协议无关交换机架构
技术与应用白皮书

联合编写单位：

中国联通网络技术研究院

网络通信与安全紫金山实验室

北京邮电大学

Barefoot Networks

2019 年 10 月

目 录

1 序言

Networking has changed enormously in the past 10-15 years. In 2005, every hyperscale datacenter used closed proprietary vendor software to run their network. Today, all of the top 10 hyperscalers worldwide write their own software to control their network (e.g. SONiC, FBOSS, Stratum), or use open-source. Next, the big telco operators will do the same, with ONAP and DANOS and ONF's ONOS, Trellis and SEBA. Why has this happened? Basically, the industry model was wrong: Only network operators know best how to control their network, so they need to write their own software. We are in the middle of a big industry transition in which they are taking over the control plane software. Nowadays we take it for granted that they can write, commission or download software and tailor it to optimize their network and make it better than their competition. Basically, they have taken charge of the software that controls their networks, which is what needed to happen - it was inevitable. I call this the First Stage of SDN.

In the Second Stage of SDN, the hyperscalers are starting to take control of how packets are processed too. After all, a network is merely a method to transfer packets from a source to a destination and process them along the way. If we are not in charge of how the packets are processed, we are not really in charge of our network. And so, there was a big need for switches to become programmable too. We started thinking about this in about 2010 and we started a project between my group at Stanford University and TI to see how we could do it. By 2013, we learned we could build a programmable switch with the same power, performance and cost as the traditional fixed-function switches. So we started Barefoot in 2013 to go make it happen. The P4 language and the Tofino switch were born soon after. Today, Tofino in 16nm has the same, or better, power, performance and cost as the leading fixed-function switches in 16nm. This is huge and very surprising to most people. There is no turning back and future switches will be programmable. With Intel's acquisition of Barefoot, the two biggest chip companies - Intel and Broadcom – are now building programmable switch chips. It makes me very happy to see.

We all want to help improve the Internet, to allow it to evolve and improve faster, to make it more reliable, more secure and easier to manage. SDN is helping network owners improve their networks faster than ever before. By allowing software developers to write the programs to decide how their networks behave, we are unleashing a “Cambrian explosion” of beautiful new ideas which will keep improving the network, making it more reliable, more secure and gradually work together with CPUs and accelerators to make our applications run faster. It is a very exciting time for networking.

Professor of Stanford University
Member of the US National Academy of Engineering
Co-founder of Barefoot
Nick McKeown

在过去的 10 到 15 年里，网络世界发生了巨大变革。2005 年的时候，每个超大规模数据中心的建立所采用的都是封闭的专有供应商软件。而今天，全球十大超规模数据中心，诸如 SONiC, FBOSS, Stratum 等，都通过自己编写的软件或是开源软件来控制网络。不久的将来，大型电信运营商也将通过 ONAP, DANOS，ONOS，Trellis 和 SEBA 来实现自主掌控。为何会发生如此巨大的变革呢？从根本上来说，传统的行业模式自身就存在着一定的问题：只有网络运营商最了解如何控制自己的网络，因此最好的实现方式本就应该是他们自己编写软件。我们正处于一个网络运营商尝试接管控制平面软件的比较关键的行业转型期。网络运营者理应可以编写，或者加载软件，并对其进行控制，以优化其网络，使之优于竞争对手。网络运营者在控制网络软件方面已经可以独当一面，这是大势所趋，也无从避免，此为 SDN 的第一阶段。

而在 SDN 的第二阶段，运营者要开始去学习控制数据包的处理方式。毕竟，网络仅仅是将数据包从源节点传输到目的节点并在传输过程中对其进行处理的一种方法。如果运营者连数据包的处理过程都不能控制的话，又何谈控制整个网络？因此，交换机支持可编程性的需求也变得十分迫切。2010 年前后，我所在的斯坦福大学的团队以及 TI 基于此课题展开了探索之旅。直到 2013 年，我们才初步确定，构建一个同功能固定的传统交换机具有相同的功能、性能和成本的可编程交换机是可行的，同年，我们创立了 Barefoot 来具体实现这一构想。不久之后，P4 语言以及 Tofino 交换机就诞生了。如今，16nm 的 Tofino 在功率、性能和成本方面与 16nm 最优的固定功能交换机相比，有过之而无不及。于我们而言，这是意想不到的惊喜。随着 Intel 对 Barefoot 的收购，业内最大的两家芯片厂商——Intel 和 Broadcom，都参与到了研制可编程交换芯片的工作中。我本人对此十分高兴。

我们都希望能为改善互联网贡献一己之力，让它得以快速发展和演进，也变得更安全可靠、更易于管理。而 SDN 正在用其特有的方式帮助网络所有者以空前的速度改善网络。随着 P4 的诞生，一系列新奇瑰丽的想法接踵而至，这些想法将不断改进网络，逐步完成与 CPU、加速器等的协同工作，从而使我们的应用程序高速运行，某种程度上可以说，我们释放了又一次的“寒武纪大爆发”。网络时代的高光时刻已然到来！

--斯坦福大学教授 & 美国国家工程院院士 & Barefoot Networks 联合创始人

Nick McKeown

2006 年，SDN 诞生之初，提出了控制平面与转发平面相分离的思想，为人们打开了认识和定义网络世界的全新窗口，OpenFlow 进入大众视野，提供控制器和数据面的动态交互，开启了网络可编程时代。然而随着网络运营者对可编程需求度的持续升高，OpenFlow 所能提供的目标无关可编程性已经远远不够，更深层次的诉求是实现协议无关可编程性，基于此，一种全新的高级编程语言——P4 应运而生。

正所谓一石激起千层浪，P4 的出现，很大程度上缓解了网络运营者的痛点，给了他们自己定义数据面的权限，因此而拓展出来的可应用场景也覆盖了网络的多个层面，我们很乐于看到这类颠覆性的新型语言和架构所带来的一系列惊喜，这无疑是具有里程碑意义的。

本册白皮书从简单介绍 P4 语言的诞生以及 PISA 架构的特点入手，继而引出可编程交换芯片的可应用领域，系统地总结了目前已知的可应用场景，很适合想要了解 P4 和 PISA 的业界同仁参阅。当然，我们相信随着网络开发者对这一领域的持续关注，更多元更广阔的应用空间还有待探索。随着白盒化工作在运营商网络中的逐步推进，未来 P4 必然会在可编程网络的舞台上一展风采。

