文章编号:1671 - 6833(2019)02 - 0012 - 06

支持 IPv6 试验和部署的新型数据平面结构研究

黄万伟[1], 杜春锋[2], 张建伟[1], 段 通[3]

(1. 郑州轻工业学院 软件学院,河南 郑州 450002; 2. 郑州轻工业学院 计算机与通信工程学院,河南 郑州 450002; 3. 国家数字交换系统工程技术研究中心,河南 郑州 450002)

摘 要: 为了解决 IPv4 网络体系结构和设备难以满足 IPv6 网络在数据包解析、匹配、动作执行等方面的问题,提出了一种支持 IPv6 试验和部署的新型数据平面结构. 该结构包含一种同时支持策略和功能的数据平面抽象机制以及一种应用于 IPv6 下一代互联网的数据平面硬件结构,支持多种创新网络体系结构在同一网络中共存,实现对新型协议的试验和验证,支持可定制的协议解析、灵活可编程的分组处理以及内部资源的动态组合,以支撑未来网络功能的试验、部署和评估. 通过系统试验和分析,验证了该结构在可接受资源开销的情况下具有较高的转发性能.

关键词: 规模部署; 数据平面; 抽象机制; 硬件结构; 可编程

中图分类号: TP393.2 **文献标志码:** A **doi:**10.13705/j.issn.1671 - 6833.2019.02.021

0 引言

随着信息化和网络化的不断发展,互联网已成为人们日常工作、学习和生活必不可少的基础设施. 传统互联网基于 IPv4 协议构建,由于 IP 地址匮乏和服务质量难以保证等问题,已严重制约互联网的进一步应用和发展. IPv6 协议以其海量的地址空间、完善的服务质量保证机制和巨大的创新空间,成为公认的构建下一代互联网的解决方案. 前期,我国在 IPv6 试验和应用方面做了大量的工作,取得了一些重要成果,已具备大规模部署的条件[1].

IPv6 的部署和应用必将重塑网络体系结构,对网络信息技术、产业、应用的创新及变革产生深刻影响. IPv6 丰富的地址空间和头部字段对网络节点在数据包解析、查找匹配以及动作执行等方面的处理能力提出了更高的要求[2]. 传统基于 IPv4 的网络体系和设备难以适应 IPv6 的规模部署和应用. 如果未来网络设备的数据平面能够支持用户定制,那么试验和部署新型网络协议和网络功能将变得十分便利,未来网络也将变得十分开放.

新型网络协议(如 IPv6 协议)和创新网络功能的试验和验证,需要网络设备能够实现可定制可编程的数据包解析,以便能够按照新协议格式提取匹配域. 为实现用户可编程可定制的匹配域提取方式,文献[3]和文献[4]分别提出了 CAFE 和 SwitchBlade,通过在数据包头部解析模块中设计任意比特抽取器,CAFE 和 SwitchBlade 实现了数据包头部任意比特自由组合,但数据包解析性能十分受限. 为实现高性能和高灵活性之间的折中,Kangaroo 结构[5]利用可编程协议树在实现多种数据包解析的同时,达到了 40 Gbps 的线速解析能力.

网络设备数据平面的处理包含解析、查找、匹配和执行等一系列动作,以上研究工作属于数据包解析方面的研究成果,其他方面的研究工作也取得了一定的进展. 软件定义网络(software defined networking, SDN)[6]实现了控制与转发的分离,可以支持用户定制网络功能,但目前 SDN 仅开放了控制平面,数据平面只能支持 MPLS 和 TCP/IP 协议,对 IPv6 等新型的网络协议数据包并不支持. 未来网络创新功能和协议需要数据平面的开放能力,为实现数据平面可编程,以支持对

收稿日期:2018 - 05 - 17;**修订日期:**2018 - 08 - 11

基金项目:国家自然科学基金资助项目(61672471);赛尔网络资助项目(NGII20160103);郑州轻工业学院校博士基金(2016BSJJ041)

作者简介:黄万伟(1979—),男,江苏盐城人,郑州轻工业学院讲师,博士,主要研究领域为人机智能交互系统、宽带信息网和硬件系统开发,E-mail:huangww79@163.com.

