Nature Flow：新转发架构赋能未来数据中心网络

Nature Flow: A New Forwarding Architecture Improves Future Data Center Network

摘要：提出一种基于端口地势值比较的数据转发新技术——Nature Flow。该技术不仅能有效确保二层数据无环路转发，而且能提升数据中心网络开放能力。新转发架构的价值在于构建大规模二层拓扑存环网络的无环转发能力、对应用程序开放网络端到端的距离感知能力、网络故障快速收敛和自愈能力、网络拥塞时的流量自主调优能力等。新转发架构有望变革现有技术，助力未来数据中心网络建设。

关键词：Nature Flow；端口地势值比较；无环路转发；自愈能力；端到端距离感知；流量自主调优

Abstract: A new data forwarding technology Nature Flow based on the comparison of port terrain values is proposed, which can effectively ensure the no-loop forwarding of layer 2 data and improve the network opening ability of the data center. The value of the new forwarding architecture lies in the following aspects: the acyclic forwarding ability of large-scale two-layer network with topological rings, the end-to-end distance perception ability for the application in open network, the rapid convergence and self-healing ability of network failure, and self-tuning ability of traffic in network congestion. The new forwarding architecture is expected to change the existing technology and help the future data center network construction.

Keywords: Nature Flow; port terrain value comparison; no-loop forwarding; self-healing ability; end-to-end distance sensing; traffic self-tuning

商志彪 /SHANG Zhibiao[1]
雷波 /LEI Bo[2]
郭茜 /GUO Xi[3,4]
（1. 中兴通讯股份有限公司，中国 深圳 518057；
2. 中国电信股份有限公司研究院，中国 北京 102209；
3. 北京科技大学，中国 北京 100083；
4. 北京市材料知识工程重点实验室，中国 北京 100083）
(1. ZTE Corporation, Shenzhen 518057, China;
2. Research Institute of China Telecommunications Corporation, Beijing 102209, China;
3. University of Science and Technology Beijing, Beijing 100083, China;
4. Beijing Key Laboratory of Knowledge Engineering for Materials, Beijing 100083, China)

DOI：10.12142/ZTETJ.202104012
网络出版地址：https://kns.cnki.net/kcms/detail/34.1228.tn.20210421.1751.002.html

网络出版日期：2021-04-22
收稿日期：2020-09-26

1. 新转发架构 Nature Flow 概述

1.1 传统 L2 转发原理

总体上，数据中心需要一个二层网络。传统的二层转发通过报文中的目的媒体接入控制（DMAC）、虚拟局域网（VLAN）、PORT（计算机端口）信息查表来确定报文的出端口，并完成源媒体接入控制（SMAC）的学习。除了出端口信息外，所查表项结果几乎不包含其他可有效指导网络报文转发的全局性信息。二层数据报文一旦遇到拓扑环路，将会造成“环路风暴”，整个报文转发系统将面临崩溃。

1.2 Nature Flow 转发架构

Nature Flow 转发架构是一种全新的转发与控制体系。它在每个转发设备端口上设置逻辑地势值，并在报文转发时通过比较该值来判断和选择转发出口。Nature Flow 可实现二层数据包无环路转发。这种在转发中去除环路的方式与现有的生成树协议（STP）[1]完全不同。STP 的目标是建立拓扑无环网络，而新转发架构的目标是在拓扑存环的网络中完成无环路的二层数据转发。拓扑环路可以有效提升整体网络的可靠性，而转发环路的存在是导致网络中出现“环路风暴”的根本原因。

三层路由转发可实现数据流在拓扑存环网络上的无环路转发。Nature Flow 转发架构也是一种新型的二层路由转发协议，在一定程度上可以通过对现有地址解析协议（ARP）/ 邻居发现协议（ND）等的改造来初步实现。

Nature Flow 转发体系的构建大致包括两个阶段：

（1）分布式地势图的构建

媒体接入控制（MAC）地址的拥有端通过一种全新方式向整个网络发布该 MAC 地址的网络转发地势值，该值被记录在途经的每个网络设备的每个端口上。针对固定的某个端口，该值等于端口到 MAC 地址所需经历的端

到端的网络距离。当该 MAC 地址的地势值在全网发布完成时，一个类似“等高线”形态的全网络的分布式地势值分布图将会形成。

（2）数据转发依地势高低进行自然流动

在地势分布构建完成之后，转发规则的设计变得非常简单。参照自然界中水自然流动的原理，建立只允许数据报文从高地势值向更低地势值的转发规则，以确保每转发一次的地势值都比之前路径节点的地势值低。在这种条件的约束下，整个转发路径中的环路就不会形成。如果转发设备中存在多条更低地势的转发路径，就选取对应地势最低的那一条。

