第 1 章　引　　言

本章首先介绍数据中心网络（datacenter networks，DCN）的相关研究背景，包括数据中心的基本概念、网络架构和业务流量特征，并概述在网络领域应用数据驱动方法的动机。随后，介绍利用数据驱动方法对数据中心网络进行建模的主要研究内容及面临的挑战。最后，介绍本书的研究成果和后续章节的内容框架。

1.1　研 究 背 景

1.1.1　数据中心网络

1. 数据中心基本概念

近年来，数据中心作为云计算的关键基础设施迅猛发展，为许多在线云服务提供了重要支撑，如视频点播、存储和文件共享、网络搜索、社交网络、金融服务、推荐系统等。通过对计算、存储和网络资源进行池化和统计复用，大量服务和应用可以共享数据中心基础设施并根据需求在数据中心内部动态扩展，从而实现更高的资源利用率。国内外各大互联网公司（如谷歌[1]、亚马逊、微软[2-3]、脸书[4]、阿里巴巴[5]、华为[6]、百度[7] 等）都逐渐发展出了大规模的数据中心网络，用于支持其自身庞大的业务或作为公有云为其他服务提供商提供支持。例如，著名的搜索引擎公司谷歌拥有遍布全球四个大洲（北美洲、南美洲、欧洲和亚洲）的 23 个数据中心[8]。思科全球云指数（Cisco Global Cloud Index）显示[9]，全球的超大规模数据中心（hyperscale data centers）数量已从 2016 年年底的 338 个增长至 2021 年年底的 628 个，同时全球每年的数据中心流量

将超过 20ZB（1ZB=10^9TB）。2021 年发布的中国互联网数据中心发展研究报告[10] 显示，2020 年中国互联网数据中心市场规模已达 1563 亿元，同比增长 27.2%。2021 年 3 月发布的“十四五”规划纲要[11] 明确提出要“加快构建全国一体化大数据中心体系”，数据中心的发展对于我国数字经济建设具有重大价值。

2. 数据中心网络架构

现在的大规模数据中心通常包含数十万台服务器，它们之间的连通和高效信息交换依赖于底层的数据中心网络。如图 1.1 所示，现行的数据中心网络主要采用三层的树形网络架构（如 Fat-Tree[12]）。服务器以机柜为单位进行组织，每个机柜中的服务器通过柜顶交换机（top-of-rack，ToR）相连，以中转机柜内和机柜外的流量。柜顶交换机连接若干汇聚层交换机，以实现柜顶交换机的互联互通。一般将由若干汇聚层交换机连接的机柜合称为一个集群（point-of-delivery，PoD），不同的集群间通过多个具有相同功能的核心层交换机连接。这样的网络拓扑结构为不同的服务器节点间提供了多条传输路径，有助于对流量进行负载均衡并提供故障容错能力。