--中国联通智能网络中心总架构师，中国联通网络技术研究院首席科学家 唐雄燕

在过去十年中，随着 SDN、NFV、云计算等技术的深入发展，SDN 技术已逐步实现了大规模应用，并推动了网络控制方式的变革。协议无关交换技术的出现，则从数据转发行为的方面对未来网络提出新的变革思路，预计将会进一步提升网络的开放性和可编程能力，对传统网络向未来网络的演进具有重要意义。与传统交换机相比，协议无关交换机通过采用更加开放的架构，可实现交换机软件与硬件的解耦，使网络设备更加通用化，从而一定程度上降低设备的硬件成本。另一方面，协议无关交换机可提供完全开放的可编程样本模型与设计思路，便于用户实现网络控制应用的快速迭代，进而实现网络的深度定制与协议优化，满足不断变化的网络功能与需求。

当然，协议无关交换技术的未来发展也还面临一些挑战。首先，软硬件解耦固然增强了灵活性，但同时对设备稳定性也带来了极大挑战，如何构建电信级高性能网络交换设备成为重要难题。其次，目前协议无关交换机主要应用于数据中心、接入网络等场景，未来如何面向骨干网、城域网、5G 核心网等场景开展应用也成为需要进一步探索的议题。此外，协议无关交换技术还将涉及到整个产业生态发展的问题，如何构建产业链上下游企业、高校、科研机构各方积极投入、蓬勃发展的产业生态，对于未来网络的发展与演进具有重要意义。

总体来看，随着面向 2030 年“万亿级、人机物、全时空、安全、智能”的未来网络发展愿景，网络技术正在快速创新和迭代。我认为从技术进步的角度来看，协议无关交换技术作为未来网络领域的重要方向之一，非常值得业界积极尝试与探索。技术的进步将会推动整个行业的发展，通过对该项技术的研究，有可能掀起新一轮运营商网络的技术创新与架构变革，进而带动整个网络产业的创新与发展。

——北京邮电大学教授，紫金山实验室未来网络中心主任 黄韬

2 云网业务与 SDN 发展趋势

SDN 最初起源于美国斯坦福大学的实验室，2008 年，McKeown 教授等人在 ACM SIGCOMM 发表文献首次详细地介绍了 OpenFlow 的概念。基于 OpenFlow 为网络带来的可编程特性，McKeown 教授提出了 SDN 的概念。SDN 是一种数据平面与控制平面分离，并可直接对控制平面编程的新型网络架构。数控分离将有助于底层网络设施资源的抽象和管理视图的集中，从而以虚拟资源的形式支持上层应用与服务，实现更好的灵活性与可控性。

SDN 自提出以来，一直受到来自各界的关注，许多标准化组织，如 ONF 和 IETF 等，都围绕 SDN 开展了相关工作，讨论 SDN 在各自相关领域的发展及应用。当前，SDN 技术的发展趋向于更加开放灵活的数据平面、更高性能的开源网络硬件、更加智能的网络操作系统、功能虚拟化的网络设备、高度自动化的业务编排等五个方面。SDN 产业发展趋势主要趋向数据中心场景下的创新应用、运营商场景下的创新应用、产业界大规模的商用部署等三个方面。近年来，产生了众多与 SDN 相关的网络新技术，包括 SR、IBN、P4 技术、SD-WAN 技术、基于 SDN 的 IP+光技术、软件定义光网络技术、智能网卡技术等。近年，SDN 与多种网络架构融合，在内容中心网络、移动边缘计算、IBN、P4、SD-WAN 等领域开展了广泛的研究，也得到了持续的发展。

2013 年，McKeown 教授和一些研究 P4 的同事成立了 Barefoot Networks 公司（已于 2019 年 7 月被 Intel 收购），致力于开发基于 P4 的网络芯片 Tofino 和软件开发套件（现更名为 P4 Studio），并帮助 P4 社区发展壮大。P4 是对数据包进行处理的编程语言，帮助网络用户摆脱来自芯片硬件厂商的各种协议制约。未来 SDN 的研究与应用仍有很大的空间。根据 IDC 预测， SDN 应用预计到 2020 年将实现 66%的年复合增长率，届时市场规模将超过 35 亿美元。

接下来，本白皮书将在 3 至 6 章依次介绍 P4 可编程语言、协议无关可编程芯片设计理念与架构、可编程芯片的主要应用领域、行业生态合作等内容，并在第 7 章进行总结与展望。

3 P4 可编程语言

3.1. P4 的诞生和概念

2014 年，由 McKeown 教授等联合发布了一篇论文《P4: Programming Protocol-Independent Packet Processors》，该论文在 SDN 界引起了极大的反响和关注度。随后，Nick 教授等人又发布了《The P4 Language Specification》、《Barefoot 白皮书》等文件。目前，P4 已经在国外引起了足够的重视，ONF 成立了协议无关转发的开源项目，该项目目前的工作重点就是为 P4 提供配套的 IR，而项目的工作成果也将被用来设计下一代的 OpenFlow 协议。

P4 是一种专用的编程语言，其目标为协议无关性、目标无关性以及现场可重配置能力，它能够解决 OpenFlow 编程能力不足以及其设计本身所带来的可拓展性差的难题。首先 P4 定义数据包的处理流程，然后利用编译器在不受限于具体协议的交换机或网卡上生成具体的配置，从而实现用 P4 表达的数据包处理逻辑。

$P4_{14}$ 和 $P4_{16}$

截至目前，P4 语言分为 $P4_{14}$ 和 $P4_{16}$ 两个大的版本。其中，2017 年 5 月 P4 社区（p4.org）发布的 $P4_{16}$ 语言是当前 P4 语言的最新版本。一个 $P4_{16}$ 结构文件应包含类型声明，常量声明以及用户需要的控制和解析模块的说明。有些数据包处理任务无法在 P4 中表达，$P4_{16}$ 支持外部功能或方法来解决这个问题，即在 P4 之外实现计算功能，可以从 P4 程序中调用。和 $P4_{14}$ 相比，$P4_{16}$ 的语言风格发生了较大的变化，它在整体语言风格上向 C++语言进行了借鉴和学习。此外，$P4_{16}$ 允许程序在任意目标上执行，对于执行的包处理类型和自定义功能的目标，$P4_{16}$ 都提供了表达的语言机制。P4 语言的开源项目都托管在 GitHub 中，在 p4lang 组织的仓库中还有很多开源项目。

3.2. P4 的创新点及优势

软件定义网络的可编程性目前仅局限于网络控制平面，其转发平面在很大程度上受制于功能固定的包处理硬件。新一代高性能可编程数据包处理芯片加上“P4”高级语言的出现，让网络拥有者、工程师、架构师及管理员可以自上而下地定义数据包的完整处理流程。这种可编程数据平面有助于网络系统供应商进行更快速的迭代开发，甚至直接通过打补丁的方式修复现有产品中发现的数据平面程

