽高可靠的互连，基于系统区域网的路由器能够利用计算资源和存储资源对 IP 报文进行深度处理，加速网络应用，增强路由器服务可扩展性。

关键词：系统区域网；计算资源；存储资源；统一交换
中图分类号：TP311.134.3 **文献标识码**：A **文章编号**：1000-436X(2012)07-0049-10

CSRouter: router architecture for service extensibility

LV Gao-feng[1], SUN Zhi-gang[1], LIN Yu-xian[2], CHEN Yi-jiao[1], LI Tao[1]
(1. School of Computer Science, National University of Defense Technology, Changsha 410073, China;
2. Unit 73232, Zhoushan 316200, China)

Abstract: The development of network applications , network end-end “channel” for data transmission impeded applications, network “platform” integrating computing and storing resource with routers is a considerable efficient means to enhance the efficiency of the network transport. System area network of routers based on high performance unified switching technology is introduced to connect forwarding elements, computing elements and storing elements. The prototype were designed and implemented based on the NetMagic platform of network innovation and the IEF switch system. The testing of packets caching proves that system area network provides high performance and high reliability network bandwidths between forwarding elements and computing elements, and the novel router architecture based on system area network could utilize additional computing and storing resources to process IP packet in depth to accelerate network applications.

Key words: system area network; computing elements; storing elements; unified switching

1 引言

随着“三网融合”和“物联网”等新型网络应用研究和部署的全面展开，应用对互联网的需求由单纯的端—端数据“传输”功能逐渐演化为多样化的“服务”。如大规模内容分发网络中边缘存储，延迟容忍网络中数据处理与融合，IPTV 中大规模流媒体组播传输和终端智能无线接入等网络应用

收稿日期：2011-11-16；**修回日期**：2012-06-10
基金项目：国家重点基础研究发展计划（“973”计划）基金资助项目(2009CB320503)；国家高技术研究发展计划（“863”计划）基金资助项目(2008AA01A323, 2008AA01A325, 2009AA01A334)
Foundation Items: The National Basic Research Program of China (973 Program) (2009CB320503); The National High Technology Research and Development Program of China (863 Program) (2008AA01A323, 2008AA01A325, 2009AA01A334)

都需要网络能够提供强大的计算能力和丰富的存储资源，提供多种网络加速服务。为此，互联网的服务架构只有从支持端—端数据传输的“管道”变为融合大量通信带宽、计算和存储资源的支持分布式应用运行的“平台”，才能够有效地丰富网络功能，促进网络应用的部署和发展。

国外对下一代互联网平台化已经展开了深入研究。FIA4 的 Nebula[1]工程将下一代互联网看作一个超级数据中心，整个互联网平台支持虚拟化、节能、VM（virtual machine）的迁移等特性。用户的服务请求通常不会进入网络核心，而是由在网络边缘数据中心虚拟机中运行的形式分布式程序提供，基于网络“平台”的服务模式优势在于可以有效支持新型业务和服务的快速部署。然而，基于数据中心增加网络计算和存储资源的方式，存在很大的局限性。数据中心中计算和存储服务器通过专用链路直连到核心路由器，但是和普通端系统一样都处于网络的边缘，并采用端到端的方式通信。对于流量测量等应用，需要将报文流镜像到数据中心服务器，对于位于数据处理路径上的应用，也需要将报文流转发到数据中心服务器，这种方式增加了网络传输开销。另外，松耦合的关系使得数据中心服务器无法使用网络拓扑信息，优化报文流传输，如进行负载均衡等。

基于路由器集成计算和存储资源的紧耦合方式成为一种有效的解决方法，但仍然存在许多问题。传统路由器转发板和控制板通过内部交换开关互连，转发板主要负责报文头校验和 IP 查找转发等功能，集成的计算和存储资源有限。由于转发板采用紧耦合的设计方式，增强其计算和存储能力的一种方式是直接通过升级扩充转发引擎相应的资源实现，该方式成本高，不利于投资保护。另一种方式是，基于松耦合方式提供计算和存储资源扩展，例如通过路由器网络接口连接提供计算和存储资源的服务器。在该方式下，服务器可以提供强大的计算和存储能力，但是数据输入速度受限于互连接口带宽，若路由器大量的接口用于连接服务器，又会减少网络连接度。另外，通过普通网络接口访问服务器，端系统访问模式没有改变，仍然需要处理 TCP/IP 协议，协议处理复杂，延迟大。

针对上述问题，本文提出了基于系统区域网的新型路由器体系结构 CSRouter（computing and storage router）。以高性能可扩展的统一交换技术构建路由器系统区域网，连接转发引擎、服务和存储资源。基于系统区域网的新型路由器体系结构 CSRouter 以紧耦合方式向路由器中引入计算和存储资源，构建支持分布式应用运行的下一代互联网“平台”。目前，基于系统区域网的新型路由器体系结构中报文转发流程、转发引擎控制方式、服务平面服务单元与数据平面转发引擎交互机制和报文调度等关键问题需要解决，是本文研究重点。

本文组织结构如下，第 2 节分析比较紧耦合片上网络连接方式、松耦合系统总线连接方式和松耦合网络连接方式等计算和存储扩展方式；第 3 节提出基于系统区域网的新型路由器体系结构 CSRouter，描述系统区域网构建、数据处理流程和控制流程等；第 4 节对系统原型系统和性能分析；第 5 节是结束语。