新型协议的适配,Nick McKeown 提出了 P4(programming protocol-independent packet processors)[6]. P4 是一种对底层设备数据处理行为进行编程的高级语言,用户可以直接使用 P4 语言编写网络应用,之后经编译对底层设备进行配置进而完成用户的需求. 与此类似的还有华为提出的 POF[7](protocol oblivious forwarding),这两种实现方式虽然提高了底层转发设备的可编程性,但需要专门的编译系统或解释系统,实现比较复杂. 为降低实现复杂度,提出基于通用 FPGA 的可编程多级流表架构[8],可编程流表架构为各级流表分配匹配、查找和动作等资源,各级流表之间可动态组合,灵活度较高,但实现难度较大. 针对现有 SDN 转发平面的不足,国防科大提出了一种普适的 SDN 转发平面抽象 LabelCast[9],能够对新型网络协议的转发行为进行抽象,因为是依赖软件实现故转发性能有待进一步提高. 文献[10]基于 FPGA 提出了一种支持网络功能演进的新型数据平面结构,该结构通过可编程的数据包解析和数据包处理达到内部逻辑可重构[11],从而实现了用户功能的可定制,对数据平面的设计和发展具有一定的参考意义.

为了支持 IPv6 和不断涌现的新兴网络协议和功能的试验和部署,结合以上分析,笔者提出了一种支持 IPv6 试验和部署的新型网络数据平面结构,该结构包含一种同时支持策略和功能的数据平面抽象机制和应用于 IPv6 下一代互联网的数据平面硬件实现结构,支持多种创新网络体系结构在同一网络中共存,实现对新型协议的试验和验证,支持可定制的协议解析、灵活可编程的分组处理以及内部资源的动态组合,以支撑未来网络功能的试验、部署和评估.

1　新型数据平面虚拟化抽象机制及实现结构

1.1　数据平面虚拟化抽象机制

网络数据平面是承载和实现网络功能和协议的重要载体,灵活支持各种新型的网络功能和协议是未来网络对数据平面属性的基本要求[12]. 设计灵活可编程的网络数据平面,需要对网络功能和协议进行建模. 当前针对网络功能和协议建模主要有两种模型:①逻辑功能元素模型,典型的功能模型如 OpenFlow 定义的流表、防火墙的 ACL 模型;②策略驱动处理模型,典型的策略模型如 ForCES 中定义的 FE 模型、SNMP 中定义的 MIB 模型. 逻辑功能模型从内部功能进行描述,目标功能从粗到细进行分解,便于功能实现;策略驱动处理模型则从外部功能进行描述,目标功能通过外部策略条件驱动内部功能执行,便于外部控制.

从以上网络功能和协议的抽象模型可以看出,不同的网络功能和协议适用于不同的抽象模型. 为了支持现有多种传输模式与复杂化的网络功能,笔者提出一种同时支持策略和功能的数据平面虚拟化抽象机制. 基于策略与功能驱动的数据平面虚拟化框架如图 1 所示,该框架基于统一的硬件抽象层,给网络应用提供标准的可编程接口,在同一网络中通过虚拟化技术支持多种网络体系共存,可以方便地试验新型协议和网络功能.

数据平面虚拟化抽象机制的设计核心可分为位于接口交换部件的策略映射表和位于数据处理部件的功能映射表. 策略映射表的设计目标是易于可编程硬件或者可配置的硬件实现. 为支持多种网络体系共存,可设计网络内容与定长策略的映射,以实现对报文转发层的统一抽象,进而实现对不同体系结构类型、不同网络业务类型报文基于策略的高速转发. 为简化硬件层设计,笔者在硬件层没有定义除了转发之外的操作,例如丢弃、存储、服务等,若要对报文深层次处理,则需要上送至数据处理部件进行进一步处理. 在实现模型中,功能映射表和数据平面抽象的控制部分位于后端的数据处理部件,统一采用软件实现. 控制部分通过标准协议配置策略映射表和功能映射表,并对表象进行修改、增加、删除、读取等操作.

基于上述模块,可设计出相应的网络应用接口,以实现初始化、模块报文处理启动、模块卸载、虚拟节点注册处理、虚拟节点注销处理、策略请求、策略通知、策略作废等功能. 用户不需要关心底层实现细节就可以通过调用这些接口函数,将新的网络协议嵌入到控制层中.

