云计算研究白皮书

2024 年

中国电信云计算研究院

2024 年 12 月 31 日

前言

云计算研究院计划在每年的年底发布云计算研究白皮书，包含对云计算领域研究进展的持续总结沉淀，也包含对中国电信云计算研究新的展望和建议。本文是年度白皮书的开篇，形成于云计算研究院的研究团队初具雏形之际，将主要呈现云计算研究院对于云计算相关研究方向的研究图谱、行业背景、研究趋势、研究热点和研究难题的分析洞察。

云计算研究院主要布局四个研究方向：面向下一代云计算的研究、围绕云计算的云网融合研究、智能算法赋能的研究和面向新兴技术的研究。本文内容按照四个研究方向分为四个章节。每章第一节介绍该研究方向的研究图谱，通过分析国内外产业、产品以及关键技术来阐述研究图谱的产生思路。第一节内容除了用研究图谱的方式给出该研究方向的全局概览，也提供了大量行业数据和案例，包含了很多有用的行业参考信息。每章第二节聚焦在研究洞察，通过对大量高水平论文的深入分析，总结提炼出当前的研究热点和研究难题。第二节内容可以为研究人员的研究选题提供重要输入，也可以为研发工程师等其他岗位的同事提供全面的技术趋势解读以及大量技术问题和解决方案的参考。每章第三节先借用 Gartner 技术成熟度曲线的形式，总结呈现前两节讨论的技术点和应用的现状和趋势，最后对该研究方向提出一些展望和发展建议。特别指出，在当前的智算时代背景下，本文包含一个贯穿四个章节的话题，即智能技术与各个研究方向的结合，这个话题也引申出和智能技术息息相关的数据问题：智能技术的研究和应用需要依赖贯穿研究、开发和运营（RDO）的数据体系。下面简要概括四个研究方向的主要洞察。

第一章面向下一代云计算的研究探讨中，首先基于全球市场与国内市场的比较，得出国内平台即服务（PaaS）具备最大潜力的结论。之后观察到云计算产业目前是头部厂商产品和开源组织共同引领，行业标准化尚处于早期。然后发现近年来 1/3 的高水平学术成果都有企业参与，特别是头部厂商贡献突出，中国电信有必要进一步提升高水平研究能力，加强前沿技术探索，提升技术影响力。最后通过对近几年数百篇高水平学术论文的深度分析，总结出数据中心网络、数据库、云存储、资源管理与 QoS 保障等热点研究问题，并重点讨论智能技术与云计算的紧密关系，特别是智能技术对于传统云计算的提升。基于以上分析洞察，本章提出一些发展建议，例如构建数据体系，智能技术与云计算技术深度融合、持续增加对 PaaS 层的投入、持续探索底层架构创新等。

第二章围绕云计算的云网融合研究探讨中，除了聚焦在云计算及算力领域，也涉及部分云网融合的核心研究。具体来讲，本章围绕云网融合的三项重要技术展开探讨，分别是云网一体化调度、算力网络平台和网络云化。云网一体化调度的理论难点是最优调度与计算复杂度的弹性平衡，云计算研究院在攻克理论难点方面已有一定积累并正在持续推进相关研究，当前云网一体化调度的研究热点包括算法复杂度、策略动态自适应和部署性能优化。算力网络是云网融合的关键技术路径，大模型智算引领算力网络平台焕发活力，相关标准正逐步走向体系化，算力服务平台的产业化也正在实现，研究难题包括算力服务效率和性能、智算加速与分布式协同等。网络云化主要由网络功能虚拟化（NFV）系列标准引领，随着 5G 兴起和部署，研究和产业化都日趋成熟，研究难题包括网络功能云原生化和电信等级的云基础设施。