序漏洞。它也可以帮助网络拥有者实现最适合其自身需求的具体网络行为。它还能使网络芯片供应商专注于设计并改进那些可重用的数据包处理架构和基本模块，而不必纠结于特定协议里错综复杂的细节和异常行为。

4 协议无关可编程芯片设计理念与架构

4.1. PISA 模型

PISA，指协议无关交换机架构，这是一种在用户完全程序控制下以最高速度处理数据包的新范例。实践证明，PISA 用户可以使用开源编程语言自行编程网络，而不会降低其性能。PISA 体系结构把数据平面全部控制权都交给网络所有者。为了做到这一点，PISA 确定了一个用于处理数据包的小的原始指令集，以及一个非常统一的可编程流水线，用以快速连续地处理数据包头。程序是用高级域特定语言（P4）编写的，经由 P4 语言编译器进行编译，并在 PISA 设备上以全速率运行。

PISA 架构的组成部分

PISA 架构的交换机可以包含以下组件：解析器/逆解析器、匹配-动作表、元数据总线。其中除了元数据总线，其他组件都是非必须的。解析器（Parser）将分组数据转化成元数据，逆解析器（Deparser）将元数据转化成序列化的分组数据。匹配动作表（Match-action Table）用于操作元数据。元数据（Metadata）负责存储数据信息。流表整合了网络中各个层次的网络配置信息，从而在进行数据转发时可以使用更丰富的规则。流表可自定义实现，控制面下发的流表，从匹配字段(Match-field)到动作(Action)都必须与 P4 程序中定义的 Match-action Table 相吻合。

4.2. 基于 Tofino 的交换设备

Tofino 芯片

Barefoot Tofino 交换芯片是业内第一个支持 PISA 架构的以太网交换 ASIC，Tofino 芯片为网络设计者提供了协议无关交换架构的强大功能。Tofino 芯片是完全可编程的，转发逻辑是由网络运营商或交换机制造商加载到芯片上的 P4 程序决定的，而不是在硬件中固定的。Tofino 芯片是独立于协议的，由 P4 程序提供处理所有支持协议的逻辑，而芯片并不知道它需要支持的网络协议。图 4-1 为 PISA 交换机架构图，数据包可经过自定义的解析后进行匹配和动作操作，实现数据平面的协议无关转发。与此同时，可编程性不会引入更多的功耗和成本。

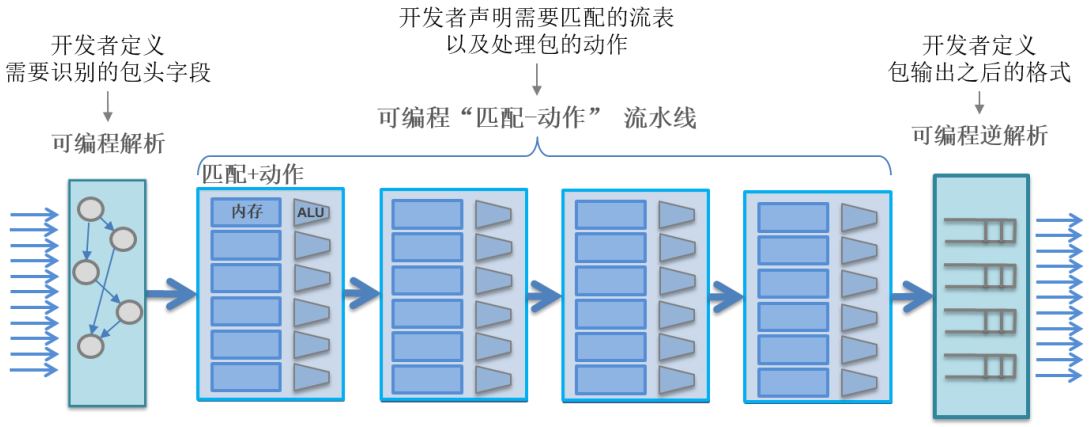

图 4-1 PISA 交换机架构

当需要支持新协议时，网络运营商或交换机制造商只需向 P4 程序添加新的逻辑。云服务提供商们的一个趋势是自产芯片，推动芯片向特定领域发展。比如 GPU 不仅仅应用于图形领域，机器学习中也有它的身影，如图 4-2 所示，他们的共同特点都是基于可编程的芯片和针对特定领域的指令集，然后将高级语言编译后实现特定功能，Tofino 芯片（及其他 PISA 架构的交换设备）和 P4 语言就是这一趋势在网络交换数据平面的体现。新一代可全编程交换机芯片——Tofino 2 同样利用 Barefoot 的 PISA 架构，并使用 P4 可编程语言实现数据包转发平面的编程，同时，SERDES 带宽相比前一代增加一倍，因此芯片的总带宽达到 12.8Tb/sec，并且拥有更多的可编程逻辑资源。

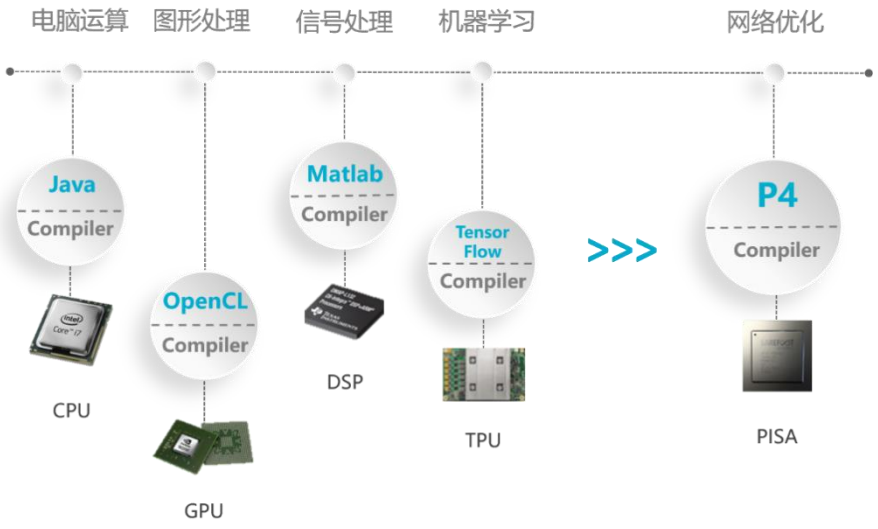

图 4-2 特定领域架构与 PISA

交换设备

固定功能交换机 ASIC 已不再是构建未来网络的可行选择。一个不影响性能

的可编程转发平面是从网络中释放革命性价值的关键。Barefoot Tofino 是世界上第一个终端用户可编程以太网交换芯片。业内已经有各种设备厂商生产/出货基于 PISA 架构设计的交换设备，如 Arista, Cisco，中国台湾的 EdgeCore, Inventec, Ingrasys 等设备厂商。

