服务感知网络

任 杰[1],王洪超[1*],王钦定[1],熊 豪[1],杨 冬[1],张宏科[1],谭 斌[2],郭 勇[2],黄光平[3]
(1. 北京交通大学电子信息工程学院,北京 100044;2. 中兴通讯股份有限公司,上海 201203;
3. 中兴通讯股份有限公司,江苏南京 210012)

摘 要: 随着人工智能、大数据、云计算等新技术和新业务的蓬勃发展,其中起支撑性作用的互联网应用间的通信范式也逐步由传统的“请求—获取”一段式演进到复杂的“请求—计算—获取”二段式. 新的通信范式不仅要求网络提供传统的数据传输与信息传递的渠道,还要求网络将计算以互联网公共服务的形式对外开放,深化计算与网络融合,促进社会的数字经济发展. 然而,在当前TCP/IP架构为代表的网络体系架构中,解析发现、网络感知、路由计算等关键机制存在一些弊端,使其难以成为面向通信计算融合新型应用的数据基座. 本文提出服务感知网络(Service Aware Network,SAN)实现架构,该架构系统性地改变了现有主机互联的设计,为算网融合等新应用场景提供全面、开放、有保障的内生服务互联支持. 在SAN中,语义服务标识使用户终端能根据服务类型发起位置无关的高效端到端连接,算网需求感知为网络主动获取算力与带宽需求提供有效途径,两级两层路由既实现了联合算网二维资源的路由编排,又保障了在不同的网络开放程度下SAN部署的可行性和有效性. SAN的重要优势是在设计之初就系统性考虑了与现有网络基础设施的兼容和共存,通过增量部署的方式避免对现有网络进行重大改造,在不影响现网已有功能的同时实现全新机制,推动网络从主机互联到服务互联的平滑演进. 实验表明SAN路由机制与基准路由算法相比分别降低了35.4%和17.5%的服务完成时间上限,同时实现了更为均衡高效的算网资源利用.

关键词: 服务感知网络;网络体系架构;服务标识;算网需求感知;路由计算
基金项目: 国家重点研发计划(No.2022YFB2901302);国家自然科学基金(No.62394325)
中图分类号: TN913.21 **文献标识码:** A **文章编号:** 0372-2112(2025)02-0371-14
电子学报URL: http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240400

Service Aware Network

REN Jie[1], WANG Hong-chao[1*], WANG Qin-ding[1], XIONG Hao[1], YANG Dong[1],
ZHANG Hong-ke[1], TAN Bin[2], GUO Yong[2], HUANG Guang-ping[3]
(1. *School of Electronic and Information Engineering*, *Beijing Jiaotong University*, *Beijing* 100044, *China*; 2. *ZTE Corporation*, *Shanghai* 201203, *China*; 3. *ZTE Corporation*, *Nanjing*, *Jiangsu* 210012, *China*)

Abstract: With the vigorous development of new technologies and services such as artificial intelligence, big data, cloud computing, etc., the communication paradigm between Internet applications, which plays a supporting role, has evolved from the traditional “request-get” one-stage to the complex “request-compute-get” two-stage. The new communication paradigm requires the network not only to provide the data transmission channels but also to assume the role of opening up computation in the form of public services on the Internet, thus deepening the integration of computing and network and promoting the development of the digital economy. However, in the network architecture represented by the current TCP/IP architecture, there are some drawbacks in the core mechanisms, including resolution and discovery, network awareness, and route computation, which make it difficult to become a foundation for new applications oriented to the convergence of communication and computing. In this paper, we propose a new implementation architecture named service aware network (SAN), which systematically changes the existing design of host interconnection to provide comprehensive, open, and guaranteed endogenous service interconnection support for new scenarios such as computing network convergence. In SAN, a semantic service identifier enables user terminals to initiate location-independent connections based on service types. Computing and network demand awareness provides a proactive method for the network to obtain the requirements of requests. “Two-stage two-layer” routing not only realizes the routing considered two-dimensional computing and bandwidth resourc-

收稿日期:2024-05-06;修回日期:2024-11-01;责任编辑:朱梅玉
*通讯作者:王洪超

es but also guarantees the feasibility and validity of SAN deployment under different degrees of network openness. To quickly implement the concept of SAN, the SAN architecture avoids major transformation of the existing network through incremental deployment. It realizes new mechanisms without affecting the functions of the existing network and promotes the smooth evolution of the network from host interconnection to service interconnection. Experiments show that the SAN routing mechanism reduces the upper bound of service completion time by 35.4% and 17.5%, respectively, compared with the benchmark routing algorithms, and achieves more balanced and efficient computing and network resource usage.

Key words: service aware network; network architecture; service identifier; network and computing demand awareness; routing

Foundation Item(s): National Key Research and Development Program of China (No.2022YFB2901302); National Natural Science Foundation of China (No.62394325)

1 引言

作为数据传输与信息传递的通道,网络在数字经济中发挥着不可或缺的基础性作用.从互联网诞生至今,网络技术的发展演进与应用侧的需求变化是密不可分的.如图1所示,早期计算机应用如电子邮件、文件传输等功能较为单一,其通信需求也只聚焦于可达性与连通性,因此网络仅为应用提供通信连接,网络与应用保持相对隔离.随着流媒体、即时通讯等新模式、新业态的出现,应用对移动性、实时性以及万物互联的需求催生了3G/4G/5G等移动互联网络技术,使得网络的通信保障功能进一步增强.近年来,人工智能、大数据、云计算等新技术和新业务蓬勃发展,其中起支撑性作用的互联网应用间的通信范式也由传统的"请求—获取"一段式演进到复杂的"请求—计算—获取"二段式[1,2].新的通信范式要求网络不仅需要提供一般性的通信渠道功能,还要将计算过程纳入管控范围[3].随着算力供给泛在化和算力架构容器化、服务化的发展,计算资源逐渐从封闭私有向开放公用转变,这需要网络将计算以互联网共性服务的形式对外开放,也指引了网络技术创新需要由主机互联向服务互联转变,深化计算与网络融合,促进社会的数字经济发展[4,5].

然而,在当前TCP/IP架构为代表的网络体系架构

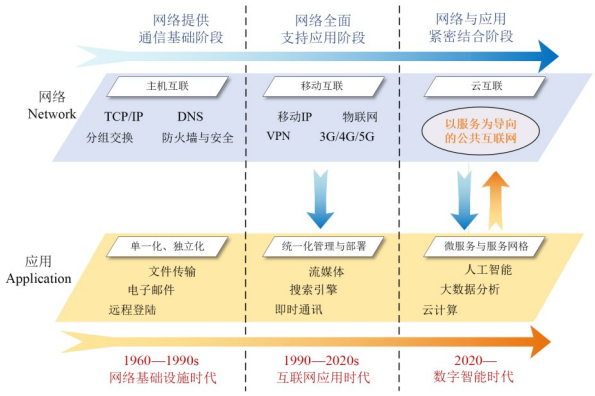

图1 网络与应用的演进

中,解析发现、网络感知、路由计算等关键机制存在着一些弊端,使其难以成为面向通信计算融合新型应用的数据基座.

(1)解析发现.当前网络使用域名解析系统(Domain Name System,DNS)解析URL标识的资源所在的IP地址,确定通信对端的物理实体及位置,用户在发起通信连接前就已经决定了通信的目的地,这种IP与位置紧密耦合的设计使得网络无法基于应用需求和网络态势的变化实现灵活的业务提供与部署[6].

(2)网络感知.现有如网络遥测、流量分析等网络感知技术的核心机理是对带宽、时延等流量特征进行统计与分析,并以此为根据改善网络性能和用户体验,然而用户的算力需求通常封装在应用层协议中,与流量特征仅存在弱相关性,现有通用的网络感知技术难以主动获取用户多样化的算力需求,为网络管理提供有效依据[7].

(3)路由计算.现有网络采用网段路由和主机路由等基于IP的路由转发模式,使用网络标识如IP五元组作为路由索引,使得网络设备难以获取应用语义和需求,调度、路由等网络基本要素难以按需动态调整与重构,阻碍了计算、网络资源的协同规划和高效适配[8].

针对上述问题,国内外学术界和产业界都开展了大量的研究工作,以增强网络对上层应用需求的理解和支持,进一步打通应用与网络的信息隔离.例如,信息中心网络作为一种新的网络架构被提出,该架构基于数据的命名描述进行数据共享和交换,用户在发起通信时无需关注通信的位置,从而有效利用网络资源,减少不必要的数据传输[9~11].Zhang等人[12]提出了基于深度学习的流量识别器,用于在接入网关侧实现实时的网络应用感知,并进一步提出了一种网络流量预测方案,为网络管理提供准确的网络状态预测.Poularakis等人[13]研究了具有多维约束的边缘计算网络中服务布局和路由的联合优化问题并提出了计算需求感知的路由算法,以最大限度地提高服务的请求数量.但是现有研究只能解决当前网络架构的部分机制弊端,未能从整体协同角度解决问题,例如信息中心网络使用数据

名称或标识符进行寻址,仅简化了解析发现中地址转换的复杂度.

为了能从根本上系统地解决上述问题,本文基于前期标识网络关于服务标识解析[14]和服务网络智慧协同映射[15]等研究基础,提出一种新型服务感知网络(Service Aware Network,SAN)实现架构,从解析发现、网络感知、路由计算等多方面机制改变现有主机互联的设计,为算网融合等新兴应用场景提供全面、开放、有保障的内生服务互联支持. 该架构主要包括图2所示的3项关键技术贡献.

