案例研究

第二代英特尔® 至强® 可扩展处理器
面向英特尔® 架构优化的 TensorFlow
智能网络
人工智能

至强® 平台集成 AI 加速 构建数据中心智慧网络

H3C 引入第二代英特尔® 至强® 可扩展处理器，提升先知网络架构的 AI 训练与推理性能，优化企业网络的智能分析和业务编排能力

H3C
数字化解决方案领导者

“SNA 通过 AI 方法来实时感知网络状态，基于网络数据分析来实现自动化部署和风险预测，从而让企业网络能更智能、更高效地为最终用户业务提供支撑。通过引入第二代英特尔® 至强® 可扩展处理器以及面向英特尔® 架构优化的 TensorFlow，SNA 的 AI 训练能力获得了大幅提升，让企业网络在应对复杂业务场景时更加游刃有余。”

敖襄桥
院长
华三 AI 研究院

概述

软件定义网络 (Software Defined Network，SDN)得益于以自动化方式对网络资源实施灵活调配的能力，已成为企业级用户部署和配置网络服务的重要选择。然而随着企业级网络应用规模的不断扩展，特别是在大规模云数据中心这种复杂度高、调整频繁的场景中，即便拥有自动化辅助手段，用户的运维和成本压力也很难得到真正缓解。

如何帮助企业级用户应对这一挑战，将网络的管理和编排从自动化进一步推向智能化？致力于企业级网络解决方案创新的新华三集团 (以下简称“H3C”)，就依托深厚的网络设备研发、制造与部署经验，推出了更具智能化属性的先知网络架构 (Seer Network Architecture，以下简称 SNA)。该架构可通过“感知-分析-决策”的模式，将丰富的网络运维数据通过人工智能 (Artificial Intelligence，AI) 的训练和推理，转化为更优的网络策略，来帮助最终用户有效提升网络智能分析和业务编排能力，同时降低运维成本。

作为 H3C 重要合作伙伴的英特尔公司，不仅提供了第二代英特尔® 至强® 可扩展处理器为 SNA 输出强大的算力支持，更结合 SNA 在不同应用场景中的需求，为其 AI 训练和推理过程提供多种优化方案。来自验证性测试的结果表明：基于英特尔® 至强® 平台集成的 AI 加速能力优化后的 SNA，可大幅提升 AI 训练能力，并在最终用户的实际部署中赢得了良好的反馈。

H3C SNA 实现的解决方案优势：

- SNA 可通过“感知-分析-决策”模式，将丰富的网络运维数据通过 AI 方法转化为自动化网络策略，目前已能提供 20 余种智能网络算法和 100 多个网络状态洞察方法；
- 英特尔® 至强® 平台集成的 AI 加速能力，包括第二代英特尔® 至强® 可扩展处理器提供的更强算力，以及与其搭配的、面向英特尔® 架构优化的 TensorFlow，可助力 SNA 大幅提升 AI 训练能力，例如可使 DNS 隧道检测模型的训练性能提升至基准值的 3.2 倍[1]；
- 英特尔® 至强® 平台集成的AI加速能力，也通过 H3C 与英特尔的合作，开始惠及其他 ICT 设备的智能化管理，例如可将服务器利用率模型的推理性能提升至基准值的10.98 倍[2]。

信息化时代的网络服务犹如水、电、煤一般，是生产生活中不可或缺的基本元素。为了向行业用户提供更为便捷、高效的网络服务，网络的建设、部署、管理和运维理念也须不断推陈出新。近年来，以 SDN 为代表的下一代网络技术，通过控制面与数据面分离的方式，让网络具备了更高的敏捷性、可扩展和可编程能力，大幅提升了网络自动化水平，并能有效降低用户在网络部署和运维方面的压力。

不过，随着用户业务与网络服务的绑定更为紧密，在许多复杂、多变和多样化的应用场景中，单一的自动化部署、运维能力已不能满足业务的需求。以企业园区为例，网络服务不仅要满足日常办公应用，还需为生产制造、移动办公、视频会议等不同应用、不同质量要求的场景提供支撑。在大规模的云数据中心内，网络也需要像服务器和存储设备一样，随时根据承载的业务、应用和数据的变化来切换自身的状态和配置。因此，网络亟需从自动化进一步向智能化演进，以更为灵活机动地支持业务运营。