1.3 新概念：端口地势值和全播过程

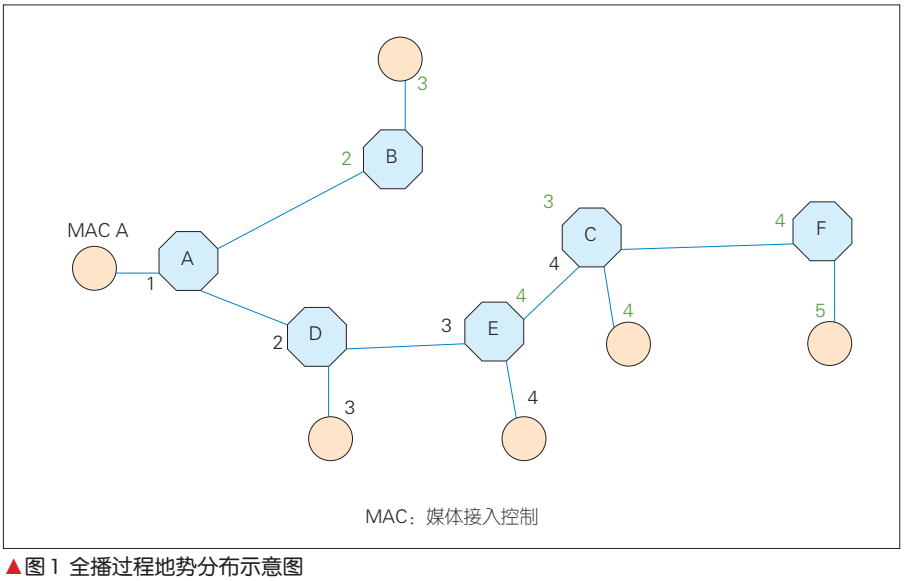
地势值是 Nature Flow 转发架构下的新概念。设备的每个转发端口都具有一个或者多个针对某个目的 MAC 的地势值。该值被记录在 MAC 查找结果表中，并在转发时被用来比较权衡。端口地势值记录的是从该端口到达特定目的 MAC 优选路径上的全部链路在某一种链路属性上的累加和。该累加和同时也表示，在某种属性下，该端口到达目的 MAC 的网络距离。以跳数（HOP）作为链路属性为例，假设某端口到达目的 MAC 的累加和为 3，则从该端口转发数据包到目的 MAC 接收端共有 3 跳的网络距离。在新转发架构下，链路属性有多种，如 HOP、时延、可靠性等。该属性需要具有随链路增加而累加增大的特征。同一个端口对不同目的 MAC 会有不同的地势值。整个网络中每个设备的每个端口针对相同目的 MAC 也会有不同的地势值。这是新架构与传统 MAC 表中数据结构最大的不同，也是新转发架构得以实现更高网络能力的基础。

全播是 Nature Flow 转发架构下构建网络地势分布的过程，也是在地势转发网络中对传统广播过程的有效替代。全播的发起者是 MAC 地址的拥有者，也是设备转发表项中的目的 MAC 端。发起者通过全播过程在全网中建立 MAC 地址的地势值分布。与广播和组播过程不同，在传播过程中全播会在报文中携带地势值，并且在每次设备转发时修正报文中的地势值。假设针对某个转发系统，即在同一个 VLAN 或者虚拟网络标识（VNI）内，存在 A、B、C、D 共 4 个端口。由 A 口收到地势值为 1 的某个 MAC 地址的全播报文，在转发时会向 B、C、D 口转发地势值为 2 的全播报文。上述转发行为是以 HOP 为地势的参考属性。针对来自相同 MAC 地址的全播报文，中间转发设备会自主记录来自不同入端口的地势值，并只会向远端传播当下最小地势值的全播报文（其他地势值对应的路径均作为本地备份路径）。这种传播方式可以有效地减少报文的传播次数，同时也避免了报文的环路传播。

以 HOP 为链路属性参考，全播过程使网络中每个端口均记录到达该 MAC 的最优“生存时间（TTL）值”。同时这个传播过程是随时可以扩展的，更适合链路的动态变化的场景。如图 1 所示，该图以 MAC 节点 A 完成基于跳数的全播过程来说明整个转发控制过程。

在 Nature Flow 转发系统中，假设网络中存在末端系统（ES）节点和中间系统（IS）节点。图 1 中橙色的圆点表示 ES 节点，即具有 MAC 地址的实体，是全播的发起者。ES 节点对应每个应用程序（APP）或者主机的 MAC 地址，具有转发表项和协议栈能力，同时也是全播的终结点。图 1 中蓝色八边形表示 IS 节点，即整个网络中的数据转发设备，如交换机或者路由器，是全播的转发节点和地势值累加节点。