(1)语义服务标识. 作为计算类应用的抽象表征,服务标识与具体计算服务之间存在语义关联,使用户终端能根据具体服务类型发起位置无关的高效端到端连接,将服务发现、服务治理和服务请求编排的能力赋予网络本身,为网络管控计算资源提供可行的机制基础.

(2)算网需求感知. 设计一种基于语义服务标识的算网需求感知方法,该方法通过提取服务标识中的计算服务性能字段进行统计分析,使SAN设备能使用通用的网络感知技术主动获取算力需求,解决了网络设备感知业务算力需求的难题.

(3)两级两层路由. 设计两级两层路由机制,“两级”路由通过复用当前服务器设备实现服务请求报文的交付,“两层”路由通过复用当前网络基础设施实现携带服务标识报文的转发,两级两层路由既实现了联合算网二维资源的路由编排,又保障了在不同的网络开放程度下SAN部署的可行性和有效性.

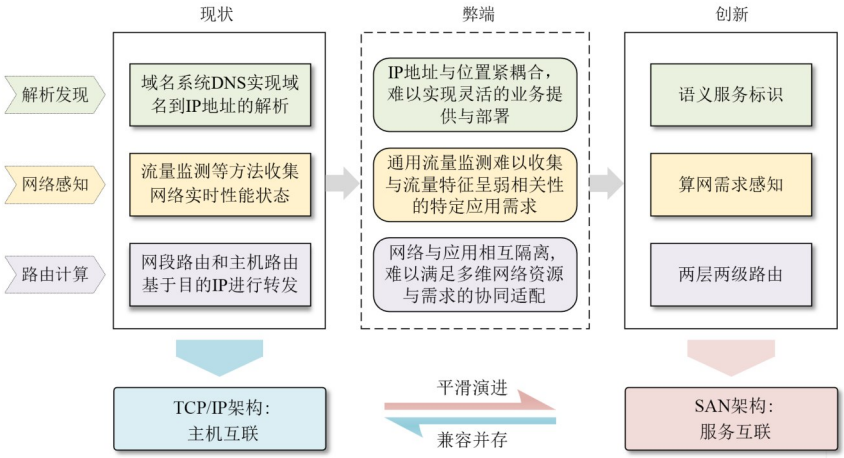

图2　网络体系架构现状与SAN创新设计

由于对现有网络基础设施改造太大,以往多种新型网络技术难以实施部署. SAN在设计之初就系统性考虑了兼容和共存难题,通过增量部署的方式避免对现有网络进行重大改造,在不影响现网已有基础设施和功能的同时实现全新网络机制,降低部署成本和时间周期. 具体而言,语义服务标识及其解析映射系统与现有DNS系统兼容并存,分别面向计算密集型应用和普通应用提供解析发现功能,引领未来多样化应用创新. 算网需求感知可深度复用现有网络感知技术,同时扩展相应技术的功能与适用范围. 两层路由避免对现有路由设备的改造,两级路由避免对现有服务器的改造. 因此,SAN的增量部署方式和与现网兼容的特点,可以最大限度地减少对现有网络设备及协议的改动,快速完成网络支持计算融合的改造,推动网络从传统的主机互联到服务互联的平滑演进.

2　服务感知网络

本文提出的服务感知网络(SAN)架构如图3所示. SAN架构在层次上可划分为基础设施层与服务层,服务层架构于基础设施层之上,二者设备相互解耦. 服务层引入服务标识实现终端用户和服务提供节点之间的高效端到端通信连接,基础设施层基于现有的网络设备转发业务数据. 服务层和基础设施层分别独立部署了层内控制器执行网络集中控制功能,控制器间建立信息通道以支持控制信令的跨层交互.

为优化基于IP的互联架构在寻址、路由、SLA保障等方面的不足,SAN引入了语义服务标识(Service IDentifier,SID)作为跨应用、跨设备、跨网络的唯一服务命名. SID用于统一描述接入SAN的各项服务,其语义仅与服务类型有关,而与服务的请求者或提供者无关. 用户在发起服务请求时,可直接复用IPv6报文生成携带SID的服务请求报文(如图4所示),其中SID可放置于IPv6报文中的目的IP字段,前64 bits为固定前缀,后64 bits基于服务类型生成.

具体地,部署于服务层的SAN设备组件包括SAN智能管控系统、SAN入口网关和SAN出口网关.

(1)SAN智能管控系统是SAN的核心组件,该组件会对本管理域内网络设备与服务进行统一管理,同时

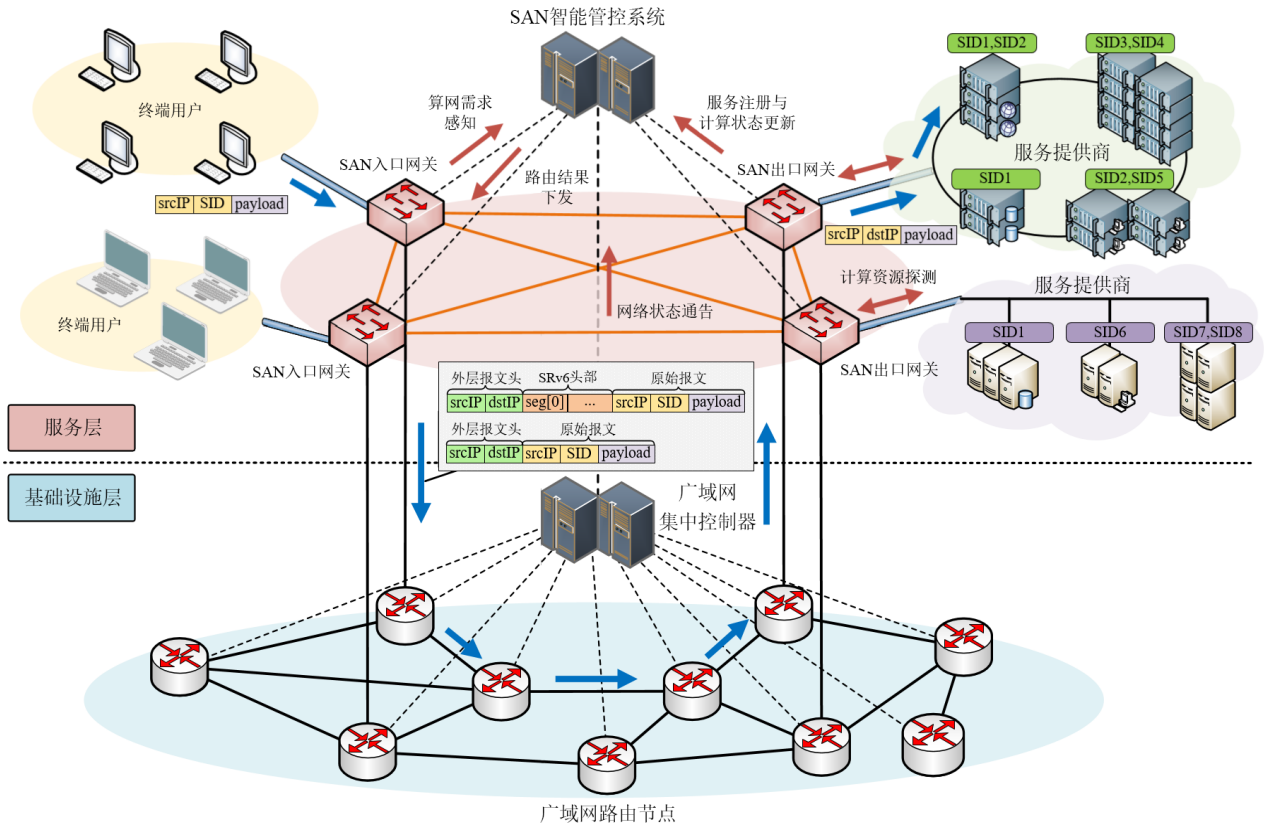

图3 服务感知网络(SAN)架构

图4 携带SID的服务请求报文结构

还负责对用户服务请求进行按需编排.SAN智能管控系统会接收并维护各个服务提供商资源池内的计算资源状态和来自广域网集中控制器定期探测的广域网路由节点的网络资源状态,同时还会定期收集每个SAN入口网关以服务为粒度的带宽和算力需求信息.基于上述2类信息,智能管控系统能够联合网络与计算资源按需为接入的服务请求编排最优的转发路径,并将编排结果通过控制通道下发到SAN入口网关.

(2)SAN入口网关负责接收和解析来自终端用户的服务请求报文,以SID为索引查询部署在本地的路由条目,并将相关路由信息封装到服务请求报文中,之后将其转发到广域网路由节点进行路由转发.在报文封装过程中,SAN入口网关不仅可以为原始报文封装新的外层IPv6报文头,还可以通过路由扩展头(Segment Routing Header,SRH)的形式显式标识逐跳的转发路径以实现流量工程,避免了现有广域网中提供流量转发的中继路由设备为解析服务请求报文所需的协议栈升级.

(3)SAN出口网关负责接收和解析来自广域网的服务请求报文,并进行2次路由.在报文解析过程中,SAN出口网关会根据报文中的SID选择适合的服务提供节点,并使用该节点的IP地址和对应应用程序的端口号替换原报文中的目的IP和目的端口号字段,以通用IP报文的形式完成服务请求的交付.