基于上述机制,还可以实现数据平面的虚拟化. 数据平面虚拟化主要是转发引擎的虚拟化,对控制平面来说,数据平面的虚拟化不需要操作系统级的支持,只需要将属于每个虚拟节点的任务封装入进程容器,并在这些容器之间流量隔离,从而实现数据平面的资源虚拟化隔离.

图 1 为数据平面虚拟处框架. 在图 1 中,每个逻辑转发引擎对应系统中的一个进程实例,通过操作系统的亲核机制映射到不同的处理器核心上. 基于内存为每个虚拟节点容器初始化若干

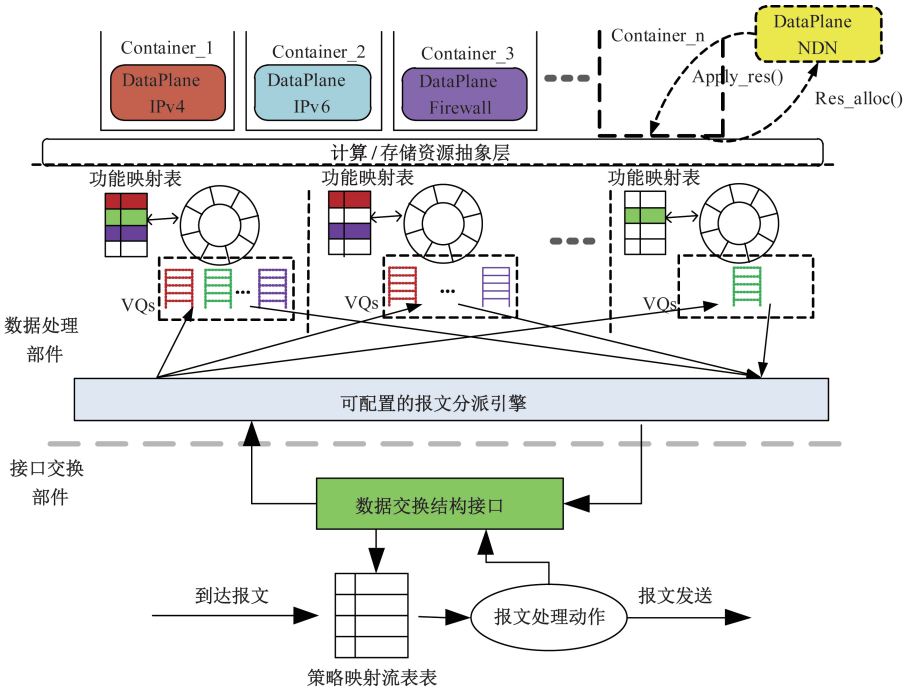

图1 基于策略与功能驱动的数据平面虚拟化框架

Fig. 1 Based on policy and feature-driven data plane virtualization framework

条虚拟队列,将收到报文的描述符送到对应的虚拟队列(virtual queues,VQs)中,所有逻辑转发引擎通过虚拟队列的消息队列收发报文.

数据处理部件的物理接口由可配置报文分派引擎统一接管.报文分派引擎利用抽象机制中的功能映射表确定报文所属的虚拟节点,将其送入目标虚拟节点所对应的虚拟队列.虚拟节点完成对报文的处理之后,通过虚拟队列区发回报文描述符,等待报文分派引擎发送.通过调整分派引擎的交付速率可以控制每个虚拟节点所能享受的处理器资源和带宽资源,从而实现调度和隔离.

1.2 数据平面硬件实现结构

上节给出了一种支持IPv6试验和部署的新型数据平面抽象机制,能够很好地支持多种网络体系共存,但随着网络高速化、宽带化,部分创新的网络功能和协议迫切需要高性能的网络创新试验平台进行部署和试验.为此,必须设计一种应用于下一代IPv6互联网的数据平面硬件结构,以支持未来网络功能的试验、部署和评估.

