软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(8):3870−3890 [doi: 10.13328/j.cnki.jos.006635]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

带内网络遥测方法综述*

吕鸿润[1,2], 李 清[2], 沈耿彪[1], 周建二[3], 江 勇[1,2], 李伟超[2], 刘 凯[2], 齐竹云[2]

[1](清华大学 深圳国际研究生院, 广东 深圳 518055)

[2](鹏城实验室, 广东 深圳 518055)

[3](南方科技大学, 广东 深圳 518055)

通信作者: 李清, E-mail: liq@pcl.ac.cn

摘 要: 网络测量是网络性能监控、流量管理和故障诊断等场景的基础. 带内网络遥测由于具有实时性、准确性和扩展性等特点使其成为当前网络测量研究的热点. 随着可编程数据面的出现和发展, 丰富的信息反馈和灵活的功能部署使得国内外学者提出许多具有实用性的带内网络遥测技术方案. 首先分析了典型的带内网络遥测方案 INT 和 AM-PM 的原理和部署挑战. 根据带内网络遥测的优化措施和扩展角度, 从数据采集流程和多任务组合方面分析了优化机制的特点, 从无线网络、光网络和混合设备网络等方面分析了技术扩展的可行性. 根据带内网络遥测在典型场景的应用, 从网内性能感知、网络级遥测系统、流量调度和故障诊断几个方面对比分析其在不同场景应用特点. 最后, 对带内网络遥测研究进行总结, 展望了未来的研究方向.

关键词: 网络测量; 带内网络遥测; 网络管理; 可编程数据平面

中图法分类号: TP18

中文引用格式: 吕鸿润, 李清, 沈耿彪, 周建二, 江勇, 李伟超, 刘凯, 齐竹云. 带内网络遥测方法综述. 软件学报, 2023, 34(8): 3870–3890. http://www.jos.org.cn/1000-9825/6635.htm

英文引用格式: Lü HR, Li Q, Shen GB, Zhou JE, Jiang Y, Li WC, Liu K, Qi ZY. Survey on In-band Network Telemetry. Ruan Jian Xue Bao/Journal of Software, 2023, 34(8): 3870–3890 (in Chinese). http://www.jos.org.cn/1000-9825/6635.htm

Survey on In-band Network Telemetry

LÜ Hong-Run[1,2], LI Qing[2], SHEN Geng-Biao[1], ZHOU Jian-Er[3], JIANG Yong[1,2], LI Wei-Chao[2], LIU Kai[2], QI Zhu-Yun[2]

[1](Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China)

[2](Pengcheng Laboratory, Shenzhen 518055, China)

[3](Southern University of Science and Technology, Shenzhen 518055, China)

Abstract: Network measurement is the basis of scenes including network performance monitoring, traffic management, and fault diagnosis, and in-band network telemetry (INT) has become the focus of network measurement research due to its timeliness, accuracy, and scalability. With the emergence and development of programmable data planes, many practical INT solutions have been proposed thanks to their rich information feedback and flexible function deployment. First, this study analyzes the principles and deployment challenges of typical INT solutions INT and AM-PM. Second, according to the optimization measures and extension of INT, it studies the characteristics of the optimization mechanism from the aspects of the data collection process and multi-tasking, as well as the feasibility of technology extension in terms of wireless networks, optical networks, and hybrid networks. Third, in view of the applications of INT in typical scenes, the characteristics of these INT applications are comparatively investigated from the perspectives of in-network performance sensing, network-level telemetry systems, traffic scheduling, and fault diagnosis. Finally, a research summary of INT is made, and the future research directions are predicted.

Key words: network measurement; in-band network telemetry (INT); network management; programmable data plane

* 基金项目: 国家重点研发计划 (2020YFB1804704); 国家自然科学基金 (61972189); 深圳市软件定义网络重点实验室项目 (ZDSYS20140509172959989)
收稿时间: 2021-08-08; 修改时间: 2021-10-11, 2021-12-07; 采用时间: 2021-12-27; jos 在线出版时间: 2022-03-24
CNKI 网络首发时间: 2023-02-23

© 中国科学院软件研究所 http://www.jos.org.cn

网络测量是网络性能监控、流量管理和故障诊断等网络研究的基础, 也是“观测→决策→执行 (M-D-E)”网络控制环路中至关重要的一环[1]. 随着网络基础设施不断进步以及云数据中心等新型网络环境的出现与发展, 网络规模在不断增大, 带宽不断增加, 时延也在不断降低. 例如, 云数据中心接入带宽已达 25 Gb/s/100 Gb/s, 出口带宽更高达 10 Tb/s[2]. 据思科全球云指数[3]预测, 2021 年全球数据中心流量将达 20.6 ZB, 在过去 5 年增长 3 倍. 依赖于高性能网络基础设施的新型业务 (如云计算、大数据、人工智能) 不断涌现, 对网络时延、可用带宽或网络可用性等需求不断提高, 从而对网络测量方案性能提出更高要求. 另一方面, 近年来可编程交换机和智能网卡等新型设备兴起与发展, 网络可编程性大为提高, 这些进步极大地扩展了网络测量方案的设计空间. 因此, 新型业务和现代网络基础设施对网络测量技术提出新要求与机遇.

传统网络测量技术发展的 20 多年间, 涌现多种网络测量方案. 根据测量方式的不同, 这些测量技术可分为主动测量 (active measurement)、被动测量 (passive measurement) 与混合测量 (hybrid measurement) 技术[4]. 主动测量技术将探测数据包注入网络然后进行相关测量, 其主动性可以检测网络潜在问题, 代表方案为 Ping、Traceroute. 然而主动测量方案产生的数据包与业务数据包经历的网络环境不完全相同, 观测值不完全代表网络的真实情况. 此外, 主动测量所引入的探测数据包通常非业务流量, 会产生额外的网络带宽开销. 而被动测量技术通过记录与分析网络内已有流量从而得到测量结果, 常用的手段为流量镜像和端口转发, 代表方案为 NetFlow、sFlow、IPFIX. 然而受到交换机性能与网络带宽的限制, 被动测量方案的测量对象与交换机功能紧密耦合、测量粒度较粗, 且较难跨管理域部署, 因而对于某些端到端的应用性能较难观测, 对复杂网络的分析与应用支持造成限制. 混合测量方案使用主动与被动测量属性, 以期融合这两种测量方案的优点. 此后, 软件定义网络 (software-defined network, SDN) 的出现使得网络精细化管理成为可能. SDN 将数据平面与控制平面解耦, 实现了中心化的网络控制和基于流表的处理机制. 因此, 在软件定义网络测量领域业界进行了大量研究工作[5]. 然而 SDN 集中的控制平面限制了扩展性, 而且数据平面功能有限, 难以实现新测量功能. 因此上述传统网络测量技术不能完全满足现今大规模网络需求.

在过去几年中, 可编程数据平面 (programmable data plane, PDP)[6,7]的出现为网络遥测 (network telemetry) 方案提供全新设计空间. 网络遥测是远程收集和处理网络信息的自动化过程. PDP 允许数据平面可编程的处理数据包, 基于 P4[8]的可编程芯片和 POF[9]是 PDP 的两条技术路线. PDP 允许定制数据包操作和访问设备内部状态, 这使得直接在数据平面实现网络测量成为可能, 因而带内网络遥测 (in-band network telemetry, INT) 近年来受到了工业界和学术界的广泛关注. 不同于传统网络测量技术, 带内网络遥测将数据包转发与网络测量结合, 转发节点 (如交换机、智能网卡) 收集网络内部状态并插入遥测数据包, 因此带内网络遥测具备实时性强、测量粒度细和测量状态丰富等优点. 目前, 阿里巴巴、Arista、CableLabs、Cisco、Dell、Intel、Marvell、Netronome 和 VMware 等厂商已经参与到带内网络遥测标准的制定[10], 华为、Broadcom、Intel 和 Cisco 等厂商先后发布了支持带内网络遥测的交换设备[11−14].

带内网络遥测以其实时性、精准性、无须控制平面参与等特性给网络测量技术带来了新方向, 基于带内网络遥测的测量方案与应用成为当前网络运营、管理与维护的研究热点. 近年来, 业界提出了很多创新性的测量框架与机制, 其应用范围已经涉及了多个领域. 但是目前针对网络遥测的综述并不多[15−17], Yu 等人[15]从自顶向下的设计方法角度分析了网络遥测系统的实时性和细粒度等问题; Tan 等人[16]重点对网络遥测的系统层面所涉及的相关技术进行研究, 如可编程数据平面和遥测数据的查询和检索等, 并讨论了网络遥测在性能遥测、微突发检测等方面的应用; Manzanares-Lopez 等人[17]着重对被动式带内网络遥测方案进行对比分析. 上述文献的研究对象是带内网络遥测的子集或超集, 因此这些文献的不足之处在于对带内网络遥测综述与讨论不全, 或对带内网络遥测自身机制的讨论不够深入. 与上述文献相比, 本文的研究对象是以 INT 和 AM-PM 为主的带内网络遥测技术, 本文着重分析带内网络遥测自身各种机制的优化、扩展及其应用进展.

本文首先介绍代表性带内网络遥测方案的基本原理及其实现方案, 分析它们的部署挑战. 然后从带内网络遥测的机制优化和网络场景扩展角度对带内网络遥测研究成果进行系统归纳总结, 然后对带内网络遥测的应用研究从网内性能感知、网络级遥测系统、流量调度和故障诊断几个方面进行分析对比, 最后总结了带内网络遥测的研

© 中国科学院软件研究所 http://www.jos.org.cn

究工作, 并对未来研究方向进行了展望.

1 带内网络遥测原理及挑战

1.1 INT

INT 标准[10]由 P4.org 应用工作组于 2017 年发布, 目前已经更新到 2.1 版. 在 INT 框架中, 数据平面无需控制平面的参与, 能够直接收集和汇报网络状态信息, 因此 INT 具有高实时性和细粒度特点. 此外, INT 的测量对象多样, 包括队列占用、链路利用率、跳时延、时间戳和节点 ID 等. INT 相关术语及解释如表 1 所示, 本文采用此套术语进行论述.

表 1 INT 术语解释

术语	解释
监控系统	从各网络设备收集遥测数据的系统
INT头	携带INT信息的数据包头
INT数据包	包含INT头的数据包
INT源节点	创建并插入INT头到数据包的实体
INT中间节点	根据INT指令收集数据平面元数据的实体
INT宿节点	提取INT头、收集路径状态信息的实体
INT指令	指示在INT节点收集的INT元数据类型
流监控列表	位于数据平面的表, 匹配数据包头、执行INT指令
INT元数据	INT源节点和中间节点插入INT头或遥测报告的信息

INT 系统运行示例如图 1 所示. 主机 1 向主机 2 发送数据包, 交换机 1 作为 INT 源节点收到数据包后向其插入包含 INT 指令的 INT 头以及相应的遥测信息, 交换机 2 作为 INT 中间节点收到数据包后根据其中的 INT 指令向其嵌入相应遥测信息, 交换机 3 作为 INT 宿节点对收到的数据包首先插入遥测信息, 然后从数据包中提取 INT 头和 INT 元数据, 生成遥测报告转发到监控系统, 数据包继续转发到主机 2.