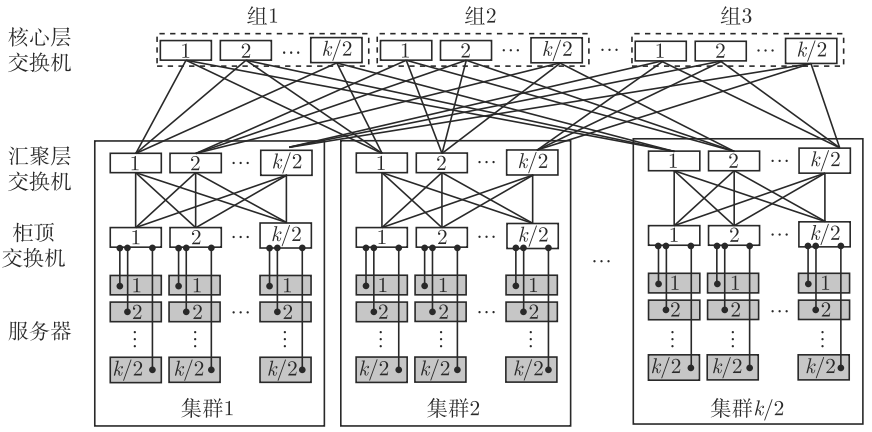

图 1.1　典型数据中心网络架构

3. 数据中心业务流量特征

随着移动互联网的发展和终端用户的增加，依赖数据中心的业务流

量需求也逐渐增加。谷歌数据中心网络的测量数据[1] 显示，数据中心网络的流量需求每 12~15 个月将翻一番，这个增长速度远高于广域网流量。同时，与先前服务器—客户端模型下“南北”流量为主的特性不同，数据中心流量中的“东西”流量，即产生并结束于数据中心内部的流量，占比超过 70%[9]。这些东西流量大部分来自数据中心网络承载的多种多样的分布式云服务应用，如网页搜索[3]、数据发掘[2]、分布式 AI 训练[13] 等。不同的应用往往拥有不同的流量通信模式和负载分布，呈现出动态性、局部性和突发性的特点[4,14-15]。在数据中心网络中，最重要的性能指标是流完成时间[16]（flow completion time，FCT），它直接反映了应用的端到端性能。然而，为了实现最小化流完成时间的目标，不同类型的流量具有不同的传输需求。对于大流而言，其传输数据量一般较大，可能多达 100MB 以上，网络带宽和吞吐是影响其完成时间的主要瓶颈。对于通常小于 10KB 的小流而言，由于传输数据量较小，网络吞吐对其影响不大，其性能主要受到网络中由于拥塞产生的排队时延影响。因此，为了满足多种流量的不同需求，数据中心网络需要在包括拓扑配置、缓存管理、拥塞控制、流量调度、负载均衡等多方面进行设计，以尽可能实现高吞吐和低时延。

1.1.2 数据驱动网络

随着过去半个世纪的发展，互联网已成为人们生产生活的必要基础设施。近年来，随着移动互联网和与云计算的发展，新兴的多种应用对网络性能的要求也逐渐提高。为了满足逐渐严格的应用需求，一方面，底层网络技术迅速发展，以数据中心网络为例，链路带宽已从 25Gb/s 逐渐增长至 100Gb/s 甚至更高。另一方面，现代的网络系统日趋复杂和异构，其高效运行依赖于大量的控制和资源管理算法。资源管理泛指各种用于确定如何将计算和通信资源（如 CPU 周期、网络带宽、数据缓存）分配给不同应用程序及管理应用之间竞争的方法。资源管理问题无处不在，存在于网络系统的各个地方和层级。典型示例既包括与用户体验直接相关的视频流媒体传输的码率自适应[17] 和网络电话的中继选择[18]、对端到端传输效率影响极大的网络拥塞控制[19-22] 和多路径分配算法[23]，也包括偏底层网络的路由策略[24-26] 和网络拓扑结构优化[27]。当然，数据中心中的

计算集群作业调度[28-29]、网络流量调度[30]、交换机参数配置[6]等也同属于此类问题。

对这类问题的研究有着悠久的历史，使用的技术手段涉及计算机科学和应用数学的许多领域和分支。然而，在实践中，真正被采用的解决方案往往会演变为精心设计的启发式方法。其典型的设计流程如下：首先，设计者根据对系统的理解构建系统优化问题的简化模型；其次，将相对抽象的高级别优化目标（如最大化用户观看视频应用的时间）分解为可以直接测量或优化的低级目标（如最小化网络数据包延迟的同时最大化网络吞吐）；最后，提出启发式方法并对其进行重复调整，使其在实际的系统中达到良好的性能。

然而，随着网络系统逐渐的复杂和异构，使用这种方法为网络系统设计高效的算法变得愈发困难。首先，现代网络系统的性能受到多种因素影响，这些因素之间往往以复杂的非线性方式交互，难以准确建模。例如，在数据中心流量优化中，流的完成时间会随着网络拓扑、负载均衡策略、流量调度策略、拥塞控制算法及其相关参数变化的影响而显著变化[3,30-31]。其次，实际可用的优化算法必须能够在各种特性迥异甚至未知的网络条件下运行，这些条件很难被固定和简化的系统模型完全覆盖。例如，交换机共享缓存管理策略必须根据当前的缓存占用和流量状态为不同端口分配缓存，并在不同的端口速度、缓存大小和流量模式下高效运行[32]。最后，很多资源管理问题在面临复杂网络结构的同时，也对需要满足的性能指标有多样化的要求。事实上，人们往往很难通过找到一组能够正确反应最终应用性能需求的低级优化目标来提高系统的性能。因此，实际的解决方案经常为多种启发式方法的组合，而当流量负载或部署的网络环境发生变化时，这个繁琐的过程可能需要被多次重复。因此，实际上部署的网络系统往往会为了简单易部署而无法完全满足性能要求，或者不得不迫使设计人员为每个可能的网络环境开发专门的启发式方法。