针对下一代IPv6互联网的需求,新型数据平面硬件结构应具有3个特征.一是支持协议解析的可定制性.网络数据包头部一般包含类型域和匹配域,类型域表示数据包协议类型,匹配域包含匹配字段.类型域和匹配域的提取是数据包解析和处理的前提,不同数据包可能具有不同的协议类型和匹配域,在进行数据包头部字段提取时可采用多叉树表示,每个类型域是一个树节点.这样就能采用解析树精确提取数据包的类型域与匹配域,实现对任意类型协议数据包的处理.二是支持分组处理灵活可编程,分组处理过程可抽象为匹配、查找、动作3个步骤,其中匹配查找是实现分组处理灵活可编程的关键.类似于数据包头部的提取过程,可以采用多叉树来表示匹配查找过程,每一个匹配域都是一个树节点,子树可以用来表示网络功能,网络功能的匹配域可以用树的匹配域节点代替,操作类型则可以用树的叶子节点表示,这样就能将分组处理过程对应到匹配树上,也就为数据处理的灵活可编程可定制找到了一种有效的解决方案.三是支持动态组合内部资源.网络资源的高效利用是未来新型数据平面的基本要求,如何基于有限的资源来满足各种各样的创新网络功能,是必须要克服的难题.网络资源灵活组合是解决上述难题的关键.网络数据以分组表示,任何协议和网络的处理过程都可看作分组的匹配、查找和动作执行过程,因此,如果能够将网络资源抽象为匹配、查找和动作等细粒度模块,那么通过灵活组合这些模块就能支持各种创新网络功能.

基于以上思想,笔者提出了一种支持IPv6试验和部署的新型网络数据平面硬件实现结构.该结构主要包括包头解析器和元处理单元,包头解

析器用来判断数据包协议类型和提取匹配域,将匹配域输送至后级元处理单元. 元处理单元是数据包"匹配 + 查找 + 动作"操作的实现,是该结构中最基本也是数量最多的数据包处理单元. 元处理单元之间的灵活组合可实现复杂的网络功能,元处理单元之间的信息传递采用元数据. 这样通过可配置的包头解析器和可灵活组合的元处理单元就可以实现数据平面内部逻辑的可扩展,从而支持用户对新型网络协议和创新功能进行试验和验证. 数据平面硬件结构如图 2 所示.

图 2　数据平面硬件实现结构

Fig. 2　Data plane hardware implementation structure

在图 2 所示的结构中,解析器负责对数据包进行解析、提取和组合,首先根据用户配置信息识别数据包的类型域,并提取该类型域,送入匹配查找模块进行匹配查找操作,根据匹配查找结果读取匹配域偏移量,并送到匹配域提取模块,匹配域提取模块根据偏移量提取匹配域字段,并将匹配域字段组合成完整的包头域送到处理单元进行处理.

在图 2 所示的硬件结构中,将数据包处理单元细分为元处理单元,每个元处理单元都是最基本的数据包处理单元,由匹配域选择器、匹配查找、动作执行等组成,元处理单元之间可以进行组合,以完成复杂的网络处理功能. 当数据包包头域到达元处理单元时,匹配域选择器会根据用户配置的匹配域选择信息提取相应的匹配域字段,送到流表匹配查找模块,根据查找结果选择相应的执行动作.

在笔者所提数据平面硬件实现结构中,包头解析器是实现对 IPv6 等新型协议支持的关键模块. 它根据用户的配置识别数据包的类型域,同时根据类型域提取相应匹配域并将其组合得到包头域向后级元处理单元输出.

包头解析器如图 3 所示,包头解析器结构包含类型域提取模块、匹配查找模块、匹配域提取模块和匹配域组合模块. 其中类型域提取模块用于识别数据包头并提取类型域,首先,类型域提取模块根据 RAM1 中的初始类型域偏移量将第一层协议包头的类型域提取出来,通过 TCAM + RAM2 匹配查找得到下一层协议包头的类型和所需匹配域的偏移量;当接收到从匹配查找模块输出的下一状态时类型域提取模块将当前状态更新至下一状态. 匹配查找模块包含一个 TCAM 单元和一个 RAM2 存储单元,其中 TCAM 中存放状态信息和用户定制的类型域信息,RAM2 中存放类型域所对应的该协议包头所需的匹配域的偏移量信息. 匹配查找模块利用 TCAM 匹配类型域和状态,根据匹配结果在 RAM2 中读取到下一状态和对应匹配域的偏移量,并分别向类型域提取模块和匹配域提取模块输出. 匹配域提取模块根据匹配域的偏移量将所需匹配域提取出来. 最后,匹配域组合模块将提取到的匹配域组合成包头域并送往后级元处理单元处理.