对于基础设施层而言,SAN复用了现有的网络基础设施,广域网路由节点可根据运营商内部的路由策略基于目的IP地址执行主机路由或网段路由.若广域网的感知能力向SAN智能管控系统开放,广域网集中控制器会定期收集基础设施层节点的网络状态,即网络资源使用情况,并将网络状态通过控制器间的信息通道通告给SAN智能管控系统.SAN智能管控系统可进行逐跳的路径编排,广域网路由节点可按照服务请求报文的SRH扩展头规定的路径进行转发.此外,若

基础设施层节点还支持特定的流量策略部署,且基础设施层管控能力向SAN智能管控系统开放,SAN智能管控系统可在路由策略的基础上增加流量策略的调度,如带宽保障策略、基于时隙的时延确定性保障策略等,并将调度结果通过广域网控制器部署到广域网路由节点.

基于上述的SAN服务层和基础设施层组件,作为服务发起方的终端用户只需要根据服务类型和需求在本地生成或通过查询得到对应的SID,并直接使用SID发起通信连接请求,而无需关心服务提供节点的位置或IP. 作为服务提供方,提供节点接收到的报文为SAN出口网关重新封装的通用IP报文且不含有SID信息,因此服务提供节点可直接进行报文的处理而无需对协议栈进行改造. 终端用户通过源IP、SID、源端口号和目的端口号(SAN出口网关的代理端口号)维持会话连接,服务提供节点通过源IP、目的IP、源端口号和目的端口号维持会话连接. 为了保持双向连接信息的一致性,SAN出口网关不仅需要在转发服务请求报文时将SID替换为IP地址实例、将目的端口号替换为对应应用程序的端口号,还需要在收到服务提供节点返回的回复报文,如传输控制协议(Transmission Control Protocol,TCP)的ACK(Acknowledgment)报文时,针对报文中的源IP字段执行由IP地址实例到SID的映射、执行应用程序端口号到原始代理端口号的映射.

SAN的基本工作流程可以分为初始化阶段、控制阶段和业务阶段(如图5所示).

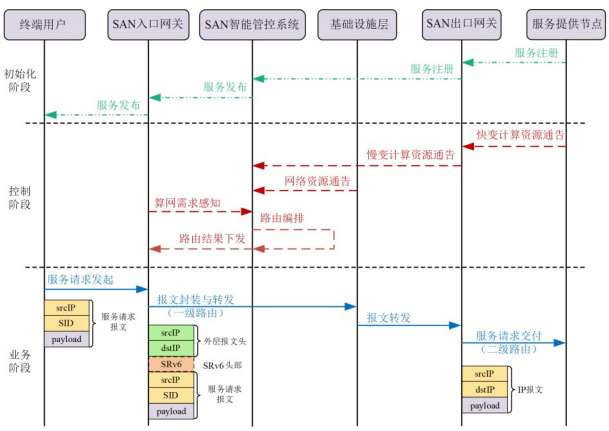

图5　SAN业务流程

(1)初始化阶段. 服务提供商将服务以统一的接口形式发布,并由资源池内部署的服务代理执行服务注册,向所连接的SAN出口网关通告本地资源池内可用服务对应的SID、节点IP地址和端口号组成的服务列表. SAN出口网关会将服务提供商资源池内所有节点的可用服务列表维护在本地,同时向SAN智能管控系统通告服务提供商本地资源池内所有可用服务的SID. SAN智能管控系统会维护SAN管控范围内的所有SAN出口网关和对应的可用服务信息,并将包含所有可用服务SID的列表下发给SAN入口网关和终端用户. 当发生服务迁移时,新旧服务提供节点都需要向SAN智能管控系统更新可用服务信息.

(2)控制阶段. 服务提供节点定期向SAN出口网关通告本节点的计算资源使用情况以及服务实例的生命周期健康状况. 由于计算资源的使用通常具有实时性强的特性,为防止该特性影响路由的收敛,SAN出口网关会将收集到的快变计算资源进行二次统计与处理,以更长的周期生成慢变计算资源向SAN智能管控系统通告. 管控系统还会接收来自广域网集中控制器的网络资源通告,包括拓扑连接、带宽占用、丢包率等信息. 同时,SAN入口网关会统计来自本局域网的终端用户所产生的全部服务请求的算力需求和带宽需求,定期将该算网需求信息通告给SAN智能管控系统. 在接收到资源使用信息和需求信息后,SAN智能管控系统会按需进行路由编排,为每一个SAN入口网关生成以SID为索引的路由条目,并将路由条目下发到SAN入口网关.

(3)业务阶段. 在发起业务时,终端用户会将服务请求的内容封装到报文中,并根据所需服务的类型在本地生成对应的SID,该SID会被填入报文的目的IP字段完成报文的封装,封装后的原始报文由终端用户转发到SAN入口网关. SAN入口网关接收到报文后,首先解析报文中的SID,并以SID为索引查询本地的路由表,然后在原始报文之外封装一层IP报文头,报文头的源IP为SAN入口网关的IP地址,目的IP为匹配到的路由表项中的目的SAN出口网关的IP地址. 若路由表项中包含逐跳的转发路径信息,该信息会以SRv6的形式封装到SRH扩展中. 封装后的IP报文会发送到广域网路由节点进行转发,最终到达对应的SAN出口网关. SAN出口网关接收到服务请求报文后去除外层IP报文头并解析报文中的SID,然后根据服务提供商局域网的路由策略选择具体的服务提供节点,使用该节点的IP地址和该服务对应的端口号替换原始服务请求报文中的SID和目的端口号字段,并将报文转发给服务提供节点完成交付. 同时SAN出口网关还会在本地生成路由表项,后续来自同一服务请求的报文可以直接基于源IP和SID匹配路由表交付到同一服务提供节点.

SAN架构通过引入SID使用户终端能发起位置和归属无关的业务连接,实现从网络视角构建一体化的服务感知与服务供给,为面向共性计算服务的全局业务编排和优化方案提供架构基础. 同时,基于当前的主机/网段路由和SRv6路由机制,SAN架构避免了在现有网络基础设施增加服务标识解析功能带来的开销,促进网络从传统模式下的主机互联到服务互联的平滑演

进. 值得说明的是,SAN所需的封装与解封包头操作往往会增加网关的处理时延,同时外层包头的存在还会损伤一定的吞吐量. 通过评估,相比于端到端时延和整体吞吐量,SAN仅会带来微小的时延与吞吐量开销,不会显著影响整个网络的运行效率. 未来,随着SAN架构的推广应用,端侧设备和交换设备会逐步向支持IP/SID双栈更新迭代,SAN将极大降低当前方案的时延与吞吐量开销. 此外,在终端移动的场景中,SAN可以通过引入连接标识来唯一标识用户终端与服务提供节点间的通信连接,支持移动场景中报文的一致交付和通信连接的维持,届时SAN将实现进一步的功能优化与增强.

3 关键技术

3.1 语义服务标识

为支撑终端用户和服务提供节点之间高效的端到端通信连接,实现计算资源的集中管控和高效请求,本文设计了一种对计算类服务进行统一语义描述的服务标识及生成规则. 基于这种标识生成规则,终端用户可以在本地生成SID来发起通信,而无需通过传统的DNS机制向集中服务器发起地址解析请求,实现了服务资源与位置的解耦.

服务标识SID定义如下:

$$\mathrm{SID} \triangleq \Psi(\mathrm{obj}, \mathrm{func}, \mathrm{meth}, \mathrm{perf}) \tag{1}$$

其中,obj表示计算服务对象,func表示计算服务功能,meth表示计算服务方法,perf表示计算服务性能,$\Psi(\cdot)$代表服务标识生成函数. 图6给出一种服务标识生成规则的实例. 在该实例中,计算服务对象按照图像、视频、语音、文本以及相应格式进行分类,计算服务功能包含对图像的去噪、识别、生成等,对文本的清洗、分析、生成等. 计算服务功能又对应具体的计算服务方法,如图像识别包括语义分割、实例分割等方法,文本分析包含聚类分析、相关分析、回归分析等方法. 特别地,计算服务性能代表计算服务所需的算力,一般由服务提供商参考对应业务的一般性需求进行分级,如图像语义分割服务的算力分级为1/2/10 GFLOPs,这3种算力会以3个不同的perf来区别.

由SID的定义可知,SID具有全局唯一语义且仅与计算服务的类型有关,SAN智能管控系统可根据标准化的服务标识生成规则与服务提供商、终端用户等一同实现SID全生命周期的统一管理,涉及注册、发布、订阅、更新与撤销等,也可以根据SID进行更灵活的服务请求编排. 对于终端用户而言,SID是用户对服务的请求接口,应用程序可根据用户需求按规则在本地生成或查询SID,以发起位置无关的服务请求. 对于SAN入口网关和出口网关而言,SID实现了语义信息的传递,网关可通过报头解析更快捷地实现算力需求感知与服务请求处理.