2 相关研究

传统路由器中实现报文转发等处理的转发板主要由网络处理器和接口等模块构成。网络处理器最快支持单向 100Gbit/s 报文的 IP 查找与转发[2]，能满足当前链路线性处理的性能要求。然而，对于协议识别、深度报文内容检测和多媒体网关等网络功能，网络处理器的性能明显不足，无法满足网络应用线速处理的性能要求。目前通过提高半导体制造工艺和采用片上多核（MPoC，multiple-processor on chip）[3]设计方法等多种方式提高网络处理器的报文处理能力，以提高路由器报文转发性能。然而，网络处理器性能升级总是滞后于网络链路速度的提升，在利用有限的计算和存储资源实现报文线速转发外，不能再执行对报文流包含的应用层请求进行深度处理等任务。

为提高路由器转发性能，并对报文流进行深度处理，国内外对于在网络边缘基于路由器引入大规模计算和存储资源展开了深入研究，并提出了多种方式，如系统总线互连方式和网络互连方式等。

在网络中引入计算资源和存储资源，通常选择以服务器等形式提供的计算和存储资源。在服务器中通过系统总线连接网络交换设备实现软件路由器，除了网络交换设备进行报文转发外，在服务器中运行的程序也执行报文转发等处理任务，以提升软件路由器转发报文的灵活性和整体性能，如 ServerSwitch[4]和 PacketShader[5]等。

在保留已有路由器资源的情况下，为了提高路

由器转发引擎报文处理的性能，ServerSwitch 提出了通过服务器系统总线连接服务器计算资源和网络交换设备的方法。在 ServerSwitch 中，网络交换设备以服务器网卡的形式连接到系统中，服务器对网络交换设备进行控制，另一方面处理网络交换设备转发的报文。网络交换设备实现路由器数据平面 IP 报文查找与转发，将需要深度处理的 IP 报文通过系统总线转发给服务器。在服务器中对报文进行深度内容检测等处理，同时实现路由计算等网络控制。ServerSwitch 采用系统总线连接网络交换设备和服务器 CPU，延迟小，然而这种紧耦合的集成方式，不利于系统的扩展。

基于 GPU 加速的软件路由器 PacketShader 也采用了服务器系统总线连接服务器计算资源和网络接口的方式。PacketShader 软件路由器安装了 8 个 10Gbit/s 以太网接口，2 个 NVIDIA GTX480，它们通过 PCI Express 系统总线连接。PacketShader 充分利用 GPU 并行处理能力，实现了 IPv4 转发、IPv6 转发、Openflow 交换[6]和 IPSec 隧道等高速报文处理引擎，实现了比基于通用 CPU 的软件路由器更高的报文转发能力。在基于系统总线连接方式下，也只有升级 CPU 处理能力才能够提高路由器转发性能。

另一种在路由器中引入计算和存储资源的方式是通过路由器网络接口直接连接提供计算和存储资源的服务器，如 CLARA(cluster-based active router architecture)[7] 、 GBP(GENI backbone platform)[8]和 Juniper 路由器[9]等。

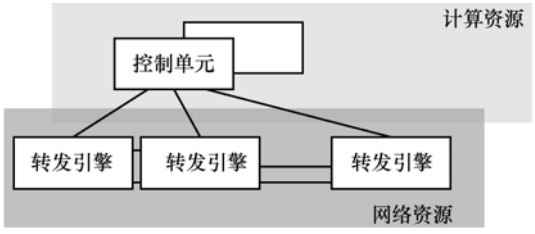
CLARA 目标是建立一个与传统路由器性能相当，而且具备计算资源的基于集群的主动路由器体系结构。CLARA 是通用 PC 用网络互连而成的集群，其节点分为路由节点和计算节点，分别完成报文的转发和面向网络应用的报文处理。集群中 PC 被配置为转发报文的传统 IP 路由器，称为路由节点，其他的 PC 为相应的网络定制服务提供计算资源，称为计算节点。CLARA 是 PC 集群路由器，其中组件可以动态加入或退出。为了支持集群路由器动态管理功能，集群管理者收集集群单元处理器、内存的使用情况等，并制定调度策略。

GBP 是 GENI 骨干网平台的原型系统，主要由路由器和刀片服务器构成，它们通过吉比特以太网连接。路由器采用了 ATCA 机箱，集成多个网络接口处理单元和报文处理单元，实现了报文转发等基本功能。刀片服务器也实现了报文转发处理等基本功能，以增强路由器报文转发的能力。GBP 主要目的通过虚拟化方式，允许多个实验网络共享 GENI 网络基础设施。

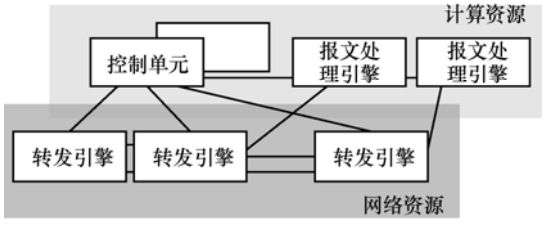
Juniper 采用了网络连接的方式向路由器增加计算和存储资源。Juniper 集群路由器通过网络连接多功能服务机柜 MultiService，MultiService 集成大量的刀片服务器。利用 Juniper SDK，Juniper 的合作伙伴可以开发网络数据协处理的加速应用，如流属性统计、报文分类、负载均衡等，在多功能服务机柜中运行。Juniper 网络集成方式具有较好的扩展性，可以根据应用需要很容易向集群路由器增加多功能服务机柜。然而，网络互连方式占用了宝贵的路由器转发引擎端口资源，成本高。另外，网络互连方式下多功能服务机柜与普通端系统相比，只是与路由器的距离更近，是一种松散的耦合方式，延迟大。

