硬件加速在核心网转发面应用的思考和实践

Thinking and Practice of Hardware Acceleration in Core Network Forwarding Application

王升 /WANG Sheng, 班有容 /BAN Yourong, 陈佳媛 /CHEN Jiayuan, 张昊 /ZHANG Hao
（中国移动研究院，北京，100053）
(China Mobile Research Institute, Beijing 100053)

摘要： 5G、边缘计算新型业务对带宽和时延的需求要求核心网用户面功能（UPF）具备低时延、高转发、零丢包的能力，对动态调整、定制化以及切片的需求又要求 UPF 进行虚拟化部署。通过对网络功能虚拟化（NFV）不匹配三角的深层次分析，找到 UPF 应用硬件加速的切入点，并通过加速比公式衡量硬件加速的效能，为后续性能提升指明方向。认为运营商可以通过定制化、标准化服务器和智能网卡的选型与规格，实现 UPF 软硬解耦，打造通用、开放的网络资源池，充分利用通用硬件的池化效应，降本增效，在提升资源利用率的同时加强运营商对网络的自主掌控。

关键词： UPF；智能网卡；硬件加速；软硬解耦

Abstract: The high demand of bandwidth and delay from new services of 5G and edge computing requires user plane function (UPF) should have the ability of low delay, high forwarding and zero packet-loss. And virtualized deployment is also needed by dynamic adjustment, customization and network slicing function. Through the deep analysis of network function virtualization (NFV) mismatched triangle, the pointcut of hardware acceleration in the UPF is found. The acceleration ratio formula can not only measure the performance of hardware acceleration, but also indicate the direction of subsequent improvement. Now the UPF equipment in the industry is proprietary to manufacturers, and the combination of software and hardware is highly enclosed. Operators can realize the decoupling of software and hardware of UPF through customized and standardized server and SmartNIC selection and specification. The general and open network resource pool makes full use of the pooling effect of general hardware, reduces costs, increases efficiency and improves resource utilization while strengthening the operator's autonomous control of the network.

Keywords: user plane function; SmartNIC; hardware acceleration; hardware-software decoupling

DOI：10.12142/ZTETJ.202003007
网络出版地址：https://kns.cnki.net/kcms/detail/34.1228.TN.20200622.1001.004.html

网络出版日期：2020-06-22
收稿日期：2020-04-16

1 硬件加速起源于不匹配三角

如同经济学领域有蒙代尔的不可能三角、分布式计算有 CAP（指一致性、可用性、分区容忍性）公理一样，网络功能虚拟化（NFV）也存在不匹配三角：计算、带宽和存储三者中总会有一方发展较慢。木桶理论中提到，最短木板决定了系统性能，因此，解决 NFV 不匹配三角问题，是硬件加速在 NFV 领域存在的基石。

随着虚拟化和微服务架构的兴起，完成一个业务所需的东西向流量急剧增加。伴随众多网络业务发展、4G 不限量套餐普及及 5G 的兴起，南北向业务流量也在急速增长。这是近两三年来，数据中心机房迅速从 10 G 网卡提升到 25 G 光纤网卡并向 100 G 网卡演进的深层次原因。随着网络带宽增长势头加剧，计算处理能力的短板逐渐凸显；因此，人们急需一种技

术方案来弥补这个短板。

硬件加速即利用中央处理器（CPU）、片上系统（SoC）、图形处理器（GPU）、数字信号处理器（ASIC）、现场可编程门阵列（FPGA）等使用不同类型指令集和不同体系架构的计算单元，组成一个混合的计算系统，通过将处理工作分配给加速硬件以减轻CPU负荷的技术，从而实现性能提升、成本优化的目标。当前，业界为了解决算力短板、满足业务密集计算需求、提升业务处理性价比，广泛使用各种加速硬件。例如，Azure、AWS等公有云推出的FPGA、GPU实例，Google推出全新架构的张量处理器（TPU）芯片，京东云、阿里云为提升网络性能使用的开放虚拟交换（OvS）卸载智能网卡等。