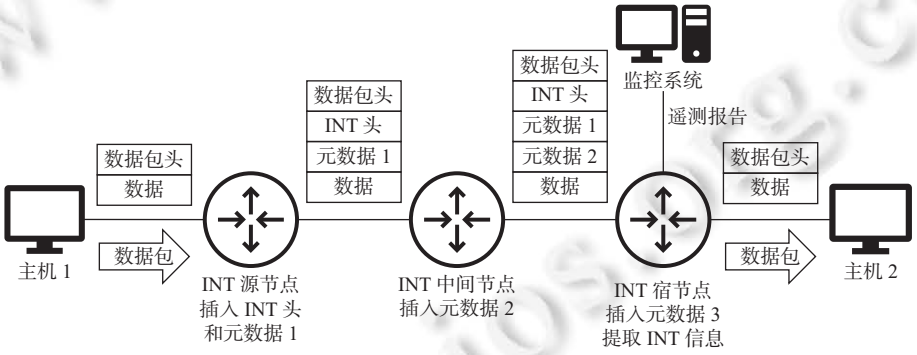

图 1 INT 系统

INT 系统配置具有极高的灵活性. INT 源节点可以为应用程序、端侧协议栈、虚拟机监视器、网卡、ToR 交换机等; 触发执行 INT 的数据包可以是业务数据包、业务数据包的副本或专用的探测数据包; 此外, 关于遥测指令和遥测信息的传递, INT spec v2.1[10]中提出了 3 种操作模式: INT-XD、INT-MX 和 INT-MD. INT-XD (INT export data) 模式又称为明信片模式, 每个 INT 节点收到数据包后根据流监控列表收集 INT 元数据并直接发送到监控系统, 因此无需修改数据包; INT-MX (INT eMbed instruct(X)ions) 模式将 INT 指令嵌入数据包, INT 节点根据 INT 指令收集 INT 元数据, 并将 INT 元数据直接发送到监控系统, 因此对数据包修改较少; INT-MD (INT eMbed data) 模式将 INT 指令和 INT 元数据均嵌入到数据包, 因此对数据包的修改程度最大. 这 3 种操作模式之间存在数

© 中国科学院软件研究所 http://www.jos.org.cn

据包容量限制、带宽开销、实时性和测量灵活性的权衡. 对于 INT-MD, 数据包携带的遥测数据随着网络跳数和遥测指令数的增加而增加, 由于网络 MTU 限制, INT-MD 可能会出现数据包剩余容量不足的问题; 而 INT-XD 和 INT-MX 直接将遥测数据发到遥测系统, 因此不存在 INT-MD 面临的上述问题. 而且这两种模式不必等到数据包到达宿节点后再处理, 因此实时性更强. 然而在这两种模式中, 每个遥测数据需要构造一个数据包进行传输, INT 报告数据包的数量大大增加, 从而导致带宽开销增加. 此外, 相比于 INT-XD, INT-MX 和 INT-MD 由数据包携带遥测指令, 因此更容易调整测量对象, 能够实现更为灵活的遥测. 后文中除非明确说明, INT 均指 INT-MD.

INT 标准文档定义了节点操作逻辑、监控对象以及 INT 报文格式等接口说明, 然而没有说明实现方式. 因此为了便于后续的研究, 一些学者以不尽相同的方式实现了 INT 机制.

Kim 等人[18]在 2015 年发表了第 1 个 INT 原型实现. 该原型利用 P4 语言在软件交换机上实现. 在软件 P4 交换机搭建的测试网络中, 每个软件交换机部署了实现 INT 逻辑的 P4 程序. 为展示 INT 应用潜力, 在周期性端到端突发流量模式下, 作者利用 INT 收集各交换机队列长度, 从而诊断造成周期性 HTTP 时延尖峰的原因.

Tu 等人[19]基于 P4 实现了 INT 机制. 在 INT 标准基础上, 作者使用 UDP 封装, 并自定义了 4 个字段: 目的端口号改写为预定义的 INT 端口号, 代表该数据包为 INT 数据包, 而真实目的端口号保存在 INT 头的 original dest port 字段中; O 字段标识该 INT 包为业务数据包或探测数据包, INT 宿节点对探测数据包转发到 ONOS 控制器进行分析, 而业务数据包进行正常转发; INT Len 字段记录 INT 头及遥测数据的总长度. 为了实现 INT 操作逻辑, 根据 P4 交换机架构, 程序在解析阶段判断数据包类型, 在入口匹配行动表中判断交换机是否为源节点或宿节点并执行相关操作, 在出口匹配行动表中根据当前 INT 节点类型添加元数据及 INT 宿处理逻辑. 实验测试表明 ONOS 控制器的 CPU 利用率与收到的 INT 包数成正比, CPU 利用率和带宽开销是 INT 大规模应用的瓶颈.

除基于 P4, 一些工作通过扩展 Open vSwitch (OVS) 实现了 INT 机制. Gulenko 等人[20]在 OVS 上实现 INT 的核心思路是扩展 OVS 的内核和用户空间模块从而支持 INT 定义的各种操作. 当入端口收到数据包后, 交换机将 INT 相关元数据如时间戳分配到该数据包; OpenFlow 流表中添加了自定义的 INT 相关规则, 数据包经过流表时添加 INT 标记; 在出端口, 带有 INT 标记的数据包的 INT 头被解析, 然后将相应的 INT 元数据嵌入 INT 头. 实验测试表明当不触发 int_transit 时, CPU 开销忽略不计; 当触发 int_transit, 内核模块的 CPU 开销约为 0.3% (1 Mb/s) 和 1% (100 Mb/s), CPU 开销与瞬时吞吐为亚线性关系.

Tang 等人[21]通过扩展 OVS 实现了 POF, 进而实现了运行时可编程的 INT 方案 Sel-INT. 为了扩展 OVS 实现 POF, 作者通过 OVS 二元组<offset, length>实现通用的数据包解析, 并根据 POF 动作重定义 OVS 动作空间. 为了在 OVS-POF 上实现 INT, 作者通过选择组表实现 INT 头插入, 利用 POF 动作定义 INT 各操作. 此外作者利用 OVS 的快速路径实现高速处理吞吐. 实验测试表明当采样率低于 20%, Sel-INT 在最坏情况下的吞吐损失低于 11.2%.

1.2 AM-PM

另一种带内网络遥测方法是 AM-PM (alternate marking performance measurement)[22,23]. AM-PM 是一种被动测量方法, 它只需数据包头的 1–2 位即可测量链路或端到端丢包、时延和抖动. 其优势包括易于实现、带宽运算存储开销低、测量准确、适用网络类型广、抵抗乱序、时间戳灵活和无互操作等.

AM-PM 测量丢包的基本原理是将流量分块, 两测量点分别对同一流量块的数据包进行计数, 通过计算两计数器的差值获得丢包数. 为了同步各测量点对同一流量块的识别, AM-PM 对同一流量块内所有数据包标记相同颜色, 相邻流量块使用不同颜色, 颜色的改变相当于自同步信号, 这确保了路径上不同设备的测量一致性, 这种方法称为标记染色法. 通过设置测量点的位置, AM-PM 能够实现对链路、节点或端到端的监控.

AM-PM 测量单程时延的基本方法是各设备分别记录同一流量块第 1 个数据包到达的时间, 时间差为时延, 此方法称为单标记方法. 为了保证测量的准确性, 单标记方法要求各设备时间同步, 以及流量块没有丢包和乱序. 为了降低对丢包和乱序的敏感性, 遥测节点可测量流量块的平均时延. 单标记方法对每个流量块只能给出一个测量值, 无法获得最大、最小、平均时延等统计信息. 在单标记方法基础上提出的双标记方法可以获得更多信息, 而且对丢包和乱序更健壮. 双标记方法的第 1 类标记与单标记方法相同, 用于流量的分块; 第 2 类标记用于时延和抖

© 中国科学院软件研究所 http://www.jos.org.cn

动的测量, 具体方法是标记一组数据包, 在各网络设备中记录其到达时间, 利用相同数据包在不同设备的时间戳进行数据包时延的计算. 提高第 2 类标记的频率可以获得大量测量数据, 但为了避免乱序的影响标记频率不能太高.

Riesenberg 等人[24]提出了用于实现 AM-PM 的时间复用解析方法. 该方法从芯片和可编程数据平面两个角度分析了实现 AM-PM 所需的抽象, 并在 P4 和 Marvell 交换机芯片上完成了原型实现. 实验数据表明 AM-PM 时延测量误差低于 100 ns, 丢包测量误差小于 0.000 1%.

1.3 分析与挑战

INT 和 AM-PM 是两类主流的带内网络遥测方法. INT 直接读取交换机内部状态, 测量对象更为丰富, 测量层次更为多样; 由于带内测量的特点, INT 的实时性好, 测量粒度细. 但是 INT 要求遥测数据包携带遥测指令和遥测数据等信息, 带宽开销较高. 此外, 对比各方案可以发现基于 P4 的 INT 实现相对简单. 再加之 P4 的语义表达能力强、协议无关和平台无关等特点[6], 后续对 INT 的研究与应用大多基于 P4 进行. 与 INT 相比, AM-PM 的测量原理简单, 测量精度高, 适用网络场景广泛, 实现和部署容易. 而且 AM-PM 无需在数据包中嵌入数据, 占用带宽资源少. 然而受限于测量原理, AM-PM 可直接测量的对象仅有丢包、时延和抖动, 测量对象有限, 不能采集测量设备内部状态或自定义测量对象.

表 2 对带内网络遥测的原型实现进行了全面对比, 包括遥测类型、遥测粒度、带宽开销、资源开销和实现方式. 带宽开销指遥测引入的额外网络带宽占用, 带宽开销的大小会直接影响流完成时间和应用级性能[25]. 由于 AM-PM 只需利用原有业务数据包的头部的 1 到 2 位进行标记, 无须嵌入其他数据, 因此 AM-PM 的带宽开销基本为零, AM-PM 具有极低的带宽开销; 对于 INT, 说明遥测需求的 INT 头占 8 字节, 每个遥测值为 4 字节. 由于 INT 采集每跳信息, 总体带宽开销随遥测需求数和跳数线性增长. 以常见的 5 跳数据中心拓扑为例, 每个数据包最少携带 28 字节 INT 数据 (每跳采集 1 个数据), 这占用了 1 000 字节数据包的 2.8%, 因此带宽开销可高达 2.8%. 而且带宽开销随着遥测需求的增加而线性增长. 而 Sel-INT 在 INT 基础上, 通过恰当的采样方法降低了高达 80% 以上的带宽开销.

表 2 带内网络遥测原型实现的对比

方案	遥测类型	细粒度	带宽开销	采集开销	实现平台
Kim等人[18]	INT	√	高	高	P4
Tu等人[19]	INT	√	高	高	P4
Gulenko等人[20]	INT	√	高	高	扩展Open vSwitch
Sel-INT[21]	INT	√	低	高	POF+扩展Open vSwitch
Riesenberg等人[24]	AM-PM	×	极低	高	P4

监控系统的计算资源开销与遥测数据包接收速率成正比[19]. 因此为了排除网络规模的影响, 采集开销定义为监控系统单位时间收集单位遥测数据所需的计算资源. 监控系统的采集开销决定了可支持网络规模的上限. 采集开销的决定因素包括收集器算法和收集方式. 各原型实现均没有对收集器算法或收集方式进行优化, 因此这些方案都具有较高的采集开销, 据 Tu 等人[26]的测试, 以 IntMon[19]为代表的未优化监控系统采集开销 (以 CPU 使用率计算) 高达 9.71% 每 Kpps, 采集开销优化代表方案 INTCollector 为 0.006 46% 每 Kpps.

带内网络遥测作为网内性能感知、网络级遥测系统、流量调度和故障诊断等重要网络应用的基础, 其测量性能的好坏直接决定了上层应用的服务能力. 由于当前网络环境的复杂性和多样性, 设计满足要求的带内网络遥测方案具有一定的挑战.

(1) 资源开销优化. 在 INT 机制中, 数据包逐跳采集路径上各交换机的状态数据, 遥测数据嵌入数据包或直接转发到遥测系统. 尽管此机制能够精确实时的采集网络内部状态, 然而随着网络规模与带宽的增长, 原有 INT 机制不但会加大网络带宽开销和网络设备资源开销, 存在扩展性问题. 而且 INT 头的体积较大, 导致流完成时间和应用级性能受到影响[25]. 此外, 在实际网络中往往需要运行多种网络任务, 不同的网络任务对遥测信息的需求不

© 中国科学院软件研究所 http://www.jos.org.cn

尽相同. 直接将多个遥测任务简单视为几个普通遥测任务的集合, 不但会由于忽略某些任务的相似性而导致重复测量, 而且在设备资源有限的条件下可部署的遥测任务数不足, 难以支撑上层应用需求.

(2) 网络场景扩展. INT 标准的服务对象是有线网络, 其监控对象和包头格式是针对有线网络设计和定义的. 然而在诸如无线网络和光网络等其他网络场景中, 网络设备与有线网络不同, 需要测量的对象也不尽相同. 在这些网络场景中实现 INT 的细粒度和高实时性的测量同样重要. 此外, 在不支持 INT 的传统网络设备和支持 INT 的网络设备共存的异质网络中, 如何尽可能保留 INT 机制原有的优势也是需要考虑的问题.

2 带内网络遥测优化与扩展

相对于 AM-PM, INT 的测量对象更为多样, 测量机制也更为复杂, 所以目前大多数带内网络遥测优化与扩展研究针对 INT 研究. 在 INT 框架下, 每个数据包都可以携带 INT 头进行遥测, 而且每次遥测都可以采集多种遥测信息. 尽管这使得 INT 获得了细粒度和测量丰富性, 但是如果不加限制, INT 将消耗过多的网络带宽和遥测系统资源, 而且随着网络规模和带宽的增长愈发严重. 此外, INT 主要针对有线网络设计测量机制和测量对象, 这使得 INT 难以直接用于诸如无线网络、光网络和异质设备网络等其他网络环境. 但是在这些网络环境中, 精确实时的测量网络关键数据对于网络运维和各种上层应用都具有重要意义. 因此针对带内网络遥测存在的上述问题, 各研究提出了不同方案, 本节按图 2 所示的分类进行分析.

