中国移动NFV

硬件加速技术白皮书

（2020年）

中国移动通信集团有限公司研究院

2020年6月

前 言

运营商网络功能虚拟化（Network Function Virtualization, NFV）首先考虑业务发展和云化的需求。电信业务种类丰富，包括核心网、无线接入网、承载网、传输网、业务平台等多个领域的电信网元都可以采用云化方式部署，各业务在部署位置和资源需求等方面往往存在较大差异，因此业务云化的趋势也是多样化的。

考虑到电信业务云化的效益、难度、可靠性要求，以及云化后的稳定性，目前各大运营商均优先从业务需求迫切、云化难度较小、易于集中化和规模化部署、资源需求较为一致和易于资源共享的业务入手推进业务云化。这类业务主要集中在核心网控制面，如IMS、EPC控制面、PCC、短信、智能网、能力开放等领域，此类业务对云资源池的要求较为一致，适合集中化部署在大规模的网络云中，通过云化实现业务的灵活部署和资源最大程度的共享。

但伴随着5G技术的快速发展和边缘业务的兴起，业务边缘部署需求日益增加，在智能制造、智慧城市、车联网、云游戏、AR/VR等各个垂直领域，时延与带宽成为此类边缘业务的核心关注点。以UPF为例，边缘业务对UPF的要求为在承载百万级用户量的前提下，其端到端转发时延要求不超过10ms，带宽要求在50Gbps以上，核心网UPF带宽甚至要达到300Gbps。同时，边缘云节点在空间和供电、承重等方面存在着很强物理约束，这个特点使得原本核心云端的无限资源模式、大规模部署成本效应淡化，因此在边缘云具体部署实现时，必须需要考虑在有限空间、供电资源、承重能力这些前提下，如何满足业务的性能需求。

为满足5G uRLLC、eMBB以及网络切片场景下边缘多样化业务的部署需求，在边缘节点有限的资源环境下实现大带宽、低时延、高可靠的网络功能，硬件加速技术应运而生。

本白皮书以边缘计算发展理念为基础，以网络转型需求和业务发展趋势为指引，向业界系统阐释中国移动NFV硬件加速的整体架构、关键技术及发展方向。中国移动倡议业界联合对硬件加速技术的方案、架构、演进路线等进行深入研究和实践，共同推进硬件加速技术的成熟，更好的支持边缘场景的业务发展和网络转型。

目 录

1 硬件加速技术概述

1.1 NFV硬件加速产生的背景

硬件加速（Hardware Acceleration）是指将处理工作分配给加速硬件以减轻中央处理器负荷的技术，其利用硬件模块来替代软件算法以充分利用硬件所固有的快速特性（硬件加速通常比软件算法的效率要高），从而实现性能提升、成本优化的目的。

引入硬件加速的计算架构又称为异构计算（Heterogeneous Computing），相对于通用计算（又称同构计算）来说，所谓的异构，就是CPU、SoC、GPU、ASIC、FPGA等各种使用不同类型指令集、不同体系架构的计算单元，组成一个混合的系统，执行计算的特殊方式。

技术的发展如同历史的发展一样，总是螺旋式上升的。NFV通过使用基于X86 CPU或ARM等的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，并基于实际业务需求进行自动部署、弹性伸缩、故障隔离和自愈等。然而，随着5G和边缘计算的兴起，面向5G UPF、新兴边缘业务（AR/VR、云游戏、AI）等计算、IO、网络密集型应用时，却发现COTS硬件并不能满足这些应用对大带宽、低时延、高可靠（低抖动，低丢包）的网络要求与并行计算算力的要求。以下聚焦于X86 COTS硬件对加速技术进行讨论。

- X86 CPU为满足通用性，需要支持复杂的逻辑判断，这样会引入大量的分支跳转和中断的处理，使得CPU的内部的控制器和内存的占比较大，计算单元的比重较低。这种架构决定了CPU擅长统领全局的调度、管理、协调等复杂逻辑处理。同时，CPU指令采用取指令、指令译码、内存取

数、指令执行、结果写回一系列串行处理方式，对并行数据处理的效率相对较低；

- X86 CPU的性能提升节奏放缓，而新兴业务特性对计算能力的要求超过了“后摩尔定律时代”CPU性能增长的速度，需求与能力之间存在鸿沟。如图1-1所示；

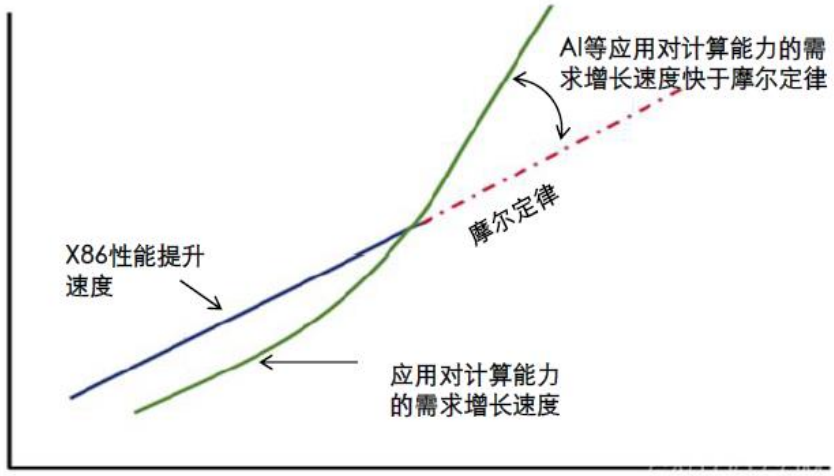


图1-1 应用对计算能力的需求增长速度超过了摩尔定律

- 边缘计算场景中，边缘机房空间小、电力少，非地面机房承重低于核心机房承重能力的1/3，而边缘业务如AI、AR/VR等对算力的需求呈指数型增长，对5G UPF吞吐和时延的要求提高，无法单纯依靠CPU堆叠算力的方式来满足边缘业务的需求。

NFV所采用的通用处理器在特定任务处理上的性能或成本方面存在不足，使得通用处理器配备FPGA、GPU等协处理器（加速卡）的硬件加速方案出现在NFV架构中，可以说电信网络经历了从专用硬件到通用COTS硬件再到通用COTS硬件+加速硬件的螺旋式发展历程。

1.2 硬件加速技术在产业的应用

当前应用和网络对性能加速的需求强烈，公有云售卖加速能力、私有云加速内部网络、设备厂商提升设备能力，均采用了硬件加速；芯片厂商也在广泛布局加速硬件。

微软的Azure catapult项目历经三代FPGA架构，除提供网络和存储虚拟化加速，还可用于加速Bing搜索、深度神经网络（DNN）等计算任务。Azure利用FPGA解决网络和存储虚拟化带来的开销，使虚拟机网络性能由25G达到接近40G线速，数据中心内虚机通信网络延迟降低10倍，多余FPGA资源还用于Bing搜索和神经网络加速。在Bing搜索业务中，采用FPGA加速使系统性能提升两倍，服务器数量投入减少一半。在MICRO'16会议上，微软提出了Hardware as a Service（HaaS）的概念，即把硬件作为一种可调度的云服务，使得 FPGA 服务的集中调度、管理和大规模部署成为可能。

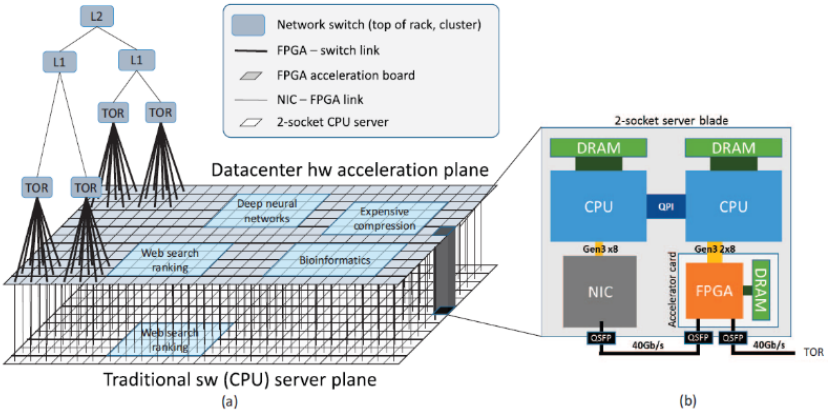


图1-2　Azure的FPGA架构

亚马逊AWS云早在16年就推出了配备FPGA硬件的EC2计算实例F1，用户可编程，为应用程序创建自定义硬件加速能力。自此，公有云服务厂商开始纷纷部署FPGA云加速业务，形成了FPGA结合云计算的FPGA-as-a-service计算平台。