技术的发展如同历史的发展一样，总是螺旋式上升的。在CT领域，NFV通过使用X86等通用性商用货架产品（COTS）硬件以及虚拟化技术来承载网络功能的软件处理，使网络设备功能不再依赖于专用硬件、资源可以充分灵活共享，实现新业务的快速开发和上线，并基于实际业务例如，需求进行自动部署、弹性伸缩、故障隔离和自愈等；然而，面向5G、边缘云移动边缘计算（MEC）新兴业务如增强现实（AR）/虚拟现实（VR）、云游戏、人工智能（AI）等计算、输入/输出（I/O）、网络密集型应用时，单纯使用COTS硬件并不能满足这些应用对低时延、高可靠的网络要求与并行计算的算力要求。如果采用服务器堆叠方式解决以上问题，总体上将增加资本支出（CAPEX）和运营成本（OPEX）压力。在一些边缘计算场景，机房有限的空间、承重、电力、散热条件制约着可承载服务器的数量。本文中，我们的研究重点是针对负责网络转发的用户面功能（UPF），提升其单位空间、能耗下的转发性能，打破计算与带宽的不匹配三角，实现通用X86服务器架构下的更高转发性能。

2 核心网网关转发的瓶颈与引入100 G网卡的优势

随着后摩尔定律时代到来，CPU制程迭代变缓，主频和单位面积芯片中可容纳的核/缓存数量提升变得困难。目前，CPU三级缓存的存取效率已经从30 ns提升到10 ns左右，将共享三级缓存近核本地化和按需分配仅可以有限地提升缓存利用效率，性能进一步提升难度较大。

在核心网网关UPF中，对一个报文的处理至少需要读（查找转发表）、写（计费）缓存各一次。CPU缓存是最大的I/O瓶颈，过多缓存丢失引起的读写内存会引发转发能力螺旋式下降[1]。I/O效率在100 G线速下几乎是不可逾越的瓶颈，因此，如何减少业务处理逻辑对CPU缓存的访问、将流表卸载至加速硬件中，是产业界尝试打破转发瓶颈的一个方向。

在提速降费、不限量套餐普及以及5G业务发展的大背景下，核心网中数据流量剧增。在4G话务模型下，虚拟化核心网网关用户面（GW-U）部署的普通双路服务器，一般会配置两块25 G网卡——不跨非统一内存访问架构（NUMA）节点。在实际商用部署中考虑到CPU毛刺等因素，理想状态下一台服务器的最大安全吞吐量约40 G。5G增强移动宽带（eMBB）场景下，单局容量远超4G。提高单服务器转发能力，降低服务器总量从而降低能耗和管理成本是当务之急；因此，网卡向100 G发展是必然趋势。若使用100 G智能网卡，由于转发流量卸载到智能网卡，CPU冲高影响降低，在确定的话务模型下，理想最大安全吞吐量可达95 G，折扣大大降低，使总体转发能力提升约4~5倍。同时，针对5G的超可靠低时延通信（URLLC）场景，智能网卡转发处理的平均时延约为10 us，较之NFV软件处理的平均时延100~200 us，可降低一个量级。100 G智能网卡在4G核心网（EPC）、5G eMBB和5G URLLC场景下，成本和时延优势明显。

3 UPF应用智能网卡的切入点

如图1所示，5G采用控制面与用户面（C-U）分离架构，UPF作为U面对外接口是无线侧N3和互联网侧N6，其中N3接口采用GPRS隧道协议（GTP）协议封装。

业界一度对核心网NFV的U面是否需要加速持怀疑态度[2]，认为：

1）通用硬件平台虚拟化是大势所趋，运营商刚从专用设备中转型脱身，智能网卡似乎又回到了熟悉的专用硬件，这是倒退；

2）硬件加速效能比达不到预期。

专用设备被诟病的主要原因在于设备商垄断造成了高昂成本。当前，核心网硬件加速的成熟应用主要聚焦在加解密、编解码等领域。性能提升和成本下降有限，同时引入加速硬件可能带来的硬件绑定问题，使运营商难以下定决心；因此，在加速硬件技术方案的选择上，需要平衡当前通用与专用之间的矛盾。业界常见的加速硬件主要有5类，表1在成本、功耗、开发难度和重用性以及适合的数据处理类型等方面对这5类加速硬件进行了对比。