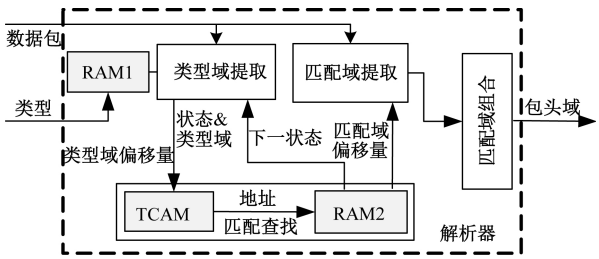

图 3　包头解析器结构

Fig. 3　Packet header parser structure

2　性能仿真与分析

笔者基于 x86 多核服务器和 NetFPGA - 10G[13] 板卡构建了一个支持 IPv6 试验和部署的新型数据平面原型系统,并验证了数据平面的性能. 数据平面原型系统由配置单元、收发单元和处理单元组成,配置单元主要用于接收用户配置信息并将配置表项送到处理单元;4 个 10 Gbps 的物理端口和 1 个虚拟端口组成收发单元,虚拟端口通过 DMA 与服务器虚拟网卡相连,物理端口作为收发包端口与外部网络相连;处理单元是新型数据平面的核心,由数据包头部解析器和 4 个元处理单元组成.

本节从资源开销和转发性能两个方面对新型数据平面结构进行性能仿真和分析. 首先对包头解析器和动作处理器占用的资源开销和性能进行分析;然后,从整体上对新型数据平面结构的转发性能进行验证和分析.

2.1　资源开销与性能分析

相比于传统的数据平面结构和其他可编程数

据平面结构,笔者提出的新型数据平面结构主要在包头解析器部分增加了资源开销.本节利用FPGA仿真工具对包头解析器的资源开销和性能进行了仿真分析,假设布局布线时钟为178.6 MHz(理论转发速率可达182.8 Gbps),数据总线位宽为1 024 bits,与EPC和Kangaroo的资源开销和性能作了比较如表1所示.

表1 包头解析器资源与性能对比

Tab.1 Header parser resource and performance comparison

属性	Slice资源	BRAM资源	转发速率/Gbps
笔者	3 098	21	183
Kangaroo	2 500	51	40
EPC	4 013	29	203

从表1可以看出,与Kangaroo相比,虽然笔者提出的新型数据平面结构Slice资源开销提高了24%,但转发速率提高了4倍左右,同时BRAM资源降低了约58%;与EPC相比,资源开销降低了约24%.综合来看,笔者所提新型数据平面结构性价比最优.

2.2 转发性能分析

对整体转发性能进行了实验验证,并与LabelCast线程数进行对比分析.由于NetFPGA资源有限,因此,本节在实现笔者所提数据平面结构时,数据总线位宽和元处理单元的处理域宽度都设定为64 bits,布局布线时钟分别设置为:172.6 MHz[使用1级元处理单元(记作RHS1)]和163.8 MHz[使用4级元处理单元(记作RHS4)].整体转发性能对比如图4所示.

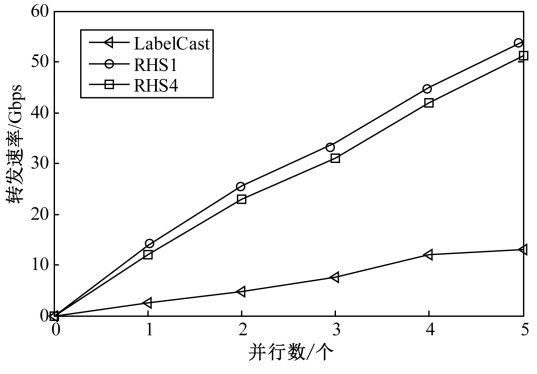

图4 整体转发性能对比

Fig.4 Overall forwarding performance comparison

图4为LableCast和笔者所提数据平面结构的转发速率对比情况.从图4中可以看出,相比于Labelcast,不论采用几级的元处理单元,笔者所提新型数据平面结构的转发速率都提升了4倍左右;同时,随着并行数增加,笔者所提数据平面结构转发速率能够接近线性增长,而Labelcast则增长缓慢.需要说明的是,元处理单元增多资源开销也会增大,因此,RHS4的转发速率相比RSH1有所下降.