第三章智能算法赋能的研究探讨中，围绕优化理论、图算法、博弈论、深度学习、强化学习和大模型六类算法展开。针对云计算和云网融合中的广泛应用，本章提炼出五大场景，包括数据管理、负载预测和负载均衡、参数调优、调度编排和故障诊断。通过对大量高水平论文的深度分析，本文总结提炼出六类算法和五大场景组合中的研究热点和研究难题。本章提出的发展建议包括大模型和深度学习助力云计算智能化变革升级、图算法赋能云计算稳定高效发展、智能优化及决策赋予云计算可解释性。

第四章面向新兴技术的研究探讨中，围绕工业互联网、智慧交通、智慧医疗、智慧政企、智慧教育等新兴技术领域对云计算和云网融合的需求展开。本章整理了国内外云厂商的相关案例，总结提炼出协同性、移动性、智能性、安全性、可靠性五个方面的挑战。通过对大量高水平论文的深度分析，本章总结了面对每个方面挑战的研究热点和研究难题。本章提出的发展建议包括推动云边协同化和智能化发展、强化安全性和合规性、推动智能化运维和管理等。

目录

第一章

面向下一代云计算的研究

在全球范围内，目前各国正在加速推动云计算技术的创新与应用，以应对日益复杂的数字化需求和全球竞争。云计算不仅为大数据、人工智能、物联网等技术的快速发展提供了底层支撑，也成为了国家战略的重要组成部分，影响着全球产业格局与经济结构的变革。过去一年，大模型应用，低空经济互联网等业务场景呈现井喷式发展，云计算作为大模型的底层算力支撑，已进一步深刻影响人类的生产生活方式和全球产业格局。

在全球云计算产业中，基础设施即服务（IaaS）、平台即服务（PaaS）和软件即服务（SaaS）构成了云服务的三大核心模型。各大云厂商围绕这三大服务层次，不断加大技术投入，提供多层次的解决方案，以满足企业对基础设施、应用开发平台和软件应用的需求。这种多层次服务模式能够灵活地满足企业的用云需求，推动了云计算的快速普及，同时也进一步加速各行业数字化和智能化的发展。本章将从这三大服务模型所呈现的产品出发，探讨全球云计算技术的发展现状以及产业界与学术界的技术演进趋势，重点分析头部云厂商在云计算领域的战略布局、技术创新投入及其市场动态。本章还将结合国内当前的云计算发展状况，分析我国在全球云计算竞争中的优势与挑战，探讨下一代云计算的发展方向。