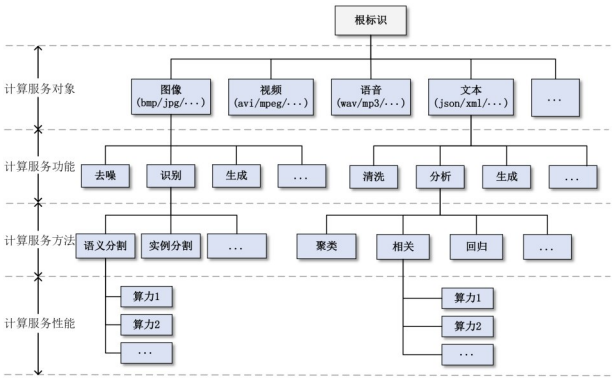

图6 服务标识生成规则实例

相比现有网络基于IP的寻址方式,SAN基于语义服务标识的寻址方式在便捷性、灵活性方面存在显著优势:(1)编码稳定,减少开销. SID仅与服务的类别语义相关,具备长期稳定不变的特点. 相较于DNS机制的多次重定向,基于语义服务标识的寻址方式避免了发起服务时寻址的时延开销,提升了服务请求的响应速度和网络运行的效率. (2)数量有限,部署方便. 由于互联网上的服务类别是一个有限的集合,SID的数量也远小于IP地址的数量. 因此,SID可以通过表项或生成规则的形式缓存在设备本地,用户在发起服务请求时可直接查询匹配或生成,既简化了网络配置又便于部署应用. (3)位置无关,编排灵活. 相较于现有网络基于IP的寻址必须在用户发起通信连接前确定目标主机或服务器的网络地址,基于语义服务标识的寻址方式使用户无需关心服务提供方的位置,只需按服务类型向SAN发送服务请求. 这种与位置无关的连接方式赋予了网络更大的管理灵活性和自主性,同时配合新型上层协议设计,可以实现服务迁移时用户侧弱感知,在动态变化的网络资源状态下更有效地满足用户需求.

此外,SAN中基于语义服务标识的寻址方式和DNS兼容并存,前者面向计算密集型应用,实现更高效灵活的服务请求编排,后者面向普通应用,实现广泛支持的域名解析和资源访问,两者并行运转不影响现有互联网的功能,为实现网络体系架构的平滑演进提供有效途径.

3.2 算网需求感知

在传统的网络架构中,业务需求一般指针对带宽的网络需求. 而在算网深度融合场景中,需求除带宽需求外还包含算力需求. 如何获取服务请求的带宽和算力二维资源需求是SAN实现跨网、跨云一体化编排的基础. 然而,在传统计算应用的通信模式中,算力需求通常封装于应用层协议中,且算力需求与流量特征之

间仅存在弱相关关系,例如,流量的多少无法准确反映算力需求的大小. 因此,基于TCP/IP协议栈的网络架构难以借助报文解析或流量监测等现有网络感知技术主动地获取算力需求. 为解决该问题,本文基于提出的语义服务标识设计了一种算网需求感知方法,使得SAN设备能够通过通用的报文解析和流量监测方法主动获取算力需求,解决网络设备感知业务算力需求的难题.

在传统网络感知技术中,节点间的流量分布一般是通过流量矩阵进行描述,流量矩阵中的每个元素表示从网络中的任意2个节点间的单向流量所需的带宽大小. 在SAN中,流量矩阵可以被扩展用于描述服务请求的带宽需求. 因此,SAN带宽需求矩阵 $\boldsymbol{D}_{\mathrm{B}}$ 可以被定义为

$$\boldsymbol{D}_{\mathrm{B}}=\begin{bmatrix} b_{i_1,\mathrm{SID}_1} & b_{i_1,\mathrm{SID}_2} & b_{i_1,\mathrm{SID}_3} & \cdots & b_{i_1,\mathrm{SID}_n} \\ b_{i_2,\mathrm{SID}_1} & b_{i_2,\mathrm{SID}_2} & b_{i_2,\mathrm{SID}_3} & \cdots & b_{i_2,\mathrm{SID}_n} \\ b_{i_3,\mathrm{SID}_1} & b_{i_3,\mathrm{SID}_2} & b_{i_3,\mathrm{SID}_3} & \cdots & b_{i_3,\mathrm{SID}_n} \\ \vdots & \vdots & \vdots & & \vdots \\ b_{i_m,\mathrm{SID}_1} & b_{i_m,\mathrm{SID}_2} & b_{i_m,\mathrm{SID}_3} & \cdots & b_{i_m,\mathrm{SID}_n} \end{bmatrix} \tag{2}$$

其中,b_{i_m,SID_n} 表示由SAN入口网关 i_m 接入且请求 SID_n 服务的所有服务请求的带宽需求之和. 为获得准确的带宽需求描述,每个SAN入口网关需要在本地进行流量监测,即统计本网关收到的所有服务请求的流量大小并按照SID进行划分. 之后,SAN入口网关周期性地将本地的带宽需求监测结果向SAN智能管控系统通告,该监测结果会用于更新带宽需求矩阵 $\boldsymbol{D}_{\mathrm{B}}$ 中对应行的所有元素.

对于算力需求而言,由于SID中的计算服务性能字段perf明确了服务请求对于算力的需求,SID的引入使得算力需求显式地存在于数据报文头部中. 因此,SAN入口网关可以通过报文解析统计服务请求的算力需求. 参考带宽需求矩阵,算力需求矩阵 $\boldsymbol{D}_{\mathrm{C}}$ 可以被定义为

$$\boldsymbol{D}_{\mathrm{C}}=\begin{bmatrix} c_{i_1,\mathrm{SID}_1} & c_{i_1,\mathrm{SID}_2} & c_{i_1,\mathrm{SID}_3} & \cdots & c_{i_1,\mathrm{SID}_n} \\ c_{i_2,\mathrm{SID}_1} & c_{i_2,\mathrm{SID}_2} & c_{i_2,\mathrm{SID}_3} & \cdots & c_{i_2,\mathrm{SID}_n} \\ c_{i_3,\mathrm{SID}_1} & c_{i_3,\mathrm{SID}_2} & c_{i_3,\mathrm{SID}_3} & \cdots & c_{i_3,\mathrm{SID}_n} \\ \vdots & \vdots & \vdots & & \vdots \\ c_{i_m,\mathrm{SID}_1} & c_{i_m,\mathrm{SID}_2} & c_{i_m,\mathrm{SID}_3} & \cdots & c_{i_m,\mathrm{SID}_n} \end{bmatrix} \tag{3}$$

其中,c_{i_m,SID_n} 表示由SAN入口网关 i_m 接入且请求 SID_n 服务的算力需求之和. 为获得准确的算力需求描述,SAN入口网关需要根据报文的源IP、源端口号等字段识别出每类服务请求包含的会话连接数,基于会话连接数和SID表征的计算服务性能,SAN入口网关能够统计出周期内每类服务请求的算力需求. 之后,SAN入口网关周期性地将本地算力需求统计结果向SAN智能管控系统通告,该统计结果会用于更新算力需求矩阵 $\boldsymbol{D}_{\mathrm{C}}$ 中对应行的所有元素.

上述算网需求感知方法借助SID的设计攻克了传统TCP/IP架构下难以主动获取算力需求的难题,显著提升了SAN对于多维资源需求的感知能力. 同时,该方法使得目前被广泛研究的网络感知技术能够被SAN快速复用,实现对现有网络技术创新的继承与发展.

3.3　两级两层路由

新型网络技术的部署通常面临着重构和复用2种选择. 重构需要对所有设备的网络协议栈进行升级以支持特定功能,而复用则会尽可能利用现有网络功能来实现预期的结果. 多年实践证明,重构所需的巨大成本使得新技术难以落地部署. 对于SAN而言,存量的路由转发设备和服务器设备都难以在短时间内支持基于SID的转发和解析,推进SAN部署需要确保与现有的IP网络架构和基础设施的平滑兼容.

为此,本文提出了如图7所示的两级两层路由机制. 两级路由中,一级路由用于编排SAN入口网关到SAN出口网关的路由,二级路由用于编排SAN出口网关到具体服务提供节点的路由,二者结合起来为服务请求提供完整的端到端路由. 两层路由中,服务层路由在考虑计算资源的条件下完成路由编排,“服务层+基础设施层”协同路由通过联合考虑计算资源与网络资源完成逐跳的路由编排. 从机制上来说,两级路由复用了存量服务器设备,保护了服务提供商内部网络的自治性和隐私性,降低了控制面和转发面的压力. 两层路由复用了存量路由转发设备,同时提供了不同网络开放程度下的路由方案. 从资源适配上来说,两级路由实现了计算资源的按需适配,两层路由实现了算网资源协同的按需适配. 二者互为补充,为SAN的规模化部署提供了可行途径.

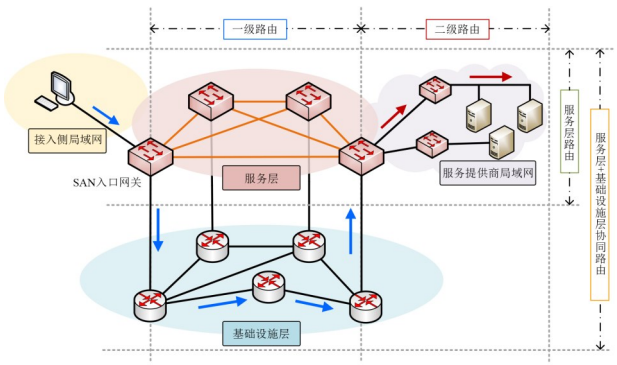

图7　两级两层路由

3.3.1　两级路由

在两级路由中,一级路由由SAN智能管控系统计

算得到. 当基础设施层的网络状态对SAN智能管控系统不可见时,管控系统仅需要完成在服务层视图的路由编排. 由于服务层是架构于基础设施层之上的虚拟网络层,从服务层视角来看,终端用户与服务提供商间只要存在可达路径就可以被视为直连的节点,从而屏蔽基础设施层实际的拓扑连接情况. 在这种视角下,SAN智能管控平台仅需要考虑服务请求的算力需求及服务提供商的计算资源状态进行路由编排,路由决策也简化为服务提供商出口网关的优化选择问题. 此时的路由决策模型可以被称为服务层路由模型.