4.3. SDK vs SDE

SDK，即软件开发工具包，指软件工程师为特定的软件包、软件框架、硬件平台、操作系统等建立应用软件时的开发工具的集合。SDE，即软件开发环境，指在基本硬件和宿主软件的基础上，为支持系统软件和应用软件的工程化开发和维护而使用的一组软件。

与提供固定功能 ASIC 的 SDK 相比，可编程芯片 SDE 提供了更多的新增功能，如 P4 语言编译器、调试工具等。它提供了一整套用于开发、调试和优化 P4 应用程序的工具，允许数据层面的自定义功能。此外，设备和抽象 API 允许开发人员轻松地将 P4 应用程序与本地或远程控制平面集成。这些工具和 API 使原始设备制造商、云运营商、电信运营商和生态系统合作伙伴能够构建具有高度差异性的适用性强的网络解决方案。

5 可编程芯片的主要应用领域

5.1. 带内网络遥测

INT，即带内网络遥测，可以直接在数据路径中收集端到端的实时状态信息。源端点在包中嵌入指令，列出要从网络元素收集的网络状态类型。每个网络元素在数据包通过网络时在数据包中插入请求的网络状态。P4 程序可以作为一种自然的方式来表示 INT 所需的包头解析和修改。

PISA 架构的 P4 可编程交换机使用 INT 标准和 INT 遥测报告可以将感兴趣的元数据嵌入到每个数据包中，或通过单独的数据包传递到像 Deep Insight 这样的分析引擎，如图 5-1 所示，从而对每个数据包提供深入的可视性。运营商可以利用 Deep Insight 来检测网络中几乎所有的异常情况，包括微突发、拥塞问题和负载均衡问题。这允许用户发现网络中每个包的四个基本事实：1）经过了哪条路径；2）传输规则是什么；3）在每个节点延时了多久；4）一同排队的有哪些包。利用 P4 可编程性和 INT 技术可以提供对网络问题，如数据包丢失和延迟问题等的实时检测，从而提高应用程序性能。

图 5-1 INT 端到端视图

5.2. 云网性能优化

传输层负载均衡

基于 PISA 架构的 P4 可编程交换机（以下简称 P4 交换机）的网络第 4 层负载平衡弥补了多太比特交换机与千兆服务器和设备之间的性能差距。P4 交换机可在其内部实现负载平衡，为第 3 层和第 4 层服务和应用程序提供太比特流量分

布。

传输层负载平衡基本上是从连接（即源和 VIP IP 地址、协议类型和 L4 端口号）到服务器 DIP 的映射功能，为每个 VIP 管理一个 DIP 池，在专用 VIP 表中维护 VIP 到 DIP 池映射。在传统部署中，当需求增加时，物理和虚拟负载均衡器通常会与性能发生冲突。尤其具有挑战性的是在多租户数据中心中启用和扩展 LB 资源。而通过 P4 交换机，大量基于软件的负载均衡器、服务器可以被一个基于 PISA 架构的现代交换机所取代，通过分布式架构和优化的流量路径，将负载平衡的成本降低了多个数量级，P4 交换机负载均衡架构如图 5-2 所示。

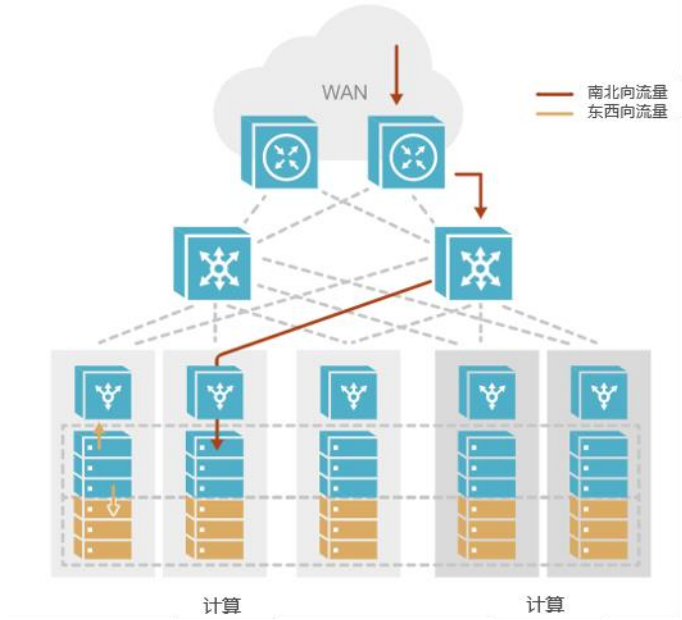

图 5-2 Tofino 负载均衡架构

即使服务器池发生变化，或者负载在池中的分布改变，P4 交换机负载均衡器仍可以实现将连接始终映射到同一个服务器上，即保证连接的相关性。传统商用交换 ASIC 不能实现具有连接相关性保证的负载均衡器，因为高性能交换 ASIC 通常无法提供具有连接相关性保证的 L4 连接状态。P4 交换机可以确保即使在同时发生 DIP 池更改和数百万连接的情况下，也能提供连接相关性保证，同时允许低延迟和 6.4Tbps 线速转发能力。

P4 交换机传输层负载均衡器有多种用例，如：扩展专用 Web 服务，如 SSL 加速器、HTTP 压缩和其他；扩展安全服务，如入侵预防系统、入侵检测系统、Web 应用防火墙等；高性能视频分发/缓存。

DDoS

DDoS，即分布式拒绝服务，通过使用分布式源生成的请求使合法用户的资源（网络、CPU、内存等）过载，使其无法访问服务、机器或网络。通过降低关键服务并导致其他攻击。来自物联网设备、路由器或云的 DDoS 威胁的复杂性和规模每年都在增加。

由于流量模式类似于合法数据包，因此 DDoS 检测本身就很难实现。因此，基于签名的检测是无效的。请求来自许多源 IP，因此基于速率和源的过滤也不起作用。这种攻击的一个显著特征是存在来自多个源 IP 的多个连接，每个连接发送一个或多个数据包。