1.1 研究图谱及其产生：云计算产业和技术分析

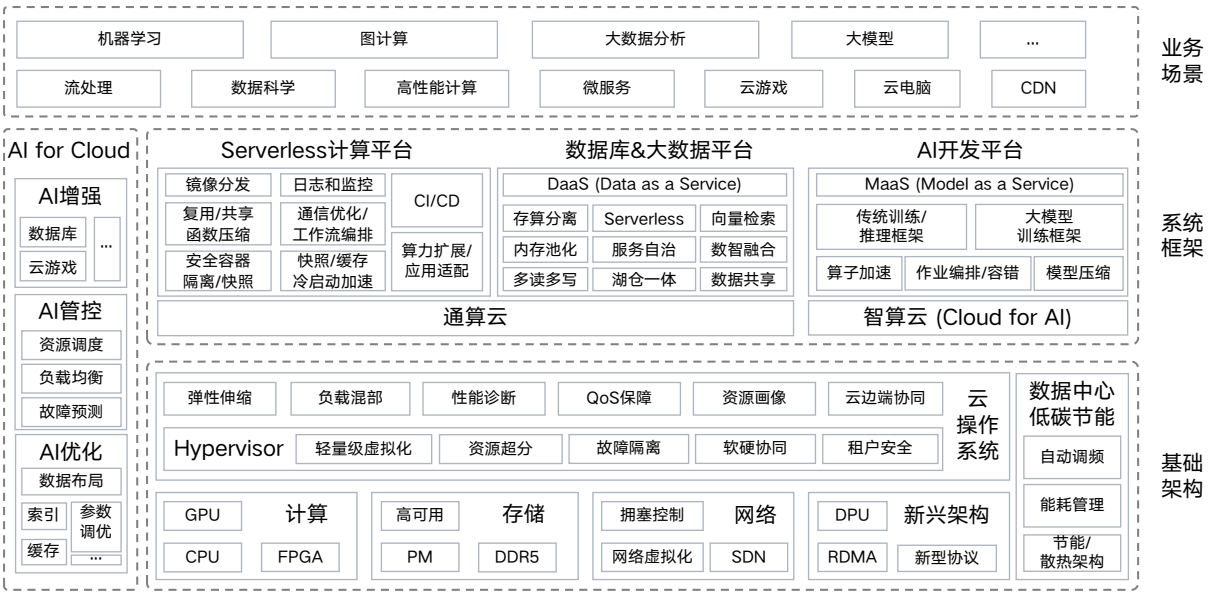


图 1.1: 云计算研究图谱 (由云计算研究院总结形成）

在当前 AI 浪潮的驱动下，云计算技术的发展也朝着更智能化、灵活化、多样化的方向迈进，一方面，基础架构层和系统框架层围绕新硬件和新业务场景带来的机遇和挑战，持续地进行深度优化以满足包括 AI 在内的多样化应用需求。另一方面，AI 技术也被广泛应用于基础架构和系统框架的优化设计，不断提高云计算的能力。图 1.1列举了当前阶段云计算领域的技术研究图谱[1]，这些研究也已经或正在助力云计

[1]按照业界共识，本文在系统框架层将现有主流云平台定义为通用计算平台和专用计算平台（以智算云最具代表性）两类。

算行业迎来新一轮的技术变革。总的来说，在基础架构层，基于新型高速互联技术的内存池化架构（例如 CXL）有助于提高算力调度的灵活性和资源利用率；存储技术与分布式系统结合，推动数据管理更加高效；新兴架构技术（例如 DPU、RDMA）则大幅提高了云平台的数据传输和控制能力；而云操作系统则结合数据中心节能技术更好地协调组织计算、存储和网络资源的分配，通过调度优化资源使用，减少碳排放。在系统框架层，Serverless 架构简化了开发者操作，帮助用户实现业务高度弹性和可扩展性。面向垂直领域（例如数据库，AI 开发）的深度优化平台则可以实现一站式开发流程并优化业务的运行性能。AI for Cloud 则贯穿所有层级，通过自动化、智能化的运维技术提升云计算系统的运行效率与服务质量。

本节余下的内容将结合研究知识图谱，通过公开资料的整理讨论云计算行业国内外市场和产品情况和发展趋势。

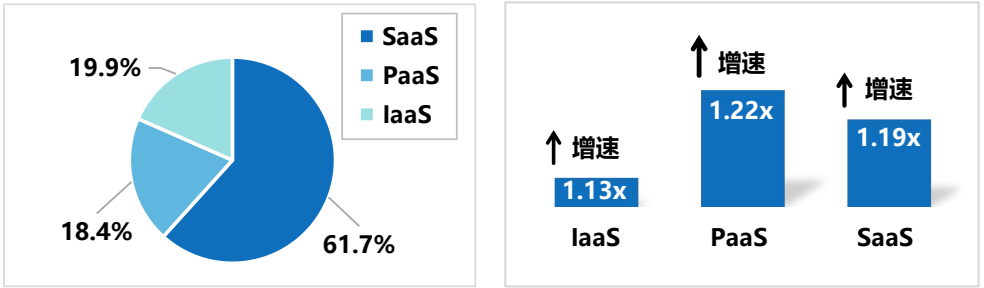
1.1.1 云计算市场规模与发展趋势

从全球云市场的发展态势来看，全球市场正迎来新一轮的增长。在国际市场方面，预计在未来几年内，全球云计算市场的规模将持续扩大，同时保持较高的增长率。新技术的应用以及市场需求的变化，正推动着云计算行业不断创新与变革，接下来将从 IaaS，PaaS 和 SaaS 市场三个方面对国内外云计算产业进行详细的剖析。