当前，主流公有云厂家均部署了GPU、FPGA计算实例为用户提供服务。如阿里云的异构计算加速引擎，涵盖了GPU、FPGA在内的多款异构计算实例，可

满足从图形渲染到高性能计算及人工智能等复杂应用的计算需求。特别是在人工智能领域，可将深度学习成本缩减一半，大幅降低人工智能计算门槛；而基于阿里云异构平台的全新高性能计算实例E-HPC，可一键部署获得媲美大型超算集群环境的“云上超算中心”。

图1-3　阿里云的异构实例

京东云、Verizon、Ucloud、Flipkart等私有云厂商也普遍采用硬件加速技术来降低成本，降低网络延时，提高云计算中心的网络质量。

Ucloud在2017年开始研究部署基于硬件卸载的虚拟网络方案，以应对25G网卡大规模部署带来的虚拟交换机资源绑定、性能不足的问题，2018年与Mellanox合作，规模商用了Connect-X5智能网卡硬件卸载OvS，在保持简单组网、灵活迁移的同时，提升网络能力4倍，时延降低3倍。

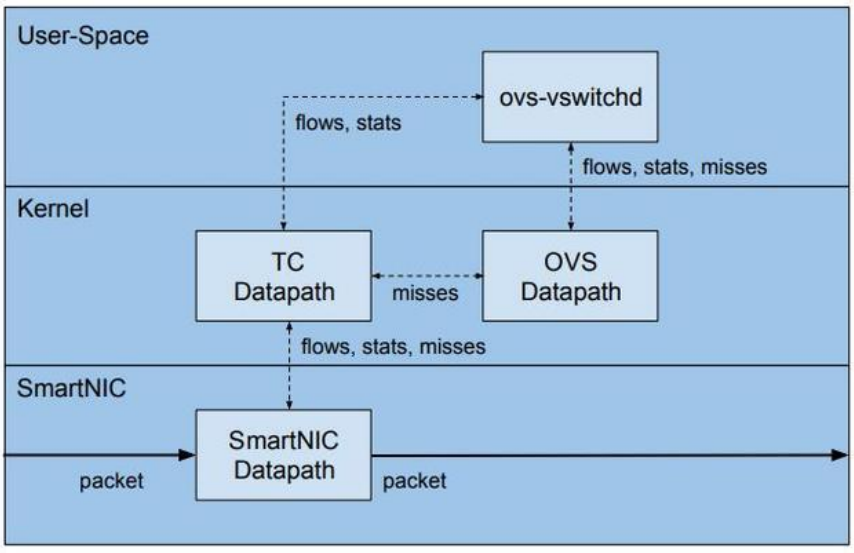


图1-4　Ucloud OvS卸载硬件加速架构

JD云数据中心普遍采用25G、100G网卡，基于虚拟机管理程序的虚拟交换和路由软件实现了敏捷性和灵活性，但面临服务器内性能低下、可扩展性差和CPU开销较高的困扰，2018年采用基于ASIC芯片的OvS卸载智能网卡，降低成本、提高网络吞吐能力，有效解决电商等业务高峰期的稳定性问题。

设备商也采用COTS+硬件加速技术构建新型网络产品，降低成本、提升性能。如A10的一款用于数据中心的负载均衡器设备，其加速硬件使用扩展LSW ASIC完成L2/L3组网出接口，扩展FPGA完成L4/L7转发卸载，扩展Crypto ASIC完成SSL功能卸载。

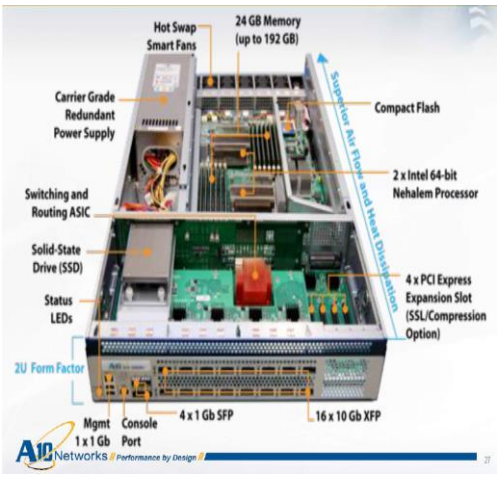


图1-5　A10的一款使用硬件加速技术的LB设备

主流芯片厂商在自身主营的计算芯片之外，也开始引入其他计算架构的加速硬件，扩大硬件布局。英特尔陆续收购Altera、eASIC等芯片厂家，19年发布面向神经网络和视觉处理的AI芯片NNP和VPU，最新推出的AI平台，包含了CPU、GPU、DSP、NNP、FPGA等一系列不同的处理核心。英伟达的机器人平台Jetson Xavier也包含了GPU、CPU、NPU、NVDLA等6种处理器。

1.3 NFV硬件加速标准和开源进展

本章节主要介绍ETSI和OpenStack对硬件加速的研究现状。

在标准方面，最新的ETSI NFV架构引入了硬件加速技术。对NFVI进行了增强，增加了加速资源虚拟化能力：将加速器进行抽象，以逻辑加速资源的方式呈现，统一提供全面的加速服务；虚拟化层提供统一接口，适配不同形态的加速硬件形态，如FPGA、ASIC、SoC等。

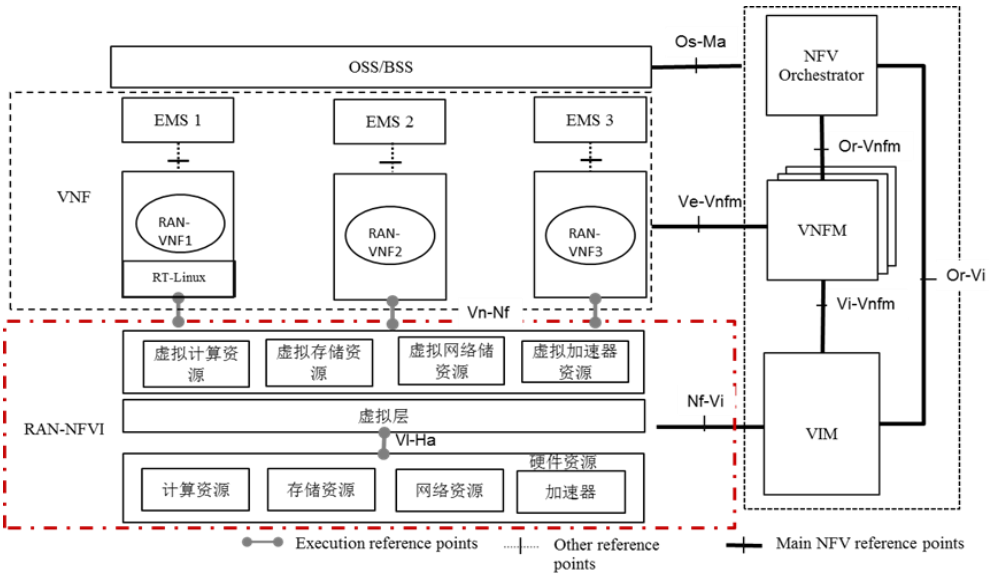


图1-6　ETSI NFV参考架构

ETSI除了在NFV架构中定义了硬件加速模块功能外，也定义了硬件加速的两种实现方案——Pass-through方案和抽象模型方案。

- Pass-through方案，即PCI/PCIe-pass-through，它将PCI插槽上的硬件加速卡直接pass through给某个特定的虚拟机（VM）使用。Pass-through方案是目前最通用的方案。其缺点是硬件被虚拟机独占，上层应用或者虚拟机需要维护不同加速卡的硬件驱动等。

- 抽象模型方案：在NFVI中，也就是hypervisor中维护“Backend/HW Driver”模块；在VNF层，在VNFC中维护“Generic Driver”模块。NFVI负责加速卡的扫描和驱动加载，加速卡硬件虚拟化管理，并将虚拟加速卡挂载到虚拟机上。其优点是一个加速卡资源可以被多个虚拟机使用，加速资

源可以被加载或者释放，并且VM针对各种加速卡，只使用一个通用的加速卡驱动程序，虚拟机维护简单。

在ETSI定义硬件加速框架及实现方案的同时，开源社区OpenStack也启动了Cyborg项目，其主要目标是管理各种加速器的安装驱动程序、依赖关系、安装和卸载。它能够将加速器和nova创建的虚拟机实例建立连接，旨在提供通用的硬件加速管理框架。OpenStack主要面向基础设施中对加速硬件的驱动集成和VIM对加速硬件的感知，不涉及上层MANO。截止到T版本，Cyborg项目已经陆续支持Xilinx、Intel的FPGA，NVIDIA的GPU卡等加速硬件的生命周期管理，已经实现了基本的管理功能，但距离商用部署还仍需继续优化。