在成本、功耗和开发难度上，数字信号处理器（ASIC）方案具有绝对优势；但是其支持的加速功能固化，芯片不可重用，灵活性低，更适合成熟稳定的算法类应用。

GPU是面向视频处理等大规模并

行计算类型领域的成熟方案，软件生态强大。边缘云业务中涉及到视频数据处理（渲染、转码）以及 AI 的推理、训练处理都采用 GPU 实现。

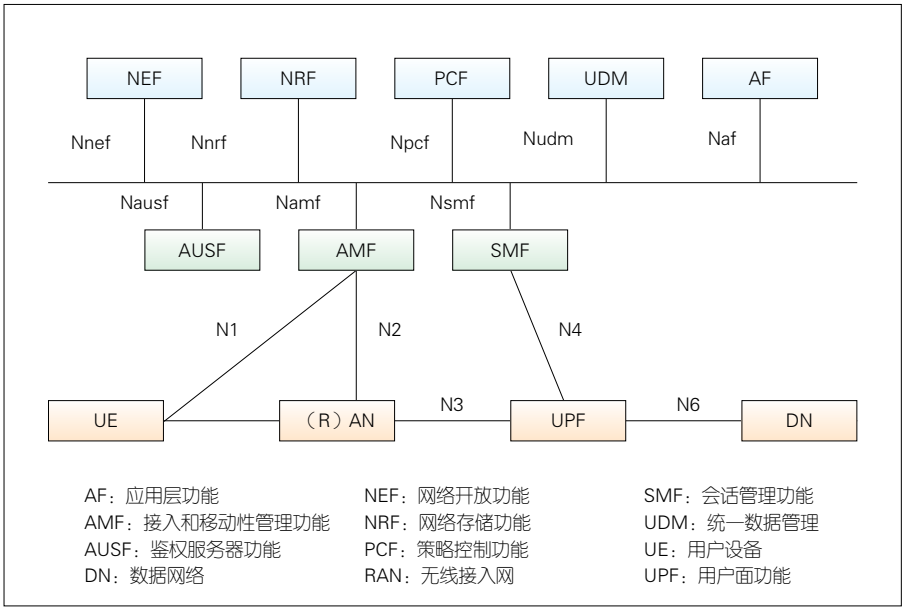
NPU 提供一定的转发规则可配置能力，通过对数据报文转发处理主要过程的固化，实现高性能数据转发，是高性能路由平台的主要方案。

SoC 具有可编程、可升级、支持热补丁特点，多为进阶精简指令集机器（ARM）架构，一般配合 ASIC 定制化使用以保证性能，适合较成熟稳定的算法类应用。

FPGA 性价比介于通用处理器和 ASIC 之间，同时处理时延低，契合 5G URLLC 场景。EPGA 灵活性高，可在线重加载配置软件以实现不同的功能特性，并且片上资源可灵活划分。随着 FPGA 处理能力的提升，FPGA 的部分重配置（PR）技术可以将一块 FPGA 配置为多块功能单元，比如各单元分别支持流量卸载、视频编解码和机器学习。同时每一个功能单元，还可以通过单引导 I/O 虚拟化（SR-IOV）方式提供给多个上层业务来使用，以充分发挥 FPGA 设备性能。实现网络加速的智能网卡是 FPGA 芯片的一种典型应用，这也是数据转发类加速硬件的一个重要形态。

UPF 加速要想做到硬件资源池化，须面临软软解耦（即网元通过应用程序编程接口调用加速硬件的加速功能）或软硬解耦（网元软件在统一的加速硬件上进行功能迭代）的选择。目前 UPF 硬件加速产业并不成熟，各厂家网元处理流程设计不同，加速卸载方案多样。在这一阶段我们选择软软解耦方案，需要统一卸载功能模块及处理流程，打开业务接口。这样的话，一方面难以发挥各种加速硬件优势，另一方面当软件功能升级时，拆分到硬件和上层软件的功能协同升级也较为复杂，同时功能模块的拆分也会为运维、故障定位带来困难；因此，软软解耦是未来产业成熟后的远期目标。针对软硬解耦方案，需要选择一类加速硬件由 UPF 厂家适配开发。5G 业务对高吞吐、低时延的需求分析，与卸载功能和流程需要不断优化演进的需求，都要求加速芯片在保证并行处理能力和低时延性能的基础上具备高度灵活性。综上所述，FPGA 芯片是一种更为灵活、成熟、可通用化部署的选择。

在图 2 所示的 UPF 业务处理模型中，GTP 封装 / 解封装、规则查找、DPI、服务质量（QoS）、计费是关键业务处理路径。如果加速硬件仅处理转发动作，所有报文仍需 CPU 处理