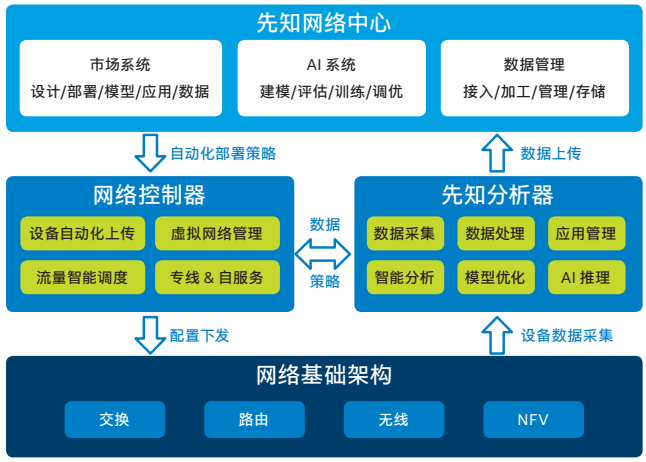
虽然各类网络设备已经可为智能运维提供丰沛的数据，但企业级智能网络的落地依然困难重重。究其原因，首先是缺乏完善的“数据分析-决策反馈”机制，其次是没有统一的 AI 平台来将数据训练转化为决策模型。为应对这一挑战，H3C 与英特尔开展深度技术合作，利用英特尔先进产品与技术，构建了 SNA 这一智慧网络架构，以 AI 方法将网络数据转化为更优的网络策略，实现了降本增效，且助力最终用户获得更强劲的网络服务能力。

H3C 全新 SNA 解析

在由交换机、路由器、无线 AP 等各类基础设备构成的网络中，时时刻刻都奔流着事务日志、易损件状态、异常告警等过程信息，这些纷繁复杂的时序性信息在传统网络维护中，基本都会变成“冷数据”被存储起来。但在 H3C 看来，这些数据与网络运维密切相关，正是把握网络脉络与构建智能网络架构的基础。

为了有效利用这些数据，H3C 将网络数据的利用归纳为感知、分析和决策三个阶段。在感知阶段，通过高性能的毫秒级采集技术将脱敏的数据从边缘端收集起来；在分析阶段，系统在云端利用数据进行 AI 建模和训练，不断优化各类网络模型；而在决策阶段，系统可在边缘或云端根据网络的当前状况，通过模型推理来生成新的策略并下发执行。

基于这一模式，SNA 如图一所示，由先知网络中心 (SNA Center)、先知分析器 (SeerAnalyzer) 以及网络控制器 (Seer Engine) 三大模块组成。其中，部署在数据中心的先知网络中心作为网络智能管理、控制和编排的核心，不仅可根据用户的业务需求实现智能编排、业务协同和资源调度，更能汇总各个网络设备的数据，在其内部 AI 平台中进行建模、评估、训练和调优，并形成一系列智能网络模型供调用。



图一 SNA 总体架构

如果把先知网络中心比作新架构的“头脑”，那么先知分析器和网络控制器则是架构的“手和眼”。通过 Telemetry 等毫秒级采集技术，先知分析器可从网络基础架构中感知和采集各类数据，并经清洗、抽取、转换等处理，上传到先知网络中心。

经过先知网络中心训练和调优的模型，在云端或先知分析器中进行 AI 推理后，可形成有效的自动化网络部署和调优策略。而网络控制器则会根据这些策略，以可编程的方式对网络基础架构实施管理，从而实现业务的自动化部署，以及网络资源的最优路径调度与网络故障的预测告警和快速排除。

依托上述流程和机制，SNA 就能将来自网络的海量数据以智能分析方法转化为各类 AI 模型，并结合用户的实际应用场景，例如园区、数据中心等，产生最终的决策应用。目前，该架构已经能为用户提供 20 余种智能网络算法以及 100 多个网络状态洞察方法。