传统的 DDoS 检测和缓解/解决方案往往利用许多带外 DDoS 检测设备监视流量, 并将可疑流量重定向到状态防火墙。在这种设计中，由于规模和成本的原因，DDoS 检测无法监控所有入网流量。因此，操作员必须在边缘路由器上设置一些静态镜像规则来镜像部分流量。因此，对于高度分布式和复杂的攻击，设备要么能够扩展以检测每秒数兆比特的流量和数百万个连接，要么能够以较低的准确度监视小规模流量。

Tofino 芯片可以使用一种被称为近似基数计数器的方法计算跨越每个基于 PISA 架构的设备的唯一连接（即 5 元组流）的数量。具体的检测方法有以下两种：1.控制平面可以通过轮询的方式，在指定的时间间隔内定期向硬件获取和估计唯一连接的数量。如果估计的唯一连接数量超过了预定义的阈值，控制平面可以认为检测到一个 DDoS 攻击，并将流量镜像到分布式防火墙；2.为了在不给控制平面增加开销的前提下提高检测速度，Tofino 芯片还可以直接在数据平面中估计和比较阈值。当数据平面检测到一个 DDoS 攻击时，会向控制平面发送信号。

基于 Tofino 芯片的 DDoS 方案的优势主要有：在任何类型的攻击下都能够利用最少的内存和资源消耗来实现高度扩展性和线速性能；具有很高的准确性和几乎可忽略的误报概率；P4 可编程性允许客户灵活定制 DDoS 检测方法和缓解措施；细粒度的统计数据允许客户快速识别受到攻击的应用程序和服务；NetFlow 解决方案的检测耗时以几十秒为数量级，而 Tofino 芯片的解决方案则以几十毫秒为数量级。

NetCache

基于 P4 可编程交换机的功能和灵活性而构建的 NetCache 键值存储架构，可以实现对热点项目的查询，并均衡存储节点间的负载。与传统的缓存和存储服务器相比，可编程交换机针对数据输入输出进行了优化，并提供了更好的性能，使

之成为构建高性能内存键值存储的路径缓存层的理想场所。NetCache 使用交换机作为负载平衡缓存，理论上只需要有限的缓存项（约 O(N*log(N))，其中 N 为存储服务器数量）就可以得到很好的平衡效果。针对服务过程中出现的热点访问项目进行负载均衡，可以使服务器上的剩余负载更加统一，即使在快速变化的工作负载下，NetCache 也提供了较高吞吐量和较低延迟。

NetCache 的核心是一个包处理流水线，它利用现代可编程交换 ASIC 的能力，有效地检测、索引、缓存和服务交换数据平面中的热键值项。使用匹配动作表来对包头中携带的关键字进行分类，并在可编程交换机中使用作为片上存储器的寄存器阵列来存储键值。图 5-3 显示了如何用匹配动作表和寄存器数组构造一个简单的键值存储。该表使用精确匹配来匹配数据包头中的关键字字段，并为每个匹配的关键字提供一个索引作为操作数据。

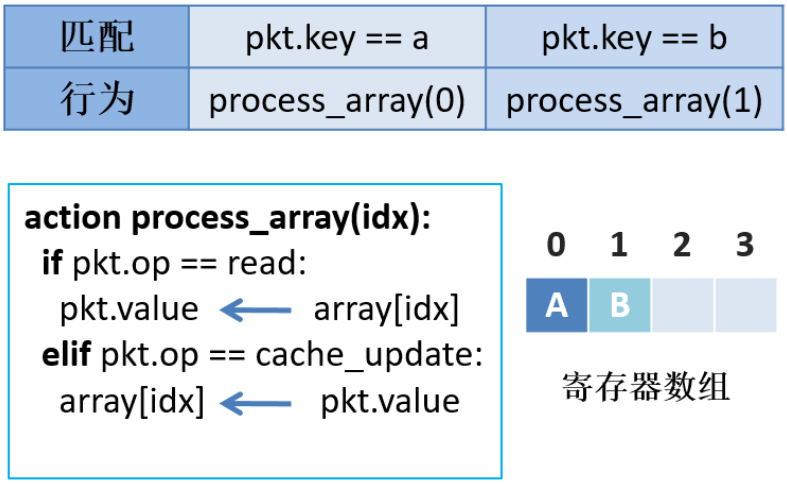

图 5-3 Netcache 数据平面键值存储

NetCache 中缓存项的读取查询由交换机直接处理，无需访问存储服务器。对缓存键执行写入查询，会使到达存储服务器路径沿途的所有交换机上的副本失效，继而服务器会相应地采用新的参数更新交换机。因此，NetCache 可以自动实现以较低开销来保证缓存的一致性。

5.3. 5G 承载

随着 5G 部署进程的加快，随之而来的竞争也在加剧，商业竞争的成败很大程度上取决于各大运营商能否为新一代服务与应用提供高性能及全面可视化的网络基础设施。5G 可适用于汽车、通信、国防、医疗保健、工业生产以及媒体等一系列丰富的应用场景。为了实现相关业务的高质量承载，5G 需要满足大带

宽、超低时延、高可靠、高精度同步、灵活性、网络切片、智能协同等诸多需求。

在现网部署时，引入协议无关可编程交换机能够带来转发时延优化，以及边缘云卸载等优势。

转发时延优化

对于超低时延的标准，对比 4G 时代而言，5G 有了更高的要求。而拥塞控制是高速网络实现超低延迟、高带宽和网络稳定性的关键。阿里巴巴在 2019 年的 SIGCOMM 上提出了 HPCC 机制，通过引入可编程网卡和基于 P4 的可编程交换机同时满足了高速网络对延迟，带宽和稳定性的要求。

P4 编程可以简化交换机很多不必要的功能，从而带来一定的时延优化，此外，HPCC 利用 P4 强大的 INT 功能可以获得精确的链路负载信息，并精确控制流量。通过解决网络拥塞过程中的延迟 INT 信息和对 INT 信息的过度反应等问题，HPCC 可以在避免拥塞的同时快速收敛以利用空闲带宽，并能在网络队列中保持接近零的超低延迟，完全满足 5G 在时延方面的高要求。通过与 DCQCN 和 TIMELY 进行数据对比，HPCC 将流量处理时间缩短了 95%，即使在大规模 Incast 的情况下也不会造成太大的拥塞。

除此之外，如本章第一节所述，通过 INT.P4，运维人员可以精确地知道数据包通过每一个交换机的确切时延。而在未来的 5G 网络中，获得精确的时延数据，有利于快速定位网络时延的瓶颈节点，从而重点部署优化方案，以期为客户提供更优质的服务。