▲图 1 5G 核心网基础架构

▼表 1 常见加速硬件对比

加速硬件	FPGA	GPU	NPU	ASIC	SoC
成本	中	高	低	低	中
功耗	中	高	低	低	中
开发难度	高	中	中	低	高
可重用性	支持	支持	不支持	不支持	支持
典型应用	计算密集算法；规模并行处理；大容量数据转发	视频数据处理；规模并行处理	大容量数据转发	计算密集算法	计算密集算法；大容量数据转发
主流厂商	Xilinx/Intel	NVIDIA/AMD	华为	Intel/Mellanox	Marvell/Mellanox/ 华为
主流形态	按需定义	PCIe 标卡	主板级	按需定义	按需定义

ASIC：数字信号处理器
FPGA：现场可编程门阵列
GPU：图形处理器
NPU：网络处理器
PCIe：外设部件互连标准
SoC：片上系统

GTP 协议、QoS、计费等业务，这一加速应用模型对于 UPF 性能的提升有限；因此，UPF 加速模型须考虑尽可能实现报文的全业务处理卸载。

ETSI NFV001 定义了硬件加速的 3 种主要模式[3]，如图 3 所示。

1）Look-Aside：旁路模式，类似协处理器的应答模式，不改变现有软件流程；

2）In-Line：随路模式，嵌入到软件的包处理过程中，是一种紧耦合模型；

3）Fast-Path：快路径模式，报文不经过主机处理。

对于 UPF 这类转发面网元，Look-Aside 模式中数据包要在加速卡和中央处理器之间多次传递，对总线带宽和处理时延均有影响。In-Line 模式和 Fast-Path 模式更适合业务功能的有效卸载。因此数据包由网卡接收后可以直接在本地处理的智能网卡比旁路外设部件互联标准（PCIe）加速卡更适合用于 UPF 加速。

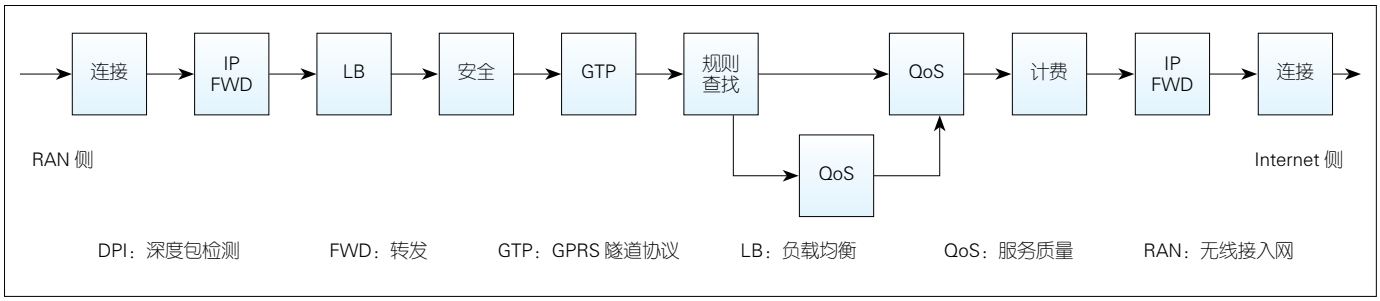
使用 FPGA 智能网卡对现有 NFV 架构的主要影响包括网元适配开发和管理和编排（MANO）纳管。FPGA 开发基于硬件编程语言 VHDL 或 Verilog，与硬件紧耦合。在 NFV 模式下，多 UPF 厂商多智能网卡配对，UPF 厂家适配开发工作量需要收敛。在 OpenStack 社区，Cyborg 组件可以实现 FPGA 智能网卡的发现、管理以及加速功能加载。FPGA 智能网卡需要支持通过 Cyborg 实现自动化在线重配置。基于降低适配开发工作量、UPF 加速业务快速上线、满足在线自动重配的需求，FPGA 智能网卡需要支持静态 - 动态区域模式，并需要运营商对智能网卡进行统一定制化设计。

在 FPGA 智能网卡的静态 - 动态区域模式中，静态区域需要封装 PCIe 接口、双倍数据传输速率（DDR）控制器等通用 IP，面向动态区域提供调用接口，由硬件厂家预先完成开发调试，UPF 厂家在动态区域进行功能开发时可以直接获得硬件平台能力。由硬件厂家提供静态区域，用户则无法修改，这为设备稳定、可靠提供保障，也可形成 FPGA 用户到服务器的隔离，提供安全保证。动态区域部分加载的 UPF 加速逻辑，由网元厂家开发设计，可动态更新，使网元加速功能开发更专注于业务逻辑，也便于后续网元的功能迭代。这一模式为 FPGA 的安全、可靠提供保证，同时使 FPGA 使用者专注于业务逻辑开发，降低了 FPGA 开发难度。