至强® 平台集成 AI 加速，助力 SNA 优化 AI 性能

除了采用创新的方法，H3C 也希望通过引入更强劲的硬件基础设施来为 SNA 加速，要求与新架构配套的计算平台，不仅要满足用户从数据采集处理、网络编排部署、资源调度优化到可视化界面的长流程需求，也要为高负载的 AI 模型训练与推理过程提供可靠的支撑。

基于英特尔® 架构的硬件基础设施无疑是满足以上需求的优选，这些基础设施组件包括：

- 第二代英特尔® 至强® 可扩展处理器。该处理器除了在数据分析、科学计算、音视频处理等通用计算领域有着显著的优势，还可凭借自身集成的多种硬件级 AI 加速能力，如英特尔®

高级矢量扩展 512 (英特尔® AVX-512)、英特尔® 深度学习加速 (Deep Learning Boost) 为广泛的 AI 应用，包括机器学习和深度学习提供加速能力；

- 英特尔® 傲腾™ 持久内存。这种性能与 DRAM 内存相近，成本、容量上更优，且具备数据非易失性的新型内存可让用户将更多分析用或训练用数据缓存在距处理器更近的内存子系统中，从而满足这些数据密集型应用对于数据访问 I/O 和时延的严苛要求；

- 英特尔为 AI 应用提供开发、部署和性能调优的一系列软件工具，如面向英特尔® 架构优化的 AI 框架 (TensorFlow、Caffe、MXnet、BigDL 等)、面向深度神经网络的英特尔® 数学核心函数库 (英特尔® MKL-DNN)、英特尔® 数据分析加速库 (英特尔® DAAL) 和 OpenVINO™ 开发套件等。

H3C 与英特尔这次技术合作的重心，就落在向 SNA 的多个模块中导入第二代英特尔® 至强® 可扩展处理器和面向英特尔® 架构优化的 AI 框架上，这一举措可为其多种 AI 模型的训练和推理加装强劲的算力引擎。

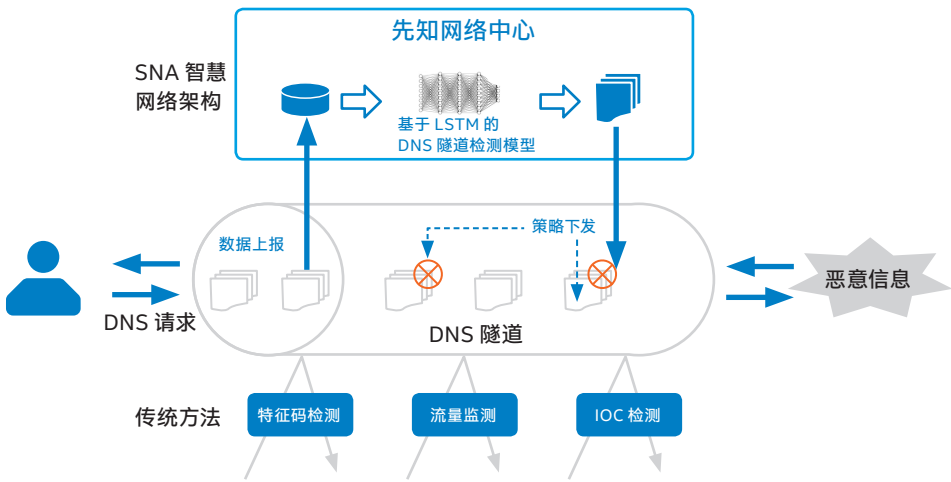
为了考察这一引擎的运转状态，H3C 进行了一系列面向实际应用场景的测试，DNS 隧道检测模型的测试就是其中之一。作为网络应用中重要的基础协议之一，域名系统 (Domain Name System，DNS) 协议主要用于将 IP 地址转化为域名供访问。因此，恶意程序经常会利用 DNS 请求时形成的 DNS 隧道来对网络实施攻击，例如将数据封装在 DNS 请求数据包中，从而绕开内外网隔离等防御措施，造成企业内网的关键信息被透传。