目前硬件加速方案，尤其是硬件加速抽象方案尚不成熟。加速硬件的使用涉及加速卡厂家、云平台厂家和网元厂家的配合与联动，这需要针对相应的加速卡产品，在云平台层集成驱动并虚拟化、提供相应的加速库或SDK，网元层进行调用。

1.4 NFV业界硬件加速应用现状

当前NFV领域的硬件加速方案通常为各厂商私有，与业务或云平台紧耦合，无法解耦，与NFV提倡的硬件通用性背道而驰。硬件加速方案在NFV中的应用现状，可以总结如下，其中存在的问题需要在技术方案中进一步明确。

- 使用方式简单

当前主要使用方式为虚拟机直接使用相应加速硬件，加速硬件无法弹性或多虚拟机之间进行共享使用，导致资源利用不均衡。

- 加速硬件和接口无规范

每个硬件厂家的加速硬件都是特定的加速驱动，云平台厂家提供的SDK差异较大，需要进行适配开发，规范化程度低。

- 业务加速硬件通用性较差

业务加速硬件与上层应用绑定，厂商锁定导致加速硬件成为专用硬件，不符合NFV分层采购建设的模式。加速资源无法池化，难以发挥规模采购、降低成本的优势。且加速资源无法共享，资源利用率难以提升。

- 需要对MANO进行扩展

硬件加速的使用可以分为感知、分配、调度、释放等四个阶段。感知需要云平台对硬件类型进行识别，分配需要VNFM和NFVO支持网元对加速硬件资源请求的解析，调度需要云平台进行加速资源的监控和部署，释放则需要云平台对加速硬件资源进行重新编程。这些都需要对MANO进行扩展。

2 中国移动NFV硬件加速需求分析

随着IT/CT融合场景的不断显现，尤其在边缘计算情景下，业务种类不断丰富，针对新业务场景的计算特点和性能需求，需要考虑引入合适的硬件加速技术。

根据边缘计算数据处理的类型特点，可以将数据处理分为三大类型：

1. 计算密集型：需要进行大量计算，消耗计算资源，比如安全加解密、视频渲染；
2. 通信密集型：涉及大量网络数据操作的任务；
3. IO密集型：涉及磁盘IO操作的任务。

从边缘云平台及其承载的业务角度，存在不同的加速需求：

- 云平台层面：虚机与网卡交互的数据通路需要灵活、高效；

- UPF业务：转发面网元需要实现大带宽、低时延转发性能，需要对其业务处理进行加速，如报文协议处理（GTP、SCTP协议加速）、防欺诈、深度包检测等；
- AI/图像处理类业务：AR\VR、云游戏、人脸识别、自动驾驶等业务领域需要处理大量数据，对并行计算能力有较高要求。

边缘云业务和加速类型的关联关系可以用下表来描述：

表2-1 边缘云业务和数据处理类型的关联关系

业务	计算密集型	通信密集型	IO 密集型
基础设施	★	★	
UPF	★	★	
AI/图像	★		★

2.1 数据通路硬件加速需求分析

当前，在NFV场景下，虚拟机与网卡之间数据交换的通路主要使用两大类技术：软交换（OvS或OvS+DPDK）或硬直通的方式（SR-IOV），如图2-1所示。

1. 软交换：
   - OvS：VM通过虚拟层实现的一层虚拟交换机（OvS）与网卡交互，实现数据收发。OvS占用CPU核资源，并且，其通过内核中断收发报文，在内核态到用户态又存在内存拷贝，性能较低。
   - OvS+DPDK：DPDK提供了用户态驱动接管内核态驱动工作，通过轮询和共享内存等技术实现OvS性能的提升。
2. SR-IOV方式：通过硬件设置将物理网卡映射成多个虚拟网卡（VF）供VM使用，虚拟机可以直接连接到物理网卡上，报文可跨过虚拟层直接到VM，这种方式不再需要虚拟交换功能，但VM需适配网卡的VF驱动，

SR-IOV在实际使用中的通信性能基本消除I层带来的性能开销。

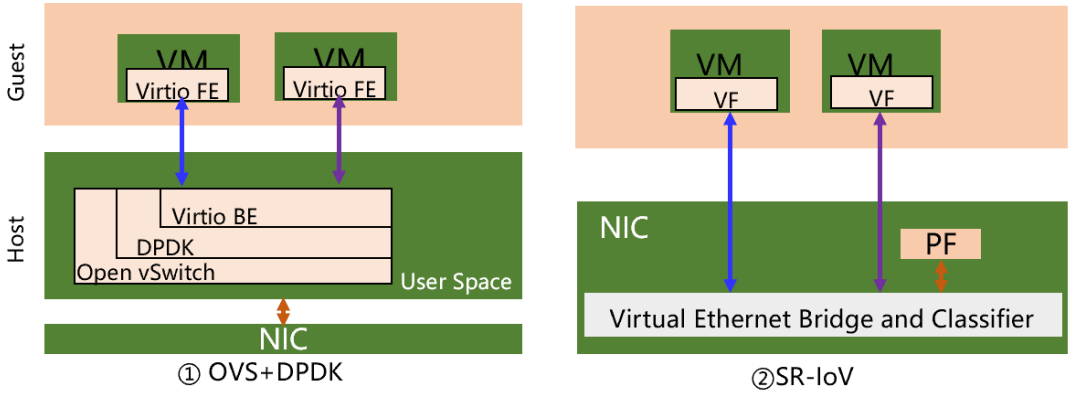


图2-1　OvS+DPDK、SR-IOV

两类数据通路的技术方式各有优劣，OvS或OvS+DPDK技术作为软件交换机实现，有可灵活配置各种流控策略（安全组等）、可作为VXLAN VTEP点、与虚拟机无绑定、端口数量无限制、支持虚拟机热迁移等优势。然而OvS或OvS+DPDK方式小包（64B、128B）转发能力差，并且存在CPU消耗问题。当前针对10G网卡，OvS绑核为2个物理核，随着数据中心向25G、100G网卡发展，届时OvS绑核的数目会大幅增加，占用原本提供给上层应用的宝贵计算资源。而SR-IOV跨过虚拟层，网卡能力直通虚拟机，性能较高，不占用主机的计算资源。但VF驱动与虚拟机紧耦合、虚拟机热迁移方案不完善等导致的灵活性差、组网复杂也一直为业界所诟病。

因此，为了在边缘场景下支持简单组网，为应用提供灵活迁移能力，可以考虑引入OvS卸载智能网卡。OvS卸载智能网卡将OvS相关功能卸载至智能网卡，利用加速硬件提高转发能力，释放软件实现占用的CPU资源。OvS智能网卡在业界已有成熟应用，可以提升数据中心的网络质量，降低成本。

表2-1在多个方面列举了OvS+DPDK技术、SR-IOV技术与OvS卸载智能网卡的技术现状。

表2-1 不同数据通路的技术现状

	OVS+DPDK	SR-IOV	OvS硬件卸载
驱动	通用驱动	VM需适配VF驱动 虚拟层需适配PF驱动 VNF与虚拟层解耦后可能版本不匹配	通用驱动（virtio场景） VF驱动（VF直通场景）
功能	支持SDN vtep、流镜像等功能	仅支持二层转发、不支持安全组、组播模式下MAC混杂性能差	支持SDN vtep、流镜像等功能
热迁移	支持	不支持	支持
资源消耗	10G：2core，核数随着带宽增加	不占用额外资源	少量CPU核，各方案不一
端口数	无限制	受限（一般为64）	无限制
性能	较差，小包及多流性能差	近线速	二者之间，尚在优化
SDN/NFV融合方案	东西向流量：Vlan Trunk 南北向流量：Vlan Trunk	东西向流量：父port Vlan转Vxlan 南北向流量：QINQ终结到VXLAN	东西向流量：Vlan Trunk 南北向流量：Vlan Trunk

2.2 UPF硬件加速技术需求分析

如图2-2所示，在5G核心网架构中，转发面采用C-U分离架构，C面网元和U面网元分别为SMF和UPF，其中SMF面负责承载建立、信令分析等控制消息处理，UPF面主要支持用户业务数据的路由和转发、业务识别、动作和策略执行等。5G uRLLC和eMBB场景，对UPF的处理时延、带宽、抖动和丢包率等性能提出了更高要求。