每个 ES 节点需要向全网全播自己的 MAC 地址，以使得网络中其他节点获得相应的的网络距离和出端口信息。反之，在接收其他节点发送过来的全播报文时，ES 节点也获得去往该节点的出口信息和网络距离信息（记录在全播报文中的经过无数次累加之后的地势值）。作为全播报文的发起者，ES 节点发送出去的初始全播报文的地势值是最低的。以 HOP 为例，如果在



▲图 1 全播过程地势分布示意图

全播报文中设置 HOP=1，那么以后每被设备转发一次则加 1。

假设图 1 中所有的节点和路径都处于同一个转发系统（即 VLAN）中，以 IS 节点 A 为例，当 IS 收到 HOP=1 的来自 MAC A 的全播地势信息时，转发系统内有 3 个端口：连接 MAC A 的端口、连接节点 B 的端口、连接节点 D 的端口。根据水平分割原理，出端口为节点 B 和节点 D 的方向。网络设备在向下转发全播时，需要在报文的当前地势值中累积增加从节点 B 到节点 A（或者从节点 D 到节点 A）的地势差值，然后将修改后的全播报文发送给节点 B 和节点 D。图 1 中的数字表示的是，在整个全播过程完成后各个节点的端口以跳数为参考的地势值分布，其中节点 C 和节点 E 都会收到两个地势值。然而，设备只把最低的地势值 3 累加后向外传播，并将更大的地势值作为本地备份链路使用。同一个 MAC 地址通过全播的方式不断扩散，并在整个网络中形成一种类似“等高线”的地势分布。该地势分布为反方向的数据转发以提供路径指导。图 1 的拓扑结构存在环路。节点 C 和节点 E 可以同时存在两个转发地势值，并形成转发出口的主备关系。不同的出口对应不同的到达 MAC A 的路径（图 1 中我们以绿色字体和黑色字体进行区别）。

1.4 Nature Flow 架构下的设备转发规则

Nature Flow 转发系统与传统的转发规则完全不同。传统转发规则中的 MAC 结果表中不记录地势值信息，只记录出接口信息。传统转发规则只能查找到特定的出口，进而完成数据的转发，并不适应网络的拓扑变化。Nature Flow 通过全播建立基于自身 MAC 的整网出端口地势值分布，使得转发数据流量可以像自然界中水流一样在整个基于地势分布的网络内流动。这也是新的转发架构被命名为 Nature Flow 的原因。

假设存在网络转发设备 M，从 A 端口进入的目的地址为 MAC X，转发系统为 VLAN Y 数据流量，那么设备转发规则为：

（1）FIND 端口组 {O} IN VLAN Y where DMAC=MAC X 且端口 i 的地势小于端口 A 的地势；

（2）最优出口 i=MIN{端口 i 地势} where i in 端口组 {O}。

依据转发规则，系统在第 1 步寻找全部可用的无环路转发端口组，在第 2 步寻找端口组中最优转发路径的出端口，以实现到达目标节点网络距离最小的出口路径转发。在链路发生变化时，这种转发方式可以有更多的转发路径选择，并具有更高的鲁棒性。除此之外，第 2 步的最优出口策略也可以进行调节。比如，在发生出口拥塞时，如果所转发的报文没有保序要求，那么第 2 步就可变更为寻找最大剩余带宽的路径出口，以更好地自主规避网络拥塞。

2 Nature Flow 架构网络的潜在应用价值

2.1 大规模网络二层数据转发中去除环路的能力

在数据网络中，无论是三层路由协议还是二层转发都面临环路转发问题。以开放式最短路径优先（OSPF）[2] 和边界网关协议（BGP）[3] 为例，OSPF 区域内通过最短路径优先（SPF）算法实现无环路路由，OSPF 区域间通过强制与骨干区域连接实现去环。外部边界网关协议（EBGP）通过自治区域路径信息（AS-PATH）属性的序列检查来实现防环，内部边界网关协议（IBGP）通过限制路由学习来实现无环。Nature Flow 以基础的二层转发架构为起点，它去除环路的原理主要通过转发地势值的持续递减来实现，即数据流的每次转发行为都会使该地势值降低一次。这样地势值就不可能回到原来的高度，也就无法形成闭合的网络转发环路。该技术打破当前二层网络必须部署在树形网络拓扑上的限制，可以实现规模更大、拓扑更加复杂的二层网络。在超大规模数据中心组网实践中，基于距离向量的路由算法具有更小的网络状态同步需求，并逐渐在诸如 FaceBook 设计的 F4[4] 和 F16 数据中心 Fabric 架构下使用。同时为了方便大规模网络的运维和管理，超大规模数据中心更倾向于使用单一的路由协议 [5]。Nature Flow 能够很好地满足上述条件。当二层网络不再受广播风暴、规模等问题限制时，数据中心网络的发展将迎来新的机遇。