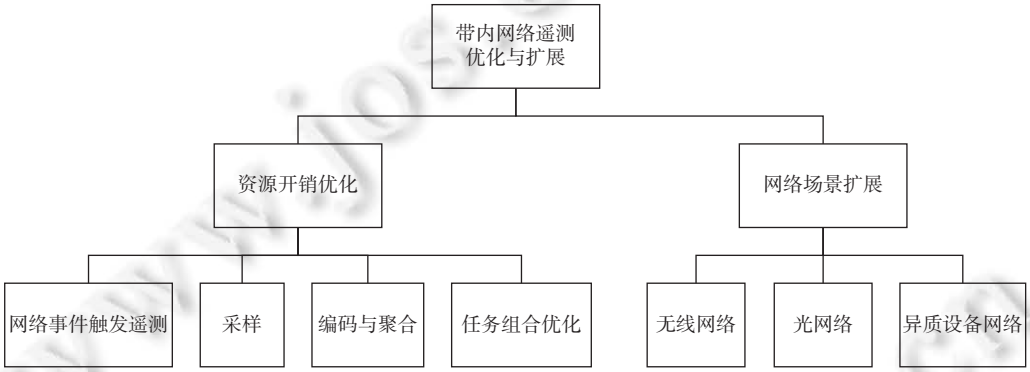

图 2 带内网络遥测优化与扩展机制分类

2.1 资源开销优化

INT 数据包与路由路径上各个 INT 节点依次交互, 收集该路径的遥测数据, 遥测信息由 INT 数据包携带或由 INT 节点直接发到遥测系统. 尽管这种经典 INT 模式能够有效采集网络内部实时状态, 然而随着网络规模与带宽的增长, INT 将导致巨大的带宽开销, 而且所添加的 INT 头将影响流完成时间和应用级性能[25]; 另一方面, 遥测系统需要处理的 INT 数据随着网络规模和带宽的增长随之增加[19], 因此遥测系统成为遥测系统的瓶颈, 影响扩展性. 对于带内网络遥测存在的上述挑战, 许多学者从不同层次与角度提出带内网络遥测优化方案, 如图 2 所示, 主要可分为网络事件触发遥测、采样机制、编码与聚合机制和任务组合优化.

2.1.1 网络事件触发遥测

在经典 INT 框架下, INT 节点根据 INT 指令收集遥测信息并插入到 INT 数据包, INT 宿节点提取 INT 数据包中的 INT 元数据, 生成 INT 报告并发送到遥测系统, 遥测系统根据具体应用对 INT 报告进行分析处理. 由于大多数应用更为关心网络中突发的事件, 所以没有发生网络事件时的遥测信息作用相对不大. 因此一类带内网络遥测的优化思路是在信息采集与传输的各个环节检测与提取网络事件.

Tu 等人[26]设计实现了高性能 INT 收集器 INTCollector. INTCollector 接收来自 INT 宿节点的 INT 报告并提取遥测数据, 然后与上一次遥测数据对比从而判断是否发生网络事件. 当发生网络事件时, INTCollector 保存相关遥测数据, 从而捕获所有重要网络事件, 同时降低所需存储与计算开销. INTCollector 在系统设计层面设计了两条

© 中国科学院软件研究所 http://www.jos.org.cn

数据处理路径, 快速路径利用 Linux 内核中快速数据包处理框架 XDP 加速数量众多的 INT 报告处理流程, 实现网络事件提取; 普通路径对数量相对较少的网络事件进行处理并存储到数据库. 实验数据表明, INTCollector 利用软件网卡能够在 8%CPU 利用率下实现 1.2 Mpps 的 INT 报告处理速度. 然而由于网络中产生的 INT 报告数量没有减少, 因此没有解决 INT 带宽开销问题.

Vestin 等人[27]提出了可编程的事件检测框架. 控制平面配置 INT 节点和流级事件检测算法等数据平面参数. 流级事件检测分为快速检测和复杂检测, 快速检测根据元数据的每跳、累加或平均值检测事件; 复杂检测根据 FastReact 格式定制复杂的检测逻辑检测事件. 基于 P4 的可编程数据平面解析 INT 数据包, 执行相应流的事件检测算法, 如果检测到网络事件则标记该数据包; INT 宿节点将检测到网络事件的 INT 报告发送到 INT 收集器, 丢弃其余报告. 此外, 与 INTCollector[26]类似, INT 收集器利用 XDP 旁路内核协议栈加速 INT 报告处理. 实验数据表明开启事件过滤机制能够提高 INT 收集器的容量 10–15 倍.

Wang 等人[28]提出了用于追踪数据包匹配规则的扩展 INT 机制方案. 为了减少 INT 报告的流量, 该方案将网络状态分为可数网络状态和基于阈值网络状态两类. 可数网络状态包括交换机 ID、出口端口 ID 等, 这类状态发生任何变化则意味着发生网络事件; 基于阈值网络状态包括跳时延、队列占用等, 这类状态的变化超过阈值则认为发生网络事件. INT 节点进行流级事件检测, 当检测到网络事件则将 INT 元数据嵌入数据包. 实验数据表明该机制对于不同的网络状态和阈值能够降低 INT 报告量 39 到 12 500 倍.

Suh 等人[29]提出基于采样的 INT 方案 FS-INT. 与 Wang 等人[28]提出的机制相似, FS-INT 将测量策略分为基于速率和基于事件两类. 基于速率策略与 sINT[30]相似, INT 源每隔 R 个数据包插入一个 INT 头; 基于事件策略与文献 [27,28] 相似, INT 中间节点根据给定准则决定是否向 INT 包插入元数据.

2.1.2 采样机制

尽管 INT 能够提供详尽的每数据包每跳遥测信息, 但是许多应用对于测量的精确性和实时性要求没有那么高. 因此一类优化思路是在 INT 框架内引入恰当合理的采样机制, 满足应用需求的同时降低带宽和处理开销.

sINT[30]根据网络状态变化的频率调整 INT 数据包的比例, 从而降低网络开销. INT 源节点维护一个监控速率表, 记录每个流 f 对应的 INT 头插入速率 r_f. INT 源节点对到达的数据包判断所属流, 每隔 $1/r_f$ 个数据包插入 INT 头. INT 监控系统运行插入速率决定算法, 该算法计算每条流前后两次遥测数据变化的显著程度, 当变化超过阈值时提高插入速率; 当超过一段时间遥测数据没有发生显著变化, 则将插入速率重置为最小值. 模拟实验发现 sINT 相对经典 INT 降低了 37% 的网络开销. 然而由于 sINT 采用反应式机制, 在网络状态迅速变化的场景下可能难以及时捕获重要的遥测数据; 此外监控系统需要收集每个流的数据并计算每个流的收集速率, 将成为系统瓶颈.

Niu 等人[31]设计实现了基于 P4 的多层 INT 系统 ML-INT. ML-INT 根据预设的采样率选择流的一小部分数据包携带 INT 头, 而且每个 INT 头仅携带路径上所有电/光网络设备的部分统计信息, 从而极大降低了 INT 头尺寸与总带宽开销. 实验数据表明 ML-INT 的数据包处理速度可达 2 Mpps.

运行时可编程的网络遥测是实现动态高效调整遥测对象和采样率的基础, 是闭环网络控制管理的重要一环, 然而现有 INT 系统大多不支持运行时编程. 为此, Tang 等人[21]提出了基于 POF 的运行时可编程采样性遥测系统 Sel-INT. 作者通过扩展 OVS 实现 POF, 然后在 OVS-POF 上实现基于流表的 INT 机制和基于选择组表的选择性 INT 头插入机制. 流表由 POF 控制器在运行时安装, 因此 Sel-INT 实现了运行时编程能力. SDN 控制器执行采样率计算算法, 通过快速傅立叶变换分析 INT 历史数据, 根据奈奎斯特采样定理确定新的采样频率. 实验数据表明 Sel-INT 能够有效实现运行时采样率和采样数据类型调整; 在快速变化的网络下采用 7.1% 的采样率仍具有较高的准确性.

Pan 等人[32]提出了轻量网络遥测框架 INT-label. INT-label 避免使用探测包降低带宽开销. INT-label 网络设备以预设的周期 T 向经过的业务数据包嵌入设备状态, 因此实现了遥测全网覆盖而且仅引入很小的带宽开销. 由于各交换机独立进行数据包采样, 因此 INT-label 网络设备是无状态的, 能够无缝适应拓扑的变化.

Chowdhury 等人[33]提出了精度自适应轻量级 INT 框架 LINT. 受到传感器网络中模型驱动数据获取的启发, 当 INT 数据包到达时, LINT 设备预测不嵌入遥测数据所造成控制器误差, 如果误差超过阈值则嵌入遥测数据. 此

© 中国科学院软件研究所 http://www.jos.org.cn

外, 基于同一条流的数据包的行为相似性, 作者提出了流上下文感知的 LINT-Flow, LINT-Flow 在预测函数中考虑流上下文以降低预测误差. 实验数据表明在识别拥塞流和拥塞交换机应用中, LINT 降低数据平面开销约 25%, 同时实现召回率在 0.9 以上.

2.1.3 编码与聚合机制

带内网络遥测能够获取准确详细的路径遥测信息, 然而多数上层应用对于数据不要求完全精确, 而且常仅需路径遥测数据序列经某种运算后的结果. 因此一类优化方法是数据编码和数据聚合机制.

Castanheira 等人[34]提出了高效数据收集系统, 用于 SDN 控制器收集交换机数据平面数据. 该系统根据网络各节点的紧邻度将网络细分为集群, 对各集群网络拓扑运行 DFS 算法获得遍历该集群的路由路径. 收集网络状态时, 控制器向各集群发送收集数据包, 该数据包按先前生成的路由遍历集群, 收集各交换机的遥测数据, 再返回控制器. 与传统的控制器轮询策略相比该机制不但降低了遥测数据收集时间, 而且减轻了控制通道的带宽压力.

Hohemberger 等人[35]提出了优化 NFV 服务链分布式监控方法 DNM. DNM 根据整数线性规划, 协调各个 INT 宿节点的布局和网络监控流量的转发, 在有限的开销下实现细粒度网络监控. 实验数据表明 DNM 能够减少 80% 的遥测数据转发量.

Basat 等人[25]提出了概率版本的 INT 框架 PINT, PINT 根据用户设定限制每数据包开销, 同时提供与 INT 相似的可见性. 由于大多数应用无需所有的 INT 每数据包每跳信息, 而且收集所有遥测信息造成开销太大也没有必要, 因此作者通过分析多种应用的需求, 提出了 3 种聚合操作: 每数据包聚合、静态每流聚合和动态每流聚合, 并针对每类聚合提出了降低开销机制. 实验数据表明利用 PINT 实现的拥塞控制、路径追踪和尾时延计算等应用, 仅使用每数据包的 16 bit 即可达到目前最优方案的性能.

2.1.4 遥测任务组合

同一网络内常需要部署多种网络应用, 各种应用的遥测信息需求不尽相同, 因此除了需要单一遥测任务高效执行, 还需要研究如何在有限的带宽和计算等资源限制下高效地完成各个或大部分遥测任务. 遥测任务组合问题的优化目标包括探测流量最小化、可满足遥测任务数量最大化等, 限制条件包括网络带宽、截止时间、网络设备资源和 MTU 限制等. 评价指标包括准确性、实时性、覆盖性、灵活性、遥测粒度、资源开销等.

Marques 等人[36]提出了基于优化的 INT 任务组合方法. 作者首先将带内网络遥测组合 (INTO) 问题分为两类: INTO 聚集优化问题和 INTO 平衡优化问题. INTO 聚集优化问题的目标是最小化遥测流的数量, 从而避免过多的数据包使交换机资源耗尽; INTO 平衡优化问题的目标是最小化单个遥测流的最大遥测项目数, 使得遥测项目在多个流之间平衡, 避免链路带宽耗尽. 作者证明了这两类 INTO 问题均为 NP 完全的, 并设计了启发式算法解决 INTO 问题. 实验数据表明, 在真实 WAN 拓扑下, 启发式算法能够在 1 s 内生成接近最优的方案, 而且在多数网络内能够将开销降低到每接口一条流.

上述 INTO 问题主要考虑遥测任务组合优化问题的形式化模型和求解算法, 然而没有考虑如何动态协调收集网络信息. Hohemberger 等人[37]提出了基于机器学习的动态遥测任务组合规划方案. 为了利用机器学习算法动态调整带内信息的获取, 该方案将 INT 组合规划问题 (INTOPP) 形式化为固定整数线性规划 (MILP) 模型. 与 INTO 问题出发点不同, INTOPP 主要考虑如何尽可能满足时间空间依赖组合. 作者提出了基于机器学习的方法动态指导遥测数据的收集. 实验数据表明, 对于网络异常识别数量, 该方法比目前最优的启发式算法好 8 倍.