本节引入算力状态矩阵$\boldsymbol{S}_{\mathrm{C}}$,用于描述服务提供节点实时占用的计算资源状态,算力状态矩阵$\boldsymbol{S}_{\mathrm{C}}$定义为

$$\boldsymbol{S}_{\mathrm{C}}=\begin{bmatrix} c^{*}_{e_1,\mathrm{SID}_1} & c^{*}_{e_1,\mathrm{SID}_2} & c^{*}_{e_1,\mathrm{SID}_3} & \cdots & c^{*}_{e_1,\mathrm{SID}_n} \\ c^{*}_{e_2,\mathrm{SID}_1} & c^{*}_{e_2,\mathrm{SID}_2} & c^{*}_{e_2,\mathrm{SID}_3} & \cdots & c^{*}_{e_2,\mathrm{SID}_n} \\ c^{*}_{e_3,\mathrm{SID}_1} & c^{*}_{e_3,\mathrm{SID}_2} & c^{*}_{e_3,\mathrm{SID}_3} & \cdots & c^{*}_{e_3,\mathrm{SID}_n} \\ \vdots & \vdots & \vdots & & \vdots \\ c^{*}_{e_m,\mathrm{SID}_1} & c^{*}_{e_m,\mathrm{SID}_2} & c^{*}_{e_m,\mathrm{SID}_3} & \cdots & c^{*}_{e_m,\mathrm{SID}_n} \end{bmatrix} \tag{4}$$

其中,$c^{*}_{e_m,\mathrm{SID}_n}$表示SAN出口网关$e_m$所在的服务提供商局域网内所有节点提供$\mathrm{SID}_n$对应服务所占用的计算资源之和. 使用函数$f(\boldsymbol{S}_{\mathrm{C}})$表征对路由编排结果的评价指标,使用$\boldsymbol{I},\boldsymbol{E},\boldsymbol{S}$分别表示SAN入口网关集合、SAN出口网关集合、SID集合,那么服务层路由编排问题可以被建模为

$$\max(\min)\{f(\boldsymbol{S}_{\mathrm{C}})\} \tag{5}$$

$$\text{s.t. } \forall i\in\boldsymbol{I},\ \mathrm{SID}\in\boldsymbol{S},\ e\in\boldsymbol{E}\colon \delta^{e}_{i,\mathrm{SID}}\in\{0,1\} \tag{6}$$

$$\forall i\in\boldsymbol{I},\ \mathrm{SID}\in\boldsymbol{S}\colon \sum_{e\in\boldsymbol{E}}\delta^{e}_{i,\mathrm{SID}}=1 \tag{7}$$

$$\forall e\in\boldsymbol{E},\ \mathrm{SID}\in\boldsymbol{S}\colon c^{*}_{e,\mathrm{SID}}=\sum_{i\in\boldsymbol{I}}\delta^{e}_{i,\mathrm{SID}}c_{i,\mathrm{SID}} \tag{8}$$

$$\forall e\in\boldsymbol{E},\ \mathrm{SID}\in\boldsymbol{S}\colon c^{*}_{e,\mathrm{SID}}\leqslant c^{\mathrm{Lim}}_{e,\mathrm{SID}} \tag{9}$$

其中,式(5)可表示通用的优化目标,如实现算力负载均衡等;式(6)指示对于从SAN入口网关i接入的请求SID对应服务的服务请求是否选择SAN出口网关e,$\delta^{e}_{i,\mathrm{SID}}$为示性函数;式(7)约束了从同一个SAN入口网关接入请求同一种类型服务的服务请求只能选择一个目的SAN出口网关;式(8)计算了SAN出口网关所在服务提供商提供的所有对应服务的算力需求之和;式(9)约束了服务提供商提供的每类服务的算力需求不能超过为该服务预留的计算资源上限,$c^{\mathrm{Lim}}_{e,\mathrm{SID}}$表示SAN出口网关$e$所在的服务提供商局域网内所有节点为SID对应服务预留的计算资源上限.

基于上述模型,SAN智能管控系统可以通过设计调度算法进行算力资源优化的服务请求编排. SAN入口网关在接到SAN智能管控系统下发的路由条目并部署后,会根据目的SAN出口网关的IP地址为每个服务请求报文封装新的IP头部,并将其交由广域网路由设备按照运营商规定的路由策略进行转发,最终到达SAN出口网关,完成一级路由过程.

SAN出口网关在收到服务请求报文后进行二级路由决策. 由于服务提供商局域网内部一般存在预置的负载均衡策略,SAN出口网关可根据该策略及服务请求的需求从可提供SID对应服务的节点集合中选择最佳节点作为服务请求的目的地,同时,使用服务列表中记录的对应节点的IP地址和服务端口号替换原始服务请求报文中的SID和目的端口号字段,之后将去除掉SAN入口网关封装的外层报文头的报文进行局域网内转发以完成交付. 通过二级路由,所有携带SID的服务请求报文被恢复为普通的IP报文,从而能够通过现存局域网设备无障碍转发. 此外,两级路由的设计还解决了计算资源快变特性导致的路由表振荡问题. 由于计算资源尤其是服务实例资源状态高度动态,极端情况下可能出现毫秒级的状态变更频率,过快的状态更新会影响SAN智能管控系统的路由收敛. 因此,SAN出口网关会将收集到的快变计算资源进行2次统计生成慢变计算资源向SAN智能管控系统通告进行一级路由,在服务请求报文到达后再根据实时的快变计算资源状态来选择具体的服务提供节点,确保了计算资源状态的时效性.

综上所述,两级路由的设计在实际部署中具有如下的优势:(1)服务提供商内部服务器设备不需要进行协议栈升级以支持携带SID报文的解析,为基于现有设备的快速部署提供了途径;(2)服务提供商可根据策略自主进行局域网内的路由,既保障了服务提供商服务的灵活部署和管控能力,又实现了广域网和局域网相对的分域隔离,保障了安全性和隐私性且降低了网络管控压力;(3)SAN出口网关可以仅向SAN智能管控系统通告慢变的计算资源状态,而将快变的计算资源状态维护在本地,既保障了SAN智能管控系统进行编排时路由收敛的稳定性,又确保了SAN出口网关在二级路由时度量信息的时效性.

3.3.2 两层路由

当SAN智能管控系统拥有对基础设施层的感知能力后,SAN可使用两层路由机制,即联合考虑服务层的计算资源与基础设施层的网络资源及服务请求对应的算力需求与带宽需求,为服务请求编排逐跳路由. 此时的路由决策模型可以被称为“服务层+基础设施层”协同路由模型. 引入带宽状态矩阵$\boldsymbol{S}_{\mathrm{B}}$用于描述基础设施层链路实时占用的带宽资源状态,带宽状态矩阵$\boldsymbol{S}_{\mathrm{B}}$定义为

$$\boldsymbol{S}_{\mathrm{B}}=\begin{bmatrix} 0 & b^{*}_{(r_1,r_2)} & b^{*}_{(r_1,r_3)} & \cdots & b^{*}_{(r_1,r_n)} \\ b^{*}_{(r_2,r_1)} & 0 & b^{*}_{(r_2,r_3)} & \cdots & b^{*}_{(r_2,r_n)} \\ b^{*}_{(r_3,r_1)} & b^{*}_{(r_3,r_2)} & 0 & \cdots & b^{*}_{(r_3,r_n)} \\ \vdots & \vdots & \vdots & & \vdots \\ b^{*}_{(r_n,r_1)} & b^{*}_{(r_n,r_2)} & b^{*}_{(r_n,r_3)} & \cdots & 0 \end{bmatrix} \tag{10}$$

其中,$b^{*}_{(r_m,r_n)}$表示基础设施层路由节点r_m到r_n的单向链路占用的带宽,若节点间不存在直连链路则元素值为0.使用函数$g(\boldsymbol{S}_{\mathrm{C}},\boldsymbol{S}_{\mathrm{B}})$表征对路由编排结果的评价指标,则“服务层+基础设施层”协同路由编排问题可以被建模为

$$\max(\min)\{g(\boldsymbol{S}_{\mathrm{C}},\boldsymbol{S}_{\mathrm{B}})\} \tag{11}$$

$$\text{s.t. } \forall i\in\boldsymbol{I},\ \mathrm{SID}\in\boldsymbol{S},\ p\in\boldsymbol{P}_{i,\mathrm{SID}}\colon \phi^{p}_{i,\mathrm{SID}}\in\{0,1\} \tag{12}$$

$$\forall i\in\boldsymbol{I},\ \mathrm{SID}\in\boldsymbol{S}\colon \sum_{p\in\boldsymbol{P}_{i,\mathrm{SID}}}\phi^{p}_{i,\mathrm{SID}}=1 \tag{13}$$

$$\forall (r_m,r_n)\in\boldsymbol{L}\colon b^{*}_{(r_m,r_n)}=\sum_{i\in\boldsymbol{I}}\sum_{\mathrm{SID}\in\boldsymbol{S}}\phi^{p}_{i,\mathrm{SID}}\varphi^{p}_{(r_m,r_n)}b_{i,\mathrm{SID}} \tag{14}$$

$$\forall (r_m,r_n)\in\boldsymbol{L}\colon b^{*}_{(r_m,r_n)}\leqslant b^{\mathrm{Lim}}_{(r_m,r_n)} \tag{15}$$

$$\forall i\in\boldsymbol{I},\ \mathrm{SID}\in\boldsymbol{S},\ p\in\boldsymbol{P}_{i,\mathrm{SID}}\colon \delta^{e_p}_{i,\mathrm{SID}}=\phi^{p}_{i,\mathrm{SID}} \tag{16}$$

其中,式(11)可表示通用的优化目标,如实现带宽和计算资源的联合均衡等;式(12)指示对于从SAN入口网关i接入的请求SID对应服务的服务请求是否选择合法路径集中的路径p,$\boldsymbol{P}_{i,\mathrm{SID}}$表示从入口网关$i$到可提供SID对应服务的出口网关的全部合法路径集合,$\phi^{p}_{i,\mathrm{SID}}$为示性函数;式(13)约束了从同一个SAN入口网关接入请求同一种类型服务的服务请求只能选择一条合法路径;式(14)计算了基础设施层每条链路承载的服务请求的带宽需求之和,$\boldsymbol{L}$为链路集合,$\varphi^{p}_{(r_m,r_n)}$为示性函数指示链路(r_m,r_n)是否在路径p上;式(15)约束了基础设施层每条链路承载的带宽需求不能超过该链路的带宽资源上限,$b^{\mathrm{Lim}}_{(r_m,r_n)}$表示节点$r_m$到节点$r_n$的单向链路的带宽上限;式(16)表示是否选择路径$p$关联了是否选择路径$p$上的目的SAN出口网关.