3 结论

针对IPv6大规模试验和部署对数据平面带来的挑战和问题,笔者基于多核服务器和NetFPGA平台提出并实现了一种支持IPv6试验和部署的新型数据平面结构,该结构包含一种同时支持策略和功能的数据平面抽象机制和一种应用于IPv6下一代互联网的数据平面硬件结构.与其他方案相比,笔者所提方案具有更低的资源开销和更高的转发速率,对IPv6下一代互联网数据平面的设计具有一定的参考意义.

参考文献:

[1] 中共中央办公厅.推进互联网协议第六版(IPv6)规模部署行动计划[EB/OL].[2018-06-11].http://www.gov.cn/zhengce/2017-11/26/content_5242389.htm.

[2] 马芳,吉星.一种优化分层式移动IPv6路由算法研究[J].郑州大学学报(工学版),2015,36(3):125-128.

[3] LU G, SHI Y, GUO C, et al. CAFE: a configurable packet forwarding engine for data center networks[C]// Proceedings of the 2nd ACM SIGCOMM workshop on programmable routers for extensible services of tomorrow (PRESTO), 2009: 25-30.

[4] ANWER M B, MOTIWALA M, TARIQ M B, et al. SwitchBlade: a platform for rapid deployment of network protocols on programmable hardware[J]. ACM SIGCOMM computer communication review, 2010, 40(4): 183-194.

[5] KOZANITIS K, HUBER J, SINGH S, et al. Leaping multiple headers in a single bound: wire-speed parsing using the Kangaroo system[C]// Proceedings of INFOCOM 2010, 2010: 1-9.

[6] MCKEOWN N. Keynote talk: software-defined networking[C]// Proceedings of the IEEE INFOCOM, 2009: 1-11.

[7] BOSSHART P, GIBB G, KIM H S, et al. Forwarding metamorphosis: fast programmable match-action processing in hardware for SDN[C]// Proceedings of the ACM SIGCOMM'13 Conference, 2013:99-110.

[8] LV G F, SUN Z G, LI T. LabelCast: A general abstraction for the forwarding plane of SDN [J]. Chinese journal of computers, 2012, 35(10):

2037 – 2047.
[9] 吕高峰, 孙志刚, 李韬. LabelCast: 一种普适的 SDN 转发平面抽象[J]. 计算机学报, 2012, 35(10): 2037 – 2047.
[10] DUAN T, LAN J L, HU Y X, et al. A reconfigurable dataplane enabling network function evolution [J]. Chinese ACTA electronica sinica, 2016, 44(7): 1721 – 1727.
[11] 马丁,庄雷,兰巨龙,等. 可重构网络中的一种新型端到端服务供应模型[J]. 郑州大学学报(工学版),2017,38(6):11 – 16.
[12] ZHANG Y, LAN J L, HU Y X, et al. A polymorphic routing system providing flexible customization for service[J]. Chinese ACTA electronica sinica, 2016, 44(4):988 – 994.
[13] ADALL Hunt. Howe-Net FPGA 10G [EB/OL]. https://github. com/NetFPGA/NetFPGA-public/wiki, 2014. [2018.07.01]

Research on New Data Plane Structure Supporting IPv6 Test and Deployment

HUANG Wanwei[1], DU Chunfeng[2], ZHANG Jianwei[1], DUAN Tong[3]

(1. Software Engineering College, Zhengzhou University of Light Industry, Zhengzhou 450002, China; 2. School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China; 3. National Digital Switching System Engineering Technology Research Center, Zhengzhou, Henan 450002, China)

Abstract: In order to solve the problem that the IPv4 network structure and equipment are difficult to meet the IPv6 network in terms of packet parsing, matching, and action execution, a new data plane structure supporting IPv6 experiment and deployment was proposed. The structure included a data plane abstraction mechanism that could support both policies and functions, and a data plane hardware structure that applied to the IPv6 Next Generation Internet. It could support the coexistence of multiple innovative network architectures in the same network, enabling the testing and verification of new protocols, Simultaneously it supports customizable protocol resolution, flexible and programmable pooket processing, and dyramic combination of internal resources, so as to support the testing, deployment and evolution of future network functions. Through system experiments and analysis, it was verified that the structure had high forwarding performance under the condition of acceptable resource overhead.

Key words: scale deployment; data plane; abstract mechanism; hardware structure; programmable