2.2 应用程序对网络端到端距离的感知能力

在现有的数据中心网络系统中，信息技术（IT）系统负责发送和接收数据报文，通信技术（CT）系统负责转发数据报文。然而，IT 系统和 CT 系统之间的深度交互却是有限的。这给整个系统业务的故障定位带来很大的困难，比如涉及业务软件系统的传输控制协议（TCP）时间超时等故障问题。

在 Nature Flow 转发系统中，每个 ES 节点都会记录目的 MAC 端的累积地势值，即到达目的 MAC 的网络距离。如图 2 所示，以跳数为例，假设数据通信发生在 MAC A 和 MAC B 之间，那么在故障发生之前，MAC A 设备上记录 MAC B 的以 HOP 为参考的地势值为 3。当故障发生后，网络会动态收敛并更新地势值，MAC A 设备上记录 MAC B 的以 HOP 为参考的地

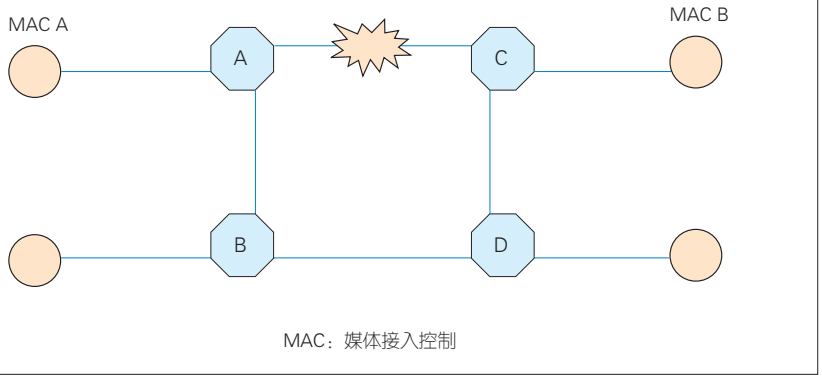
势值会变成5。此时，MAC A设备可以通过这种变化感知到网络状态也发生了变化。如果HOP变成5之后不可接受，那么应用软件可以更灵活地进行判别与处理。在新架构下，应用程序可通过对本地目的MAC表中的地势值的查询来发现网络的变化，以提升应用程序对网络的感知能力。与此同时，网络也向应用开放端到端的网络距离感知能力。在传统的网络中，应用无法感知网络的变化和具体状态，只能通过应用层的超时异常来报告网络故障。类似Ping的运维手段也无法反映图2中的网络变化过程差异。新转发架构的这一开放能力，将在云计算与网络技术的融合中带来巨大的商业价值。IT系统可以查询到达目标网络的端到端距离，可更好地感知网络的变化，从而更好地规划如何使用网络来打造更优质的云平台，实现网端入云和云端知网的信息通信技术（ICT）融合，助力运营商打造更优质、更开放的网络新平台。以当前的内容分发网络（CDN）业务为例，新转发架构可以使业务通过判断不同缓存节点到达目标互联网协议（IP）的网络距离来选择最近的缓存节点，而不是只能依靠固定物理地址与固定IP的对应关系来计算远近距离。其中，后者只是算法的能力，而不是网络的开放能力。

2.3 网络弹缩的快速收敛及自愈能力

任何转发和路由算法都需要面对网络中设备和链路的动态增删。在动态增删过程中，快速收敛特性是整个算法的核心优势。在Nature Flow转发系统中，链路的增删会带来局部多个MAC地址的地势值变化，并需要触发对前期全播过程的扩展。当整个扩展的全播过程完成时，新的转发地势分布就会形成，整个网络的快速收敛也将实现。

网络的变动情况大体上分为两类：网络链路的增加和网络链路的删除。网络中整台网络设备的增减可以映射为多条网络链路的变动。

在Nature Flow系统中，如果增加新的链路，新链路两端的节点设备在感知到网络发生变化后，会针对本地MAC表中具有相同转发系统（即VLAN）的条目，在新链路上启动新的地势分布的全播。该过程不仅实现向新增链路的两个端口发送本地最优地势值的全播报文，还实现新增接口针对转发系统内全部MAC的地势值的分布。此外，如果出现新增链路接口的地势值低于设备原有地势值的情况，就需要把新的最小的地势值继续通过同时后者仅粗略地估计距离，在时延方面的应用比较有限。