Bhamare 等人[38]提出了虚拟网络功能服务链 (SFC) 监控框架 IntOpt. IntOpt 允许为各个服务链定制监控需求, 然后将监控需求转换为遥测任务从而获取网络内部信息. 在遥测任务组合问题中, IntOpt 的优化目标是在满足监控需求下最小化监控流数量, 为此作者提出了基于模拟退火的随机贪心元启发 (SARG) 算法. 模拟实验表明, IntOpt 降低了 39% 监控开销与 57% 总时延.

2.1.5 分析与对比

总结上述各种优化机制能够发现, 对于单一遥测任务的资源开销优化主要围绕数据采集的流程进行, 对于多个遥测任务, 采用恰当的组合方法能有效降低总体资源开销. (1) 采样和事件触发机制从源头减少了数据采集量. 采样机制直接限定 INT 数据包所占比例, 从而降低带宽和设备资源开销; 事件触发机制则是针对上层应用需求,

© 中国科学院软件研究所 http://www.jos.org.cn

自定义一些网络事件, 当发生网络事件时进行数据采集. 前者虽然能够确定性的降低资源开销, 但是面临测量精确性与实时性损失, 可以通过检测到异常后加大采样频率来减轻此问题; 后者尽管能够有针对性地进行数据采集, 同时具有较高的精确性, 但是这种机制依赖于事件的准确定义与检测, 降低了遥测的通用性. 而且对事件精确而复杂的定义会导致设备资源开销问题甚至可实现性问题, 而设置粗略的事件定义则不能有效降低遥测开销. (2) 在遥测数据传输过程中, 编码机制通过对遥测数据的精巧编码降低带宽开销, 但是有损压缩会导致精确性下降, 将一份遥测数据分散到多个数据包传输则面临实时性权衡; 数据聚合机制将上层应用对遥测数据序列的处理过程提前到数据采集时进行, 从而降低遥测数据传输量, 但这种机制要求数据处理过程不能太复杂, 需要能够在交换机上执行. (3) 遥测任务组合优化的基本思路是多遥测任务的建模与优化问题的求解. 如何建模各种限制因素和优化目标是遥测任务组合优化的关键. 对优化问题的快速求解是遥测任务组合优化面临的挑战.

带内网络遥测优化方案优化数据采集过程和多任务组合, 降低了带内网络遥测的开销. 分析现有带内网络遥测优化方案, 从方案的主要优化机制、优化位置、带宽开销、采集开销、精度、粒度和通用性对各方案进行对比, 结果如表 3 所示, 各指标描述各优化方案相对于未优化的 INT-MD 的优化效果, 遥测任务组合优化方案由于部分指标直接对比没有意义, 故这部分用斜杠填充. 带宽开销和采集开销的定义与表 2 相同.

表 3 带内网络遥测资源开销优化方案的对比

方案	优化对象	优化机制	优化位置	带宽开销	采集开销	精度	粒度	通用性
INTCollector[26]	遥测机制	事件触发; 处理加速	遥测系统	高	低	高	细	中
Vestin等人[27]		事件触发; 处理加速	INT宿节点; 遥测系统	高	低	高	细	中
Wang等人[28]		事件触发	INT节点	低	高	高	细	中
FS-INT[29]		事件触发	INT节点	低	高	高	细	中
sINT[30]		采样	INT源节点	低	高	中	中	高
ML-INT[31]		采样	INT节点	低	高	中	中	高
Sel-INT[21]		采样; 运行时可编程	INT节点	低	高	中	中	高
INT-label[25]		采样	INT节点	低	高	中	粗	高
PINT[25]		编码与聚合	INT节点; INT收集器	低	高	中	中	高
LINT[33]		采样	INT节点	低	高	中	中	高
Castanheira等人[34]		聚合	INT报告收集策略	高	低	高	细	中
Marques等人[36]	多任务组合	优化目标下任务组合	遥测任务组合	—	—	—	—	高
INTOPP[37]		机器学习规划任务组合	遥测信息收集协调	—	—	—	—	高
IntOpt[38]		最小化监控流数量	SFC监控任务组合	—	—	—	—	低

2.2 网络场景扩展

INT 标准主要针对有线网络制定, 其遥测行为、测量对象和 INT 头定义适用于有线网络场景. 然而在各种网络环境中, 精确实时的测量网络关键数据对于网络运维和各种上层应用都具有重要意义. 因此一些学者提出了将 INT 扩展到不同的网络场景, 如无线网络、数据包光网络和混合设备网络等.

数据包光网络的层次结构与部分设备不可编程性导致 INT 机制的实现存在挑战. Anand 等人[39]提出了用于数据包光网络的意图驱动遥测框架 POINT. POINT 将查询指令转化为设备相关动作. 对于不支持 INT 的设备, POINT 利用现有的测量机制执行查询指令. 网络设备将查询意图及其响应在不同层之间映射, 从而解决了跨层的数据传递问题. 此外, POINT 通过响应推迟和聚合解决了在层间传递遥测数据存在尺寸限制的问题.

Niu 等人[31]提出了基于 P4 的 IP-over-Optical 网络遥测系统 ML-INT. ML-INT 对 IP 层和光层同时进行实时监控. 光 INT 模块收集来自光信道监视器的光路参数, 数据包 INT 模块收集 IP 层的流遥测数据, INT 代理聚合来自光层和 IP 层的遥测数据. 此外, ML-INT 通过流采样和数据包采样降低遥测开销. ML-INT 的高性能数据分析器提取和分析高速 IP 流携带的 INT 数据. 实验数据表明, ML-INT 能够提高 IP-over-Optical 网络实时可见性, 其数据分析器的处理速度可达 2 Mpps. 在后续工作[40]中作者提出了 ProML-INT, ProML-INT 将 AI 数据分析功能与 ML-INT

© 中国科学院软件研究所 http://www.jos.org.cn

结合, 提高了监控和故障排除能力.

INT 机制的另一扩展场景是无线网络. Karaagac 等人[41]首次将 INT 机制扩展到工业无线传感器网络 6TiSCH. 该方案提出了机会捎带机制, 利用 802.15.4e 帧中的剩余空间携带 INT 数据, 从而不会对业务流量造成影响, 而且无需预留任何资源. 各 INT 节点允许采取不同的初始化、添加和编码策略, 允许中间节点初始化 INT, 允许决定添加或跳过遥测, 从而形成了自组织和分布式的遥测方案. 该遥测方案可用于网络性能监测、拥塞控制、路由和调度管理, 也可应用于类 802.15.4e 和基于 TSCH 的网络.

Haxhibeqiri 等人[42]验证了 INT 机制扩展到 IEEE 802.11 无线网络的可行性与优势. 作者根据无线网络的特性添加了无线链路遥测选项, 如 RSSI、数据率和信道信息等. INT 头嵌入在 L2 和 L3 头之间. 作者利用基于 Linux 的 WiFi 设备验证了 INT 机制在无线网络的可行性. 在后续工作[43]中, 作者扩展和统一了 INT 节点框架, 在面向应用层的 API 中添加了实时在线的 INT 参数配置功能, 在面向物理层的 API 中添加了更多的无线参数支持, 并引入了每跳可靠性确定技术作为新 INT 选项. 此外, 作者将 INT 节点框架与无线 SDN 框架深度融合, 为网络重配置和验证等应用提供了支撑. 方案在 WiFi 硬件上实现, 实验数据表明该方案在单跳链路开销降低 6 倍, 此外基于监控数据的重配置能够满足应用需求.

多数 INT 遥测系统在设计时默认了网络内的交换机均支持带内网络遥测. 然而在实际部署时, 直接将全体网络设备迁移到下一代设备的资本支出和运营支出均过于昂贵. 因此常采用过渡方案, 即逐步更换网络设备, 从而造成了混合网络的出现. 混合网络中包含传统设备和可编程设备. 由于仅有部分设备支持 INT, 因此混合网络对 INT 的部署造成挑战. Zaballa 等人[44]提出了用于 NGS 和 MPLS 混合网络的 INT 方案. 作者首先讨论了混合网络存在的若干限制因素, 如 INT 头的位置与尺寸限制、交换机位置和同步等, 然后研究了通过设计 P4 交换机布局最大化混合网络中的交换机性能与可用性. 此外还设计了必要的控制平面应用和数据平面配置, 并添加了 MPLS 标签验证等新遥测选项. 实验数据验证了混合网络中实现 INT 的可行性, 并进一步证明了将遥测信息反馈到流量工程算法从而实时调整转发规则的可行性.

2.2.1 分析与对比

表 4 对网络场景扩展方案进行了全面对比, 包括网络场景、遥测类型、实现平台和主要特征. 带内网络遥测网络场景扩展方案的扩展对象主要是 INT, 这是因为 INT 的机制较为复杂, 需要考虑适合于目标网络的测量对象和测量方法, 而 AM-PM 原理简单, 容易扩展到其他网络. 总结 INT 在以上各种网络场景的实现可以发现, 将 INT 迁移到其他网络场景, 需要考虑目标网络中需要测量的对象, 例如无线网络对于 RSSI 和数据率等数据特别关心. 不同网络场景下网络设备的可扩展性和灵活性不同, 因此需要根据具体网络情况设计 INT 的各种机制的实现方式. 如果网络设备的可编程性较强, 则 INT 相关操作可以直接在网络设备上实现, 否则可以通过添加代理或智能设备等方式在目标网络上实现. 此外, 尽管针对特定网络场景研究人员设计了相应的遥测方案, 但是缺乏对通用性遥测框架的研究.

表 4 网络场景扩展方案的对比

方案	网络场景	遥测类型	实现平台	主要特征
POINT[39]	数据包光网络	INT	DWDM设备	查询意图及其响应在不同层之间映射
ML-INT[31]	IP-over-Optical网络	INT	P4	INT代理聚合光层和IP层遥测数据
Karaagac等人[41]	6TiSCH网络	INT	Contiki-NG	自组织和分布式的遥测
Haxhibeqiri等人[42,43]	IEEE 802.11无线网络	INT	Linux WiFi设备	统一的无线网络INT节点框架
Zaballa等人[44]	异质设备网络	INT	P4	布局P4交换机最大化性能与可用性

3 带内网络遥测应用

带内网络遥测能够采集丰富的网络内部状态信息, 这不但能够提高基于测量的网络应用性能, 而且为创造新应用提供了可能. 带内网络遥测的典型应用包括网络故障诊断、拥塞控制、先进路由和数据平面验证等.

© 中国科学院软件研究所 http://www.jos.org.cn

以拥塞控制和故障诊断为例说明带内网络遥测应用于网络应用的典型模式. 拥塞控制方案 HPCC[45]的框架如图 3 所示, HPCC 发送端利用 INT 获取流的路径上各交换机精确的拥塞信息, 如队列长度、已传输流量数和时间戳. HPCC 发送端通过遥测数据计算出精确的新窗口大小, 从而在极短时间内占用可用带宽同时避免拥塞. SDN 故障定位系统 PAINT[46]框架如图 4 所示, PAINT 根据发现的故障症状推断可能的原因, 然后部署相应的 INT 任务, 直接从相关网络设备收集信息, 从而确定故障原因.

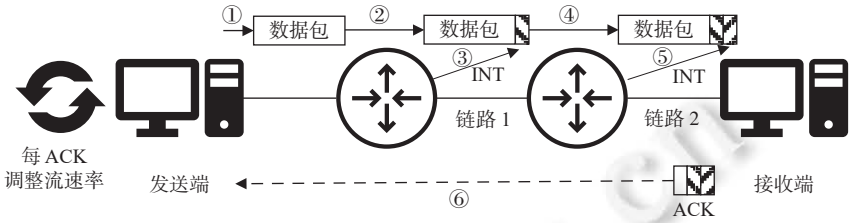

图 3 HPCC 框架[45]

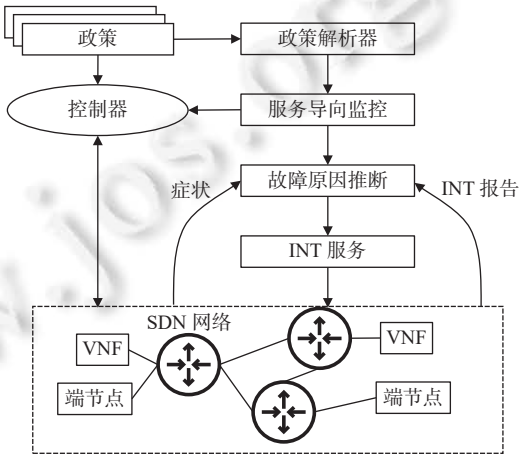

图 4 PAINT 框架[46]

3.1 网内性能感知

网络运维人员和控制系统需要获知及时准确的网络性能相关信息, 以确保网络安全可靠运行以及获得决策依据. 常用的网络性能指标包括时延、丢包、可用带宽、QoS 等.