边缘云卸载

P4 在边缘云卸载方面的应用目前主要体现在 VNF 卸载以及 UPF 卸载上。

VNF 卸载方面：基于 P4 的数据层面可编程性提供了很强大的灵活性，使得数据层面不仅仅是一个传输管道。以 P4 具备的可编程能力出发，可以构建相应的 VNF 模块，如表 5-1 所示。VNF 卸载后，对于 DC 交换机，VNF 可以工作在 Tbit/s 为数量级的线速，性能得到了很大提升；由于避免了 x86 存在的不确定性处理等问题，从而降低了时延和拥塞；相较传统模式，处理数据包所占用的 CPU 资源变少，从而降低了整体能耗。

表 5-1 基于 P4 可编程能力构建的 VNF 模块列表

可编程能力	VNF 构建模块
任意包头解析/逆解析	域内特定封装/解封装(如：PPPoE 终端，GTP 等)
状态记忆	TCP 连接跟踪(如：L4 负载均衡，NAT,防火墙等)
计算能力	计费

通过大约 300 行 $P4_{16}$ 代码构建的 spgw.p4 可以实现在服务和分组网关用户平面对 PoC 的 P4 部署，该脚本支持 GTP 封装/解封装、过滤、计费功能，足以很好地描述端到端连接。不过当前的版本还不支持 QoS 以及切换期间的下行缓冲等，未来工作中可根据需要进行开发。

目前，ONF 正在同国外某运营商公司合作推进基于 P4 的住宅服务边缘/BNG 生产级部署的开源化，基于此可以快速地为住宅接入获得相应路径。目前计划支持的特性有 PPPoE 终端、预留路径过滤（MAC, IPv4/v6）、测量、TR-101 双 VLAN 终端以及两层标签的 MPLS 终端等。

UPF 卸载方面: 某创新解决方案提供商专门为联通 5G 核心网 CUPS 架构定制了基于 Barefoot 交换机的 UPF 解决方案，方案中采用了完全分离部署的处理模式，控制平面和用户平面均可以按需扩展，同时也提供灵活处理方式，既可以独立部署，也可以集成在 Fabric 中进行部署。目前，单个 UPF 节点最大可支持 3Tbps 的吞吐率，未来随着 PISA 架构交换机的发展，性能可以进一步提升，此外，诸如负载均衡，QoS，DPI 等相关的 L4-L7 功能也会相继引入。

5.4. 网络自动化测试

P4 等编程语言可以对网络进行编程，这赋予了网络测试新的能力。当前网络测试主要依靠专用的网络测试仪表来实现网络流量测试、数据拦截、IP 查询、流量分析等功能。然而，传统的测试仪表价格昂贵，软硬件成本较高，例如，8 端口 100G 的高端测试仪表售价可高达百万，而且测试仪端口密度较小，难以实现高端口密度。并且，现有测试仪协议升级困难，用户难以自定义新的测试协议，需要依靠测试仪表厂商来提供新的协议测试功能，给网络新协议与新功能的测试带来了挑战。通过 P4 可以实现新型网络测试仪表，利用 P4 强大的可编程与转发能力，可以实现低成本高端口密度的网络测试仪，并且能够对用户开放灵活的可编程接口，支持用户自定义一些新的测试协议，实现更加灵活的网络功能测试。

此外，当前大部分网络自动化测试仅能实现简单的端到端连接测试，无法发

现交换机的问题或者单个数据包经历的拥塞过程。利用 P4 等编程语言，用户可以在可编程转发平面快速部署测试和探测，通过将探测添加到数据平面数据包去识别拥塞链路，或者以高速率生成测试数据包以测试许多不同路径的状态。根据网络的工作原理和测试需要，程序员只需编写简单的测试程序就可以实现网络自动化测试。

6 行业生态合作

6.1. 开源生态

Stratum

2018 年 3 月，ONF 发布了下一代 SDN 接口战略，并在谷歌的支持下推出了 Stratum 项目，同 Openflow 仅仅定义控制转发的行为机制不同，该开源项目的目标是提供一个白盒交换机和开放软件系统，通过使用可编程芯片以及包含 P4 和 P4 Runtime 的工具箱，来实现真正的软件定义的数据平面参考平台，并基于此支持包括配置、控制、操作、可选流水线可编程性等在内的全生命周期的控制和管理。与此同时 Stratum 创始成员计划采用尽可能广泛的网络芯片以及来自多厂商的白盒交换机来提供 Stratum 解决方案，作为 Stratum 首个版本的代码贡献者，谷歌即将在其生产网络中部署 Stratum。2019 年 9 月，ONF 宣布 Stratum 项目正式开源化，目前已获得 Apache 2.0 开源许可证。

除了谷歌之外，其他参与 Stratum 项目的成员包括：

- ✧ 电信运营商：中国联通、NTT、土耳其电信；
- ✧ 网络厂商：Big Switch Networks、锐捷网络、VMware；
- ✧ 白盒 ODM 厂商：Delta、Edgecore Networks、QCT；
- ✧ 芯片厂商：Barefoot、Broadcom、Cavium、Mellanox、Xilinx。

支持 P4 的 FPGA 和 NIC

由于 P4 的可编程性不仅可应用在网络交换机上，同时也可以应用于服务器中的智能网卡，基于此，Barefoot Networks 在演示 P4 的可编程能力方面已经同某些业界知名厂商达成了合作。

5G 网络中，为了支持高带宽需求，NFV 处理数据包的线路速率需要达到 100G/200G，在此基础上才能提供 vOLT、vBNG 等服务。以实现上述特性为目标，网络设备需要支持 L2/L3 处理、流分类、流缓冲、OvS、TCP/IP 以及 100G/200G 数据包速率下的 VXLAN/NVGRE 卸载功能等。

基于某厂商 FPGA 的 P4 可编程智能网卡来实现上述功能，在增强网络性能

的同时，还可以获得 Barefoot 深度分析功能带来的可视性优势，在此基础上，网络运营商能够使用这种功能来精准定位数据包在通过网络中整条路径时的问题所在。为进一步优化网络性能提供强有力的保障。

P4 开源平台

P4.org 成立于 2015 年 5 月，作为一个非营利组织，其已经建立起了一个蓬勃发展的开源社区，致力于 P4 语言的使用和改进，同时让参与者能够开发符合 P4 语言规范的新技术。P4.org 拥有来自工业界和学术界的 100 多名成员，自诞生以来，相继成立了五个工作组，分别是：语言设计工作组、API 工作组、整体架构工作组、应用工作组和教育工作组。P4 工作组负责处理与编程语言相关的所有技术活动，P4 组织的成员可以通过登录邮件列表自由参加任何工作组，相关的邮件列表可以供公众使用。