过去一年里，全球 IaaS 市场持续增长，但增速有所放缓。根据 IDC 报告，全球 IaaS 市场呈现出持续增长且竞争激烈的态势。从收入规模来看，2021 年至 2023 年期间，全球 IaaS 市场规模从 2021 年的 915 亿美元不断攀升至 2023 年的 1334 亿美元，其同比增长率虽有所下降，但仍保持着相对可观的增长速度（见表 1.1）。究其背后原因，以生成式 AI（GenAI）为代表的业务场景在很大程度上助力了 IaaS 市场的增长 [1]。例如在医疗、生物、制造业等领域，生成式 AI 往往依托于 IaaS 云服务来获取相应资源，间接带动了 IaaS 市场的增长。算力需求的扩大也推动了智算型数据中心的建设和发展。全球以 Amazon、Microsoft、Google 为主体的头部云计算厂商持续增加 IaaS 层的投入。与此同时，国内的云计算厂商也每年投入大量资金用于建设智能云数据中心（例如百度智能云，天翼智算云），不断提高数据中心在 GPU 算力提供、IaaS 服务优化等方面的能力。

表 1.1: IDC2023 年度全球公共云服务收入及同比增长统计 [2]（单位：十亿美元）

云服务	2021	份额	2022	份额	2023	份额	21-22 增长率	22-23 增长率
IaaS	91.5	20.6%	115.5	20.7%	133.4	19.9%	26.2%	13.4%
PaaS	70.1	15.8%	95.4	17.1%	123.3	18.4%	32.1%	22.6%
SaaS	282.7	63.6%	347.5	62.2%	412.5	61.7%	40.1%	18.7%



(a) 2023 全球 IaaS，PaaS 以及 SaaS 市场份额　　(b) 2023 全球 IaaS，PaaS 以及 SaaS 市场增速

图 1.2: 2023 全球 IaaS，PaaS 以及 SaaS 市场份额分析 [2]

全球 PaaS 市场份额与 IaaS 市场份额持平，同时市场增速超过 SaaS 成为第一。如图 1.2(a)所示，2023 年全球 PaaS 市场规模达到 1230 亿美元，相比 2022 年市场份额上升 22.6%。尽管相比 2022 年的市场份额

增长率有明显的下降，但相比同期的 IaaS 和 SaaS 市场，PaaS 市场的增长率在 2023 年超过 SaaS，成为第一（图 1.2(b)）。与 IaaS 市场增长类似，当前 PaaS 服务迅速增长很大程度上归结于以生成式 AI、大模型为代表的新兴应用场景的发展。生成式 AI 和大模型等新技术的爆发，促使众多开发者和中小型企业需要简便高效的一站式模型开发平台来支持其 AI 应用的构建与部署，PaaS 平台能够很好地满足这些需求，从而推动了市场增长。随着未来云技术的不断演进，作为云平台架构中承上启下的关键中间层，PaaS 层所面临的市场需求将会持续攀升，相应的功能也必将得到进一步的强化与拓展。

目前全球 SaaS 市场规模最大，各大云厂商均在 SaaS 领域大量进行布局，为云厂商带来高额利润。据统计，2023 年全球云计算市场中，SaaS 市场贡献了超过 60% 的市场份额，例如国外云厂商的 Google 地图服务，Microsoft Office 365 协作文档，国内以阿里云钉钉办公软件为代表的办公软件等均是当前流行的 SaaS 服务，收获了大批忠实用户。AI 大模型的井喷式爆发为 SaaS 带来新的契机，但尚未进入盈利期。其原因主要有两点：一方面，AI 技术研发成本极高，从算法优化到模型训练都需要大量资金与人力投入，如购买昂贵的 GPU 设备、聘请顶尖 AI 人才等。另一方面，目前市场处于培育阶段，许多 AI SaaS 产品为吸引用户，采用低价甚至免费策略，如一些智能客服服务提供免费试用期且基础功能免费，依靠增值服务收费，但增值服务转化率尚低，导致整体盈利困难，不过随着技术成熟与市场拓展，未来盈利潜力巨大。