3　服务可扩展的路由器体系结构

在路由器发展过程中转发与控制分离的结构使得路由器可以使用额外的计算资源处理路由协议，增强路由器控制功能[10]，如图 1(a)所示。新型路由器系统结构中引入服务器计算和存储资源使得路由器可以使用强大的计算资源主要实现报文转发，增强路由器报文转发性能[11]，如图 1(b)所示。以上计算资源的引入是从提高路由器性能的角度出发，并没有增强路由器协议处理的灵活性，不能满足应用发展对网络功能多样化的需求。



(a)面向控制的计算资源



(b)面向转发功能扩展的计算资源

图 1　路由器中计算和存储资源的引入

只有将大规模计算资源、存储资源和路由器网络资源高效融合，并改变报文处理模式，才能够实现基于平台的边缘—边缘的服务模式。基于系统区域网的路由器 CSRouter 以高性能可扩展统一交换技术为核心构建路由器系统区域网，连接大规模计算资源、存储资源和网络资源，具有高性能和高扩展性等优点。CSRouter 利用计算能力和大规模存储能力，对于报文的处理将由以往的存储—转发模式转变为计算—存储—转发模式，为实现网络“平台”模式奠定基础。

3.1　路由器体系结构

基于系统区域网的路由器 CSRouter 可以抽象为数据平面、服务平面和控制平面，如图 2 所示。数据平面中报文转发引擎（PFE, packet forwarding engine）实现基于流表报文转发等功能；服务平面的服务单元 SE 和存储单元 Storage，不仅支持网络数据协处理的加速应用，还提供网络应用层协议深度处理，如 Web cache 等网络服务；控制平面的控制器（CE, controlling engine）实现路由协议和路由器的管理和配置。在 CSRouter 中以路由器系统区域网为核心，高速互连大规模计算资源、存储资源和网络资源等，为路由器服务的扩展提供开放的运行平台。

数据平面转发引擎 PFE 负责 IP 报文的接收、发送和查表转发等。根据流表查找结果，转发引擎将报文发送到服务平面的服务单元 SE 进行深度处理，发送到服务平面的存储单元 Storage 进行缓存，发送到数据平面的其他转发引擎 PFE 通过目标端口发送。

控制平面控制器 CE 负责对转发引擎 FE、服务单元 SE 和存储单元 Storage 等资源进行管理和配置。控制器发送网络探测报文，获得网络拓扑后，计算节点路由信息。控制器根据路由器内部拓扑关系和路由信息表，为路由器中数据平面转发单元 PFE 和服务平面服务单元 SE 分别计算路由表，并将结果下载到各单元的转发表中。采用路由器分布式控制机制[12]能进一步增强控制器 CE 的可靠性。

服务平面由提供计算资源的服务器和提供存储资源的盘阵控制器等构成，提供了大规模的计算资源和存储资源。根据报文处理策略，需要进行应用层深度处理的报文由转发引擎发送到空闲的服务单元进行处理。服务单元解析报文携带的网络应用请求，从存储服务器读取相应的请求数据再返回给转发引擎。

3.2　路由器服务—转发模型

传统路由器主要实现报文的转发，对接收到的报文先排队缓存，再查表和转发，是一种存储—转发模型[16]。路由器存储—转发模型适合端—端网络应用，却不能适应新型网络应用的需求。面向下一代互联网的基于系统区域网的路由器 CSRouter 不仅实现报文转发，还能根据报文流内容，对进入的报文流进行协议分析等应用层深度处理，实现一种

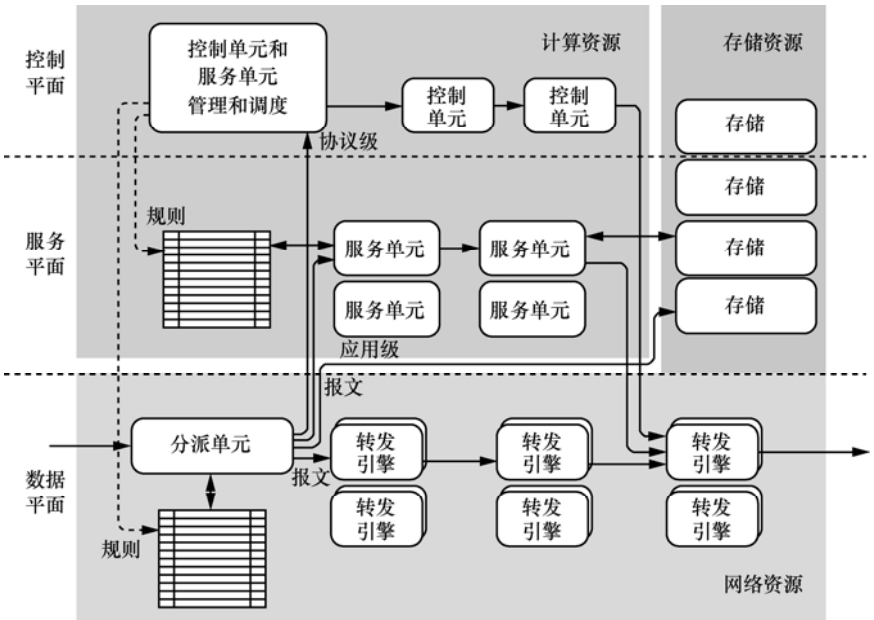


图 2　服务可扩展的路由器体系结构

服务—转发模型。报文处理模式的改变，使得报文处理流程更加灵活，支持更复杂的处理，从而实现网络服务扩展，这也是 CSRouter 区别以往路由器的显著特点。

在新型路由器体系结构中，路由器转发引擎 PFE 从网络端口接收报文，根据处理策略，查找流表后从其他端口转发。对于满足网络应用层协议深度处理条件的 IP 报文，转发引擎将报文转发给服务单元进行协议识别和处理，服务单元处理完后，再发回到原转发引擎 PFE 进行处理，或者再转交给其他服务单元进行处理后返回。对于与缓存数据相关的 IP 报文，转发引擎将报文转发给服务单元，服务单元缓存到存储单元，或从存储单元读取数据，将返回结果给转发引擎，如图 2 所示。