▲图2 末端系统节点对目标节点的地势感知示意图

全播的方式向远端传递。

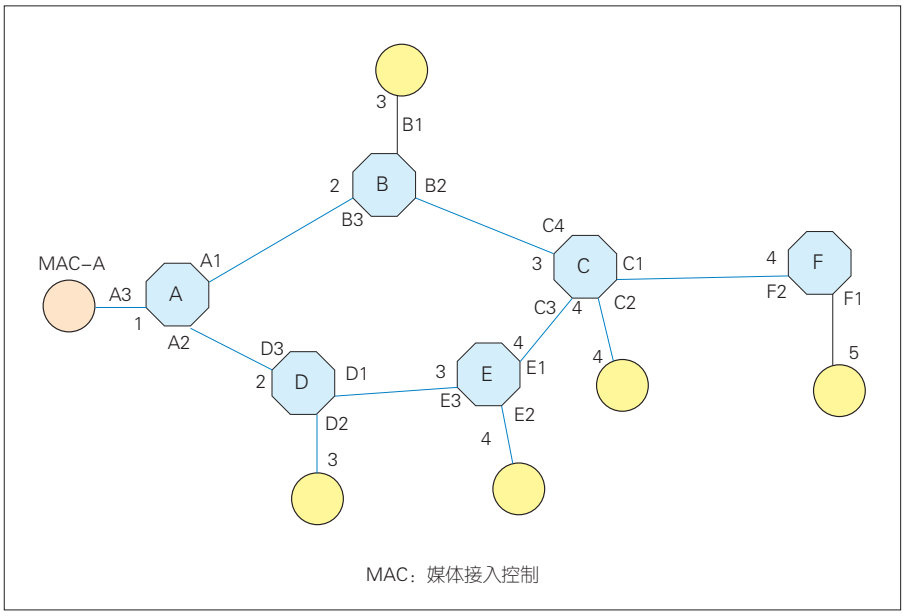
在Nature Flow系统中，假设原有链路可被删除，包括链路故障或者节点故障等情况。在删除前的全播过程中，如果该链路作为最优路径被选中，则需要向原来的该链路全播方向发送一种全新的全播链路删除报文，以告知整个路径中的节点删除转发表项中早期通告过的针对某个MAC地址的地势值，并重新选择最小的地势值路径。全播链路删除报文需要扩散至整个故障链路以下的全部网络节点和主机节点。如果在前期的全播传播中，被删除的链路只作为备份路径使用，那么只需要在MAC转发表中删除原有的备份表项，同时通过全播扩展过程只在备份链路上通告删除备份路径的相关表项。

在整体算法设计上，链路的动态增删只涉及原有全播过程的扩展和修正。全播报文传递完成意味着对应的网络收敛过程的完成。与传统网络中的双向转发检测（BFD）和快速重路由（FRR）过程相比，新的转发框架可以有效实现网络的自愈，能够更好地应对网络的故障收敛。

Nature Flow转发框架是为未来数据中心动态网络而设计的。如果网络中的某条链路属性发生变化，整个网络中基于该属性的地势分布的变化也可能会被触发。通常情况下，以HOP为参考的属性不易发生变化，可以作为Nature Flow的基础属性。然而，以时延为参考的链路属性却常常是动态变化的。如果某一条链路时延属性的变动超出一定范围，就需要通过全播来重新发送到全网。该实现过程与链路的增删类似。如果某条链路的时延属性变大，那么需要删除原来的低时延全播通告，同时完成新的更大时延的全播通告；如果某条链路的时延属性变小，就需要删除原来较大的时延

全播通告，同时完成新的较小时延的全播通告。

如图 3 所示，我们对每个 IS 设备的端口进行详细命名。命名规则为：以向上为 1 开始，并沿顺时针方向递增。在全播过程完成后，各设备的转发表项状态如表 1 所示。转发设备可以根据目的 MAC、转发系统、入端口来查找整个表项，并找到有效的出端口。例如，当主用出口发生故障时，C1 和 C2 这类具有备用出口的转发就需要网络设备感知到故障，并删除当前的最优路径。主路径删除后，再次查询时备用路径就会被作为最优的转发出接口，完全不需要 FRR 过程和拓扑无环备份（LFA）保护，这具有在大规模数据中心网络中应用的潜力。