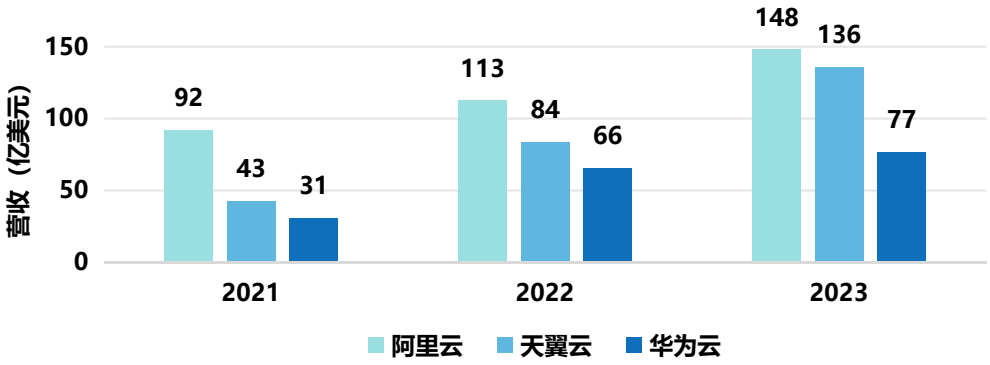


图 1.3: 国内主要云厂商云计算业务营收（亿美元）

在国内云计算市场方面，产品服务竞争激烈、需求多样化带来新市场渗透，主要云厂商的业务营收普遍呈现上升趋势。如图 1.3，据有关公开资料整理[2]，阿里云、天翼云、华为云这三家国内云厂商在近三年间的云计算业务营收均呈上升趋势。阿里云的营收始终保持领先，不过天翼云在 2023 年与阿里云的营收差距有所缩小，而华为云的营收增长相对较为缓慢。国内云服务提供商间的激烈竞争以及针对运营商云的政策倾斜等方面的因素，可能是引起国内市场份额重新调整的主要原因。此外，随着行业需求的多样化，以及各大云厂商在市场竞争中不断优化服务和降低价格，更加促进了客户的迁移和新客户的加入。这种价格和服务的优化，使得更多企业愿意将传统 IT 基础设施转向云计算平台，同时推动了云计算市场的进一步渗透和市场竞争格局的变化。

PaaS 将成为未来云计算行业核心增长动力已是行业共识，但国内 PaaS 市场仍处于起步阶段，与全球云计算产业布局存在较大差距。如图 1.4 所示，2023 年国内云计算市场份额中，IaaS，PaaS 和 SaaS 市场分别占据 57.8%，17.6% 和 24.6% 的市场份额比例 [3, 4]。与同期全球云计算行业市场对比，国内仍处于以“售卖基础算力原材料”的 IaaS 主导型市场阶段，PaaS 市场份额远远小于当前 IaaS 市场的体量。一方面是由于国内云计算市场起步较晚，很多国内企业尚未或正在进行数字化转型。另一方面，国内云计算

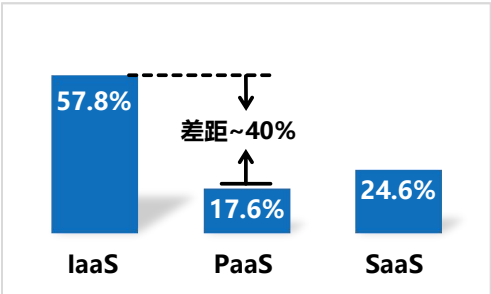


图 1.4: 2023 国内云计算三大市场份额占比

[2]部分数据来源于中国信通院《云计算白皮书 (2024)》。

服务供应商相较于国际竞争者，在 PaaS 平台的技术创新、生态建设和行业深度应用方面存在差距。国内的 PaaS 平台虽然在基础设施层面逐渐追赶上来，但在平台的集成能力、可定制化服务和行业解决方案的深度挖掘上，仍缺乏足够的技术积累和市场经验。