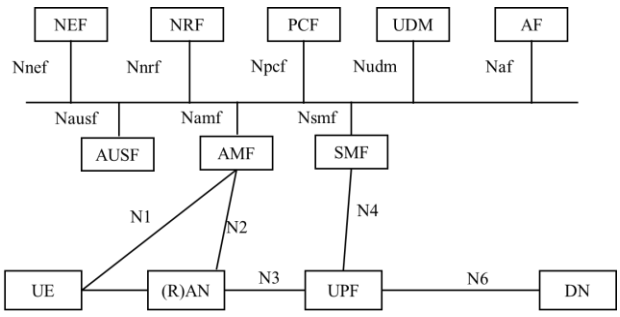


图2-2 5G核心网基础架构

5G uRLLC场景业务如车联网，对端到端处理时延要求低至5ms。而目前转发面网元仅U面转发处理时延已在0.5ms~0.8ms左右，转发处理时延存在下降空间。通用服务器采用X86 CPU架构，CPU的中断、分支预判等处理机制限制了处理速度的提高。因此，通过软件优化降低处理时延无法从根本解决时延问题。

随着提速降费以及无限量套餐的普及，核心网数据转发量迅猛上升，如图2-3所示，2019年一季度，中国移动用户上网数据流量同比增长达270%。伴随5G发展，AR/VR、4K/8K高清视频、3D游戏等大带宽应用将进一步提升用户上网数据总流量。目前，4G GW-U的服务器已经使用25G网卡取代10G网卡，随着带宽需求增加，5G UPF使用100G网卡是未来的发展趋势。高吞吐代表大量并行数据包处理工作，CPU本身的串行处理机制在大吞吐场景下存在局限性，加速硬件的并行处理能力在高吞吐需求下具备更强优势。

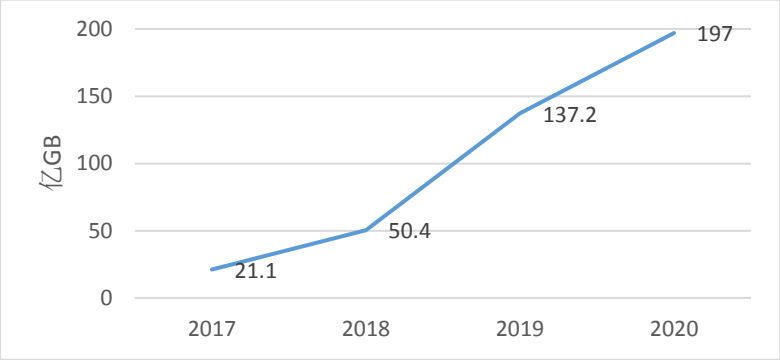


图2-3 中国移动各年一季度用户上网总流量增长趋势

为满足业务的低时延需求，UPF将下沉部署到边缘机房，机房的电力、承重无法按照核心机房要求建设且空间较小，引入硬件加速可提高单设备转发能力，通过减少服务器数量有效降低设备占地、功耗。同时，也有利于降低边缘节点运维难度。

结合UPF的强计算、强转发的业务需求和部署机房的环境限制，硬件加速技术的引入可以提供提高单位空间下网元的转发能力的解决方案。

NFV硬件的统一、池化的理念有利于网元的灵活扩容、快速部署，并且也便于统一运维、提升资源利用率。在地市等条件较好的边缘机房，资源池化具备可

行性。边缘硬件资源池的统一也有利于减少边缘应用的适配难度，因此，UPF加速硬件需要支持池化部署。

2.3 边缘应用硬件加速技术需求分析

随着边缘计算的发展，边缘云需要承载的业务呈现多样性，其中以AI及图形图像处理类业务为主。人工智能是第四次工业革命的核心驱动力，在全球范围内，人工智能正在上升为国家战略，成为推进经济转型和产业升级的战略制高点。人工智能将对所有传统产业的演进产生深远影响，预计在2030年，全球经济16%的产值将受益于人工智能的技术推动。边缘计算+人工智能的场景必将加速人工智能技术在各行各业中的落地速度。依赖边缘计算提供的低延迟、高带宽的能力，人工智能技术，尤其是基于GPU加速的机器学习、深度学习及相关神经网络算法模型等技术更能够突破想象，创造无限可能。在边缘计算领域，无论是智慧城市、智能制造、自动驾驶、智能机器人、AR/VR，边缘计算+GPU加速的能力更是将人工智能的能力赋予了城市中的各行各业。因此，在边缘云中引入GPU的加速能力，是业务驱动的基本要求。

技术的进步是无止境的，中国移动也在积极探究未来新兴技术的加速能力，新兴业务的加速需求，如存储面加速及AI专用芯片的加速方案等。

3 中国移动NFV硬件加速整体架构

中国移动硬件加速整体架构当前主要覆盖数据通路加速（OvS）、转发面网元加速（UPF）和边缘业务加速（AR/VR/AI App）三类场景，与NFV传统三层一域架构类似，分为异构计算硬件层、虚拟层、应用层三层与加速硬件管理编排域，如图3-1所示。

由插入OvS加速网卡、FPGA智能网卡或GPU加速卡的通用COTS服务器构成异构计算硬件资源池，对虚拟化层提供基础异构计算算力；虚拟层主要实现加速硬件的虚拟化，依靠OpenStack Cyborg组件管理硬件和软件加速资源，实现列出、

识别、发现和上报加速器，挂载、卸载加速器实例的功能；上层应用层基于特定的业务需求，部署在特定的异构计算服务器上（当前OvS加速服务器可支持网络加速，GPU服务器支持边缘计算AI、AR/VR、云游戏等应用，FPGA服务器支持UPF加速或其他应用的加速应用）。MANO作为管理编排域，实现标准化硬件管理接口、VNFD模板等加速器管理端到端流程的实现。

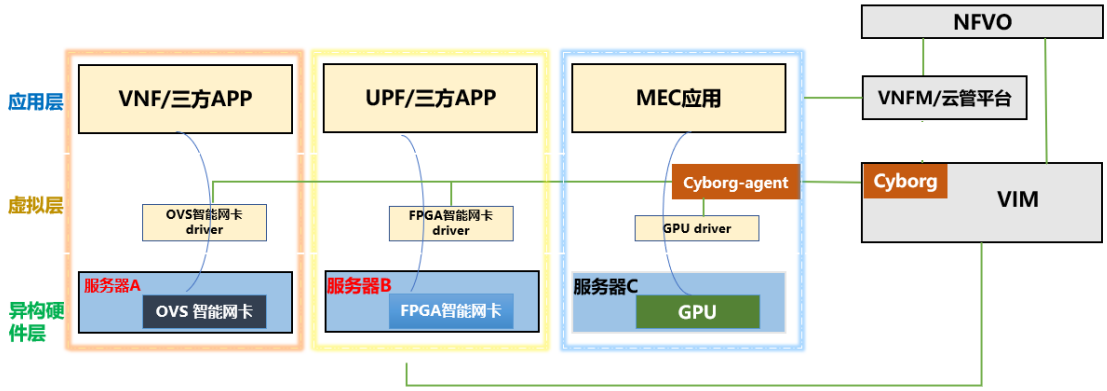


图3-1 中国移动NFV硬件加速整体架构

4 中国移动NFV硬件加速关键技术

4.1 数据通路硬件加速关键技术

数据通路加速技术即OvS卸载，其核心是将OvS的功能模块卸载到智能网卡，通过智能网卡上的加速芯片协助CPU处理虚拟网络负载，提升虚拟转发网络吞吐及时延性能。OvS智能网卡要求至少支持诸如组播、混杂模式、虚机热迁移、VLAN透传、QoS等OvS主要功能的卸载。目前业界OvS智能网卡卸载能力各不相同，有OvS控制面转发面全卸载和OvS转发面卸载两种方式。

OvS智能网卡北向对接虚拟层。若虚拟层与OvS全解耦，则需OvS智能网卡厂商与虚拟层完成兼容性适配。若OvS控制面仍由虚拟层实现，OvS硬件加速厂商仅负责OvS转发面功能，需标准化OvS控制面和转发面协议，目前OvS控制面和转发面协议有rte_flow、TC flower和私有协议三种实现方式，其中，rte_flow为最优