2019 年 4 月，ONF 宣布已经完成与 P4.org 的合并，此后将主持所有 P4 活动和工作组。ONF 的运营商成员包括 AT&T、中国电信、中国联通、康卡斯特、德国电信、谷歌、NTT 和土耳其电信。ONF 将把 P4 工作置于运营商主导的开源 SDN 项目之下，同其他所有的 ONF 项目一样，ONF 旨在战略性地将 P4 活动与 Linux 基金会结合起来，继而推动开源的发展。ONF 同 P4.org 合并之后，P4 项目将继续按照之前的方式运作，同时又可以受益于与 ONF 的战略联盟以及 ONF 的运营基础设施和专业知识，以期推动 P4 的蓬勃发展。

6.2. 产业生态

运营商网络

P4 和 Tofino 在运营商网络中拥有广泛的使用案例。早在 2017 年 3 月，AT&T 就宣布已在部分现有的基于 MPLS 的网络中安装了基于 Tofino 的白盒。AT&T 还利用该芯片的可编程性来改进网络遥测，在旧金山和华盛顿特区分别布置了一个基于 Tofino 的交换机，利用 Tofino 全面了解他们在全美的流量。2018 年，AT&T 开源 dNOS，它支持现有的网络协议，也能提供扩展功能，以支持新工具，如开源编程语言 P4。

除此之外，在许多传统网络中，运营商仅仅使用了中间设备中的很少部分功能，现在通过 Tofino 和 P4 的集成，可以直接将需要的功能编程到交换机上，从而淘汰掉大量昂贵的中间设备。目前已经有成功部署 Barefoot Tofino 交换机的实

际案例，显著的降低了成本，并且大概率上，性能有所提升，因为相较于传统解决方案，Tofino 实现了全线速转发，在一个应用场景中，将 L4 负载均衡集成到 Tofino 交换机上，可以维持十万台服务器的数千万个连接，DIP 池可以调整大小，而无需中断此时的连接。以上这些功能，只需要几百行 P4 代码就可以实现。相似的方法也可以集成其他中间设备，比如防火墙，入侵检测系统，地址接口转换器等等。

科研试验网

未来网络试验设施（CENI）作为我国在通信与信息工程领域的唯一的国家重大科技基础设施，是国家重大科技基础设施建设中长期规划（2012—2030 年）优先安排的项目。设施覆盖全国 40 个主要城市，并分别在南京、北京、合肥、深圳建设“一总三分”运行管控中心及四个创新实验中心。实现与国内互联网骨干网络及国际试验床的互联互通，支持服务定制网络、一体化智能网络等的创新研究，满足天地一体化、空间卫星网络协议、5G 创新环境、网络攻防与评测、智能电网、大规模高清视频等多种试验验证需求。2019 年初，项目完成初步设计方案批复，正式启动建设。四家共建单位通力协作，在与中国联通、中交信通等单位的密切合作下，设施首批 12 个主干网络节点，北京、太原、郑州、西安、合肥、南京、武汉、成都、杭州、南昌、广州、深圳率先开通，在教产学研等方面的集聚成果初显。

CENI 网络的创新实验中心构建了协议无关转发试验软件平台，该平台是基于 P4 语言的体系结构来实现的。P4 语言支持对交换机处理逻辑进行编程定义，从而使得协议版本在更新迭代时无需购买新设备，只需通过控制器编程更新交换机处理逻辑即可。该试验平台设计了一种适应于网络试验平台的协议无关软件的技术方案，该方案可大大增强基于协议无关数据包处理程序在试验平台上的可用性和便利性。

试验系统可以自行生成直接基于协议无关数据包处理网络的试验环境。用户可以根据自己的需求在试验环境中装载各种 P4 应用，并通过远程访问的形式试验其功能和性能，并且用户也可以在试验环境中运行一些与 P4 应用配套的软件，比如 SDN 控制器、SQL 数据库、数据包发送接收程序以及许多流量测试工具。试验环境中也自带了许多这样配套的插件，用户可以根据需求简单配置后直接使用。这样，用户无需对众多组件进行编译和安装调试，可以直接在系统中进行带内遥测、流量控制、负载均衡、流量检测等等试验，大大降低了协议无关转发试

验的难度。

同时，试验环境也进一步降低了用户对 P4 应用进行测试仿真的时间成本和经济成本，用户无需购买任何支持 P4 的网络设备，也无需自建网络并连接众多服务器来自己搭建试验环境，只需要申请资源并远程连入即可在已经搭建好的试验平台中进行相关试验。但由于 P4 平台现在暂时不支持虚拟化，所以同时只能允许一位用户在试验平台中进行相关的试验，不过可以通过“时分复用”的方式按照不同的时间点租借给不同的用户使用，以满足多个用户同时使用的需求。

互联网数据中心

互联网数据中心（IDC）是指一种拥有完善的设备（包括高速互联网接入带宽、高性能局域网络、安全可靠的机房环境等）、专业化的管理、完善的应用服务平台。在这个平台基础上，IDC 服务商为客户提供互联网基础平台服务（服务器托管、虚拟主机、邮件缓存、虚拟邮件等）以及各种增值服务（场地的租用服务、域名系统服务、负载均衡系统、数据库系统、数据备份服务等）。

P4 其活力与价值已吸引 80 多个企业会员和 20 家高校科研机构加入开源社区，大家共同推动 P4 语言生态的建设和商业应用。在产业界，P4 创始成员之一的 Barefoot Networks 也已经在 2016 年发布了世界上第一个可编程 6.5Tbps 交换芯片“Tofino”， 之后又于 2019 年发布后续产品 12.8Tbps 交换芯片 Tofino2。在 Tofino 上运行的默认“switch.p4”程序将 Wedge 100B 交换机转换为机架顶交换机，具有数据中心所需的所有标准功能。用户可以根据自己的选择增加或删除功能、增加新协议、更改流表大小，提供更多的可视化和中间设备。同时涌现了 INT、SRv6 等一批典型网络应用，应用在数据中心等场景。

2016 年 3 月，微软正式发布了 SONiC 网络操作系统。SONiC 的所有软件功能模块都是开源的，推动了 OCP 社区以及其他厂商在开放网络方面的创新。OCP 主导的交换机抽象接口（SAI），向上为 SONiC 提供了一套统一的 API 接口，向下则对接不同的芯片设备。SONiC 大量使用了现有的开源项目和开源技术，如 Docker，Redis，Quagga 和 LLDPD 以及自动化配置工具 Ansible、Puppet 和 Chef 等，支持 P4 可编程交换机。由于 SONiC 的网络应用都是基于容器构建的，因此可以非常方便的在生产环境实现不停机部署或升级应用。