服务单元的计算能力是对转发引擎中报文处理能力的扩展，存储单元的存储资源可以作为转发引擎存储空间的一部分。基于服务—转发模型的体系结构，可以根据服务策略，对报文流进行多次迭代处理，实现新型网络服务。报文流在转发引擎、服务单元和存储单元之间流转，实现用户定义的处理流程，增强了路由器协议处理的灵活性和服务可扩展性。

3.3　路由器系统区域网

通常路由器转发单元通过交换开关连接，实现多个网络接口卡、转发引擎和控制引擎的高速互连。交换开关紧耦合系统结构，能够减小报文转发延迟，但限制了系统的可扩展性。采用了普通以太网连接计算资源、存储资源和网络资源是一种最简单的方法，能够直接利用路由器提供的网络接口。通过网络接口直连的松耦合方式协议处理复杂，会增加报文转发时延，还造成网络连接度的下降。下一代路由器采用路由器系统区域网，连接转发引擎、服务单元和存储单元等组件，如图 3 所示，系统区域网具有高性能高可靠、支持多路径传输和支持异构网络互连的特征。

1) 高性能可扩展

交换开关采用简单的查找机制，交换性能高。接口通常为缓冲区读写接口，不适合连接计算和存储资源。传统以太网为了在全球唯一标识设备，将地址长度设为 48bit，实现线速查找难度大。另外，传统以太网在查表失败等情况下采取泛洪机制，大大降低了网络性能。路由器系统区域网中节点命名只要求在系统区域内统一，空间小，能有效减小交换机转发表项数目以提高交换机查表性能。另外，路由器系统区域网具有良好的扩展性，能够支持路由器中计算和存储资源等灵活升级。

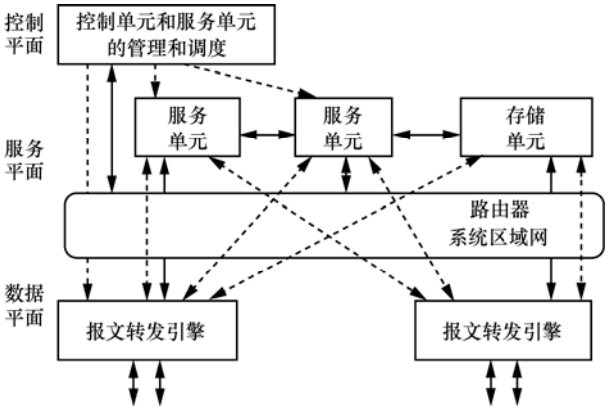


图 3　CSRouter 组成和控制

2) 支持多路径传输

传统以太网在转发报文过程中为避免转发环路，构造生成树，删除冗余路径，选择转发路径。传统以太网生成树路由算法不支持多路径传输，网络利用率低。为提高节点间聚合带宽，增强网络传输可靠性，路由器系统区域网应支持转发引擎和服务单元之间的多路径传输。路由器系统区域网络采用集中式控制方式计算转发路径，可有效避免路径冲突。

3) 支持异构网络

路由器系统区域网连接路由器转发引擎、服务单元和存储单元，通常这些单元的网络接口采用多种物理介质，如面向存储的 Fibre Channel、面向高性能计算的 InfiniBand，接口不统一。路由器系统区域网应该具有承载多种业务的能力，支持不同网络类型的报文共享同一种物理介质，实现统一交换。具有统一交换能力的系统区域网络能够简化路由器系统结构，降低建设成本。

3.4　路由器资源管理控制

基于系统区域网的路由器 CSRouter 控制器对网络、计算和存储等多种资源进行集中式管理和控制。集中式控制器可以是单独的控制单元[13]，连接到路由器系统区域网中，也可以作为控制进程运行在服务单元。

控制器运行路由协议，通过转发引擎收发报文，探测网络拓扑，计算路由表，然后，将路由表下载到各个转发引擎中的转发表。控制器和转发引擎之间交互可以采用 OpenFlow 或 ForCES 协议[14]

进行交互。当转发引擎支持 OpenFlow 交换模型，可以采用 OpenFlow 控制协议进行通信，如图 3 所示。控制器通过 OpenFlow 协议对转发引擎流表进行配置，控制转发引擎处理 IP 报文的流程。

控制器将服务单元作为计算能力更强，处理更复杂的转发引擎，对服务单元的转发策略进行管理和配置。控制器发现网络拓扑后，并根据系统区域网拓扑结构，为服务单元计算最优的转发路径，并下载到服务单元，控制服务单元的转发策略。

在开放控制体系结构支持下，路由器系统区域网连接的服务单元中运行的路由器控制器对系统区域网连接的路由器转发引擎、服务单元和存储单元等多种资源进行统一管理和控制。在集中式控制下，路由器系统区域网连接的转发引擎、服务单元和存储单元等构成了集成计算和存储资源紧耦合的新型路由器。