方案。在当前控制面与转发面协议实现有差异的情况下，建议OvS整体打包由智能网卡厂家提供，与虚拟层解耦。

SDN场景下，OvS控制面北向与SDNC对接，SDN与OvS控制面为异厂商实现，需标准化OpenFlow协议和OvSDB协议。

OvS智能网卡与虚拟层、SDNC的解耦方式如下图4-1所示：

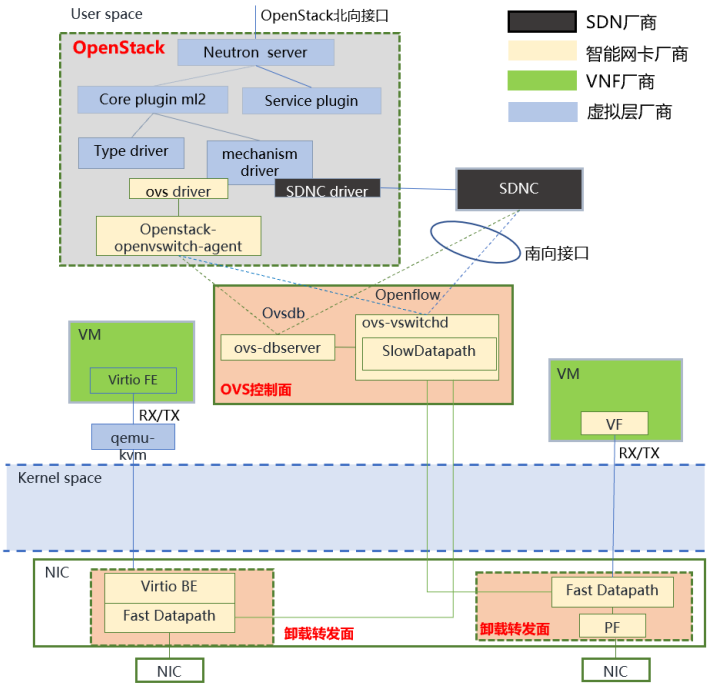


图4-1　OvS加速网卡与hypervisor、SDNC解耦架构图

中国移动分别对中兴、Mellanox、华为的OvS智能网卡进行了25G网络加速测试，测试数据表明，三家厂商的智能网卡性能对比软件OvS方案均有不同程度的提升，Mellanox的智能网卡1W条流的小包处理吞吐能力较为突出，中兴和华为的智能网卡在吞吐和时延方面的性能均优于纯软OvS方案。不止于此，OvS智能网卡无论在功能实现、性能调优和解耦规划上均在持续推进与优化。

4.2 UPF引入硬件加速关键技术

UPF的业务处理流程如图4-2所示，主要功能模块包含负载均衡、安全校验、GTP协议处理、DPI、QoS、Charging等，数据包被网卡接收后发送到CPU处理，再从网卡转发。

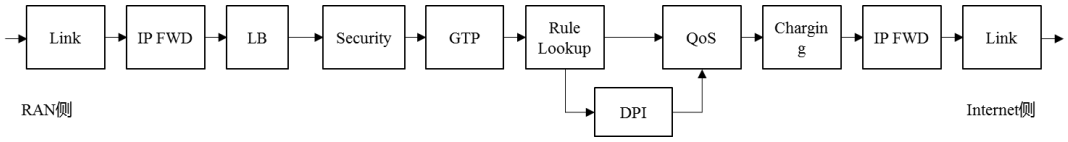
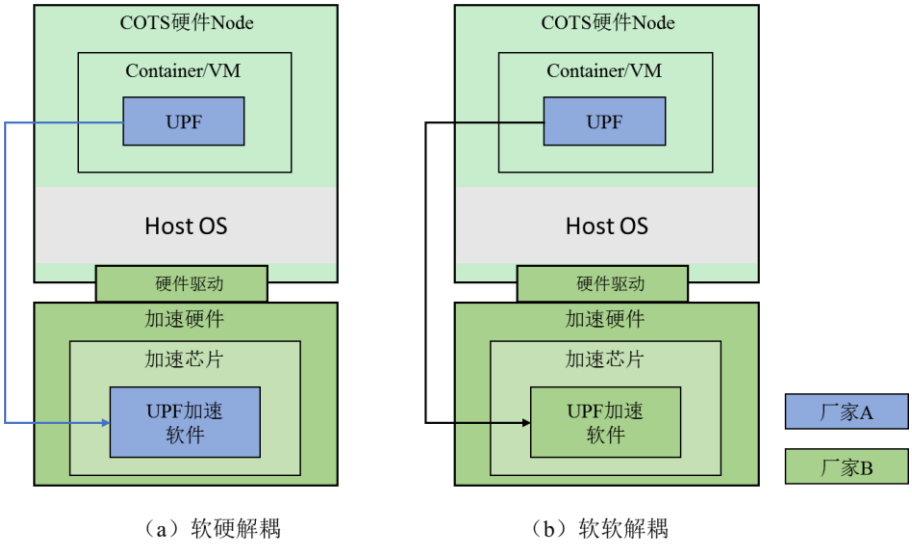


图4-2 UPF转发处理流程及主要功能点分析

随着通用网卡技术发展，如Intel 700系列网卡和Mellanox Connect-X 5网卡均可提供一定数据包处理能力。以Intel xxv710的DDP功能为例，可以识别GTP内层报文，按照配置规则完成数据包分发，这一类通用网卡可在一定程度上提高转发性能。由于不引入新型硬件，不增加建设成本，因此在硬件加速部署初期，可以考虑利用通用网卡的数据包处理能力，实现UPF转发能力的提升。

通用网卡仅能完成UPF个别功能的卸载处理，性能提升有限，UPF加速需要考虑卸载更多功能。UPF业务是一种计算密集、网络密集的应用，与Azure加速网络和业务的场景类似，智能网卡这类直接与网络连接的加速硬件应用形式更加适合UPF加速场景。采用智能网卡加速，可将UPF业务功能组合卸载到智能网卡处理，实现大部分数据包以fast-path形式直接由智能网卡处理转发，只有少部分数据包上送CPU，以in-line形式处理，从而实现更高处理性能。

智能网卡资源通用要实现与软件的解耦。卸载到智能网卡的业务功能组合是UPF业务加速软件。针对UPF硬件加速的解耦方式有两种。第一种是软硬解耦，如图4-3（a）即UPF加速软件与UPF软件同厂商，与加速硬件异厂商；第二种是软软解耦，如图4-3（b），UPF加速软件与与加速硬件同厂商，与UPF软件异厂商。



（a）软硬解耦　　（b）软软解耦

图4-3 解耦方式

软硬解耦时，业务相关软件由UPF厂家提供，便于性能调优，无性能损失。Openstack社区Cyborg组件已经支持软硬解耦方案中加速硬件的管理编排，采购模式沿用NFV分层采购模式不变，运维权责清晰，不同厂家的不同应用都可基于同款池化加速硬件开发加速逻辑。

软软解耦时加速硬件厂商可以自由选择加速芯片类型，但在应用中存在以下问题：

- UPF软件与业务加速软件间功能割裂，难以调优，有性能损失
- 产业不成熟，采购评价标准制定困难，运维权责不明
- UPF升级需与加速软件升级同步，上线周期拉长
- 功能固定，无法应对上层业务的更新迭代

在解耦策略上，由于各UPF厂商软件设计不同，软软解耦短期难以实现，并面临测试、采购的一系列问题，因此建议采用软硬解耦，规范智能网卡硬件要求。

4.2.1 通用网卡加速技术

随着技术的发展，通用网卡的高级特性也可完成部分简单的数据包处理工作，实现业务加速。以下以Intel 700系列网卡的DDP功能为例进行介绍。DDP（Dynamic Device Personalization）又称动态设备个性化设置，此功能通过加载固件配置文件（profile）动态地实现重新配置数据包处理流水线，以满足特定的场景需求。即英特尔700系列网卡具有部分可编程能力，通过加载特定的固件配置文件，可以为用户提供特定通信网络协议的解析支持，结合网卡的FDIR（流引导）和RSS（散列技术）特性，实现网络报文解析和分发的硬件卸载，从而提高网络性能。英特尔当前提供的工业级配置文件（Profile），已涵盖多种协议类型，如PPPoE、GTP-U/C、L2TP等。这些配置文件可以通过通用的 Ethtool 或者 DPDK 驱动进行便捷的加载。

在 5G UPF 的应用场景中，主要处理的报文类型是上行链路（从 UE/eNB 到 PDN）的 GTP 报文和下行链路（从 PDN 到 eNB/UE）的 TCP/IP 报文。在没有 DDP 特性的时候，GTP 的内层报头不能被网卡解析，导致网卡只能缺省的判断其为普通四层报文，从而无法识别包含用户信息的内层报文，这样 GTP 报文就无法利用网卡的 FDIR 或者 RSS 分流功能散列到不同队列并绑定到不同的处理器核上做并行处理。为了解决这个问题，传统方案只能特别地使用若干处理器核用软件的方式来解析 GTP 内层报头分并发到不同的处理核上。这种场景下，软件分发核的性能常常会成为整个 UPF 处理的性能瓶颈，并且软件分发增加了系统整体处理的延迟开销。