▲图 3 网络中各设备节点接口分布

▼表 1 全播过程完成后，各设备的转发表项状态

目的节点	转发设备	出端口	出口地势	入端口	转发系统
MAC-A	SW-C	C4	3	C1	VLAN1
MAC-A	SW-C	C4	3	C2	VLAN1
MAC-A	SW-C	C4	3	C3	VLAN1
MAC-A	SW-C	C3	4	C4	VLAN1
MAC-A	SW-C	C3	4	C1	VLAN1
MAC-A	SW-C	C3	4	C2	VLAN1

MAC-A：媒体接入控制 A　SW-C：转发设备 C　VLAN：虚拟局域网

2.4 路径资源占用的拥塞调度与自主调优能力

针对数据中心网络，网络流量模型具有一定的突发性和动态性特征[6]。网络链路利用率的不均衡调节和拥塞控制调节就变得极为重要。在高性能计算（HPC）网络和分布式存储网络中，应用对丢包极为敏感。例如，1% 的丢包可能会造成极为重大的性能损失。这就需要网络在整体上构建路径拥塞控制调节机制，以尽可能减少网络拥塞造成的丢包[7]。继 DCQCN（数据中心 QCN）之后，2019 年阿里巴巴集团提出高精度拥塞控制（HPCC）算法[8]，通过在 TCP 协议的确认字符（ACK）报文中增加拥塞控制标识来完成有效的端点发送流量控制，并通过有效调节应用发送端的流量大小来规避进一步的拥塞。但是对于网络内部由其他流量对共享链路资源的争用所带来的拥塞，仅通过该算法在发送端调节流量是不能彻底地解决这一问题的。本文所提出的 Nature Flow 算法可以有效地解决此类链路拥塞问题，并实现对拥塞控制的自主优化调度。

Nature Flow 架构的转发设计在拥塞控制方面具有如下可行方案：

（1）当整体网络中存在软件定义网络（SDN）控制器等全局统一管控平台时，如果某条链路发生拥塞，就可以针对该链路中占比较大的拥塞流量，调高该端口针对该条流量的地势值。该地势值的变动会重新触发整个网络的地势变化和局部重新选路，使得部分流量绕行拥塞链路。

（2）如果在最优路径转发时仍然出现链路拥塞，Nature Flow 则有能力调节转发选路策略。比如，针对无严格保序要求的报文，如用户数据报协议（UDP）报文，Nature Flow 不再按照最优路径转发，而是在全部无环转发路径组中选择当前剩余带宽最大的路径，以避免进一步加剧拥塞。

（3）当某条链路发生拥塞时，基于 Nature Flow 的转发架构具有链路增删的快速收敛能力，可以在拥塞链路的局部增加对应转发系统的链路。其他转发系统的链路也可以被临时借用到拥塞链路的流量转发上，并在拥塞解除后被重新还原，以实现网络架构对拥塞的动态应对。

与传统路由协议和二层转发相比，Nature Flow 转发架构在全网络所有设备的 MAC 表项中分布式地记录网络距离（即地势值）的全局性信息。相比于当前的链路状态算法，如 OSPF 和

ISIS 等，该架构使用全局性信息来指导网络流量转发，具有更优的网络动态适应性。

Nature Flow 在网络流量工程调节方面也具有综合优势，尤其是在与未来网络 SDN 控制器及人工智能（AI）技术的结合方面。全播过程使每个 MAC 地址都有一张网络地势分布图，可以有效指导网络路径转发，规避无环路和拥塞。更重要的是，通过 SDN 控制器或者 AI 技术来优化和调节这些地势值，可以实现对整个网络流量的精准调度与控制。

3 对新转发架构的思考

3.1 Nature Flow 新架构给现有设备带来的改变

新转发架构改变了整个二层 MAC 数据流的转发规则，给整个网络能力的开放带来新的机会与挑战。新架构可以解决当前网络所面临的诸多难题，但同时也对转发设备提出新的要求。Nature Flow 新转发架构的实现会给网络设备带来如下需求：

（1）MAC 转发表项的数据结构变化

Nature Flow 转发架构改变了底层目的 MAC 转发表的数据结构，在 MAC 转发表中增加了一个或多个基于属性的地势值。这种改变增加了 MAC 转发表项的大小，但并未增加 MAC 表项的条目需求。MAC 表项的条目增加仅仅是备份链路的增加，它可以解决传统 MAC 飘逸等带来的相关问题。MAC 学习和 MAC 老化都是由整个全播过程来完成的。在当前的网络协议中增加全播能力并不是一件困难的事情。具体的全播过程可以在现有网络上通过免费 ARP 等相关技术的改造来实现。