对于时延测量, INT 机制能够直接利用现有指令完成跳时延的测量; AM-PM 机制能够完成两个测量点之间的时延测量. Riesenberg 等人[24]提出了时间复用解析方法, 在 P4 和 Marvell 交换机芯片上原型实现了 AM-PM. 实验数据表明 AM-PM 数据平面开销低, 时延测量误差低于 100 ns, 丢包测量误差小于 0.000 1%. Karaagac 等人[47]提出了基于 AM-PM 的工业无线传感器网络 (IWSN) 遥测方案, 用于测量关键流的端到端和每跳可靠性和时延.

对于丢包测量, AM-PM 机制对比不同测量点对相同流量的计数值, 直接计算出该时间段两测量点间的路径丢包数. 由于测量原理限制, INT 机制难以直接完成丢包的测量.

对于可用带宽和链路容量的测量, Kagami 等人[48]提出了一种被动测量方法 CAPEST. CAPEST 交换机收集业务数据包长度和散布等统计数据, 当收到 INT 探测包时进行统计分析从而获得容量和可用带宽. 容量和可用带宽分析分为 3 步, CAPEST 首先生成用于分析数据包散布的直方图, 然后利用 reverse histogram 和自相关减轻干扰的影响, 最后根据数据包尺寸和时间戳估计利用率和可用带宽. 实验数据表明 CAPEST 减少 80% 侵入性, 同时提升 10% 测量精度和一个数量级的实时性.

对于 QoS 测量, Liang 等人[49]提出了基于黑盒的轻量 NFV 遥测框架 NFT. NFT 部署于 NFV 平台, 无需修改

© 中国科学院软件研究所 http://www.jos.org.cn

待测 NF. NFT 包含 3 类遥测节点: NFT 分类节点根据特定政策创建和插入遥测头; NFT 转发节点根据遥测头对经过 NF 前后的数据包收集数据, 将遥测数据嵌入数据包或生成明信片; NFT 宿节点提取元数据并复原业务数据包. 实验数据表明 NFT 分类节点根据简单规则可实现高速分类, 但是 NFT 转发节点写入遥测数据的开销显著.

与 NFT 提出的 NFV 监控框架不同, IntOpt[38]和 DNM[35]从优化的角度提出了 NFV 服务链监控框架. IntOpt 控制器将服务流遥测项目和频率需求映射到各物理链路, 在覆盖所有服务流的条件下, 找出开销最低的探测频率、遥测对象和监控流集合. IntOpt 通过周期性发送 INT 探测包测量每条监控流. 控制器收集遥测信息, 检查每个服务流是否违背 SLA. 模拟实验表明 IntOpt 降低 39% 监控开销, 降低 57% 总时延. DNM 基于整数线性规划, 协调各个 INT 宿节点的布局, 实现网络监控流量的高效转发. DNM 在有限的开销下实现细粒度网络监控. 实验数据表明 DNM 减少了 80% 遥测数据转发量.

除上述性能指标和测量方案, Wang 等人[28]提出了数据包规则匹配追踪机制. 规则管理器为安装到交换机的每条规则分配全局唯一 ID, 规则数据库记录安装到交换机的所有规则. 当 INT 包在交换机上匹配了某条规则, 该规则的动作函数将规则 ID 复制到 INT 头的相应字段, 从而完成该跳的匹配规则收集.

INT 机制不但能够直接采集每跳时延, 而且通过访问交换机内部数据, 能够采集队列长度和已传输流量数等更细致的网络信息. 然而由于测量原理的限制, INT 测量丢包是一个挑战; AM-PM 机制天然适合测量时延和丢包信息, 测量粒度取决于测量周期和计数对象. 此外, INT 的测量框架比 AM-PM 更灵活, 因此 INT 相对更容易扩展新遥测指标.

3.2 网络级遥测系统

INT 标准详细定义了用于监控路径状态的 INT 操作原语, 然而如何实现全网或特定范围的遥测以及完整的测量系统没有定义. 网络级遥测系统可帮助运维人员轻松获悉实时网络状态, 也可通过 API 为上层应用提供所需的网络内部信息. 一些学者提出了基于 INT 的网络级遥测系统, 用于高效的监控部分或全局网络的状态信息.

一类扩展方法是利用网络内已有的业务流携带 INT 头, 从而实现对所关注流量或路径的测量. 由于该方法直接利用现有流量, 因此属于被动测量. Hyun 等人[50]结合先前关于 INT 实现[19]和 INT 报告收集优化[26]等工作, 提出了基于 INT 的实时细粒度网络监控框架. 该框架包含 INT 使能数据平面[19]、INT 管理系统和 INT 报告收集器[26]. INT 管理系统运行于 SDN 控制器, 通过通用接口控制异质的 INT 设备; INT 收集器通过 eXpress 数据路径和事件检测机制, 实现高效的 INT 数据收集. 实验数据结果表明 INT 收集速度提高了 27 倍.

尽管 Hyun 等人[50]方案能有效收集感兴趣流的遥测数据, 然而如何确定网络内哪些流量需要关注仍待解决. Castanheira 等人[34]提出基于可编程数据平面的综合流量监控系统 FlowStalker. FlowStalker 首先寻找需要关注的大流, 然后重点进行测量. 在主动阶段, 为了检测大流, 交换机检查每个经过的数据包所属流的数据包计数器是否超过预设阈值, 发现超过阈值的大流则进入反应阶段; 在反应阶段, 数据平面提取和存储目标流的每流和每数据包数据, 同时继续计数数据包数, 超过预设阈值后通知控制器进入收集过程. FlowStalker 的数据收集系统利用网络分治的方法实现高效聚合数据平面信息收集 (见第 2.1 节). 模拟实验表明 FlowStalker 仅降低 12% 网络吞吐, 而且监控项目的数量不影响带宽开销.

直接利用网络内已有的业务流携带 INT 头进行遥测能有效采集关注流量的遥测数据, 然而对于关注的路径未必存在遍历该路径的业务流量. 因此该方法难以实现对任意路径的灵活测量, 从而难以高效实现全网范围的遥测. 因此一类遥测系统的构建思路是采用主动测量, 将 INT 与路径规划结合, 从而实现灵活的任意路径网络遥测和全网范围遥测.

Liu 等人[51,52]提出了基于 P4 的主动探测 INT 系统 NetVision. NetVision 采用主动探测方法, 按需生成适当数量的探测包; 利用段路由实现简单灵活的路由控制和运行时探测包路由定制. NetVision 提出了基于 Hierholzer 算法的环形路由路径生成算法, INT 节点同时充当 INT 源节点和 INT 宿节点, 从而降低不同节点的同步开销, 减少探测包开销与对业务数据包的影响. 实验表明, 在 HTTP 延迟测试、交换机负载均衡测试和路由黑洞测试等故障场景下, NetVision 能快速有效的定位故障原因.

Pan 等人[53]提出了全网遥测框架 INT-path. INT-path 能够生成覆盖全网的最小路径数量的非重叠 INT 路径集

© 中国科学院软件研究所 http://www.jos.org.cn

合. 与 NetVision[51,52]相似, INT-path 将 INT 测量系统解耦为 INT 探测包路由机制和路由路径生成两部分. INT 探测包路由机制采用源路由, 从而实现测量路径的完全可控. 为了实现遍历全网的路由路径集合的高效生成, 并且路由集合能够尽可能降低带宽开销、处理开销并提高测量实时性, 作者提出了两种路由路径生成机制: 基于深度优先搜索, 算法简单明了而且时间效率高; 基于欧拉迹, 能够生成最小路径数量的非重叠 INT 路径. 实验表明, 基于欧拉迹的路由生成相对于深度优先搜索能有效降低 INT 路径数量, 降低 INT 遥测开销.

NetVision、INT-path 等方案从不同角度对 INT 探测包进行路经规划, 然而这些方案没有协调探测包生成过程, 而且忽视设备容量限制, 导致 INT 收集器的开销增加. 为此 Castro 等人[54]提出了 INT 探测包规划方案 P^{2}INT. P^{2}INT 协调探测包的生成和路由过程, 确保遍历所有链路且收集所有所需 INT 数据. P^{2}INT 采用整数线性规划作为优化模型. 作者提出了基于数学规划的启发式算法指导模型寻找高质量解. 实验结果表明, 与同类工具相比, P^{2}INT 生成探测包数量改善达 6 倍, 资源利用率提高多达 3 倍, 传输开销降低 2 倍.

INT 与主动探测的结合还可用于对高负载网络设备的重点监控. Yang 等人[55]提出了轻量高效的 INT 框架 Fast-INT. Fast-INT 利用强化学习算法 DDQN 根据网络状态变化动态调整 INT 监控任务, 从而减少目标点的监控时间、提高监控效率. Fast-INT 将网络遥测转化为强化学习任务, 设定强化学习的状态为 INT 元数据序列; 行动为标记高负载、需要进一步监控的网络设备, 而后 INT 包聚焦监控被标记设备; 奖励由代理根据行动后的网络状态计算. 实验数据表明, Fast-INT 能够更加高效的利用网络资源、更智能的部署遥测任务.

网络级遥测系统研究如何实现高效的全网覆盖遥测, 或对指定范围或目标进行高效的遥测. 对于全网覆盖性的方案, 网络内已有的业务流量随机性较大, 难以覆盖所有链路. 因此全网覆盖性的方案均采样主动探测方法, 主动生成探测数据包执行遥测, 再加之与段路由的结合, 探测包的路径完全可控. 为了降低带宽开销、提高实时性, 如何设计高效的路由路径生成算法是全网覆盖性的方案主要关注的问题. 对于遥测特定范围或目标的方案, 采用主动或被动测量取决于具体的遥测目标, 例如指定流量的遥测自然适合将这些业务数据包作为 INT 包, 而对于特定设备的遥测, 为了保证确定性, 采用主动探测更为适合.

3.3 流量调度

Katta 等人[56]提出了基于虚拟机监视器的拥塞感知负载均衡方案 Clove. Clove 部署于虚拟机监视器, 网络内采用 ECMP, 因此无需修改网络硬件或租户虚拟机协议栈. Clove 使用路由追踪机制发现路径, 利用 ECN 机制 (Clove-ECN) 或 INT 机制 (Clove-INT) 获取路径粗粒度的拥塞情况或精确的路径利用率, 然后根据反馈的网络状态更新路径权重表. Clove 通过修改数据包头字段控制路由路径, 以加权轮转的方式以 Flowlet 粒度进行负载均衡调度. 实验数据表明 Clove 在 70% 网络负载下降低流完成时间 1.5–7 倍.

与 Clove 基于虚拟机监视器的负载均衡不同, Lim 等人[57]提出了网内最佳下一跳负载均衡算法. 该算法分为 3 个阶段, 在网络监控阶段, 交换机利用 INT 测量并记录此交换机视角下的最大跳时延、队列占用和链路利用率以及对应目的 IP 和出口端口号; 在路由决策阶段, 交换机计算各源 IP 对应路径的拥塞程度, 记录源 IP 对应的最小拥塞程度和相应的最佳端口; 在 Flowlet 分配阶段, 交换机根据阈值将每个流划分为 Flowlet, 每个 Flowlet 的转发到路由决策阶段计算的最佳端口.

对于新型拥塞控制算法, Li 等人[45]提出了高精度拥塞控制算法 HPCC. HPCC 通过 INT 获取精确的链路负载信息, 计算精确的流上传速率, 从而克服了传统网络中缺乏细粒度网络负载信息导致的拥塞控制算法收敛慢、无法避免队列建立和复杂的参数调整等问题. 实验数据表明相对于 DCQCN 和 TIMELY, HPCC 的流完成时间降低最多 95%.

带内网络遥测能够采集大量丰富的网络数据, 为机器学习应用到网络管理与控制提供了必要条件. Hyun 等人[58]提出了自驱动网络架构, 将软件定义网络 (SDN)、网络遥测 (INT) 和知识定义网络 (KDN) 有机组合形成闭环管理. INT 用于收集数据包级遥测信息, KDN 使用遥测数据对网络进行智能的管理, SDN 根据 KDN 的决策对网络进行控制. 该框架基于 ONOS 和 BMv2 实现. 此外, 作者描述了基于 KDN 的流量工程和异常检测的实现思想.