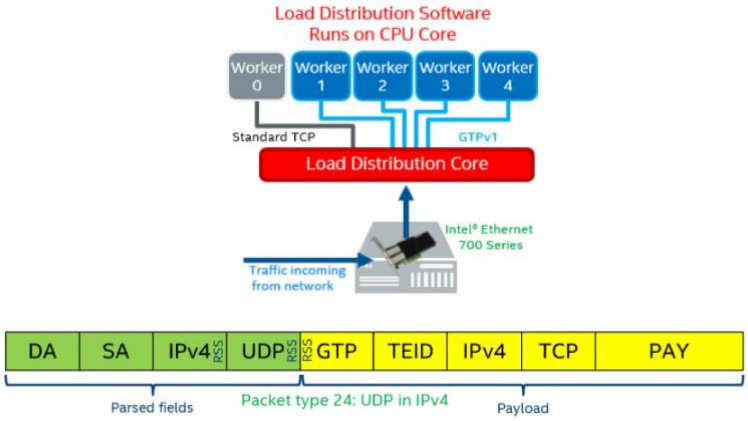


图4-4 基于没有DDP功能标卡的报文处理模式

应用了 DDP，网卡实现了报文识别和协议解析，可以将报文识别深度扩展到 GTP 内层报文的传输层，这样就能够直接在 GTP 报文上针对内层包头应用 FDIR 和 RSS 分流功能，利用网卡直接将 GTP 报文根据内层报文信息散列到不同的网卡队列并绑定到不同的核上进行并行处理，实现分发功能的硬件卸载，从而达到提高性能、减小系统整体处理延迟的效果。

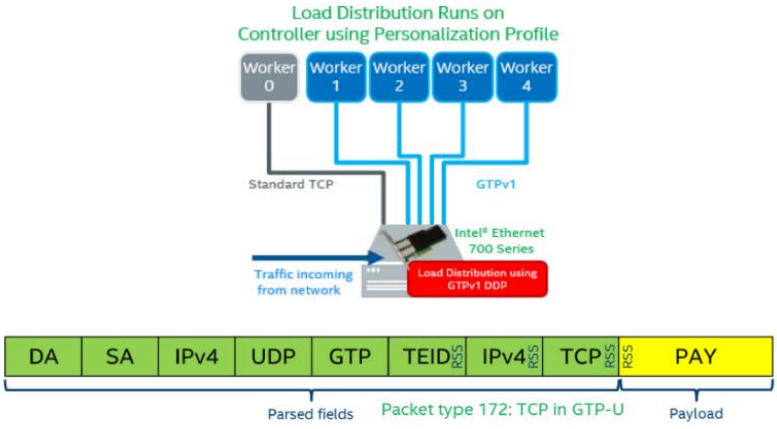


图4-5 基于DDP功能标卡的报文处理模式

详细的来说，为了UPF使获得更好的高速缓存利用率以及减少核间数据流量交互，DDP功能可以将来自相同UE IP的所有数据流量绑定到固定的工作核上。UPF使用UE IP地址作为工作核识别的key。上行链路流量为GTP-U封装的IP报文，因此从封装的IP报文中提取源地址作为UE IP地址。下行链路流量为普通IP报文，因此UE IP地址为报文的目的地IP地址。通过DDP技术，修改GTPU协议和普通IP

报文类型的匹配关键字，UPF将来自相同UE IP的所有数据流量绑定到固定的工作核上。实现了非对称报文的对称hash、双向流的亲和性，节省了专有的收发核与负载均衡核，并大大地减少了大吞吐量的核间通信，进而有效地提高了系统性能。

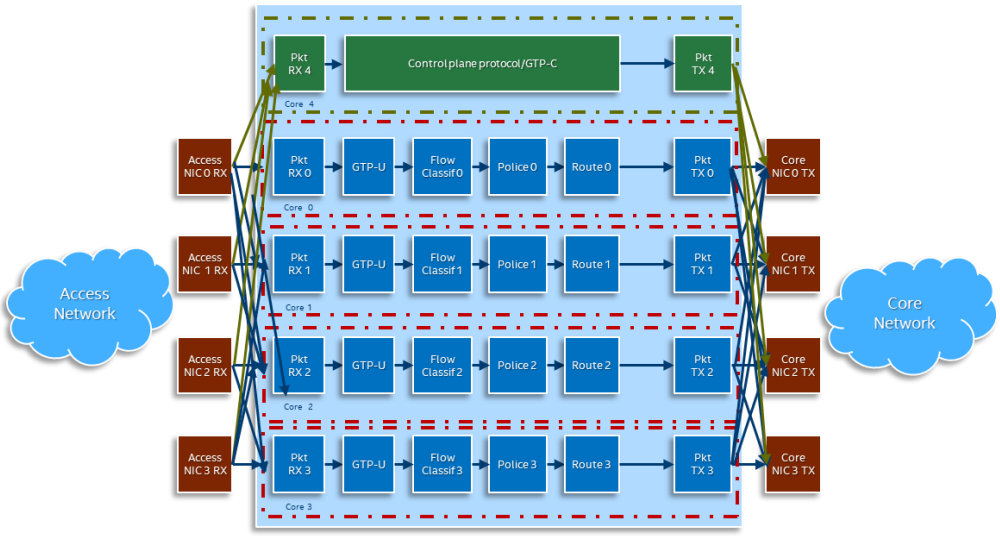


图4-6 基于DDP功能标卡的典型软件架构

4.2.2 智能网卡加速技术

规范智能网卡硬件首先需要明确加速芯片选型，需要结合业务需求选择适合的加速芯片。当前业界主流芯片包括GPU、FPGA、ASIC、SoC四种类型，我们从芯片架构、性能、功耗、开发难度等几个方面对四类加速芯片进行分析。

GPU（graphics processing unit，图形处理器）擅长处理图形、图像处理类的并行计算，这一类计算的特征为计算数据之间关联性小，计算密度高。GPU的硬件架构中多个计算单元共享一个逻辑控制单元和存储单元，GPU的指令处理与CPU相似，需要经过取指令（逻辑控制单元）->指令译码（逻辑控制单元）->指令执行（计算单元）的过程，这种结构就要求适合GPU处理的算法本身复杂度较低，对于密集型数据的并行处理可以发挥GPU的计算优势。GPU的高并行计算能力基于多计算单元和高时钟频率，由此使得GPU的功率偏高，在应用时需要考虑

供电、散热等机房条件的保障。GPU基于CUDA或OpenCL进行开发，开发环境稳定，技术普及程度高，软件开发门槛较低。

FPGA（field programming gate array，现场可编程逻辑门阵列）是由逻辑门电路组合成的可重复编程器件。CLB是FPGA的基础逻辑单元，CLB之间的PI是可编程互联线，IOB作为可编程IO模块支持FPGA直接收发网络数据包。当前一块FPGA集成的CLB数量数以万计，可以在较低功耗下实现高性能数据计算，丰富的I/O和编程器件使其在并行计算领域有广泛应用。FPGA一般通过Verilog或VHDL语言开发，硬件描述语言直接定义片上电路的组合连接，从而实现某种功能，与软件编程相比，无需经过指令处理，这使得FPGA的计算处理速度更快，适合于时延敏感业务。FPGA应用对开发人员要求较高，需要了解底层硬件知识，目前FPGA厂家为了降低开发门槛，在不断优化开发平台（如xilinx Vitis），以期使FPGA开发更加快速、便捷。

ASIC（application specific integrated circuit）是一种专门为某种特定需求定制的集成电路。ASIC芯片的计算能力、性能均可按需定制，定制化使得AISC在尺寸、功耗、性能方面具备极好的优势。ASIC硬件的定制化要求业务需求稳定、不可变更，这使得ASIC芯片更适用于成熟稳定期的应用，难以支持持续演进领域的需求。同时，定制化使得ASIC芯片在前期的设计、验证上需要花费大量时间和人力，从研发到市场应用的周期很长，这也使得ASIC芯片需要批量化大规模应用才能弥补前期的投入。

SoC（systerm on chip，片上系统）是一类将多种芯片集成在一芯片上组成有专用目标的芯片系统。通常集成了中央处理器和特定功能的ASIC芯片，以满足性能要求。当前SoC芯片主要用于终端设备，在数据中心的应用暂不成熟。由于涉及ASIC集成，硬件也需要根据客户需求定制。