3.5 CSRouter 特点

1) 加速网络应用

针对处于数据处理路径上的网络应用，CSRouter 能够缓存网络应用数据，代理网络应用服务器，提前向端系统返回应答数据，减小请求等待时间，加速网络应用。CSRouter 数据平面中报文转发引擎 PFE 根据流规则表对报文处理，对于匹配流规则表的报文流送到服务平面中服务单元处理。服务单元对报文流进行加密或解密、压缩或解压缩等数据处理。或者服务单元识别报文流应用层协议再进行深度处理，如根据缓存结果和服务策略向端系统返回报文流中应用请求的数据。CSRouter 也能够代理无线接入的移动终端处理与网络服务器交互，向移动端系统返回结果。

2) 优化数据传输

基于路由器系统区域网的紧耦合的报文处理单元、服务单元和存储单元根据流规则表顺序处理报文的同时，还能够共享上级报文处理结果和网络拓扑信息，在报文处理方式上具有更高灵活性。当 CSRouter 报文转发引擎根据流规则表报文处理策略需要将报文直接缓存到存储单元，或者从存储单元读取数据，能够利用网络拓扑和高性能系统区域网络拓扑决定目标存储单元和目标输出转发引擎，优化报文在路由器中传输路径，实现高效组播。另外，当 CSRouter 转发引擎根据流规则表需要将报文送服务单元处理时，只需通过系统区域网络将报文描述符送到服务单元处理，而服务单元根据需要通过系统区域网络访问相关报文数据，减少服务单元传输数据开销。

3) 支持第三方服务

传统路由器是封闭式开发环境，只向用户提供了简单的管理 CLI 等，用户增加新的功能十分困难。新型路由器体系结构 CSRouter 融合计算和存储资源，向用户提供开发接口，并提供大量共享组件，便于用户根据需要开发合适的功能。用户通过流规则表捕获相关报文流，进行深度处理。

4 系统实现与测试

本节基于 Netmagic 和 IEF 统一交换平台搭建了 CSRouter 原型系统，设计了转发引擎和服务单元通信、报文调度、应用开发模型等，最后对原型系统的内部带宽和延迟等进行测试。

4.1 CSRouter 原型系统

基于自主研发的网络交换平台 NetMagic[15]实现转发引擎的功能，基于统一交换网关 IEF 设计了系统区域网，连接服务器和盘阵控制器搭建了 CSRouter 原型系统，如图 4 所示。NetMagic 是自主研发的网络体系结构创新研究平台，由通用交换 ASIC 和可编程的 FPGA 构成，具有强大的交换能力和可编程的处理能力。NetMagic 支持高速的 IP 报文查找与转发，实现转发引擎的功能。IEF 网关是自主研发的支持异构网络互连的无状态协议转换网关，由 BridgeX ASIC 构成，实现协议转换，支持 InfiniBand 网络、Ethernet 网络和 FC 网络的接入和统一交换，是高性能的路由器系统区域网的核心构件。

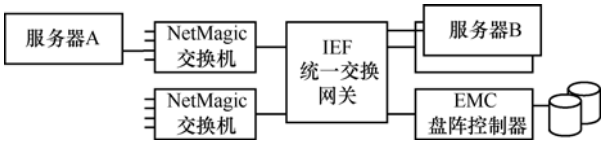


图 4 CSRouter 原型系统

通常路由器转发引擎具有 Ethernet 接口，高性能计算服务器采用 InfiniBand 接口，为了实现转发引擎与服务器的互连，需要实现 InfiniBand 协议到 Ethernet 协议的转换。存储系统采用 Fibre Channel 接口，为了实现服务器与存储系统的互连，需要实现 InfiniBand 协议到 FC 协议的转换，为了实现转发引擎与存储系统的互连，需要实现 Ethernet 协议到 FC 协议的转换。InfiniBand 交换技术具有高性能可扩展、集中式控制等特点。目前以 InfiniBand 交

换技术为核心，结合 InfiniBand 协议到 FC 协议，InfiniBand 协议到 Ethernet 协议转换网关，设计面向路由器组件互连的路由器系统区域网是一种最佳方案。

基于系统区域网路由器的集中式控制器运行在服务器 B 中，对网络进行管理，计算路由等。NetMagic 提供了 NMAC 协议，实现控制器与 NetMagic 的交互，控制器通过 NMAC 协议将转发表下载到 NetMagic。服务器 B 中运行文件系统对盘阵控制器进行管理和控制，服务器 B 向 NetMagic 提供基于文件系统的存储资源，支持 NetMagic 将数据直接写入到盘阵控制器。

4.2　基于路由器系统区域网的报文传输机制

基于系统区域网的新型路由器 CSRouter 报文处理单元包括转发引擎、服务单元和存储单元等，高效的转发引擎中 IP 报文处理进程和服务单元中服务进程进行通信，服务单元中服务进程和盘阵控制器中存储控制进程通信等是路由器性能优化的关键。

1) 转发引擎与服务单元间数据传输

在路由器中转发引擎根据报文处理策略，对报文进行报文 IP 查找后，在本地进行交换转发，从其他端口输出，或者经系统区域网络转发给其他服务单元进行处理。转发引擎中网络处理器通过 TCP/UDP Socket 向服务单元转发 IP 报文，通过 Ethernet 接口发送到 Ethernet 网络。Ethernet 数据 EthPDU 在网关处封装到 InfiniBand 链路帧 IBPDU(EthPDU)，在 InfiniBand 网络中传输，再转发到服务单元。