© 中国科学院软件研究所 http://www.jos.org.cn

Yao 等人[59]提出了自学习控制策略框架 NetworkAI. NetworkAI 通过建立网络状态上传通道和决策通道实现网络闭环控制. NetworkAI 使用 INT 收集细粒度的遥测信息, 遥测信息经由 Kafka/IPFIX 上传到数据分析平台. 然后 NetworkAI 利用 SDN 的集中视图运行深度强化学习算法, 根据网络遥测信息计算近优的决策, 实现网络的闭环控制.

除上述应用场景, Isolani 等人[60]提出了 IEEE 802.11 网络切片组合框架. 网络切片是控制网络资源精确分配的手段, 恰当的分片方法能够提高 QoS. 该方案首先获得各个应用需求 (APP ID、五元组和相应需求), 根据需求分配网络切片并安装 Open Flow 规则. 然后方案利用 INT 监控应用流量统计数据, 并与应用需求比较. 如果没有满足应用需求, 方案重新分配切片, 使得该应用获得更大份额的开始时间; 否则逐渐释放资源, 使得尽力而为的流量能够利用剩余资源.

带内网络遥测不但能够提供详细的网络状态, 如跳时延、队列大小和链路利用率等, 而且具有实时性强和粒度细的特定. 丰富及时的网络信息为新型负载均衡和拥塞控制等算法提供了新设计空间, 为基于机器学习的智能流量调度提供大量准确的数据.

3.4 故障诊断

带内网络遥测的采集数据种类多、细粒度和实时性等特点, 使得带内网络遥测适用于网络故障诊断.

Kim 等人[18]首先演示了 INT 用于故障诊断的潜力. 作者基于 P4 实现 INT 机制, 利用 INT 监控各交换机队列占用, 解释引起 HTTP 请求时延尖峰问题的原因.

Tang 等人[46]提出了基于 INT 的智能 SDN 故障定位系统 PAINT. PAINT 自动解析网络服务政策, 动态定义与部署粗粒度的端到端测量任务, 监控与收集服务等级症状, 初步推断症状原因. 当所收集的症状信息不足以准确定位故障组件时, PAINT 根据症状故障遥测模型, 将 INT 动作融入故障推理, 自动生成 INT 指令, 通知 INT 源节点执行 INT, 找出症状根源后停止 INT.

Tang 等人[21]利用所提出的遥测系统 Sel-INT 演示了路径验证. 对于交换机误配置导致的负载失衡, Sel-INT 以 10% 的采样率收集每跳交换机 ID, 分析每个交换机的流量是否过高或过低, 从而确定交换机误配置的位置. 对于恶意攻击导致的路径改变, 攻击者修改了被攻击交换机的 ID, 导致直接收集交换机 ID 不能发现异常. Sel-INT 采集入口端口号信息, 分析入口端口号的变化从而确定存在恶意攻击.

Choi 等人[61]提出了端到端服务实时性能监控验证修复系统 Verification Transverse. 该系统使用 MDL 简洁的形式化表达复杂的端到端 SLA, 利用网络测量检测即将或已经违反的 SLA, 根据相应更细粒度的 SLA 自动生成 INT 遥测监控. 当发现违反性能属性后触发实时修复动作. 为了降低 INT 流量, 该系统在大范围内采用粗粒度的监控, 在可疑位置按需采用更细粒度的监控.

Niu 等人[31]利用所提出的多层 INT 系统 ML-INT 演示了 IP-Over-Optical 的可视化与故障排除. 当发生性能问题时, ML-INT 测量 IP 层性能指标 (如交换机时延), 判断 IP 层是否存在问题 (如拥塞); 同时测量光层性能指标 (如能量、OSNR) 判断光层是否存在问题. 在后续工作中作者提出了 ProML-INT[40], ProML-INT 利用 AI 辅助分析工具分析遥测数据, 实现性能监控与故障排除.

Jia 等人[62]提出了数据中心灰色故障迅速检测和定位系统. 该系统中服务器周期发送简化的 INT 探测包, INT 探测包通过组播到达所有其他服务器, 从而覆盖源目对间的所有路径. 探测包收集器存储所有可行的路径. 一旦发生灰色故障, 相应路径的更新超时, 服务器则利用源路由主动绕开故障路径, 并上传故障信息到远程控制器, 远程控制器确定故障位置.

Nam 等人[63]提出了基于 INT 和 RNN 的网络异常检测系统. 控制平面配置 INT 节点和被监控的流, 当发现异常后执行屏蔽端口、负载均衡等缓解措施; 数据平面根据 INT 配置收集遥测信息并转发到管理平面; 管理平面根据固定的时间间隔将数据包级的遥测数据汇总为流级遥测数据, 发送到知识平面; 知识平面利用 RNN 网络将流级遥测数据作为输入, 正常状态和异常状态分数作为输出. 如果异常分数更高则认为发生异常, 根据异常分数的大小选择缓解措施.

相比于 INT, AM-PM 的原理简单、实现容易, 因此常作为大型测量系统的一个部分. Fang 等人[64]提出了用于

© 中国科学院软件研究所 http://www.jos.org.cn

云规模覆盖网络的自动化持续丢包诊断系统 VTrace. VTrace 在虚拟转发设备中对感兴趣数据包进行染色、匹配和记录日志, 记录每跳详细转发环境, 利用日志进行路径重建与分析.

3.5 分析与对比

带内网络遥测作为一类网络测量方法, 其基本用途是网内性能感知. (1) INT 和 AM-PM 作为两种典型的带内网络遥测方案, 其原理完全不同, 因此所适合的测量对象不同. INT 通过直接访问交换机内部状态, 能够直接获取队列长度、已传输流量数和时间戳等信息, 但是此机制决定了难以直接测量丢包. AM-PM 在两个或多个节点对流量进行染色计数, 能够直接计算出路径时延和丢包, 然而这种框架不利于对其他数据的采集, 测量对象的丰富性和扩展性不如 INT. (2) 在此基础上, 带内网络遥测是实现网络级遥测系统的重要组件, 网络级遥测系统的主要组件还包括段路由和路由规划算法. 在网络级遥测框架下, 主动式网络遥测和段路由的结合实现了对任意路径的遥测, 路由规划算法求解在给定条件下最优的一组覆盖全网的路由路径. (3) 由于 INT 测量数据的丰富性和反馈的实时性, INT 适合用于拥塞控制、负载均衡和故障诊断等闭环控制系统. AM-PM 由于精度高和容易实现的特点, 常常作为大型系统中的一个小模块.

带内网络遥测应用研究主要集中于网内性能感知、网络级遥测系统、流量调度和故障诊断等方面, 带内网络遥测应用总结如表 5 所示.

表 5 带内网络遥测应用总结

类别	方案	适用场景	遥测类型	控制结构	测量类型	主要特征
网内性能感知	Riesenberg等人[24]	时延测量	AM-PM	集中式	被动	AM-PM时延测量误差低于100 ns, 丢包测量误差小于0.000 1%
	Karaagac等人[47]	时延测量	AM-PM	集中式	被动	工业无线传感器网络的关键流的端到端和每跳可靠性和时延测量
	CAPEST[48]	带宽测量	INT	集中式	被动	收集业务数据包的长度和散布等统计数据, INT探测包触发统计分析获得容量和可用带宽
	NFT[49]	NFV遥测	INT	集中式	被动	基于黑盒的轻量NFV遥测框架, 部署于NFV平台, 无需修改待测NF
	IntOpt[38]	NFV服务链监控	INT	集中式	主动	将服务流遥测项目和频率需求映射到各物理链路, 找出开销最低的探测频率和遥测对象以覆盖所有服务流
	DNM[35]	NFV服务链监控	INT	集中式	被动	基于整数线性规划协调各个INT宿的位置布局和网络监控流量的高效转发
	Wang等人[28]	匹配规则追踪	INT	集中式	被动	为INT补充数据包规则匹配追踪机制
网络级遥测系统	Hyun等人[50]	报告高效收集	INT	集中式	被动	高效的INT报告收集和控制框架
	FlowStalker[34]	大流遥测	INT	集中式	被动	两阶段大流检测与测量; 基于分治的高效数据采集
	NetVision[51, 52]	网络级遥测	INT	集中式	主动	段路由控制探测包路经; 基于 Hierholzer 算法环形路由
	INT-path[53]	网络级遥测	INT	集中式	主动	源路由控制探测包路经; 基于欧拉迹的最小路径数路由
	P^2INT[54]	探测包规划	INT	集中式	主动	转化为整数线性规划模型, 基于数学规划启发式算法求解
	Fast-INT[55]	高负载设备监控	INT	集中式	主动	利用强化学习算法根据网络状态变化动态调整INT监控任务
流量调度	Clove[56]	负载均衡	INT	分布式	被动	利用INT获取精确的路径利用率, 更新路径权重表, 修改数据包头字段控制路由路径, 以加权轮转的方式在Flowlet粒度进行负载均衡
	Lim等人[57]	智能路由	INT	分布式	被动	交换机利用INT测量记录拥塞信息与对应目的IP和出口端口号, 计算源IP对应的最小拥塞程度和相应的最佳端口

© 中国科学院软件研究所 http://www.jos.org.cn

表 5 带内网络遥测应用总结 (续)

类别	方案	适用场景	遥测类型	控制结构	测量类型	主要特征
流量调度	HPCC[45]	拥塞控制	INT	分布式	被动	利用INT获取精确链路负载, 计算精确流上传速率, 克服传统拥塞控制算法收敛慢、无法避免队列建立和复杂的参数调整问题
	Hyun等人[58]	智能网络	INT	集中式	被动	INT用于收集数据包级遥测信息, KDN使用遥测数据对网络进行智能的管理, SDN根据KDN的决策对网络进行控制
	NetworkAI[59]	智能网络	INT	集中式	被动	INT收集细粒度的遥测信息, SDN的集中视图运行深度强化学习算法, 根据网络遥测信息深度强化学习算法生成近优的决策
	Isolani等人[60]	网络切片组合	INT	集中式	被动	获得各个应用需求, 根据需求分配网络切片并安装Open Flow规则; INT监控实际应用需求满足情况, 重新分配切片
故障诊断	Kim等人[18]	网络异常检测	INT	集中式	被动	监控各交换机队列占用, 解释引起HTTP请求时延尖峰问题的原因
	PAINT[46]	SDN故障定位	INT	集中式	被动	根据症状故障遥测模型, 将INT动作融入故障推理, 自动生成INT指令, 通知INT源执行INT, 找出症状根源
	Sel-INT[21]	路径验证	INT	集中式	被动	收集每跳交换机ID, 分析每个交换机的流量是否过高或过低, 确定交换机误配置的位置
	Verification Transverse[61]	端到端性能验证	INT	集中式	被动	利用网络测量发现即将或已经违反的SLA, 自动生成INT遥测监控, 监控发现违反性能属性后触发实时修复动作
	ML-INT[31], ProML-INT[40]	可视化与故障排除	INT	集中式	被动	测量IP层性能指标判断IP层是否发生问题; 测量光层性能指标判断光层是否发生问题; 引入AI辅助分析
	Jia等人[62]	灰色故障检测和定位	INT	分布式	主动	INT探测包通过组播到达所有其他服务器, 覆盖源目的之间所有路径, 服务器则利用源路由主动绕开故障路径进行重路由
	Nam等人[63]	网络异常检测	INT	集中式	被动	配置INT节点、监控的流, 发现异常后执行屏蔽端口、负载均衡等缓解措施
	VTrace[64]	丢包诊断	AM-PM	集中式	被动	在虚拟转发设备中对感兴趣数据包进行染色、匹配、日志, 记录每跳详细转发环境, 利用日志进行路径重建与分析

4 总结与展望

网络测量方案的性能直接影响网络运维水平和上层应用性能, 进而对网络性能和服务体验产生重大影响. 带内网络遥测是一种近年快速发展的新型网络测量方案. 带内网络遥测相对于传统测量方案具有高实时性、细粒度和测量对象多样等特点. 近年来网络设备快速发展, 特别是网络可编程能力的增强, 使得带内网络遥测的实用化成为可能, 因此带内网络遥测受到业界广泛研究. 现有带内网络遥测方案包括 INT 和 AM-PM, 相比于 AM-PM, INT 的测量对象更为丰富, 具体机制和实现也更为复杂, 所以多数文献以 INT 为研究对象. 经典 INT 机制不但资源开销大, 而且没有考虑网络场景变化和多任务遥测, 因此现有方案从上述 3 个角度对 INT 优化与扩展进行研究. 此外, 带内网络遥测卓越的测量性能为新型网络应用提供新设计空间. 目前主要的研究方向包括网内性能感知、遥测系统、流量调度和故障诊断等.