UPF主要面临的问题包括高并发大带宽和低时延的性能要求，下沉到边缘机房的低功耗要求以及后续持续的优化演进。因此，FPGA作为一种灵活性高，能

耗比高且处理时延更低的加速硬件，更加适合当前UPF的加速需求。

在NFV架构下，采用软硬解耦方式引入FPGA智能网卡，需要对FPGA智能网卡硬件统一要求，在定制智能网卡时需要考虑三个问题：

1、UPF厂家需要基于FPGA智能网卡开发加速功能，在NFV模式下，多UPF厂商多智能网卡配对，为保证业务快速上线，需要考虑降低适配开发工作量；

2、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实现自动化重配置，保证后续UPF的功能迭代；

3、FPGA智能网卡加载加速逻辑后，不能影响服务器的稳定运行。

针对以上问题，FPGA的shell-role技术给出了解决方案。Xilinx FPGA架构下的shell-role和Intel FPGA架构的BBS-GBS，本质上是将FPGA片上资源划分为两大区域：静态区域和动态区域。静态区域由网卡厂家预先完成开发调试，封装PCIe接口、DDR控制器等通用逻辑，对动态区域提供调用接口；动态区域加载用户的加速逻辑。

1. 由网卡厂家提供通用逻辑，减少应用的重复开发，应用直接调用硬件平台能力，可专注于业务加速逻辑的开发，降低开发工作量；

2. Shell提供支持通过cyborg自动加载的模块，同时，shell屏蔽了加速逻辑与硬件之间的关联，可实现在线的自动加载；

3. 网卡厂家提供shell，应用无法修改，可形成FPGA用户到服务器的隔离，为设备稳定、可靠提供保障，也可保证用户安全。

此外，FPGA开发是基于芯片硬件的布局，针对不同厂家、不同型号的芯片适配开发的加速bin包不同，同时，shell-role架构涉及硬件资源划分，roll的划分区域不同同样需要适配生成不同的加速bin包。厂家的适配开发压力较大，同时多个

bin包会使开发调试和上线后的软件版本管理变得复杂，因此为了降低适配和管理难度，定制智能网卡需要明确板卡芯片型号及shell-role设计。

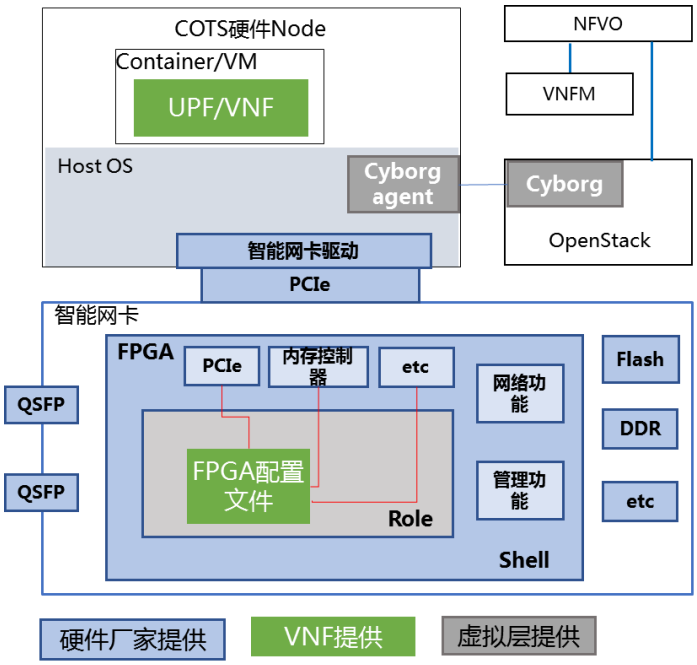


图4-7　FPGA开发流程图

综上，引入FPGA智能网卡的架构如图4-7所示。

需要定义的智能网卡要求可以总结为：

1. 通用要求，如：尺寸、功耗、环境适应性等；
2. 部件要求，如：FPGA型号、网口、内存、PCIe等；
3. 软件要求，需支持shell-role架构，并明确shell部分的功能要求：
   a) 提供基本网络功能，使智能网卡在加载加速bin包前具备网卡功能，如支持SR-IOV、VLAN透传、组播等；
   b) 提供UPF加速使用的基本IP功能，降低VNF厂家开发工作量，如PCIe、内存控制器、数据包收发端口等；
   c) 管理类功能，如role资源利用率监控、部分可重配模块实现FPGA在

线重配置等。

4. 板卡设计及shell-role设计。

中国移动引入FPGA智能网卡，目标是构建通用硬件加速平台。通用硬件加速平台包括两层含义，一是针对UPF厂家的通用，二是加速能力可向其他网元或第三方开放，构成Accelerator-aaS。采用软硬解耦方式，选择部署支持shell-role模式的FPGA智能网卡，在支持U面加速的同时，既可以给有加速需求的其他网元（如SBC），也可以以公有云模式提供FPGA-aaS，给三方应用使用，提升边缘云能力与价值。

4.3 边缘应用硬件加速关键技术

NFV边缘云针对应用需求引入GPU加速技术的整体架构如图4-8所示：

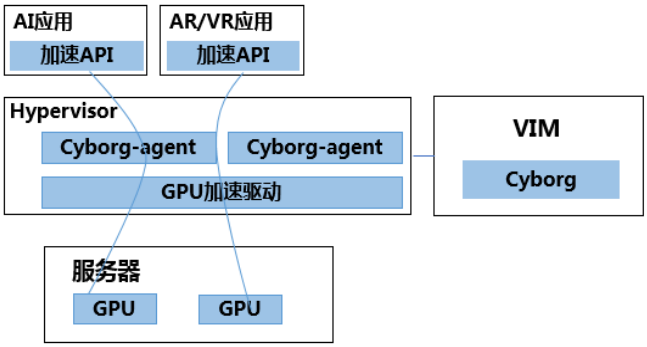


图4-8 边缘云引入GPU加速架构图

首先，在服务器层面，异构服务器需要支持不同类型的GPU卡（如Nvidia Tesla T4、Quadro RTX 6000等），并保证功耗、供电、散热的完备性。对GPU卡的管理通过OpenStack Cyborg组件统一纳管、分配、编排、挂载相应的加速器。Hypervisor需要支持GPU直通挂载与GPU虚拟化即vGPU挂载的方式，其中vGPU虚拟化的能力包括但不限定于以下三种：

1. Best effort策略：如果虚拟机没有任务或已用完其时间片，则调度程序将移至下一个虚拟机。此种策略下无法保证每个虚拟机共享GPU周期，也无

法保证每个虚拟机的性能，但能够最大化利用GPU核心资源，资源利用效率最高，没有资源浪费，且任务重的需求得到更多的资源。

2. Equal Share Scheduler策略：是一种平等共享循环调度策略，调度程序会根据当前虚拟机运行数量将GPU核心的运算时间片均等分，如果虚拟机在其时间片期间没有任何任务，GPU将处于空闲状态。每个虚拟机的GPU周期的确定份额。性能取决于运行的虚拟机数量。这个策略的好处在于资源平均分配，不会出现Best Effort策略中有些虚拟机的工作负载轻就总分不到资源的情况。缺点是性能和QoS不是稳定的，虚拟机运行中如果开启新的虚拟机，则所有虚拟机的GPU计算能力都会受到影响。

3. Fixed Share Scheduler策略：是一种固定共享循环调度策略，预先设定好最大的运行虚拟机数量，按照这个数量将GPU计算核心进行均等分，每个虚拟机拿到固定的计算核心时间片，如果虚拟机在其时间片期间没有任何任务，GPU将处于空闲状态。每个虚拟机都获得了GPU周期的确定份额，保证了每个虚拟机的GPU性能稳定。它的好处是稳定的QoS和性能指标，缺点是GPU总体利用效率比其他两种策略低。

GPU使用方式包含裸机、裸机容器、虚机及虚拟机中容器等，所有使用方式中，平台需具备识别、使用、调度、释放GPU资源的能力。

1. 识别，主机或虚拟机（包含虚拟机中容器）可以准确识别到GPU资源，主要信息包含GPU型号、规格等。

2. 使用，主机或虚拟机（包含虚拟机中容器）可以使用被分配到的GPU资源。另外，使用切片虚拟化vGPU时，虚拟机可在同资源池中不同物理机上进行热迁移；同一台虚拟机最多可配置4个1:1分片vGPU。