由于这类恶意攻击隐藏到了 DNS 隧道中，因此，如图二所示，常规的网络安全检测方法，例如特征码检测、流量监测、威胁标志 (Indicators Of Compromise，IOC) 检测等都难以对其发挥作用。为此，H3C 以海量 DNS 请求报文为基础，在先知网络中心中构建基于长短期记忆网络 (Long Short-Term Memory，LSTM) 的 DNS 隧道检测模型，来帮助用户提升网络安全等级。

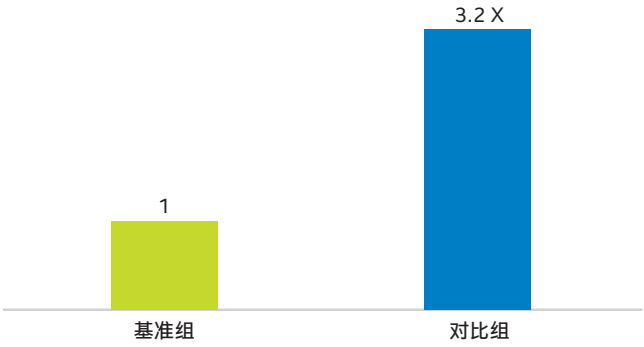
作为递归神经网络 (Recurrent Neural Networks，RNN) 的重要衍伸模型，LSTM 可以通过 3 个特别的“门”结构设计，来大幅提升模型的记忆时长，因此特别适于 DNS 请求这类典型的时序性数据。它可以围绕一段时间内的黑白名单数据集中正常和恶意请求的不同特征，例如主机名、DNS 名称、特定字符等，来预测新的请求中可能面临的安全风险。

更长的时序特征提取、更复杂的门结构，意味着模型在训练和推理中需要更大的计算量。H3C 引入第二代英特尔® 至强® 可扩展处理器，就是看中了这一平台能够依托优化的微架构、更多及更快的内核和内存通道带来计算性能提升，为 AI 训练和推理提供全面的加速能力。

同时，为使第二代英特尔® 至强® 可扩展处理器充分发挥潜能，H3C 还引入了面向英特尔® 架构优化的 TensorFlow，它的优化特性包括了对英特尔® MKL-DNN 的集成、计算图的优化以及针对 Kernel 的优化等等，从而确保 SNA 的 AI 训练和推理过程可以工作在英特尔® MKL-DNN 基元上，并最大程度地对处理器资源分配进行优化，从而在不改变神经网络模型的情况下提升整体性能。



图二 基于 LSTM 的 DNS 隧道检测模型



图三 DNS 隧道检测模型训练性能归一化对比

两组测试的硬件配置相同，但左侧基准组搭配了原生的 TensorFlow，右侧的对比组则搭配了可以更大程度发挥第二代英特尔® 至强® 可扩展处理器潜力的面向英特尔® 架构优化的 TensorFlow

在基于上述软硬件优化组合的平台上，H3C 鉴于真实应用环境的需求和配置状况，以 DNS 隧道检测模型为例进行了一系列的验证。验证采用的黑白名单数据集包含了 20,000 条黑名单样本和 30, 000 条白名单样本，并以 10, 000 条数据作为测试集。测试结果如图三所示，在基于同一硬件平台，采用面向英特尔® 架构优化的 TensorFlow 之后，训练性能可提升到基准值的 3.2 倍[3]。

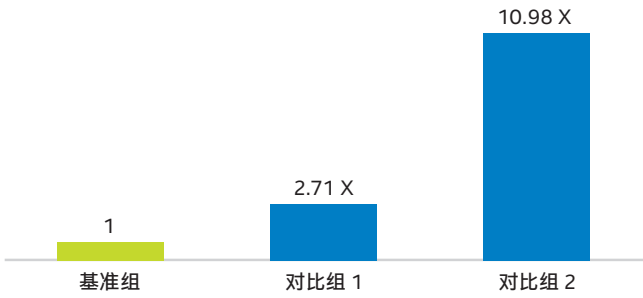
以上的测试数据，是 SNA 应用收益的量化见证，但更具说服力的，则是来自最终客户基于其应用实践的直观感受。以国内某知名高校为例，其基于 SNA 改建的校园网，不仅扭转了过去校园网“不可知、不可控、不可用、不可修”的形象，显著改善了师生的使用体验，而且还成为了科研教学中的好帮手。AI 方法的加入，使学校的关键网络服务质量得到了更好保障。以远程视频教学为例，通过新架构，视频教学过程中的网络抖动、时延、流量等数据通过不断地被采集和分析，以确保能及时下发路径调优、启动业务 QoS 保障等策略，杜绝了视频断线、卡顿等问题。