基于上述模型,SAN智能管控系统可以通过设计调度算法进行带宽和算力资源联合优化的服务请求编排. 特别地,该模型的编排结果会包含逐跳的路由信息,因此,SAN入口网关除了会根据SAN智能管控系统下发的路由条目为每个服务请求报文封装新的IP头部,还会将转发路径信息封装到服务请求报文SRH扩展头实现逐跳的转发控制. 到达SAN出口网关后,数据报文会进行二级路由完成最终的报文交付.

为支持如实时云渲染、全息通信等时延敏感类服务,未来的算网一体化架构还需要具备时延保障能力[16]. 对于SAN而言,若基础设施层的传输节点和链路具有时间确定性保障能力,“服务层+基础设施层”协同路由可扩展为“服务层+基础设施层”协同确定性路由以实现时间维度的确定性,该问题可被建模为

$$\max(\min)\{g(\boldsymbol{S}_{\mathrm{C}},\boldsymbol{S}_{\mathrm{B}})\} \tag{17}$$

$$\text{s.t. } \forall i\in\boldsymbol{I},\ \mathrm{SID}\in\boldsymbol{S},\ e\in\boldsymbol{E}\colon t_{i,\mathrm{SID}}=t^{\mathrm{T}}_{i,\mathrm{SID}}+t^{\mathrm{C}}_{e,\mathrm{SID}} \tag{18}$$

$$\forall i\in\boldsymbol{I},\ \mathrm{SID}\in\boldsymbol{S},\ p\in\boldsymbol{P}_{i,\mathrm{SID}}\colon t^{\mathrm{T}}_{i,\mathrm{SID}}=\sum_{(r_m,r_n)\in p}t^{\mathrm{T}}_{(r_m,r_n)}+h \tag{19}$$

$$\forall \mathrm{SID}\in\boldsymbol{S},\ e\in\boldsymbol{E}\colon t^{\mathrm{C}}_{e,\mathrm{SID}}=\frac{\sum_{i\in\boldsymbol{I}}\delta^{e}_{i,\mathrm{SID}}l_{i,\mathrm{SID}}}{c^{\mathrm{Lim}}_{e,\mathrm{SID}}} \tag{20}$$

$$\forall i\in\boldsymbol{I},\ \mathrm{SID}\in\boldsymbol{S}\colon t_{i,\mathrm{SID}}\leqslant t^{\mathrm{Lim}}_{i,\mathrm{SID}} \tag{21}$$

其中,式(17)表示如算网资源均衡等通用优化目标;式(18)计算了业务时延等于传输时延与计算时延之和,$t^{\mathrm{T}}_{i,\mathrm{SID}}$和$t^{\mathrm{C}}_{e,\mathrm{SID}}$分别表示从SAN入口网关$i$接入的请求SID对应服务的服务请求的传输时延和在SAN出口网关e接收SID对应服务的服务请求的计算时延,T(Ttransmission)和C(Computing)分别表示传输和计算;式(19)计算了传输时延等于路径上所有确定性链路的时延上限和服务提供商局域网内传输时延阈值之和,$t^{\mathrm{T}}_{(r_m,r_n)}$为链路$(r_m,r_n)$的确定性时延上限,$h$为预设阈值用于描述服务请求在服务提供商局域网内传输的时延上限,通过该变量可以为在服务提供商局域网内非确定性链路上转发的过程预留时延;式(20)计算了计算时延等于本局域网内提供对应服务的计算量之和除以最大运算次数,$l_{i,\mathrm{SID}}$表示从SAN入口网关i接入的请求SID对应服务的计算量;式(21)约束了业务时延不能超过对应服务请求的时延需求上限[17],$t^{\mathrm{Lim}}_{i,\mathrm{SID}}$表示从SAN入口网关$i$接入的请求SID对应服务的服务请求的时延需求上限.

两层路由虽然要求SAN管控平台具备全面的管控能力,但它能支持更加精细、灵活和可靠的端到端路由控制,进而可扩展如时延确定性保障策略等复杂流量策略,最终实现传输与计算的全过程保障和网络与计算资源的最优利用. 同时,逐跳路由信息通过SRv6的方式进行部署,也避免了存量路由设备需要进行协议栈升级以支持SID解析带来的实际部署难题.

3.3.3　算法设计

两层两级路由中的一级路由的编排问题是随机优化问题,可以利用常见的最优化方法进行算法设计. 本节给出一种基于多智能体强化学习的SAN一级路由(SAN-Routing,SAN-R)算法. 在SAN-R算法中,一级路由的过程可以转化为一个多智能体马尔可夫过程(Multi-agent Markov Decision Processes,MMDP). 在MMDP中,SAN将每个入口网关和路由节点视为一个智能体,从接入服务请求的智能体开始以逐跳的顺序决策下一跳并更新状态,在获得完整路径后环境返回奖励并更新策略. 多智能体的引入不仅使得路由编排

更加高效，还有效解决了复杂拓扑下的可扩展性问题．每个智能体的状态空间、动作空间和奖励设计如下：

（1）状态空间．状态空间被设计为包含带宽状态矩阵$\boldsymbol{S}_{\mathrm{B}}$，算力状态矩阵$\boldsymbol{S}_{\mathrm{C}}$，当前时间步$t$中由SAN入口网关$i_m$接入且请求$\mathrm{SID}_n$服务的所有服务请求的带宽和算力需求之和$b_{i_m,\mathrm{SID}_n}$，$c_{i_m,\mathrm{SID}_n}$，以及当前智能体决策前其他智能体所做出的动作$\boldsymbol{A}_t$，即已经决策出的部分路径．对于服务层路由模型而言，其状态空间不包含带宽状态矩阵$\boldsymbol{S}_{\mathrm{B}}$．

（2）动作空间．动作空间被设计为智能体的邻居节点集合．智能体的动作即从自己的邻居节点集合中选择一个节点作为下一跳．

（3）奖励．奖励的设计与模型的优化目标紧密关联．以均衡算网资源使用的优化目标为例，奖励可被设计为路由合法性奖励r_t^{G}和算网资源使用情况奖励r_t^{B}，r_t^{C}的加权和．

$$r_t=\alpha r_t^{\mathrm{G}}+\beta r_t^{\mathrm{B}}+\gamma r_t^{\mathrm{C}} \tag{22}$$

其中，路由合法性奖励r_t^{G}与转发路径是否存在环路且满足模型中的相关约束条件有关，r_t^{B}与r_t^{C}可分别用带宽状态矩阵$\boldsymbol{S}_{\mathrm{B}}$和算力状态矩阵$\boldsymbol{S}_{\mathrm{C}}$中所有元素的标准差来表征带宽和算力资源使用的均衡性．此外，各项奖励还需按照总体目标统一单调性．对于服务层路由模型而言，权重β的值为0，权重$\alpha+\gamma=1$．对于“服务层+基础设施层”协同路由模型而言，权重$\alpha+\beta+\gamma=1$．

SAN-R算法的伪代码如算法1所示．该算法采用Actor-Critic强化学习架构，每个智能体都有独立的策略网络和评价网络，网络参数θ^i和ω^i在算法开始前会被初始化．在每个时间步，从带宽需求矩阵和算力需求矩阵中提取1组服务请求进行路由决策．从代表SAN入口网关的智能体开始，依次采取行动选择下一跳并更新状态．路由完成后，检查路径是否满足合格性，在此基础上获得奖励和新状态．然后，每个智能体会存储当前时间步的状态、动作、奖励轨迹．在一个训练轮次结束后，每个智能体都会从各自存储的轨迹中随机选N个小批量样本，根据样本计算本地目标函数和损失函数，分别通过梯度上升和梯度下降更新每个智能体中策略网络和评价网络的参数．当智能体的训练过程达到收敛状态时，SAN-R一级路由算法则可以输出训练好的一级路由编排策略$\pi_{\theta^i},\forall v_i\in\boldsymbol{V}$．