SaaS 市场普遍面临行业需求定制化和技术整合等挑战，但 AI 与 SaaS 的深度融合为行业发展带来了新的机遇。与国际市场相比，国内 SaaS 市场的份额仍然较低，主要原因在于国内 PaaS 平台能力的不足以及较高的服务资源成本，特别是对于资源有限的中小型 SaaS 企业来说，这一成本压力制约了其市场拓展和服务普及。然而，这一现状也为国内 SaaS 行业提供了独特的机遇。通过借助 AI 技术与 PaaS 平台的优化，国内 SaaS 企业可以利用更加灵活的开发环境和定制化行业解决方案，降低技术门槛和基础设施成本，从而提升产品的市场竞争力。通过这些创新，国内 SaaS 市场有望加速发展，逐步接近国际市场水平，释放出巨大的潜力。

1.1.2 云计算行业开源组织与事实标准

成为“像水和电一样随取随用”的标准化服务，是云计算诞生之初的美好愿景。然而，云计算市场在各大厂商的竞争中呈现出多元化发展的态势，暂未走向统一标准。与网络、通信等依赖标准引领的成熟行业不同，新兴的云计算行业技术发展主要由开源组织引领，标准化进程相对滞后。在当前阶段，云厂商更注重围绕用户实际需求展开布局，结合开源社区项目和自研创新，赋能产品并提升影响力，实现差异化竞争优势。在此进程中，具备前沿技术优势和广泛社区影响力的众多开源项目逐渐成为了云计算行业各个领域的事实标准。

相比于网络与通信领域标准对产业发展和技术产品的引导作用，云计算领域的标准建设起步较晚，产业标准化仍面临较大难度。网路与通信领域的标准化历史悠久，其网络通信的标准化过程相对成熟，诸如 TCP/IP、LTE、5GC 等协议都是经过国际组织（如 IETF、ITU、3GPP 等）广泛认可和采用的标准。这些标准确保了不同设备和系统之间的互操作性，促进了网络的普及和发展。相较于网络通信领域，云计算是一个相对较新的领域，相关标准的建设和发展尚未成熟。一方面，由于技术和市场需求变化迅速，企业更关注产品迭代速度和差异化竞争优势，不愿过多受制于已有的标准，从而在一定程度上弱化了标准的建设。另一方面，云计算环境通常涉及多种服务模型（如 IaaS、PaaS、SaaS）和部署模型（如公有云、私有云、混合云），云服务提供商在各种应用场景中可能具有不同的侧重点和设计目标，导致业务需求相对分散，行业产品标准难以统一。

云厂商已意识到统一标准对云产业发展的重要性，当前阶段开源社区项目已在部分领域成为了云计算事实标准。当前，云计算各大厂商在技术与产品上展开激烈竞争，市场呈现出多样化的繁荣格局。但各个云厂商均在各自为战，缺乏不同云业务间的接口互操作性，用户面临云服务商锁定的问题，跨云迁移的开销巨大。为了提升自身的影响力，头部的云服务厂商致力于通过开源社区推广自身产品的技术路线，从而建立繁荣的云市场生态，而跟随者也寄希望于兼容上述生态来实现业务市场的扩展。在此过程中，国内外著名的开源社区例如 Apache 软件基金会、云原生计算基金会 (Cloud Native Computing Foundation, CNCF) 和开放原子开源基金会成功孵化了诸如 OpenStack、Kubernetes、OpenEuler 等在内的一系列开源项目，目前均已成为云计算领域的核心技术。许多当前革命性的新技术，正是在开源社区中率先被提出并进行验证，最终成为云计算行业相关领域的事实标准。