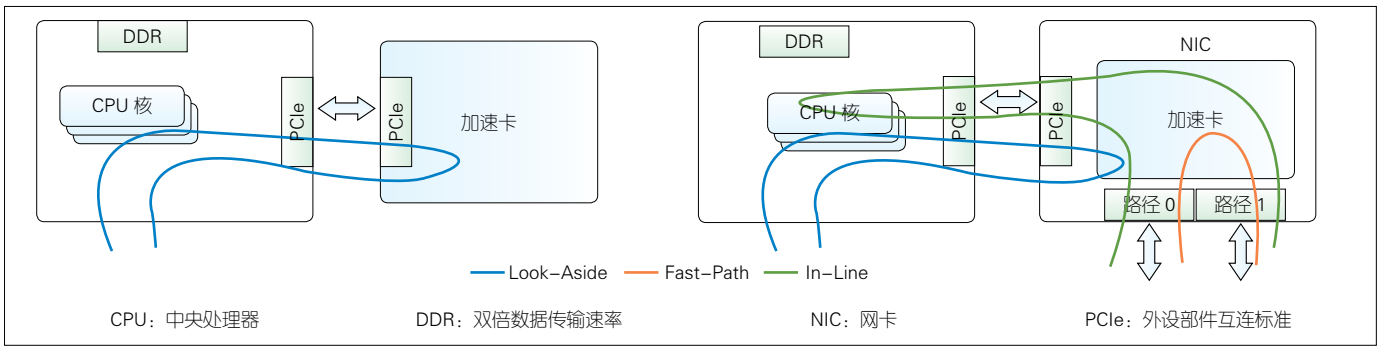
4 智能网卡业务卸载参考设计

考虑到智能网卡更适合处理逻辑简单的重复并行业务，在进行卸载功能选择时，原则上选择稳定且逻辑简单的功能卸载，卸载功能处理流程须符合 In-Line 或 Fast-Path 模式。使用 In-Line 模式时，我们需要考虑哪些功能必须由 CPU 处理，哪些适合下沉到智能网卡。对于 Fast-Path 模式，我们需要考虑满足了哪些条件后，报文可以不经过 CPU，完成正确的转发和计费。