相比于一级路由，二级路由一般由服务提供商内部的策略来决策，且仅为服务提供节点多选一的择优过程，因此二级路由可根据本局域网内快变的计算资源状态进行灵活选择．

算法1 SAN-R一级路由算法

输入：训练轮次H，更新轮次K，网络拓扑，带宽需求矩阵$\boldsymbol{D}_{\mathrm{B}}$，算力需求矩阵$\boldsymbol{D}_{\mathrm{C}}$，带宽状态矩阵$\boldsymbol{S}_{\mathrm{B}}$，算力状态矩阵$\boldsymbol{S}_{\mathrm{C}}$，初始策略网络参数$\theta^i$，初始评价网络参数$\omega^i$

输出：最优的一级路由编排策略$\pi_{\theta^i},\forall v_i\in\boldsymbol{V}$

```
1. FOR 训练轮次<H DO
2.  重置网络环境并获取初始化状态
3.  FOR 训练步数小于|D_B| DO
4.  从带宽需求矩阵和算力需求矩阵中提取新的需求元组 (b_{i_m,SID_n}, c_{i_m,SID_n})
5.  选择代表入口网关i_m的智能体为当前智能体v_c
6.  WHILE v_c没有被访问过 DO
7.  根据策略π_{θ^i}及当前状态选择下一跳，并将下一跳添加到当前路径A_t中
8.  IF 下一跳为出口网关且相连的服务提供节点能提供SID_n服务 THEN
9.  BREAK
10.  END
11.  选择代表下一跳的智能体作为当前智能体v_c
12.  END
13.  检查路由合法性，计算奖励并更新状态
14.  每个智能体将轨迹存入经验池
15.  END
16.  FOR 更新轮次<K DO
17.  FOR 拓扑中的任意智能体 DO
18.  从经验池中随机采集N个轨迹样本
19.  计算本地目标函数并梯度更新策略网络参数
20.  计算损失函数并梯度更新评价网络参数
21.  END
22.  END
23.  清空经验池
24. END
```

4 性能评估

4.1 测试环境

本节通过仿真软件和硬件原型设备分别对SAN的性能进行测试与分析．

首先，使用OMNeT++仿真软件搭建了测试环境来模拟服务请求在SAN上的传输与计算过程，该环境使用公开的运营商核心网拓扑[18]作为基础设施层拓扑，并选择其中的10个节点作为接入点分别设置与其直连的SAN入口网关，选择其中的另外4个节点作为出口点分别设置与其直连的SAN出口网关，其余节点作为中继节点．出口网关所对应的服务提供商的域内总算力为100 TFLOPS，网络内共部署20类服务，编号1~10的服务部署于4个服务提供商中的2个，编号11~20的服务部署于其余2个服务提供商．生成数据流来模拟服

务请求,每个服务请求对应1个数据流,其源节点从10个接入节点中随机选择,其SID依据概率从1~20个服务中选择,具体而言,选择编号1~10的服务概率为70%,选择编号11~20的服务概率为30%. 其余网络和服务请求的具体参数设置如表1所示. 服务层路由算法和"服务层+基础设施层"协同路由算法分别被记为SAN服务层路由算法(SAN-Service Routing,SAN-SR)、SAN"服务层+基础设施层"协同路由算法(SAN-Service-Facility Routing,SAN-SFR),并与下列路由基准算法进行比较:

(1)最短路径算法(Shortest Path, SP). 该算法用于选择源节点到目的节点间的跳数最短路径.

(2)等价多路径路由(Equal-Cost Multi-Path, ECMP). 该算法用于在从源节点到目的节点间的所有可达路径中选择多个路径进行转发以实现负载均衡,本次实验设置多路径数为3.

在每个测试例中,分别生成2 000、3 000、4 000、5 000个服务请求,通过算法对所有服务请求进行路由调度,之后将路由结果导入仿真环境中并模拟生成对应的服务请求,收集链路的带宽使用率、服务提供商的算力使用率,以及服务请求的完成时间来进行对比.

表1　网络和服务请求参数设置

网络参数值	
网络节点数	38
入节点(终端用户局域网)数	10
出节点(服务提供商)数	4
链路数	62
链路带宽/Gbps	100
链路传播时延/ms	5
服务提供商局域网内设备总算力/TFLOPS	100
服务提供商可提供的服务类型数	10
服务预留算力 $c_{e,\mathrm{SID}}^{\mathrm{Lim}}$/TFLOPS	10
服务请求参数值	
终端用户可请求的服务类型数	20
带宽需求/Mbps	20~100
算力需求/GFLOPS	10~100

其次,在硬件原型设备上部署了SAN的服务客户端应用,记录50次服务请求发起的响应时间,即从执行发起流程到用户终端开始发送相关数据包所需的时间. 同时,记录使用DNS解析的服务请求发起的响应时间作为基准项来验证SAN的响应时间性能.

4.2　测试结果与分析

(1)链路带宽使用率. 链路带宽使用率方差表征着链路带宽使用的均衡性. 如图8所示,随着网络中的服务请求实例数的增加,不同算法的链路带宽使用率方差都呈现上升趋势,这是由于服务请求的增加会在特定链路上引入更多负载,加剧了链路间带宽使用率的差异. 在同样的服务请求实例数下,SP算法的方差最大,然后依次是SAN-SR算法、SAN-SFR算法,最后是ECMP算法. 这是由于SP算法决策只考虑跳数并不考虑链路负载,使得链路间负载差距最大. ECMP算法将服务请求的流量负载分散到多个路径,因此链路间负载差距最小. 而SAN-SR算法由于考虑了计算资源的均衡性,会将相同SID的服务请求分散到不同的服务提供节点,间接均衡了链路负载. SAN-SFR算法在SAN-SR算法的基础上,增加考虑了链路负载状态,产生更好的链路负载均衡性. 相比于SP算法,SAN-SR和SAN-SFR算法在5 000个服务请求实例下分别降低了13.9%和31.9%的链路带宽使用率方差.

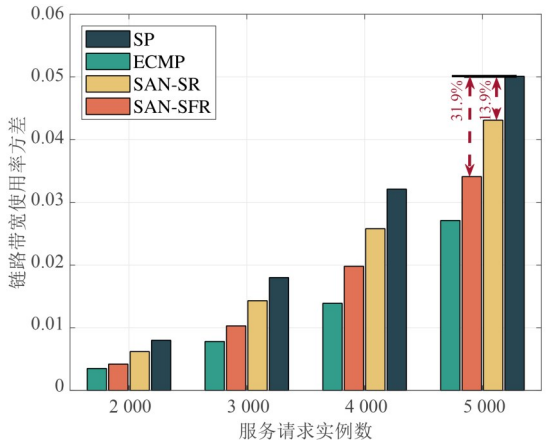

图8　链路带宽使用率测试结果

(2)算力使用率. 算力使用率方差表征着节点间算力使用的均衡性. 如图9所示,服务请求实例数的增加导致不同算法的算力使用率方差呈现上升趋势,因为服务请求的增加会在特定服务提供节点上引入计算负载,加剧了算力使用率的差异化. 由于SP算法和ECMP算法在路由决策时不考虑算力的使用情况,且通过2种算法决策出的服务提供节点基本一致,所以2种算法的算力使用率方差最高. 相比之下,SAN-SR算法由于只考虑算力负载,能够实现最均衡的算力负载分布和最低方差. 而SAN-SFR算法综合考虑算力负载与带宽负载进行决策,其算力使用率方差略高于SAN-SR算法. 相比于SP算法,SAN-SR和SAN-SFR算法在5 000个服务请求实例下分别降低了39.4%和34.3%的算力使用率方差.

(3)服务完成时间. 选择同一个服务请求作为测试对象,在不同算法调度结果下,收集了15 000个该服务请求在时间序列上不同实例的服务完成时间. 图10展示了该服务请求在不同算法下的服务完成时间的分布情况. 测试结果表明SAN-SFR算法使得该服务请求的平均完成时间和完成时间上限最短,其中完成时间上

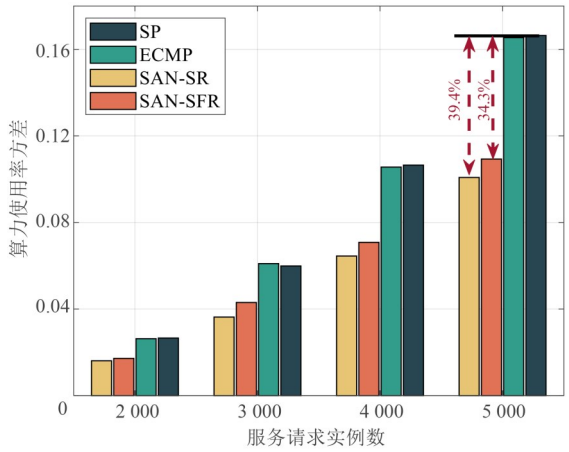

图9 算力使用率测试结果

限相比SP和ECMP算法降低了35.4%和17.5%. 这是由于SAN-SFR算法同时考虑了链路带宽使用情况和算力使用情况,优化了服务请求的传输与计算全过程. ECMP算法和SAN-SR算法由于分别只考虑了带宽负载和算力负载状况,难以联合利用算网资源,因此这2种算法调度下的服务完成时间要长于SAN-SFR算法,多出的时间分别来源于计算过程中额外的等待时延以及传输过程中额外的排队时延. 相比之下,SP算法由于难以根据服务请求的需求和算网的资源状态进行合理决策,该算法调度下的服务完成时间最长.