（2）设备转发逻辑和算法的创新

基于地势的全播过程创新地打造一组针对目的 MAC 地址的无环路转发路径。相关路径信息被分布式地记录在设备的转发表项中。由于有地势值的指导，Nature Flow 转发逻辑路径选择的空间更大，优选路径的策略更多，可以实现更高效的数据流量工程能力。此外，与实现 IP 路由的参数化模块库（LPM）查找类似，网络设备也需要比原来传统转发逻辑更加复杂的算法。新的转发逻辑虽然可以在纯软件的基础上实现，比如将地势转发逻辑构建在基于软件的 MAC 路由信息表（RIB）中，真实的报文转发依然由传统的转发芯片来承担。更近一步地，如果能够在芯片层面实现对新的转发架构逻辑的支持，就有可能打造出新的数通转发设备，如白盒设备等。

（3）全播报文的控制与对账

新转发架构建立在整个网络的全播过程上，取代了传统的泛洪式转发。由于需要建立高效的全播地势分布，整个网络中全播流量的带宽需求会比传统网络有所提升。在某些高动态网络中，新增链路和删除链路带来的全播流量会增加。当然，这种增加是相对于传统转发环境而言的。如果考虑整个网络接口的带宽，那么从最早的 1 GE 增长到目前的 10 GE 和 25 GE，带宽需求的占比可能并未增加。在理论上，如果需要构建一个高效的动态管理路由网络，控制层面的流量与接口带宽的比例必须是合理的，以避免 1 GE 带宽的网络和 10 GE 带宽的网络使用同样带宽（如 500 Mbit/s）的管理和控制流量。全播过程的安全控制最好由 SDN 控制器来完成。控制器是全局信息的拥有者，完全可以实现对整个网络地势分布的实时控制和一致性对账，并提供更高的网络稳定性和一致性，进而打造软件可控的未来数据中心网络。

3.2 Nature Flow 应用场景与未来目标

新的转发架构更适用于数据中心的超大规模组网，能够实现网络规模和链路的动态弹缩。在与云计算技术融合方面，新的转发架构可以把网络的端到端基础能力开放给软件应用，使得软件程序在通信发起时可初步预测“信息”被送达的情况，比如需要多少跳网络，或者需要多少时延等。这种开放能力不仅有助于提升软件应用感知能力和应用网络平台能力，还能提高网络运维和排障效率。网络转发端到端能力的开放更适合打造面向未来的确定性网络。与当前应用程序需要网络具有端到端确定性保障不同，端到端能力开放（或可感知网络）把整体网络视作一个动态过程，并由 IT 软件的应用程序来判断网络的确定性。例如，当信息在 3 跳之内或者 3 s 之内可达时，成功的概率在 90% 以上。不同于当前的基于报文复制和副本消除的确定性网络解决方案，新的转发架构把网络基础能力的选择权交给应用端，同时网络本身只致力于提供更低时延、更大带宽等技术指标。该转发架构更适用于网络分片技术和网络流量工程的精细化管控。当 MAC 表多记录一种不同链路属性的地势值时，整个转发层就会提供一种基于该链路属性的分片转发能力。在 5G 的切片转发应用中，带宽敏感流量可以通过基于 HOP 的地势转发实现，时间敏感流量可以在基于时延的地势转发中实现。针对同一个物理网络、同一个目标地址，当应用所需要的网络指标不同时，支持 Nature Flow 的转发系统可以实现不同路径的路由转发处理。

新转发架构的最大贡献在于从根本上解决了网络环路转发的问题。虽然在极端情况下仍然存在环路的可能，但是当新转发架构配合 SDN 控制器构建整个网络时，通过控制器层面的

基于全局算法的防环路补充机制，可以彻底地解决环路问题。虽然新转发架构的目标在于为二层网络设计，但是其防环路的原理完全可以被其他三层路由协议所借鉴，如路由信息协议（RIP）等。由于在控制和转发之间只使用全局分布式的地势值，新转发架构更适合在 SDN 控制器上引入 AI 算法，也更适合作为未来白盒设备的基础转发规则，同时还可以对原生 SDN[9] 系统的 Openflow 流表做更深入的改进。

Nature Flow 转发架构是一种新的转发与控制的框架体系。相比于传统转发架构，Nature Flow 可实现设备整体转发规则的高度统一和全网分布式差异化地势的分布，通过分布式的设备算力降低整个网络中 SDN 控制器的负担[10]，可以打造更大规模、更精准的流量控制数据中心网络，具有变革当前数据中心网络的潜力。