当服务单元从 InfiniBand 网络接口接收到 IBPDU(EthPDU)后，InfiniBand 驱动对报文进行解封装，获得以太网帧 EthPDU，转发到上层以太网驱动，以太网驱动再将请求转发给服务进程进行处理，如图 5 所示。

2) 服务单元和存储单元间数据传输

在报文深度处理过程中，在一个服务单元可能需要访问存储，对报文进行缓存，或者从存储系统读取数据。

服务单元在 IB 网络接口之上虚拟 FC 接口，将访问存储系统的 FC 协议数据封装在 IB 链路帧，在虚拟 FC 网络接口和 FC 网络间传递。

当服务单元接收到转发引擎转发的 IP 报文，需要将报文缓存，或者需要从存储系统读取报文时，通过虚拟的 FC 端口访问磁盘阵列。读写盘阵的 IO 请求封装在 IB 链路帧中，发送到 IB/FC 网关。IB/FC 网关将 IB 协议数据单元解封装，发送到 FC 网络中对应的磁盘控制器。磁盘控制器读取数据块，经 FC 链路发送。请求的结果经过 IB/FC 网关封装后发送到计算服务器，如图 5 所示。

从转发引擎与服务单元通信过程、服务单元和存储单元通信过程可以看出，在服务单元 InfiniBand 网络接口之上，不仅要虚拟 Ethernet 接口，实现与转发引擎的通信，还要虚拟 FC 网络接口，实现与存储系统的通信。服务单元是路由器系统的关键，由于采用系统区域网连接计算服务器，具有良好可扩展性，因此可以根据路由器处理 IP 报文性能需要，不断增加计算和存储资源，实现对路由器性能的升级。

4.3　基于流速—时间积的报文调度机制

转发引擎与多个服务单元通过系统区域网

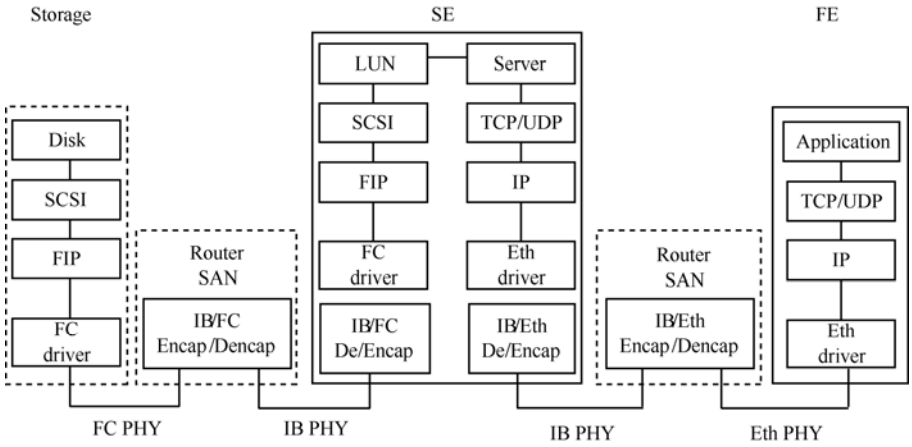


图 5　基于路由器系统区域网的传输机制

连接，转发引擎可能需要将接收到的IP报文转发给多个服务单元。为了保证报文处理的效率，转发引擎需要将报文平均分派到多个服务单元进行处理。转发引擎报文调度器有多种机制可以将接收到的报文分配到服务单元，如随机方式和轮询方式。

为了报文流一致性，转发引擎报文调度器根据报文流标识将报文流分配到合适的服务单元进行处理。当报文调度器从转发引擎接收到IP报文，计算报文流标识。根据报文流标识查找映射表，若报文流的映射存在，则将报文转发到目标服务单元进行处理；若报文流的映射不存在，则为报文流选择负载最轻的目标服务单元，并在流表中创建映射表项。

当创建新的报文流映射关系时，若流表中存在空闲的表项，则将映射关系记录在流表中；若流表中不存在空闲的表象，则将最长时间没有报文到达的报文流表项删除。

由于报文流流量的不均衡性，在固定的报文流映射关系下，可能导致服务单元处理报文的不均衡性，造成有些服务单元拥塞，而有些服务单元空闲。选择映射到负载最重的服务单元的报文流重新映射到负载最轻的服务单元，均衡服务单元负载。其中，服务单元的负载是根据映射到服务单元的报文流的流速和持续时间乘积决定的。

4.4　服务开发模型

对需要在网络边缘加速的应用，要移植到CSRouter，以利用CSRouter路由器提供的计算和存储资源。在CSRouter中服务单元采用Linux系统，通过底层驱动对存储单元管理，并连接到网络资源；同时服务单元向上层网络应用提供了共享组件，如缓冲区管理、查表、压缩和解压缩等，加速上层应用的开发，如图6所示。当客户请求到达网络边缘CSRouter节点时，CSRouter中报文转发引擎对报文进行深度处理，解析IP报文请求，查找本地缓存，在请求匹配的情况下将数据直接返回给客户端，形成基于平台的边缘—边缘的服务模式。

在融合计算和存储资源的新型路由器支持下，网络应用能够部署在网络边缘，接近网络应用的客户端。用户请求能够在网络边缘处理，这种通信模式能够减小IP请求和应答报文在网络中的转发次数，减小报文延迟，提高网络性能和利用率。