虽然现在有多种方案设计了多种机制优化与扩展带内网络遥测, 设计了多种基于带内网络遥测的应用, 但仍存在一些不足: (1) 现有 INT 机制专为有线数据包网络设计, 网络场景的可扩展性不足; (2) 带内网络遥测的测量机制不够灵活, 难以融入其他测量机制; (3) 现有的带内网络遥测没有考虑安全性问题; (4) 带内网络遥测与人工智能结合的应用不足. 鉴于现有方案的分析, 带内网络遥测未来可能的研究方向有以下几个方面.

© 中国科学院软件研究所 http://www.jos.org.cn

(1) 通用网络场景的 INT 遥测框架

在无线网络、光网络等各种网络场景下, 类似 INT 在有线网络下的高实时性、细粒度和丰富的测量对象等卓越的测量能力对于这些网络的运维和上层应用创新具有重要意义. 然而现在 INT 机制主要为有线数据包网络设计, 仅考虑了有线网络下的测量对象和测量机制等. 尽管目前有一些研究将 INT 机制扩展到了无线网络、光网络和工业传感器网络, 但是这些方案仅将 INT 移植到相应网络场景, 没有从整体的角度考虑设计通用的 INT 遥测框架. 因此设计具有扩展性的通用场景 INT 遥测框架是未来的重要研究方向.

(2) 测量机制灵活的带内网络遥测

INT 和 AM-PM 作为两种代表性带内网络遥测方案, 尽管它们的测量机制完全不同, 但是这两种方案的测量机制都相对固定, 可扩展性差. 例如, INT 的测量机制是交换机根据遥测指令嵌入相应数据, 但是在该框架下部署其他测量算法较为困难. 因此设计测量机制灵活的带内网络遥测框架是提高带内网络遥测测量能力的重要途径. 近期提出的 LightGuardian[65]便是将带内遥测和 Sketch 方法结合, 在交换机侧利用 Sketch 测量流量数据, 通过带内遥测将拆分的 Sketch 数据结构传输到端侧进行分析, 从而同时获得了 Sketch 的轻量与全流量测量和带内测量的实时性. 因此设计测量机制灵活的带内网络遥测框架是未来研究的一个热点.

(3) 安全带内网络遥测框架

带内网络遥测随着网络可编程性的提高而迅速发展. 但是网络可编程性同时导致软件脆弱性增加, 出现后门和病毒的可能性提升. 由于带内网络遥测潜在的高资源消耗性, 恶意网络节点可能通过冒充源节点执行大量的遥测任务, 消耗网络可用带宽和设备资源. 而且恶意节点可以通过对现有遥测任务作出恶意的响应, 导致依赖于遥测结果的上层应用出现异常行为, 从而扩大了受影响范围甚至导致更严重的后果. 此外, 可编程数据平面的数据包处理灵活性可能被恶意利用, 导致业务数据包的保密性和安全性受到影响. 因此设计具备安全机制的带内网络遥测框架也是未来的重要研究方向.

(4) 带内网络遥测与人工智能结合的网络控制

人工智能技术近十年来获得了巨大的成功, 特别是在计算机视觉、自动控制和自然语言处理等方面取得了重大成果. 人工智能在解决复杂问题上表现出优异的性能, 网络研究人员也对人工智能产生了浓厚的兴趣. 人工智能技术的常用模型需要大量数据进行特征学习, 然而传统网络中收集额外信息的成本高且灵活性差, 这是在计算机网络领域引入人工智能技术的障碍之一. 带内网络遥测具有测量粒度细、灵活性高和实时性强等优点, 将带内网络遥测技术与人工智能技术结合有望极大的改善现有各种控制方案的性能. 因此带内网络遥测技术与人工智能技术结合的网络控制是未来研究的一个热点.

References:

[1] Mihailovic A, Nguengang G, Kousaridas A, Israel M, Conan V, Chochliouros I, Belesioti M, Raptis T, Wagner D, Moedeker J, Gazis V, Schaffer R, Grabner B, Alonistioti N. An approach for designing cognitive self-managed future internet. In: Proc. of the 2010 Future Network & Mobile Summit. Florence: IEEE 2010. 1–9.

[2] Zhang HK. The Application of telemetry technology in the operation and maintenance of cloud data center network. China New Telecommunications, 2021, 23(3): 44–45 (in Chinese with English abstract). [doi: 10.3969/j.issn.1673-4866.2021.03.020]

[3] CISCO. The Cisco global cloud index: Forecasts and methods, 2015–2020. 2021 (in Chinese). https://www.cisco.com/c/dam/m/zh_cn/solutions/service-provider/sp_gciwhitepaper_whitepaper_cn.pdf

[4] Morton A. Active and passive metrics and methods (with hybrid types in-between). Fremont: IETF, 2016. https://www.rfc-editor.org/rfc/rfc7799.html

[5] Dai M, Cheng G, Zhou YY. Survey on measurement methods in software-defined networking. Ruan Jian Xue Bao/Journal of Software, 2019, 30(6): 1853–1874 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5832.htm [doi: 10.13328/j.cnki.jos.005832]

[6] Bifulco R, Rétvári G. A survey on the programmable data plane: Abstractions, architectures, and open problems. In: Proc. of the 19th IEEE Int'l Conf. on High Performance Switching and Routing (HPSR). Bucharest: IEEE, 2018. 1–7. [doi: 10.1109/HPSR.2018.8850761]

[7] Han S, Jang S, Choi H, Lee H, Pack S. Virtualization in programmable data plane: A survey and open challenges. IEEE Open Journal of the Communications Society, 2020, 1: 527–534. [doi: 10.1109/OJCOMS.2020.2990182]

© 中国科学院软件研究所 http://www.jos.org.cn

[8] Bosshart P, Daly D, Gibb G, Izzard M, McKeown N, Rexford J, Schlesinger C, Talayco D, Vahdat A, Varghese G, Walker D. P4: Programming protocol-independent packet processors. ACM SIGCOMM Computer Communication Review, 2014, 44(3): 87–95. [doi: 10.1145/2656877.2656890]

[9] Li SR, Hu DY, Fang WJ, Ma SJ, Chen C, Huang HB, Zhu ZQ. Protocol oblivious forwarding (POF): Software-defined networking with enhanced programmability. IEEE Network, 2017, 31(2): 58–66. [doi: 10.1109/MNET.2017.1600030NM]

[10] P4.org. In-band network telemetry (INT) dataplane specification. 2021. https://p4.org/p4-spec/docs/INT_v2_1.pdf

[11] HUAWEI. World's first ifit pilot on the 5G transport network successfully completed by Beijing Unicom and Huawei. 2021. https://www.huawei.com/en/news/2019/6/first-ifit-pilot-5g-transport-network-Beijing-unicom-Huawei

[12] BROADCOM. Broadcom's next-generation inband telemetry solution designed for hyperscale datacent. 2021. https://www.broadcom.com/blog/broadcoms-next-generation-inband-telemetry-solution-designed-for-the-hyperscale-datacenters-is-here

[13] Intel. Intel, kaloom create P4-programmable network solutions. 2021. https://networkbuilders.intel.com/solutionslibrary/intel-kaloom-create-p4-programmable-network-solutions/

[14] CISCO. Cisco nexus dashboard insights. 2021. https://www.cisco.com/c/en/us/products/data-center-analytics/nexus-insights/index.html

[15] Yu ML. Network telemetry: Towards a top-down approach. ACM SIGCOMM Computer Communication Review, 2019, 49(1): 11–17. [doi: 10.1145/3314212.3314215]

[16] Tan LZ, Su W, Zhang W, Lv JH, Zhang ZY, Miao JY, Liu XX, Li N. In-band network telemetry: A survey. Computer Networks, 2021, 186: 107763. [doi: 10.1016/j.comnet.2020.107763]

[17] Manzanares-Lopez P, Muñoz-Gea JP, Malgosa-Sanahuja J. Passive in-band network telemetry systems: The potential of programmable data plane on network-wide telemetry. IEEE Access, 2021, 9: 20391–20409. [doi: 10.1109/ACCESS.2021.3055462]

[18] Kim C, Sivaraman A, Katta N, Bas A, Dixit A, Wobker LJ. In-band network telemetry via programmable dataplanes. In: Proc. of the 2015 ACM SIGCOMM Conf. Posters and Demos. London: ACM, 2015. 15.

[19] Van Tu N, Hyun J, Hong JWK. Towards ONOS-based SDN monitoring using in-band network telemetry. In: Proc. of the 19th Asia-Pacific Network Operations and Management Symp. Seoul: IEEE, 2017. 76–81. [doi: 10.1109/APNOMS.2017.8094182]

[20] Gulenko A, Wallschläger M, Kao O. A practical implementation of in-band network telemetry in open vSwitch. In: Proc. of the 7th IEEE Int'l Conf. on Cloud Networking. Tokyo: IEEE, 2018. 1–4. [doi: 10.1109/CloudNet.2018.8549431]

[21] Tang SF, Li DY, Niu B, Peng JQ, Zhu ZQ. Sel-INT: A runtime-programmable selective in-band network telemetry system. IEEE Trans. on Network and Service Management, 2020, 17(2): 708–721. [doi: 10.1109/TNSM.2019.2953327]

[22] Fioccola G, Capello A, Cociglio M, Castaldelli L, Chen M, Zheng L, Mirsky G, Mizrahi T. Alternate-marking method for passive and hybrid performance monitoring. Fremont: IETF, 2018. https://datatracker.ietf.org/doc/draft-ietf-ippm-alt-mark/14/

[23] Mizrahi T, Navon G, Fioccola G, Cociglio M, Chen M, Mirsky G. AM-PM: Efficient network telemetry using alternate marking. IEEE Network, 2019, 33(4): 155–161. [doi: 10.1109/MNET.2019.1800152]

[24] Riesenberg A, Kirzon Y, Bunin M, Galili E, Navon G, Mizrahi T. Time-multiplexed parsing in marking-based network telemetry. In: Proc. of the 12th ACM Int'l Conf. on Systems and Storage. Haifa: ACM, 2019. 80–85. [doi: 10.1145/3319647.3325837]

[25] Basat RB, Ramanathan S, Li YL, Antichi G, Yu MN, Mitzenmacher M. PINT: Probabilistic in-band network telemetry. In: Proc. of the 2020 Annual Conf. of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication. Virtual: ACM, 2020. 662–680. [doi: 10.1145/3387514.3405894]

[26] van Tu N, Hyun J, Kim GY, Yoo JH, Hong JWK. INTCollector: A high-performance collector for in-band network telemetry. In: Proc. of the 14th Int'l Conf. on Network and Service Management. Rome: IEEE, 2018. 10–18.

[27] Vestin J, Kassler A, Bhamare D, Grinnemo KJ, Andersson JO, Pongracz G. Programmable event detection for in-band network telemetry. In: Proc. of the 8th IEEE Int'l Conf. on Cloud Networking (CloudNet). Coimbra: IEEE, 2019. 1–6. [doi: 10.1109/CloudNet47604.2019.9064137]

[28] Wang SY, Chen YR, Li JY, Hu HW, Tsai JA, Lin YB. A bandwidth-efficient INT system for tracking the rules matched by the packets of a flow. In: Proc. of the 2019 IEEE Global Communications Conf. (GLOBECOM). Waikoloa: IEEE, 2019. 1–6. [doi: 10.1109/GLOBECOM38437.2019.9013581]

[29] Suh D, Jang S, Han S, Pack S, Wang XF. Flexible sampling-based in-band network telemetry in programmable data plane. ICT Express, 2020, 6(1): 62–65. [doi: 10.1016/j.icte.2019.08.005]

[30] Kim Y, Suh D, Pack S. Selective in-band network telemetry for overhead reduction. In: Proc. of the 7th IEEE Int'l Conf. on Cloud Networking (CloudNet). Tokyo: IEEE, 2018. 1–3. [doi: 10.1109/CloudNet.2018.8549351]

[31] Niu B, Kong JW, Tang SF, Li YC, Zhu ZQ. Visualize your IP-over-optical network in realtime: A P4-based flexible multilayer in-band network telemetry (ML-INT) system. IEEE Access, 2019, 7: 82413–82423. [doi: 10.1109/ACCESS.2019.2924332]

© 中国科学院软件研究所 http://www.jos.org.cn

[32] Pan T, Song EG, Jia CH, Cao WD, Huang T, Liu B. Lightweight network-wide telemetry without explicitly using probe packets. In: Proc. of the 2020 IEEE Conf. on Computer Communications Workshops (INFOCOM WKSHPS). Toronto: IEEE, 2020. 1354–1355. [doi: 10.1109/INFOCOMWKSHPS50562.2020.9162684]

[33] Chowdhury SR, Boutaba R, François J. LINT: Accuracy-adaptive and lightweight in-band network telemetry. In: Proc. of the 2021 IFIP/IEEE Int'l Symp. on Integrated Network Management (IM). Bordeaux: IEEE, 2021. 349–357.