标准建设与开源社区相结合，二次开发与自研创新相结合，优势互补，将持续为云计算的发展注入动力。随着技术的不断进步和行业需求的变化，开源生态在云计算行业的主导地位将进一步加强。标准化和开源并非对立关系，而是互为补充、相互促进的协同发展模式。开源社区的快速迭代和创新能力，为标准化工作提供了实践验证和技术积累；而标准化则为开源技术的规模化应用提供了可靠的规范指导。在未来，云计算行业将依托开源技术实现更多突破性创新，从容器化、微服务到人工智能和边缘计算等前沿领域，并通过标准化确保技术的广泛应用和跨平台兼容性，打破技术壁垒，解决运营商锁定等问题，从而成为推动云计算行业健康发展的核心力量，引领新一代信息技术革命，推动产业升级和数字化转型。

1.1.3 头部云厂商主流产品与优势分析

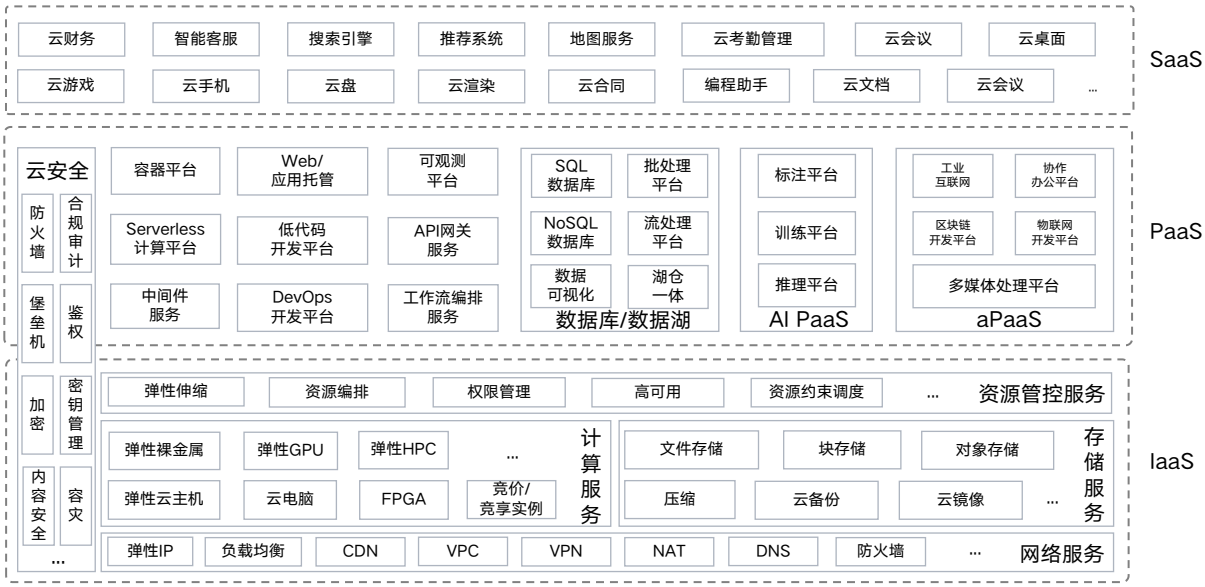


图 1.5: 主流云计算产品全景图（由云计算研究院整理形成）

目前，各大云厂商对外提供的云服务产品功能基本涵盖了 IaaS、PaaS 以及 SaaS 的主要涉及领域（如图 1.5）。不同层次的服务各有特点，具体如下：

- IaaS 产品服务聚焦提供基础的计算、存储、网络等基础设施资源，客户可在此基础上搭建自己的 IT 环境，如 Amazon 的弹性云主机、EBS、VPC、S3 等，是云计算的底层支撑。
- PaaS 产品服务为开发者提供了开发平台和软件运行环境，便于他们更高效地进行应用开发、部署和管理，通常会包含一些中间件、数据库管理系统等。
- SaaS 产品服务以软件应用的形式提供给用户使用，用户无需关心软件的开发、部署和升级维护等问题，只需使用其功能即可，像阿里云的阿里钉钉、云会议等产品就属于此类。