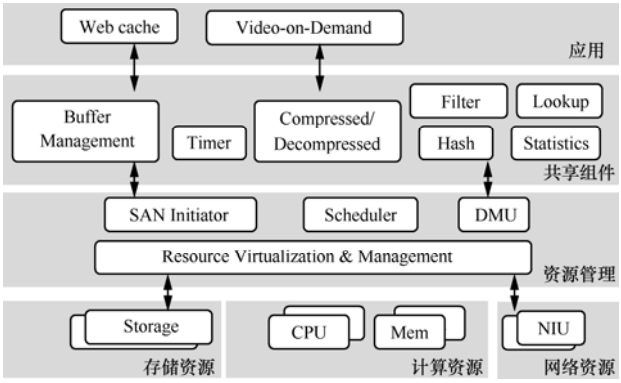


图6　网络服务开发模型

4.5　路由器系统区域网性能分析

基于系统区域网的路由器CSRouter中报文需要经过转发引擎和服务单元等处理，在处理过程中，报文需要通过系统区域网在转发引擎和服务单元间多次传输，因此，转发引擎到服务单元的系统区域网带宽和延迟对于路由器转发性能有很大的影响。在原型系统中实现了NDN中报文缓存服务[17]，对转发引擎到服务单元的带宽和延迟等参数进行了测量。

在测试环境下，服务器A连接NetMagic，NetMagic将服务器A发送的报文通过IEF转发到目标服务单元，用iperf测量服务器A与服务单元之间的带宽。在测试过程中，服务器A产生多条流发送到服务单元。测试不同报文大小的情况下网络的聚合带宽。从测试结果可以看出，随着报文大小的增加，网络的聚合带宽也在不断增加，如图7所示，最大达到956Mbyte/s，接近网络端口理论值。在新型路由器体系结构中系统区域网的带宽能够满足转发引擎和服务单元之间报文转发需要。

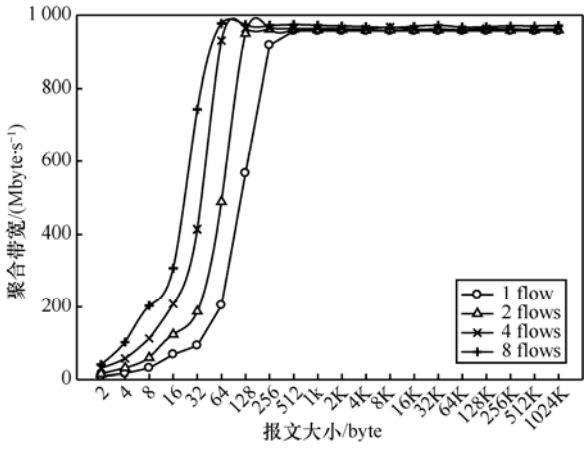


图7　CSRouter内部单端口有效带宽

在测试环境下，服务器 A 连接 NetMagic，NetMagic 将服务器 A 发送的报文通过 IEF 转发到目标服务单元，用 iperf 测量服务器 A 与服务单元之间的延迟。在测试过程中，服务器 A 产生多条流发送到服务单元。从测试结果可以看出，随着流数据的增加，平均延迟在增加，如表 1 所示，但是最大延迟也只有 0.037ms，而且没有分组丢失。基于高性能统一交换技术构建的路由器系统区域网实现了高带宽和高可靠的数据传输，支持转发引擎和服务单元间数据可靠传输。

表 1　　CSRouter 处理客户请求的延迟

流数目	平均延迟/ms	分组丢失率
10	0.001	0%
20	0.008	0%
30	0.012	0%
40	0.019	0%
50	0.024	0%
60	0.028	0%
70	0.031	0%
80	0.037	0%

4.6　增量部署

基于系统区域网络的路由器 CSRouter 融合了计算和存储等多种资源，提供了开放的服务运行平台，能够根据用户需求对 IP 报文进行深度处理，如基于应用层协议识别，代理网络服务器端系统处理请求报文，对于客户端系统是透明的，客户端系统的处理是原有的处理流程，不需要对客户端系统进行修改。另外，基于系统区域网络的路由器运行的控制协议是正常路由器的控制协议，能够与普通路由器进行交互。因此，新型路由器的部署可以增量式部署，不断地在网络边缘进行部署，提高本地网络处理性能和传输效率。

5　结束语

传统路由器是一种转发引擎紧耦合的体系结构，系统性能取决于转发引擎，而转发引擎升级空间有限，通常采用替换方式，不能保护已有路由器的投资。基于系统区域网的新型路由器体系结构 CSRouter 采用统一交换技术构建路由器系统区域网，连接转发引擎、计算资源和存储资源等，能够根据网络应用需求不断加入转发引擎、服务单元和存储单元，实现报文的服务—转发处理，提升路由器的处理能力和可扩展性。新型路由器体系结构能够利用强大的计算资源和存储资源，对网络 IP 报文进行深度处理，加速网络应用响应，提高网络传输效率。新型路由器内部采用了开放控制体系结构，在集中式管理控制下，具有独立性，支持增量式部署。

基于系统区域网络的新型路由器体系结构在普通路由器中集成了计算和存储资源，为在网络中进行网络应用开发提供了开发实验平台。下一步，将基于 NetMagic 和 IEF 构建的原型系统提供开放的网络应用开发平台和实验环境，加速网络应用创新。

参考文献：

[1] Nebula project[EB/OL]. http://nebula.cis.upenn.edu.2011.