助力更多企业 ICT 设备迎接智能化

通过与英特尔开展的合作，H3C SNA 顺利达成了目标，并初步赢得了最终用户的认可。接下来在继续推广 SNA 方案的同时，H3C 还将目光瞄向了更多企业级 ICT 设备，特别是部署在大规模云数据中心之内，在管理和运维上同样需要 AI 助力提升效率和自动化程度的设备，因为它们也面临着实现智能化变革的迫切需求。

以在数据中心中扮演计算力关键角色的服务器为例，早在多年前其运维就开始导入可感知其运行状态 (包括使用率、耗电量、散热情况以及故障反馈)的遥测技术，以及与之相匹配的“监控-学习-行动-决定“管理机制。而随着 AI 应用优势显现，包括能利用数据训练更好用和实用的管理和策略模型，并在无需或较少人工介入的情况下及时且并发响应不同事件，有的放矢地实施策略等能力日渐凸显，将 AI 方法引入 ICT 设备的管理和运维也就成为大势所趋。

基于这一认识，继成功推出 SNA，H3C 又携手英特尔启动了服务器管理领域的智能化探索，着手继续利用第二代英特尔® 至强® 可扩展处理器与面向英特尔® 架构优化的 TensorFlow 的组合，来实施服务器利用率模型的优化和测试验证等工作。H3C 近期推进的相关测试就基于真实应用环境的配置，使用了某用户过去 2 年间服务器处理器、内存和存储器的每月利用率作为数据集 (其中 20% 作为测试集，80% 作为训练集)。测试结果如图四所示，在导入面向英特尔® 架构优化的 TensorFlow 后，模型的推理性能可提升至基准值的 2.71 倍[4]；而在进一步导入并行多实例优化后，模型的推理性能还能在不影响延时的情况下，进一步提升到基准值的 10.98 倍[5]。



图四 服务器利用率模型的推理性能归一化对比测试结果

通过以上紧密技术协作和贴近实用环境的测试，在不远的将来，双方必将合力打造出一系列面向企业多样化 ICT 设备或平台的 AI 智能化方案组合，进而将更多行业用户的云平台或 ICT 基础设施带入全方位智能化的时代。

[1, 3] 数据源自 H3C 进行的、在平台配置上贴近用户真实生产环境的 DNS 隧道检测模型训练验证测试。基准组和对比组的硬件均采用双路英特尔® 至强® 金牌 6240 处理器 (工作频率 2.60GHz, 18 内核36 线程)，使用 192GB (12*16GB DDR4-2666) 内存。软件方面基准组采用 TensorFlow 1.13，对比组则采用面向英特尔® 架构优化的 TensorFlow 1.13，两组的操作系统均为 CentOS 7.6 (Kernel 版本: 3.10.0-957)。

[2, 4, 5] 数据源自 H3C 进行的、在平台配置上贴近用户真实生产环境的服务器利用率模型推理验证测试。基准组和对比组的硬件均采用双路英特尔® 至强® 金牌 6240 处理器 (工作频率2.60GHz, 18内核 36 线程)，使用 192GB (12*16GB DDR4-2666) 内存。软件方面基准组采用 TensorFlow 1.13，对比组 1 采用了面向英特尔® 架构优化的 TensorFlow 1.13 和英特尔® Python 3 发布版，对比组 2 在对比组 1 的基础上又使用了并行多实例优化 (4 stream)，三组的操作系统均为 CentOS 7.6 (Kernel 版本: 3.10.0-957)。