近年来，随着机器学习技术的迅速发展，数据驱动方法在许多涉及复杂建模和决策的领域取得了巨大的成果，如视频游戏[33]、围棋[34]和蛋白质结构预测[35]。使用深度神经网络的深度学习[36]方法是其典型代表，它可以直接从原始数据中提取特征，提供一种灵活高效的建模手段。强化学习[37]（reinforcement learning，RL）是机器学习技术的重要组成部分，它

旨在训练一个代理，通过直接与环境交互，从环境得到奖励信号，并据此学习如何做出高效决策。深度强化学习[38]（deep reinforcement learning，DRL）将强化学习与深度学习相结合，利用深度神经网络的强大建模能力，提高了强化学习代理处理信息和表达复杂策略的能力。

受到上述成功应用的启发，本书相信使用以深度学习为代表的数据驱动方法非常有助于解决传统方法难以解决的网络系统问题。首先，数据驱动方法可以直接从来自真实系统环境的数据中学习，针对实际工作负载和网络条件进行优化，无须依赖不准确的人工系统模型。其次，作为通用函数逼近器的深度神经网络具有强大的特征提取和建模能力，数据驱动方法可以使用丰富的原始观察数据（如网络状态指标、流量模式）来提取其与最终性能指标之间的关系，从而提升异构工作负载和网络环境下的决策性能。最后，数据驱动方法可以通过学习直接对高级或抽象目标进行优化（如视频用户的观看时长），无须关心其低级分解目标（如排队延迟、网络吞吐率）。

1.2 研究领域及面临的主要挑战

近年来，有大量的协议、算法和系统被提出用于提升数据中心网络的性能[2-3,5,12,32,39-78]。尽管这些方法非常创新并且取得了巨大的成功，但网络的规模和复杂性使得对这些新设计的测试和评估变得愈加困难。数据驱动方法的强大表达能力为解决这一问题提供了可能。因此，本书将对基于数据驱动的数据中心网络性能建模与优化展开研究。

1.2.1 主要研究领域

网络建模对于理解系统特性、预测未来性能、辅助方案设计具有重大价值[79-80]。网络模型需要提供针对“假想”网络场景的性能评估能力，即能够根据不同的网络配置（如拓扑结构、交换机参数等）和流量描述信息推断网络的关键性能指标（如吞吐率、时延等）。传统的网络性能建模方法一般有三种，分别是数学分析建模、离散事件仿真和系统模拟。数学分析建模（如排队论、网络演算）虽然计算速度快，但依赖于对网络系统的不真实假设，导致其无法灵活考虑多种网络影响因素，导致其对性能

评估不准确或者难以广泛应用；离散事件仿真器（如 ns-3、OMNET++等）可以提供灵活的接口对各种网络协议进行细粒度的仿真，但由于其需要对所有细粒度的包级别事件进行软件仿真，往往计算开销很大，难以对大规模网络进行快速性能评估；系统模拟（如 Mininet[81]）采用真实网络协议栈对网络系统进行模拟，准确度相对较高，但其可评估的网络环境受到采用的物理机器计算能力限制，导致应用范围有限，同时其结果的可复现性较差。可见，现有的网络仿真方案在速度、准确性和易用性等方面仍在存在问题。本书将在 2.2节对这一话题进一步讨论。

上述内容都驱使研究人员探索能够构建灵活、高效、准确的网络性能模型的其他可能路径。以深度学习为代表的数据驱动方法的快速发展为网络建模和性能优化提供了新的可能，它可以避免人工参与，从数据中自动学习网络实体间的复杂映射关系。然而，当前的数据驱动方法并非为解决网络问题所设计，面对异构的网络场景和多样的应用需求，数据驱动方法在实践中仍面临巨大挑战。

1.2.2 面临的研究挑战

1. 网络问题特性差异大

当今的数据中心网络应用需求多样，问题特性各异。使用数据驱动方法对数据中心网络进行建模和优化时，需要针对不同的网络场景对问题进行合理的规约和抽象，从而确定合适的状态输入或决策输出，不合理的抽象方式会增大数据驱动方法的学习复杂度和训练难度。例如，为了解决任务调度问题，可能需要针对任务完成时间建立性能模型，如果将输入状态抽象为可用的服务器编号，会导致随着可用服务器数量的增加，状态空间无限增长，从而严重影响学习效率[29]。因此，如何针对异构的网络系统和应用需求特性对网络问题进行规约是将数据驱动方法应用于网络建模和优化的关键挑战之一。

2. 性能影响因素广

数据中心网络的性能与多种影响因素相关。为了提高建模的准确性，网络模型必须具有足够的灵活性和表达能力，否则将会限制其适用范围。同时，不同的影响因素可能具有不同的“不变性”（invariance）[79]，对这