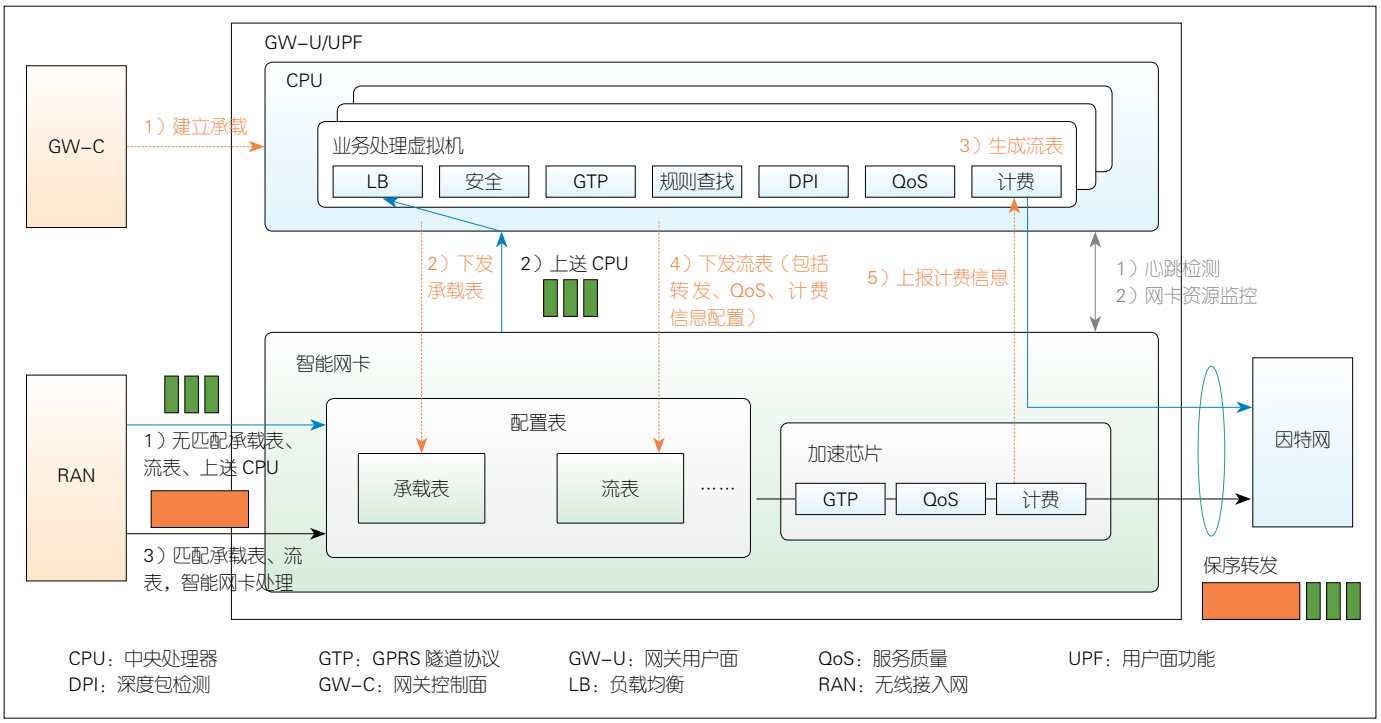
图 4 给出了一种 UPF 业务卸载的



▲图 2 用户面功能业务处理模型



▲图 3 硬件加速模式



▲图 4 UPF 硬件加速参考设计

参考设计，其中配置下发和计费等统计信息上送均通过流表完成。表 2 给出了 CPU 下发的部分配置流表设计。

1）通过首包（一个或几个）上送 CPU，CPU 生成配置流表下发给智能网卡，流表中含路由、计费策略等内容；

2）后续报文到达，智能网卡查找流表，命中则直接转发，不再经过 CPU 处理，未命中上送 CPU；

3）智能网卡实现 GTP 报文的封装 / 解封装等处理；

4）根据计费等策略，智能网卡把计费等信息上报 CPU。

5 加速比是衡量硬件加速效能的关键指标

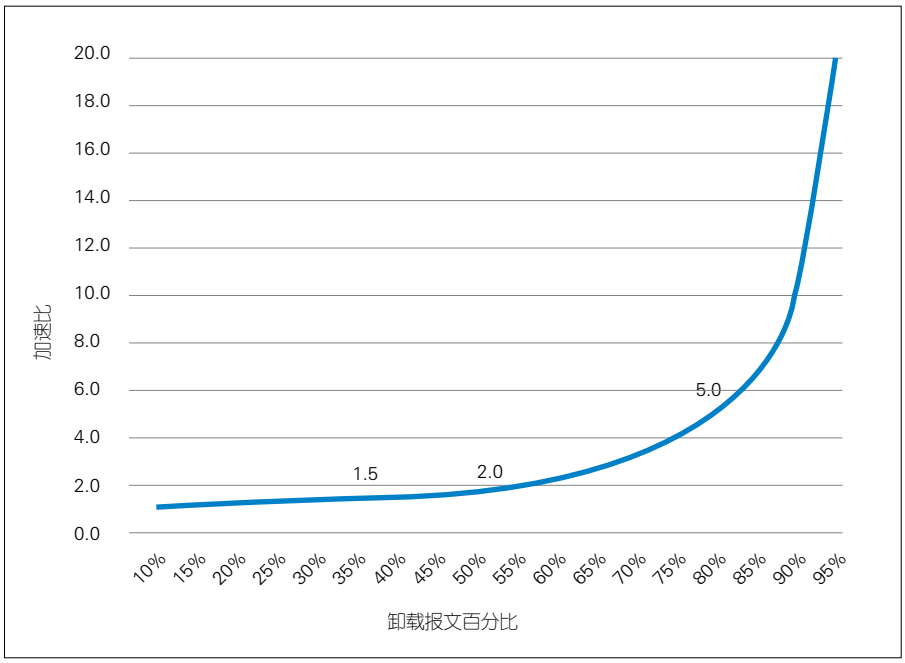
阿姆达定律定义了多核计算的加速比，其核心思想是可并行计算的模块占比与核的数量之间的关系。类似地，衡量硬件加速的效能，也可采用加速比这个概念；但不同之处在于使用了“可卸载报文比例”作为关键因子。