7 总结与展望

本白皮书由中国联通网络技术研究院, 网络通信与安全紫金山实验室, 北京邮电大学和 Barefoot Networks 联合编写。从 P4 语言的诞生伊始为出发点，依次介绍了协议无关可编程芯片的理念和架构，其在网络遥测、云网性能优化、5G 承载、网络自动化测试等方面的应用，以及开源和产业生态合作等相关内容，旨在为业界同仁提供协议无关交换机架构多角度全覆盖的技术与应用介绍。

未来，中国联通会一直秉持开放包容的态度，诚挚邀请所有运营商、相关开源社区、芯片设备厂商，软硬件厂商等一起探索协议无关交换机的更多应用可能性。相信基于 P4 语言的可编程芯片一定会在 5G 和 B5G 时代大放异彩，推动网络创新更上一层楼。

缩略语

缩略语	英文全称	中文释义
ACC	Approximate Cardinality Counters	近似基数计数器
API	Application Programming Interface	应用程序编程接口
ASIC	Application Specific Integrated Circuit	特殊应用集成电路
B5G	Beyond 5G	超 5G
BNG	Broadband Network Gateway	宽带网络网关
CUPS	Control and User Plane Separation	控制用户平面分离
DANOS	Disaggregated Network Operating System	分解式网络操作系统
DC	Data Center	数据中心
DCQCN	Data Center Quantized Congestion Notification	数据中心量化拥塞通知
DDoS	Distributed Denial of Service	分布式拒绝服务
DIP	Direct IP	直接 IP 地址
FPGA	Field Programmable Gate Array	现场可编程逻辑门阵列
GTP	GPRS Tunneling Protocol	GPRS 隧道协议
HPCC	High Precision Congestion Control	高速网络拥塞控制协议
HTTP	Hyper Text Transfer Protocol	超文本传输协议
IBN	Intent-Based Networking	基于意图的网络
IETF	Internet Engineering Task Force	国际互联网工程任务组
INT	In-band Network Telemetry	带内网络遥测
IR	Intermediate Representation	中间表示
LB	Load Balance	负载均衡
MAC	Media Access Control	介质访问控制
MPLS	Multi-Protocol Label Switching	多协议标签交换
NFV	Network Function Virtualization	网络功能虚拟化
NVGRE	Network Virtualization using Generic Routing Encapsulation	通用路由封装的网络虚拟化
OLT	optical line terminal	光线路终端
ONAP	Open Network Automation Platform	开放式网络自动化平台
ONOS	Open Network Operating System	开放式网络操作系统
P4	Programming Protocol-Independent Packet Processors	可编程的协议无关包处理器
PISA	Protocol Independent Switch Architecture	协议无关交换机架构
PoC	Push-to-talk over Cellular	无线一键通
QoS	Quality of Service	服务质量
SDE	Software Development Environment	软件开发环境
SDK	Software Development Kit	软件开发工具包
SDN	Software Defined Network	软件定义网络
SD-WAN	Software Defined Wide Area Network	软件定义广域网
SEBA	SDN Enabled Broadband Access	支持 SDN 的宽带接入

SERDES	SERializer/DESerializer	串行器/解串器
SONiC	Software for Open Networking in the Cloud	云端开放网络软件
SSL	Secure Sockets Layer	安全套接层
UPF	User Plane Function	用户面功能
VIP	Virtual IP	虚拟 IP 地址
VLAN	Virtual Local Area Network	虚拟局域网
VNF	Virtual Network Function	虚拟网络功能
VXLAN	Virtual Extensible Local Area Network	虚拟扩展局域网

白皮书联合编写单位

中国联合网络通信集团有限公司（简称“中国联通”）于 2009 年 1 月 6 日在原中国网通和原中国联通的基础上合并组建而成，是中国一家在纽约、香港、上海叁地同时上市的电信运营企业，连续多年入选“世界 500 强企业”。中国联通主要经营固定通信业务，移动通信业务，国内、国际通信设施服务业务，卫星国际专线业务、数据通信业务、网络接入业务和各类电信增值业务，与通信信息业务相关的系统集成业务等。作为中国联通的网络技术支撑专业机构，中国联通网络技术研究院于 2013 年 7 月正式成立。网研院自成立以来，通过聚焦网络技术，进行技术跟踪、标准预研、验证测试、规划编制、网络分析、网络测评等多项重大科研工程项目研究与实施，为集团公司网络运营发展提供了整体解决方案和全面技术支撑，加强了集团公司网络技术演进、建设、运行、优化等关键环节的技术研究和统筹规划能力，进一步提升了公司的竞争力。

网络通信与安全紫金山实验室是江苏省和南京市为了深入贯彻习近平新时代中国特色社会主义思想，打造具有全球影响力的创新名城，共同推进建设的重大科技创新平台。紫金山实验室以解决网络通信与安全领域国家重大战略需求、行业重大科技问题、产业重大瓶颈问题为使命，重点围绕未来网络、普适通信、内生安全等布局一批重大科研任务，开展基础性、前沿性研究，突破重大基础理论和关键核心技术，建设若干重大示范应用，促进成果在国家经济和国防建设中的落地，引领全球信息通信技术发展，建设世界一流水平的国家战略性科技创新基地，为建设世界科技强国提供强大战略支撑。

北京邮电大学（Beijing University of Posts and Telecommunications），简称北邮，位于北京市海淀区西土城路十号，是中华人民共和国教育部直属，工业和信息化部共建的一所以信息科技为特色，工学门类为主体，管理学、文学、理学等多个学科门类协调发展的全国重点大学，是北京高科大学联盟成员高校。系国家“211 工程”、“985 工程优势学科创新平台”项目重点建设，列入首批“卓越计划”、“111 计划”。被誉为“中国信息科技人才的摇篮”。

Barefoot Networks 诞生于 2016 年，总部设立在美国硅谷（已于 2019 年 7 月被 Intel 收购），在正式成立之前该公司的创始人潜心两年致力于研发一种全新的软件定义交换机技术，以期在不影响性能的前提下实现对转发平面的控制。Barefoot 赋予了网络所有者及其合作伙伴以设计、优化和创新的能力，使之能够自行按需定义网络，以此来获得更大的竞争优势。在实现 P4 编程语言与快速可编程交换机相结合的过程中，Barefoot 还为编译器、工具以及 P4 程序创建了一个良好的生态系统，使得 P4 可以为任何人所用。更多详细信息，请访问：https://barefootnetworks.com/。