3. 调度，同一个硬件资源池中，虚拟机可根据所需GPU资源的类型、规格被调度到有相应资源的主机上，无需GPU资源的虚拟机可被优先调度到不带GPU资源的主机上。

4. 释放，不再需要GPU资源的虚拟机可将GPU资源释放，并供其他需要此资源的虚拟机使用。

对于GPU资源的性能，不同上层应用有不同的需求。需选择合适的GPU资源，优化底层平台中与GPU协作的资源（如CPU，网络资源等），以发挥GPU资源的最大性能。

4.4 硬件加速管理编排关键技术

当前，在NFV加速管理编排虚拟机方面，需要MANO、OpenStack和加速器协同完成，共同打通管理编排流程。加速管理架构图如图4-9所示。

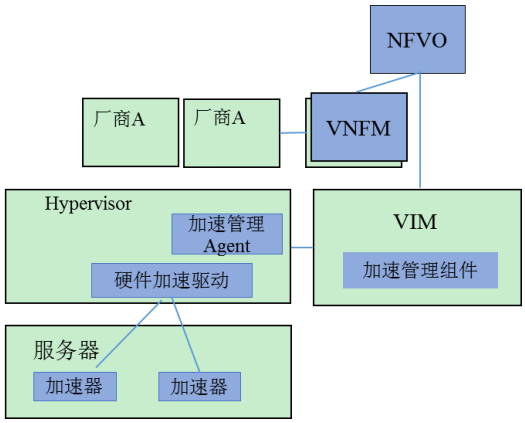


图4-9 加速管理架构图

其中，NFVO/VNFM对于加速资源管理的流程和接口要求包括：

- NFVO要求能够获取VIM上的加速资源信息
- VNFD中有对加速资源的描述信息
- VNFM能够解析VNFD中的加速资源信息
- 其他要求（可靠性、安全和兼容性要求）

Hypervisor对于加速资源的管理和兼容性要求包括：

- 加速管理Agent对加速硬件的纳管要求

- 加速镜像管理要求
- 加速数据的标准化（包括加速镜像名称，镜像UUID ,设备商，版本号，驱动，驱动版本等）
- 其他要求（可靠性、安全和兼容性要求）

VIM对于加速资源的管理和接口要求包括

- Cyborg相关组件的管理、调度和其他组件的协同要求
- VIM北向接口加速相关原生接口使用要求
- VIM需要将普通网卡与加速器信息区别上报要求
- 其他要求（可靠性、安全和兼容性要求）

当前Cyborg项目实现与需求之间还有一定的距离，需要进一步推动和加强，如表4-1所示。

表4-1 加速管理编排需求与Cyborg项目实现分析

	加速管理编排需求	Cyborg实现情况（截止到U版本）	需要推动和加强的工作
NFVO/VNFM对于加速资源的管理要求	NFVO要求能够获取VIM上的加速资源信息	不属于Cyborg范围，Cyborg只提供对上的接口，暴露相关接口和信息	需要MANO做相统一要求和实现，协同加速管理流程的端到端实现
	VNFD中要有对加速资源的描述信息		
	VNFM具有解析相应加速字段的能力		
虚拟层对于加速资源的管理和兼容性 (Hypervisor &VIM)	对加速网卡的纳管要求	目前实现的是FPGA的纳管，其他加速硬件并未有实质进展	需要考虑更多硬件的纳管，目前相对比较成熟的是FPGA
	加速镜像（FPGA）管理要求	已实现与Nova-Cyborg集成工作，已支持带加速资源的虚拟机的创建和删除等功能	1.只是实现了基本功能，尚不足以支持商用。主要表现与Openstack组件互操作。仍需要推动厂商充分参与Cyborg项目 2. 加速卡信息需与普通网卡信息区

			别上报
	加速数据的标准化问题	已经标准化	暂无
	性能和安全问题，比如迁移后，加速资源的使用问题	未考虑	只是实现了基本功能，需要从中国移动自身的需求上推动完备性考量
	告警	不涉及	可视为虚拟层告警中的一部分
VIM对于加速资源的管理和接口要求	Cyborg相关组件的管理、调度和其他组件的协同要求	已实现基本功能，尚需成熟和稳定	1.只是实现了基本功能，尚不足以支持商用。Cyborg接口是否直接对上层暴露都需要待项目成熟稳定。需要推动厂商充分参与Cyborg项目
	VIM北向接口加速相关原生接口使用要求	没有问题	推动Cyborg成熟
	VIM需要将普通网卡与加速器信息区别上报要求	没有问题	只需要vim北向接口区别上报即可
	其他要求（可靠性、安全和兼容性要求）	不涉及	只是实现了基本功能，需要从中国移动自身的需求上推动完备性考量

5 展望与呼吁

随着5G与边缘计算的快速发展，面向新场景、新架构、新需求的业务如雨后春笋般不断涌现，对网络提出了更高的要求，因此，硬件加速技术成为边缘场景下的必要条件。然而，硬件加速技术、尤其是转发面网元硬件加速技术的方案制定和产业成熟度等诸多方面尚未成熟，亟需进一步完善和推进。

当前，中国移动在OvS硬件加速、UPF硬件加速、边缘应用加速等方面开展了相关方案、架构、关键技术的研究。后续将继续在标准、开源、产业等方面，推动硬件加速尤其是转发面网元加速方案和技术的成熟，并依托中国移动新技术试验与现网试点，验证技术方案的可行性，逐步推动硬件加速技术成熟商用。

- 标准方面，中国移动将继续积极参与ETSI NFV工作组硬件加速项目IFA 001-004的制定，与业界一起提出硬件加速参考架构，其中包括加速调用方式、抽象方法、管理模式等相关要求；中国移动于2018年4月联合中国

电信和中国联通在CCSA成立“网络功能虚拟化硬件加速技术研究”项目，研究硬件加速在运营商网络架构中的需求和加速方案。我们在此也再次呼吁业界合作伙伴联合开展标准讨论制定。

- 开源方面，中国移动积极跟踪、参与OpenStack中的Cyborg项目，致力于实现加速卡在NFV管理架构下的统一管理和运维；中国移动联合华为、中兴、诺基亚、爱立信等10家企业，在OPNFV主导成立Rocket项目，旨在输出硬件加速通用API，推动由多样化加速硬件组成的通用硬件加速平台快速实现；中国移动积极跟踪、参与CNTT，致力于NFVI层标准化工作中的加速现相关模型需求、具体功能要求实现和接口的制定等。同时，中国移动也在紧密关注OHCF、DPDK等加速相关开源社区，推动整个硬件加速产业发展。

- 产业推进方面，为凝聚产业力量，中国移动积极与业界分享硬件加速技术尤其是转发面网元加速技术的相关研究和验证成果，后续将依托产业论坛、产业峰会等多种形式，进行更加深入的技术和产业合作的探讨，共同推动硬件加速产业发展和成熟。

- 试验落地方面，中国移动依托NovoNet试验网，开展面向硬件加速的技术方案验证，现已完成部分OvS硬件加速技术的测试与验证，计划针对转发面网元的加速技术进行试验验证。为寻求更加合理的解决方案，后续将吸纳更多的厂家参与到硬件加速技术尤其是转发面网元加速技术的技术攻关和试验验证之中，加速转发面网元硬件加速的商用进程。

为了推动硬件加速尤其是转发面网元硬件加速技术与产业成熟，中国移动将以开放合作、包容共赢的态度，分享最新研究进展，接收产业最新知识，愿与业界携手共同推进硬件加速技术与方案的早日落地。

附录：术语、定义和缩略语

缩略语	全称	解释
ASIC	Application Specific Integrated Circuit	专用集成电路
COTS	Commercial Off The Shelf	商用现成品
DPDK	Data Plane Development Kit	数据平面开发套件
DPI	Deep Packet Inspection	深度包检测
ECMP	Equal Cost Multipath Routing	等价路由
EPC	Evolved Packet Core	演进的分组核心
FPGA	Field Programmable Gate Array	现场可编程门阵列
IPSec	Internet Protocol Security	互联网安全协议
GPU	Graphics Processing Unit	图形处理器
GTP	GPRS Tunnelling Protocol	GPRS 隧道协议
NFV	Network Function Virtualization	网络功能虚拟化
NFVI	Network Function Virtualization Infrastructure	网络功能虚拟化基础设施
NFVO	NFV Orchestrator	NFV 编排器
SDK	Software Development Kit	软件开发工具包
SoC	System on Chip	片上系统
SR-IOV	Single Root Input/Output Virtualisation	单根输入/输出虚拟化
MANO	Management and Orchestration	管理与编排
MEC	Mobile Edge Computing	移动边缘计算
NP	Network Prossessor	网络处理器
OHCF	Openn Heterogeneous computing Foundation	开源异构计算基金会
OvS	Open OvS	虚拟交换机
QoS	Quality of Service	服务质量
RAN	Radio Access Network	无线接入网
VIM	Virtualised Infrastructure Manager	虚拟化基础设施管理器
VM	Virtual Machine	虚拟机
VNFC	VNF Component	VNF 组件
VNFM	VNF Manager	VNF 管理器

结 束 页

中国移动研究院

地址：北京市西城区宣武门西大街32号

邮政编码：100032

联系电话：15801696688

E-mail:wangshengyjy@chinamobile.com