在某种程度上，Nature Flow 是一种基于 MAC 的二层路由内部网关协议（IGP）算法，可实现对单播路由的无环路计算，并在新框架中使用全播来替代传统的广播转发。Nature Flow 的组播或可通过配合最新的 BIER（基于比特索引的显示组播复制）协议来实现。引入 Nature Flow 会给数据中心网络带来新的变化和需求，比如对带内遥测技术（INT）的需求。Nature Flow 需要 INT 来获取每条链路的不同维度的属性值，如丢包、时延等。此外，Nature Flow 可以实现高效的网络自愈和流量自主调优。在这种情况下，数据转发路径相对不完全固定，这对数据中心的运维和排障能力提出新的要求，对流量可视化、历史流量转发路径确认等的需求更为迫切。此外，当前数据中心大多使用基于 Overlay 的虚拟扩展局域网（VXLAN）等相关技术。Nature Flow 与 VXLAN 的结合必然会在 Overlay 层实现这使得 Overlay 层的应用程序可以感知到到达通信对端的“网络距离”，从而把网络层的基础能力开放给平台层和应用层，有助于实现 ICT 技术的深度融合与综合提升。诸如 VXLAN、SRv6 等 Overlay 技术本质上是基于隧道实现的远程连接。Nature Flow 虽然在设计时是将链路作为承载流量的基本元素，但是完全可以平滑扩展到向支持链路一样来支持隧道。该方法把隧道看成一种基于 Overlay 的特殊链路，实现了与当前数据中心主流协议的结合。

4 结束语

Nature Flow 转控架构基于自然界水流的无环路流淌，为每个设备的端口引入地势值的新概念。网络转发路径的构建过程以创新的全播过程来实现，转发出口选择转换为对应出口地势值的比较结果。新架构转发表中记录的地势值，为应用程序对网络的端到端距离感知提供基础能力，也为网络动态变化时路由快速收敛和拥塞控制提供指导和支撑。新架构能够有效提高现有网络的基础能力，更适用于未来大规模高动态数据中心网络的建设。我们希望产业界、学术界的研究者能够关注 Nature Flow 这一新技术，对其做进一步研究，以解决当前网络所面临的诸多问题，进而推动未来网络的变革。

参考文献

[1] IEEE. Local and metropolitan area networks: media access control (MAC) bridges: 802.1D–2004 [S]. 2004
[2] IETF. OSPF version 2: RFC 2328 [S]. 1998
[3] IETF. A border gateway protocol 4 (BGP–4): RFC 4271 [S]. 2006
[4] 马绍文 . 超大规模云网络数据中心创新 [EB/OL]. (2020–04–21)[2021–04–10]. https://www.sdn–lab.com/24039.html
[5] IETF. Use of BGP for routing in large–scale data centers: RFC 7938 [S]. 2016
[6] ROY A, ZENG H, BAGGA J, et al. Inside the social network’s (datacenter) network [C]//Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication. New York, NY, USA: ACM, 2015: 123–137
[7] 王江龙，雷波，解云鹏，等 . 云网一体化数据中心网络关键技术 [J]. 电信技术，2020, 36(4): 125–135
[8] LI Y, MIAO R, ZHANG M. Hgh precision congestion control [EB/OL]. [2021–04–10]. https://dl.acm.org/doi/pdf/10.1145/3341302.3342085
[9] MCKEOWN N, ANDERSON T, BALAKRISHNAN H, et al. OpenFlow: enabling innovation in campus networks [EB/OL]. (2008–04)[2021–04–10]. http://www.sigcomm.org/node/2683
[10] 邬贺铨 . 关于 5G 的十点思考 [J]. 中兴通讯技术，2020, 26(1): 2–4. DOI: 10.12142/ZTETJ.202001002

作 者 简 介

商志彪，中兴通讯股份有限公司运营商市场数据中心网络方案总工；曾从事网络处理器芯片开发工作，现致力于运营商 5GC NFV 云、SDN IT 云、云网融合方案，以及数据中心场景新技术的研究；获发明专利 5 项。

雷波，中国电信股份有限公司研究院未来网络研究中心主任，边缘计算产业联盟 ECNI 工作组联席主席、CCSA“网络 5.0 技术标准推进委员会”管理与运营组组长；主要研究方向为未来网络架构、新型 IP 网络技术等；发表论文数十篇，出版图书《边缘计算与算力网络》和《边缘计算 2.0：网络架构与技术体系》。

郭茜，北京科技大学计算机与通信工程学院副教授，现担任中国计算机学会（CCF）北京科技大学会员代表、数据库专委会通信委员；研究方向为数据查询处理、信息安全等；曾主持多项国家自然科学青年基金和校企合作项目；发表论文近 30 篇。