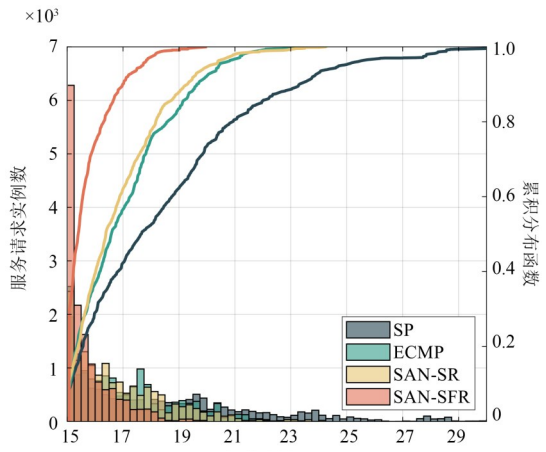

图10 服务请求完成时间测试结果

(4)服务请求发起响应时间. 服务请求发起响应时间的结果统计如图11所示. SID本地生成所需的响应时间要远小于DNS解析过程,因为SID生成仅需要在本地进行简单计算,而DNS解析涉及端到端的通信,会带来大量的通信开销. 实验表明SID生成的最大响应时间比DNS解析的平均响应时间降低了85.3%,验证了SAN机制在服务请求发起时的性能优势.

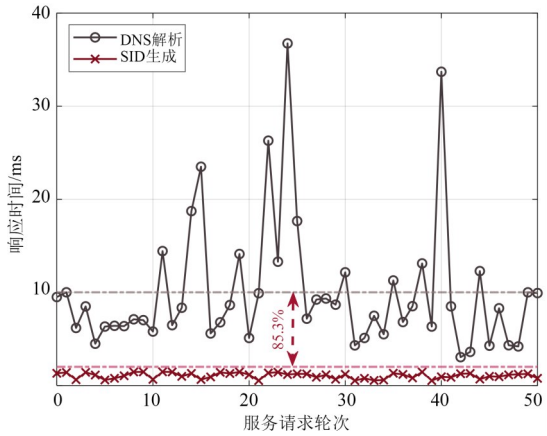

图11 服务请求发起响应时间测试结果

5 结论

本文提出服务感知网络(SAN)架构,该架构以通信过程中的解析发现为抓手改变现有网络架构主机互联的设计,为实现算网深度融合提供内生服务互联支持. 在SAN中,语义服务标识使能用户终端根据服务类型发起位置无关的高效端到端连接,算网需求感知为网络主动获取算力与带宽需求提供了有效途径,两级两层路由既实现了联合算网二维资源的路由编排,又保障了在不同的网络开放程度下SAN部署的可行性和有效性. SAN的设计秉持了通过增量设备避免对现有网络设备或协议栈进行重大改造的核心思想,推动网络从传统的主机互联到服务互联的平滑演进. 未来需要对SAN在移动性支持、传输协议设计、分布式控制等方面进行深入研究,进一步丰富完善SAN的技术框架.

参考文献

[1] TANG X Y, CAO C, WANG Y X, et al. Computing power network: The architecture of convergence of computing and networking towards 6G requirement[J]. China Communications, 2021, 18(2): 175-185.

[2] DUAN S J, WANG D, REN J, et al. Distributed artificial intelligence empowered by end-edge-cloud computing: A survey[J]. IEEE Communications Surveys & Tutorials, 2023, 25(1): 591-624.

[3] 雷波, 刘增义, 王旭亮, 等. 基于云、网、边融合的边缘计算新方案: 算力网络[J]. 电信科学, 2019, 35(9): 44-51.
LEI B, LIU Z Y, WANG X L, et al. Computing network: A new multi-access edge computing[J]. Telecommunications Science, 2019, 35(9): 44-51. (in Chinese)

[4] REN X X, QIU C, WANG X F, et al. AI-bazaar: A cloud-edge computing power trading framework for ubiquitous AI services[J]. IEEE Transactions on Cloud Computing,

2023, 11(3): 2337-2348.
[5] ZHOU Y Q, LIU L, WANG L, et al. Service-aware 6G: An intelligent and open network based on the convergence of communication, computing and caching[J]. Digital Communications and Networks, 2020, 6(3): 253-260.
[6] 张宏科, 于成晓, 权伟, 等. 融算网络体系基础研究[J]. 电子学报, 2022, 50(12): 2928-2934.
ZHANG H K, YU C X, QUAN W, et al. Fundamental research on computing integration networking[J]. Acta Electronica Sinica, 2022, 50(12): 2928-2934. (in Chinese)
[7] QI J P, SU X, WANG R. Toward distributively build time-sensitive-service coverage in compute first networking[J]. IEEE/ACM Transactions on Networking, 2024, 32(1): 582-597.
[8] TANG S J, YU Y, WANG H, et al. A survey on scheduling techniques in computing and network convergence[J]. IEEE Communications Surveys & Tutorials, 2024, 26(1): 160-195.
[9] 孙彦斌, 张宇, 张宏莉. 信息中心网络体系结构研究综述[J]. 电子学报, 2016, 44(8): 2009-2017.
SUN Y B, ZHANG Y, ZHANG H L. Survey of research on information-centric networking architecture[J]. Acta Electronica Sinica, 2016, 44(8): 2009-2017. (in Chinese)
[10] CARZANIGA A, PAPALINI M, WOLF A L, et al. Content-based publish/subscribe networking and information-centric networking[C]//Proceedings of the ACM SIGCOMM Workshop on Information-Centric Networking. New York: ACM, 2011: 56-61.
[11] AMADEO M, CAMPOLO C, QUEVEDO J, et al. Information-centric networking for the internet of things: Challenges and opportunities[J]. IEEE Network, 2016, 30(2): 92-100.
[12] ZHANG J L, YE F, QIAN Y. Intelligent and application-aware network traffic prediction in smart access gateways[J]. IEEE Network, 2020, 34(3): 264-269.
[13] POULARAKIS K, LLORCA J, TULINO A M, et al. Service placement and request routing in MEC networks with storage, computation, and communication constraints[J]. IEEE/ACM Transactions on Networking, 2020, 28(3): 1047-1060.
[14] 张宏科, 王洪超, 董平, 等. 标识网络体系及关键技术[M]. 北京: 人民邮电出版社, 2020.
ZHANG H K, WANG H C, DONG P, et al. Architecture and Key Ttechnologies of Identifier Network[M]. Beijing: Posts & Telecom Press, 2020. (in Chinese)
[15] ZHANG H K, SU W, QUAN W. Smart Collaborative Identifier Network[M]. Berlin: Springer, 2016.
[16] 杨冬, 程宗荣, 田伟康, 等. 广义确定性标识网络[J]. 电子学报, 2024, 52(1): 1-18.
YANG D, CHENG Z R, TIAN W K, et al. Generalized deterministic identification networks[J]. Acta Electronica Sinica, 2024, 52(1): 1-18. (in Chinese)
[17] YANG D, ZHANG W T, YE Q, et al. DetFed: Dynamic resource scheduling for deterministic federated learning over time-sensitive networks[J]. IEEE Transactions on Mobile Computing, 2024, 23(5): 5162-5178.
[18] KNIGHT S, NGUYEN H X, FALKNER N, et al. The internet topology zoo[J]. IEEE Journal on Selected Areas in Communications, 2011, 29(9): 1765-1775.

作者简介

任　杰　男,1996年1月出生,山东济宁人. 北京交通大学博士研究生. 主要研究方向为未来网络体系架构、确定性网络、算网融合等.
E-mail: 21111028@bjtu.edu.cn

王钦定　男,1992年8月出生,甘肃白银人. 北京交通大学博士研究生. 主要研究方向为未来网络体系架构、算力网络、算网融合等. 中国电子学会会员编号:E190131238A.
E-mail: 22110022@bjtu.edu.cn

王洪超　男,1982年12月出生,河北衡水人. 北京交通大学电子信息工程学院副教授、硕士生导师. 主要研究方向为新一代信息网络关键理论与技术、工业互联网、空天地信息网络技术等.
E-mail: hcwang@bjtu.edu.cn

熊　豪　男,2002年8月出生,江西南昌人. 北京交通大学博士研究生. 主要研究方向为未来网络体系架构、算力网络、算网融合等. 中国电子学会会员编号:E190091600A.
E-mail: 24110079@bjtu.edu.cn

杨 冬 男,1980年12月出生,山西大同人.北京交通大学电子信息工程学院教授、博士生导师.主要研究方向为新一代信息网络关键理论与技术以及工业互联网、网络智能化技术等.中国电子学会会员编号:E190035787M.
E-mail: dyang@bjtu.edu.cn

张宏科 男,1957年9月出生,山西大同人.中国工程院院士.北京交通大学电子信息工程学院教授、博士生导师.移动专用网络国家工程研究中心主任.主要研究方向为新一代信息网络理论与关键技术.中国电子学会会员编号:E190004689S.
E-mail: hkzhang@bjtu.edu.cn

谭 斌 男,1976年4月出生,浙江金华人.中兴通讯股份有限公司有线产品架构总工,未来网络技术研究项目经理.主要研究方向为IP网络、SDN系统架构与技术.
E-mail: klinzmann@hotmail.com

郭 勇 男,1976年9月出生,山东烟台人.中兴通讯标准战略规划总监.主要研究方向为人工智能基础设施、未来网络技术架构.
E-mail: guo.yong3@zte.com.cn

黄光平 男,1979年4月出生,湖北恩施人.中兴通讯股份有限公司资深架构师.主要研究方向为算力网络、确定性网络以及下一代IP网络.
E-mail: huang.guangping@zte.com.cn