种特性的合理利用能够有效提高数据驱动模型的学习效率。例如，交换机端口之间具有排列不变性，而数据流按照给定路由经过的节点的状态之间往往存在顺序依赖。因此，如何对具有不同特性的网络性能影响因素进行建模并支持领域知识嵌入是将数据驱动方法应用于网络建模和优化的另一个关键挑战。

3. 网络状态空间大

许多网络系统优化问题常常面临巨大的状态空间，尤其是对于那些涉及整网规模的问题而言。以数据中心网络中的参数配置优化为例，DCQCN[69] 是远程直接内存访问（remote direct memory access，RDMA）网络中的典型拥塞控制算法，其高效运行依赖于对端侧和交换机中十多个参数的合理配置。由于即使是中等大小的数据中心网络也可能包含几百台甚至上千台网络设备和主机，每台设备又可以根据其流量负载进行单独配置，这一问题的空间组合爆炸，很难建模和求解。因此，将数据驱动方法应用于网络系统优化的另一个关键挑战是如何设计可扩展的学习模型以进行高效的表示和决策。

1.3 研究内容与研究成果

本书以解决上述挑战、构建高效的网络性能模型为目标，研究如何利用数据驱动方法对数据中心网络进行性能建模和优化。为了使网络模型支持尽可能多的性能影响因素，书中分别从全局网络拓扑、单点网络设备和时序网络流量三个角度对影响网络性能的关键因素开展研究。首先，关注宏观的网络拓扑对于稳态流量传输性能的影响，用于初步验证使用数据驱动方法进行性能建模的可能性。其次，探索组成数据中心网络的共享缓存交换机，研究微观的交换机缓存管理行为对流量传输性能的影响。最后，利用对网络拓扑和交换机缓存管理进行建模和优化的经验，研究如何对一般的数据中心网络进行建模，同时支持多种网络配置并提供时序流级别性能预测能力。图 1.2 展示了本书的主要研究思路及主要研究内容之间的关系。本书的主要贡献如下。

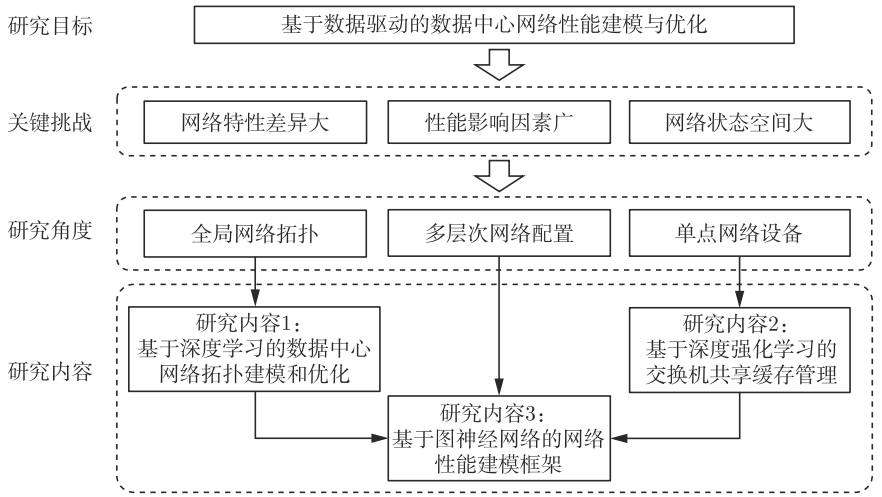

图 1.2　研究思路

1.3.1　基于深度学习的数据中心网络拓扑建模和优化方案

新兴的光/无线拓扑重构技术在提高数据中心网络性能方面展现了巨大潜力。然而，如何找到最佳拓扑配置来支持动态流量需求也为其实际应用带来了巨大挑战。本书提出了 xWeaver，这是一种流量驱动的深度学习解决方案，用于在线推断高性能网络拓扑。xWeaver 构建了一个功能强大的网络模型，可以针对不同的性能指标和网络架构进行拓扑优化。通过设计结构合理的神经网络，它可以自动从数据跟踪（trace）中推导出关键流量模式，并学习流量模式和特定于目标数据中心的拓扑配置之间的底层映射。离线训练后，xWeaver 在线生成最佳（或接近最佳）的拓扑配置，还可以平滑地更新其模型参数以适应新的流量模式。书中还构建了一个基于光电路交换机（optical circuit switch，OCS）的测试平台，以展示本书提出的解决方案的能力和传输效率。进一步地，本书进行了广泛的仿真实验，以显示 xWeaver 在支持更高的网络吞吐量和更短的流完成时间方面的显著性能提升。