[34] Castanheira L, Parizotto R, Schaeffer-Filho AE. FlowStalker: Comprehensive traffic flow monitoring on the data plane using P4. In: Proc. of the 2019 IEEE Int'l Conf. on Communications (ICC). Shanghai: IEEE, 2019. 1–6. [doi: 10.1109/ICC.2019.8761197]

[35] Hohemberger R, Lorenzon AF, Rossi F, Luizelli MC. Optimizing distributed network monitoring for NFV service chains. IEEE Communications Letters, 2019, 23(8): 1332–1336. [doi: 10.1109/LCOMM.2019.2922184]

[36] Marques JA, Luizelli MC, da Costa Filho RIT, Gaspary LP. An optimization-based approach for efficient network monitoring using in-band network telemetry. Journal of Internet Services and Applications, 2019, 10(1): 12. [doi: 10.1186/s13174-019-0112-0]

[37] Hohemberger R, Castro AG, Vogt FG, Mansilha RB, Lorenzon AF, Rossi FD, Luizelli MC. Orchestrating in-band data plane telemetry with machine learning. IEEE Communications Letters, 2019, 23(12): 2247–2251. [doi: 10.1109/LCOMM.2019.2946562]

[38] Bhamare D, Kassler A, Vestin J, Khoshkholghi MA, Taheri J. IntOpt: In-band network telemetry optimization for NFV service chain monitoring. In: Proc. of the 2019 IEEE Int'l Conf. on Communications (ICC). Shanghai: IEEE, 2019. 1–7. [doi: 10.1109/ICC.2019.8761722]

[39] Anand M, Subrahmaniam R, Valiveti R. POINT: An intent-driven framework for integrated packet-optical in-band network telemetry. In: Proc. of the 2018 IEEE Int'l Conf. on Communications (ICC). Kansas City: IEEE, 2018. 1–6. [doi: 10.1109/ICC.2018.8422785]

[40] Tang SF, Kong JW, Niu B, Zhu ZQ. Programmable multilayer INT: An enabler for AI-assisted network automation. IEEE Communications Magazine, 2020, 58(1): 26–32. [doi: 10.1109/MCOM.001.1900365]

[41] Karaagac A, De Poorter E, Hoebeke J. In-band network telemetry in industrial wireless sensor networks. IEEE Trans. on Network and Service Management, 2020, 17(1): 517–531. [doi: 10.1109/TNSM.2019.2949509]

[42] Haxhibeqiri J, Moerman I, Hoebeke J. Low overhead, fine-grained end-to-end monitoring of wireless networks using in-band telemetry. In: Proc. of the 15th Int'l Conf. on Network and Service Management. Halifax: IEEE, 2019. 1–5. [doi: 10.23919/CNSM46954.2019.9012678]

[43] Haxhibeqiri J, Isolani PH, Marquez-Barja JM, Moerman I, Hoebeke J. In-band network monitoring technique to support SDN-based wireless networks. IEEE Trans. on Network and Service Management, 2021, 18(1): 627–641. [doi: 10.1109/TNSM.2020.3044415]

[44] Zaballa EO, Franco D, Thomsen SE, Higuero M, Wessing H, Berger MS. Towards monitoring hybrid next-generation software-defined and service provider MPLS networks. Computer Networks, 2021, 191: 107960. [doi: 10.1016/j.comnet.2021.107960]

[45] Li YL, Miao R, Liu HH, Zhuang Y, Feng F, Tang LB, Cao Z, Zhang M, Kelly F, Alizadeh M, Yu ML. HPCC: High precision congestion control. In: Proc. of the 2019 ACM Special Interest Group on Data Communication. Beijing: ACM, 2019. 44–58. [doi: 10.1145/3341302.3342085]

[46] Tang YN, Wu YX, Cheng G, Xu ZW. Intelligence enabled SDN fault localization via programmable in-band network telemetry. In: Proc. of the 20th IEEE Int'l Conf. on High Performance Switching and Routing (HPSR). Xi'an: IEEE, 2019. 1–6. [doi: 10.1109/HPSR.2019.8808121]

[47] Karaagac A, De Poorter E, Hoebeke J. Alternate marking-based network telemetry for industrial WSNs. In: Proc. of the 16th IEEE Int'l Conf. on Factory Communication Systems. Porto: IEEE, 2020. 1–8. [doi: 10.1109/WFCS47810.2020.9114490]

[48] Kagami NS, Da Costa Filho RIT, Gaspary LP. CAPEST: Offloading network capacity and available bandwidth estimation to programmable data planes. IEEE Trans. on Network and Service Management, 2020, 17(1): 175–189. [doi: 10.1109/TNSM.2019.2934316]

[49] Liang JZ, Bi J, Zhou Y, Zhang C. In-band network function telemetry. In: Proc. of the 2018 ACM SIGCOMM Conf. on Posters and Demos. Budapest: ACM, 2018. 42–44. [doi: 10.1145/3234200.3234236]

[50] Hyun J, Van Tu N, Yoo JH, Hong JWK. Real-time and fine-grained network monitoring using in-band network telemetry. Int'l Journal of Network Management, 2019, 29(6): e2080. [doi: 10.1002/nem.2080]

[51] Liu ZZ, Bi J, Zhou Y, Wang YY, Lin YX. NetVision: Towards network telemetry as a service. In: Proc. of the 26th IEEE Int'l Conf. on Network Protocols (ICNP). Cambridge: IEEE, 2018. 247–248. [doi: 10.1109/ICNP.2018.00036]

[52] Liu ZZ, Bi J, Zhou Y, Wang YY, Lin YSX. Paradigm for proactive telemetry based on P4. Journal on Communications, 2018, 39(S1): 162–169 (in Chinese with English abstract). [doi: 10.11959/j.issn.1000-436x.2018181]

[53] Pan T, Song EG, Bian ZZ, Lin XC, Peng XY, Zhang J, Huang T, Liu B, Liu YJ. INT-path: Towards optimal path planning for in-band network-wide telemetry. In: Proc. of the 2019 IEEE Conf. on Computer Communications. Paris: IEEE, 2019. 487–495. [doi: 10.1109/

© 中国科学院软件研究所 http://www.jos.org.cn

INFOCOM.2019.8737529]

[54] Castro AG, Lorenzon AF, Rossi FD, Da Costa Filho RIT, Ramos FMV, Rothenberg CE, Luizelli MC. Near-optimal probing planning for in-band network telemetry. IEEE Communications Letters, 2021, 25(5): 1630–1634. [doi: 10.1109/LCOMM.2021.3053485]

[55] Yang FC, Quan W, Cheng N, Xu ZH, Zhang X, Gao DY. Fast-INT: Light-weight and efficient in-band network telemetry in programmable data plane. In: Proc. of the 92nd IEEE Vehicular Technology Conf. (VTC2020-Fall). Victoria: IEEE, 2020. 1–5. [doi: 10.1109/VTC2020-Fall49728.2020.9348823]

[56] Katta N, Ghag A, Hira M, Keslassy I, Bergman A, Kim C, Rexford J. Clove: Congestion-aware load balancing at the virtual edge. In: Proc. of the 13th Int'l Conf. on Emerging Networking Experiments and Technologies. Incheon: ACM, 2017. 323–335. [doi: 10.1145/3143361.3143401]

[57] Lim J, Nam S, Yoo JH, Hong JWK. Best nexthop load balancing algorithm with inband network telemetry. In: Proc. of the 16th Int'l Conf. on Network and Service Management (CNSM). Izmir: IEEE, 2020. 1–7. [doi: 10.23919/CNSM50824.2020.9269053]

[58] Hyun J, Van Tu N, Hong JWK. Towards knowledge-defined networking using in-band network telemetry. In: Proc. of the 2018 IEEE/IFIP Network Operations and Management Symp. Taipei: IEEE, 2018. 1–7. [doi: 10.1109/NOMS.2018.8406169]

[59] Yao HP, Mai T, Xu XB, Zhang PY, Li MZ, Liu YJ. NetworkAI: An intelligent network architecture for self-learning control strategies in software defined networks. IEEE Internet of Things Journal, 2018, 5(6): 4319–4327. [doi: 10.1109/JIOT.2018.2859480]

[60] Isolani PH, Haxhibeqiri J, Moerman I, Hoebeke J, Marquez-Barja JM, Granville LZ, Latre S. An SDN-based framework for slice orchestration using in-band network telemetry in IEEE 802.11. In: Proc. of the 6th IEEE Conf. on Network Softwarization (NetSoft). Ghent: IEEE, 2020. 344–346. [doi: 10.1109/NetSoft48620.2020.9165358]

[61] Choi N, Jagadeesan L, Jin Y, Mohanasamy NN, Rahman MR, Sabnani K, Thottan M. Run-time performance monitoring, verification, and healing of end-to-end services. In: Proc. of the 2019 IEEE Conf. on Network Softwarization (NetSoft). Paris: IEEE, 2019. 30–35. [doi: 10.1109/NETSOFT.2019.8806660]

[62] Jia CH, Pan T, Bian ZZ, Lin XC, Song EG, Xu C, Huang T, Liu YJ. Rapid detection and localization of gray failures in data centers via in-band network telemetry. In: Proc. of the 2020 IEEE/IFIP Network Operations and Management Symp. Budapest: IEEE, 2020. 1–9. [doi: 10.1109/NOMS47738.2020.9110326]

[63] Nam S, Lim J, Yoo JH, Hong JWK. Network anomaly detection based on in-band network telemetry with RNN. In: Proc. of the 2020 IEEE Int'l Conf. on Consumer Electronics—Asia (ICCE-Asia). Seoul: IEEE, 2020. 1–4. [doi: 10.1109/ICCE-Asia49877.2020.9276768]

[64] Fang CR, Liu HY, Miao M, Ye J, Wang L, Zhang WS, Kang DX, Lyv B, Cheng P, Chen JM. VTrace: Automatic diagnostic system for persistent packet loss in cloud-scale overlay network. In: Proc. of the 2020 Annual Conf. of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication. ACM, 2020. 31–43. [doi: 10.1145/3387514.3405851]

[65] Zhao YK, Yang KC, Liu ZR, Yang T, Chen L, Liu SY, Zheng NQ, Wang RX, Wu HB, Wang Y, Zhang N. LightGuardian: A full-visibility, lightweight, in-band telemetry system using sketchlets. In: Proc. of the 18th USENIX Symp. on Networked Systems Design and Implementation. USENIX Association, 2021. 991–1010.

附中文参考文献:

[2] 张洪凯. 遥测技术在云数据中心网络运维中的应用. 中国新通信, 2021, 23(3): 44–45. [doi: 10.3969/j.issn.1673-4866.2021.03.020]

[3] CISCO. 思科全球云指数: 预测和方法, 2015–2020年. 2021. https://www.cisco.com/c/dam/m/zh_cn/solutions/service-provider/sp_gciwhitepaper_whitepaper_cn.pdf

[5] 戴冕, 程光, 周余阳. 软件定义网络的测量方法研究. 软件学报, 2019, 30(6): 1853–1874. http://www.jos.org.cn/1000-9825/5832.htm [doi: 10.13328/j.cnki.jos.005832]

[52] 刘争争, 毕军, 周禹, 王旸旸, 林耘森箫. 基于P4的主动网络遥测机制. 通信学报, 2018, 39(S1): 162–169. [doi: 10.11959/j.issn.1000-436x.2018181]

© 中国科学院软件研究所 http://www.jos.org.cn

吕鸿润(1998－), 男, 硕士生, 主要研究领域为数据中心网络, 流量调度.

江勇(1975－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为计算机网络体系结构, 下一代互联网, 人工智能网络.

李清(1985－), 男, 博士, 副研究员, CCF 专业会员, 主要研究领域为网络架构, 边缘计算, 传输优化.

李伟超(1979－), 男, 博士, 副研究员, CCF 专业会员, 主要研究领域为网络性能测量.

沈耿彪(1991－), 男, 硕士生, 主要研究领域为数据中心, 负载均衡, 流量调度, 协议栈, 软件定义网络, 内容分发网络.

刘凯(1978－), 男, 工程师, 主要研究领域为路由与交换, 高性能转发面, 网络操作系统.

周建二(1986－), 男, 博士, 主要研究领域为网络测量, 传输优化, 云虚拟网络.

齐竹云(1983－), 女, 高级工程师, CCF 高级会员, 主要研究领域为软件定义网络, 区块链, 软件工程.

© 中国科学院软件研究所 http://www.jos.org.cn