▼表 2 流表参考设计

类别	字段	大小	说明
五元组	IPType	1	IPv4/IPv6
五元组	UserIP	128	用户 IP 地址
五元组	PDNIP	128	PDN IP 地址
五元组	UserPort	16	用户端口
五元组	PDNPort	16	PDN 端口
五元组	ProType	8	协议类型
五元组	innerDSCP	6	内层报头 TOS
路由	srcMac	48	源 MAC
路由	dstMac	48	目的 MAC
计费	ulAccPakNum	12	上行报文数
计费	ulACCVol	20	上行报文长度
计费	dlPakNum	12	下行报文数
计费	dlACCVol	20	下行报文长度
计费	Type	3	计费上报方式
VLAN	vlanType	1	VLAN 类型
VLAN	vlanID	12	VLAN ID
隧道	outerDSCP	6	外层隧道 TOS
隧道	tunnelType	2	默认 GTP
隧道	tunnelIPType	1	IPV4/IPV6
隧道	tunnelLocolIP	128	隧道本端地址
隧道	tunnelPeerIP	128	隧道对端地址
隧道	localTEID	32	隧道本地 ID
隧道	peerTEID	32	隧道对端 ID

注：实际实现时，流表会根据业务需求进行优化，如承载 / 流 / 计费拆分等等，此处仅供参考。

GTP：GPRS 隧道协议
ID：身份标识
MAC：介质接入控制
PDN：公用数据网
TOS：服务类型
VLAN：虚拟局域网



▲图 5 加速比变化趋势

加速比的定义如公式（1）所示：

$$Y = 1/(1-X) \quad (1)$$

其中 X 为可卸载报文比例。

例如，卸载 80% 的报文时，加速比为 1/(1–80%)，即 5 倍。这意味着一台服务器可以处理原来 5 台服务器处理的报文。

图 5 展示了加速比的变化趋势：目前多数加速应用卸载比例约为 35%，效能比低于 1.5。当卸载 50% 的报文时，加速比为 2 倍，这个数值是加速比的拐点。当卸载比例超过 50%，加速比将大幅提升。

在实际转发流量中，我们把超文本传输协议（HTTP）访问称为“短流”，把视频类流称为“长流”。“长流”持续时间长、报文数量多。显而易见，“长流”可以获得更高的加速比。随着 5G 和视频应用的普及，视频流量的比例将大幅提升，并达到个人用户上网流量的 80%、行业流量的 70%，同时智能网卡卸载加速的效果将会更加显著。

硬件加速的加速比存在极限。以 EPC 话务模型为例，以业界通常评估的平均一个流 20 个报文计算，除去必须首包学习上送 CPU 的报文，理论上剩下 19 个报文都可以被卸载，此时加速比的极限为 20 倍，这是这一话务模型下硬件加速的理想目标。

6 结束语

不匹配三角揭示的矛盾，在 5G 核心网 U 面 UPF 上体现为网卡带宽需求远超过当前主流双路服务器的 CPU 计算能力。通过 FPGA 智能网卡实现报文卸载，可有效降低 CPU 负荷实现再平衡，由此降低了每吉比特流量的设备成本。同时，FPGA 智能网卡的灵活性也可保证加速硬件资源池的通用性。

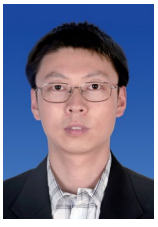
参考文献
[1] 董琳，郑胜利．高性能网关设备及服务实践 [EB/OL].(2014–11–28)[2020–06–15]. https://blog.csdn.net/yangdelong/article/details/80876784
[2] 岳青伦．NFV 硬件加速，在困窘中前行 [EB/OL].(2017–12–25)[2020–06–15]. https://www.sd-nlab.com/20374.html
[3] ETSI. Network Functions Virtualisation (NFV); Acceleration technologies; Report on acceleration technologies & use cases: GS NFV–IFA 001[S]. 2015

作 者 简 介

王升，中国移动研究院网络与 IT 技术研究所项目经理；研究方向为 NFV、虚拟化平台、异构硬件加速、泛在计算等。

班有容，中国移动研究院网络与 IT 技术研究所项目经理；研究方向为异构硬件加速，泛在计算等。

陈佳媛，中国移动研究院网络与 IT 技术研究所技术经理；研究方向为网络功能虚拟化NFV等。



张昊，中国移动研究院网络与 IT 技术研究所副所长；长期从事移动通信领域相关工作，研究方向为 EPC、5G、IMS、NFV、SDN 等。