1.3.2　基于深度强化学习的交换机共享缓存管理方案

数据中心交换机经常使用片上共享内存来提高缓存效率和吸收突发

流量。当前的缓存管理实践通常依赖于简单的启发式方法，并且对流量模式有不切实际的假设，因为制定适合每种情况的缓存管理策略是不可行的。现代机器学习技术使得自动学习有效策略成为可能。本书提出使用深度强化学习来学习缓存管理策略的神经动态阈值（neural dynamic threshold，NDT）方案，除了一个高级优化目标外无需其他人为指示。为了解决缓存管理复杂性高和规模大的问题，本书在已有的深度强化学习解决方案的基础上开发了两种具有特定领域知识的技术。首先，利用交换机端口的排列对称性设计了一个可扩展的强化学习模型。其次，设计了一个两级控制机制，以实现高效的训练和决策。具体而言，缓存分配在决策间隔期间由低层的启发式算法直接快速控制，而强化学习代理仅根据流量密度对启发式算法的控制因子进行高层的慢速控制。测试平台和仿真实验表明，即使在没有明确训练的工作负载上，NDT 也能成功泛化，并且优于手动调整的启发式策略。

1.3.3 基于图神经网络的网络性能建模框架

当今的网络系统规模巨大、复杂异构，导致其性能具有极强的不确定性。网络运营商往往依靠网络模型来实现高效的网络规划、运营和优化。网络模型负责理解网络性能指标（如延迟）和网络特征（如流量）之间的复杂关系。然而仍然缺乏一种系统的方法来开发准确和轻量级的网络模型，这些模型需要能感知网络配置的影响并提供细粒度的流级别时序预测。本书提出了 xNet，一种基于图神经网络（graph neural network，GNN）的数据驱动网络建模框架。与之前的提议不同，xNet 不是专为具有约束考虑的特定网络场景而设计的专用网络模型。相反，xNet 提供了一种使用关系图表示和可配置图神经网络模块对关注的网络特征进行建模的通用方法。xNet 学习时间步之间的状态转换函数并通过迭代以获得完整的细粒度预测轨迹。本书实现了 xNet 框架，并使用三个典型场景对其进行实例化。实验结果表明，与传统的数据包级仿真器相比，xNet 可以准确预测不同的性能指标，同时实现超过两个数量级的加速。

1.4 全书组织结构

本书的后续章节组织如下：

第 2 章对与本书相关的研究背景和相关工作进行总结。首先，对数据中心网络的 3 个重要研究领域的研究现状进行了总结，主要包括数据中心网络拓扑架构、共享缓存管理和拥塞控制。其次，对传统的网络性能建模方法进行了概述，主要包括数学分析建模、离散事件仿真和系统模拟。最后，对使用数据驱动方法的网络研究进行了综述，主要包括资源管理与性能优化、网络性能建模与评估，以及基础设施和数据驱动网络系统的解释验证。

第 3 章介绍了基于深度学习的数据中心网络拓扑建模和优化方案 xWeaver。首先，对网络拓扑和流量模式之间的关系对流完成时间的影响进行了分析，发现了关键拓扑结构的存在，验证了我们使用深度学习方法的动机。之后，提出了 xWeaver 设计的 3 个关键模块：评分模块、标记模块和映射模块。最后，通过仿真和测试平台实验说明 xWeaver 能够对数据中心网络拓扑的性能进行快速准确推断，同时生成的拓扑配置能够有效降低数据流的传输完成时间。

第 4 章介绍了基于深度强化学习的交换机共享缓存管理方案 NDT。首先，将其规约为强化学习问题，提出使用的状态和奖励函数。之后，针对两个主要挑战分别进行了详细设计：为了解决扩展性问题，观察到交换机端口具有排列不变性，设计了排列等变神经网络作为 NDT 代理的策略表示；为了满足硬件的推理能力要求，同时提高 NDT 的决策和训练效率，提出使用两级控制机制，将启发式算法的控制因子编码为动作，同时设计了累积事件触发机制减少不必要的决策。最后，通过测试平台和仿真实验说明了 NDT 的性能优势和泛化能力。

第 5 章介绍了基于图神经网络的网络性能建模框架 xNet。首先，对当前网络建模方案的局限性进行了分析，并指出这些要求对应的挑战。之