[2] SIMON H, THOMAS W, ARTHUR M, *et al*. Packet processing at 100 Gbps and beyond-challenges and perspectives[A]. ITG Symposium on Photonic Networks[C]. Leipzig, Germany, 2009.

[3] ZEPPENFELD J, HERKERSDORF A. Autonomic workload management for multi-core processor systems [J]. Architecture of Computing Systems (ARCS), Lecture Notes in Computer Science, 2010, 5974:49-60.

[4] LU G H, GUO C X, LI Y L, *et al*. Serverswitch: a programmable and high performance platform for data center networks [A]. Proceedings of 8th USENIX Symposium on Networked Systems Design and Implementation(NDSI)[C]. Boston, MA, USA, 2011.

[5] HAN S J, JANG K, PARK K S, *et al*. Packetshader: a GPU-accelerated software router[J]. ACM SIGCOMM Computer Communication Review- SIGCOMM, 2010,40(4): 195-206.

[6] MCKEOWN N, ANDERSON T, BALAKRISHNAN H, *et al*. Openflow: enabling innovation in campus networks[J]. ACM SIGCOMM Computer Communication Review, 2008, 38(2):69-74.

[7] WELLING G, OTT M, MATHUR S. A cluster-based active router architecture [J]. IEEE Micro, 2001, 2(11): 16-25.

[8] TURNER J. A proposed architecture for the GENI backbone platform [A]. Proceedings of ACM-IEEE Symposium on Architectures for Networking and Communications Systems (ANCS)[C]. San Jose, California, USA: IEEE Computer Society, 2006.

[9] Juniper routers[EB/OL]. http://www.juniper.net/de/de/products-services/ routing/mx-series, 2009.

[10] 徐恪, 吴鲲, 王青青. 可扩展路由器控制平面的高性能通信模型[J]. 软件学报, 2007, 18(9): 2205-2215.
XU K, WU K, WANG Q Q. High performance control-plane communication model for scalable routers[J]. Journal of Software, 2007, 18(9): 2205-2215.

[11] NORBERT E, GREENHALGH A, HANDLEY M, *et al*. Improved forwarding architecture and resource management for multi-core software routers [A]. IFIP Conference on Networking and Parallel Computing[J]. Gold Coast[C]. Australia: IEEE Computer Society, 2009.

[12] 徐明伟, 江学智, 陈文龙. 路由器分布式控制研究综述[J]. 电子学报, 2010,38(8): 1892-1899.
XU M W, JIANG X Z, CHEN W L. Survey on distributed control in a router[J]. Acta Eletronica Sinica, 2010, 38(8):1892-1899.

[13] MARKUS H, SJÖDIN P, HAGSAND O. Control and forwarding plane interaction in distributed routers[J]. IFIP International Federation for Information Processing Networking, 2005,3462: 1339-1342.

[14] Forwarding and control element separation (ForCES)[EB/OL]. http://www.rfc-editor.org/info/rfc5812, 2010.

[15] Netmagicresearch group. Using netmagic to accelerate network technology innovation[EB/OL]. http://www.netmagic.org/Data/documents/presentations/NetMagic%20 workshop.pdf.

[16] CHOI Y H, TIMOTHY M P. Crossbar analysis for optimal deadlock recovery router architecture[A]. Proceedings of the 10th International Parallel Processing Symposium[C]. Geneva, Switzerland, 1997. 583-588.

[17] GHODSI A, KOPONEN T, RAJAHALME J, *et al*. Naming in content-oriented architectures[A]. Proceedings of SIGCOMM Workshop on ICN[C]. Toronto, Ontario, Canada, 2011.

作者简介：

吕高锋（1980-），男，陕西扶风人，博士，国防科学技术大学助理研究员，主要研究方向为计算机网络、高性能路由器、统一交换技术等。

孙志刚（1973-），男，江苏东海人，博士，国防科学技术大学研究员，主要研究方向为网络体系结构、高速网络交换技术等。

林雨弦（1988-），男，福建福安人，国防科学技术大学硕士生，主要研究方向为计算机网络、可扩展路由器体系结构等。

陈一骄（1972-），男，湖南益阳人，博士，国防科学技术大学副研究员，主要研究方向为计算机网络、高性能路由器、网络安全等。

李韬（1983-），男，安徽萧县人，博士，国防科学技术大学助理研究员，主要研究方向为计算机网络、网络处理器、路由与交换技术。

（上接第 48 页）

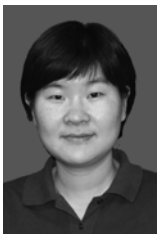
on Intelligent Information Hiding and Multimedia Signal Processing[C]. Harbin,China,2008.1392-1395.

[6] SUBHAMOY M, GOUTAM P, SHASHWAT R. Some observations on HC-128[EB/OL]. http://eprint.iacr.org 2008/499.pdf. 2010.10.

[7] GAUTHAM S, BART P. Improved distinguishing attacks on HC-256[A]. Proceedings IWSEC'09[C]. Toyama, Japan, 2009.38-52.

[8] SEKAR G, S, PRENEEL B. New weaknesses in the keystream generation algorithms of the stream ciphers Tpy and Py[A]. Proceedings of ISC'07- 10th Information Security Conference on Information Security[C]. Dresden, Germany, 2007. 249-262.

作者简介：



高海英(1978-)，女，河南沈丘人，博士，解放军信息工程大学副教授，主要研究方向为密码理论。

金晨辉（1965-），男，河南扶沟人，博士，解放军信息工程大学教授、博士生导师，主要研究方向为密码理论。