此外，在云基础设施服务广泛应用于各行各业的同时，安全问题，特别是在数据存储、网络传输、身份管理和合规性审查等方面，变得愈加重要。为了保护云环境中的敏感数据和应用，云服务提供商已采取了一系列安全措施来确保平台的安全性。然而，尽管如此，金融等对数据安全极为敏感的行业，仍面临着较大的上云难题，数据隐私和合规性的顾虑使得这些行业在迁移至云平台时显得尤为谨慎。

各大云厂商在不同的云产品服务上均有所专长和侧重。比如 Amazon 在基础设施服务（如弹性云主机、EBS、VPC、S3）上起步早，产品种类多，提供了最广泛的资源管理服务。Microsoft Azure 在基础设施服务方面也有竞争力，尤其是在与 Microsoft 系列产品（如 Windows Server、SQL Server）兼容方面做得很好。Google 则在云数据分析以及智能应用框架等方面保持着技术领先优势。阿里云在国内市场的数据库上云替换方面拥有较强的实践经验，其自研 OceanBase [5], PolarDB [6] 等云原生数据库，具有明显的技术优势和市场竞争力。除此以外，阿里云的阿里钉钉、云会议等 SaaS 产品也吸引了不少用户群体。华为云拥有较强的自主可控能力，其在弹性裸金属、弹性云主机的产品服务投入较大，具有较高的成本优势。天翼云在电信运营商基础设施的优势下，适合大规模企业和政府客户，其在 IaaS 和 PaaS 领域已有大量的技术积累，同时 SaaS 市场也保持着强有力的竞争力。例如其“桌面即服务”引领国内市场，目前在国内该领域中公有云市场份额排名第一。

总的来说，Amazon 在全球范围内依托其庞大的产品生态体系，占据了国际云计算市场的领导地位；Microsoft Azure 以企业级应用的深度集成为优势；阿里巴巴云则在中国市场具有强大的本地化优势；华为云则更多在基础设施和智能技术方面形成了特色；天翼云的定位则侧重于国内运营商市场和行业应用。

这些厂商之间的差异，既反映了它们的技术重点，也影响了它们在全球云计算市场的竞争格局。

1.2 研究洞察：当前云计算的研究热点和难题

在过去的二十年里，云计算产业经历了快速发展阶段。尽管时至今日，全球范围内的庞大云计算产业链已经形成，各大云计算厂商在云产品、云能力的建设方面也积累了大量技术和实践经验，但云计算领域仍然存在大量难题制约着云技术的进一步发展。本节通过对云计算领域的研究进行了大量的调研和分析，将现有的热点研究方向归结为七个主要方向，并总结了每个方向在当前阶段所面临的主要难题（如下表所示）。举例来说，数据中心利用率与云服务 QoS 间的矛盾，面向分离式架构的远端内存访问延迟等挑战现阶段是企业研究的热门话题。本节着重分析了近三年业界的主要研究成果，并展望了各个研究方向未来的技术热点和发展趋势。

研究热点和难题

1. **数据中心网络**：如何优化网络协议与架构以应对大规模流量调度与云服务体验挑战？
2. **数据库**：如何提升存算分离架构下多点写入冲突和数据库无状态设计以及如何高效管理非结构化数据从而以最小的查询成本服务 AI 等领域？
3. **云存储**：如何设计元数据服务，以实现高性能、低成本且语义融合特性的海量数据存储服务？
4. **资源管理与 QoS 保障**：如何解决改善数据中心资源利用率与多租负载间性能干扰的矛盾？
5. **OS 与分布式系统**：如何为 AI 提供高效云平台基础能力支持并优化分布式系统中通信开销？
6. **分离式数据中心架构**：如何解决内存池拉远导致的负载性能劣化与资源高可用问题？
7. **Serverless 计算**: 如何解决函数的长冷启动时延问题，实现极致快速的函数扩缩容？

1.2.1 主要研究分布及热点剖析

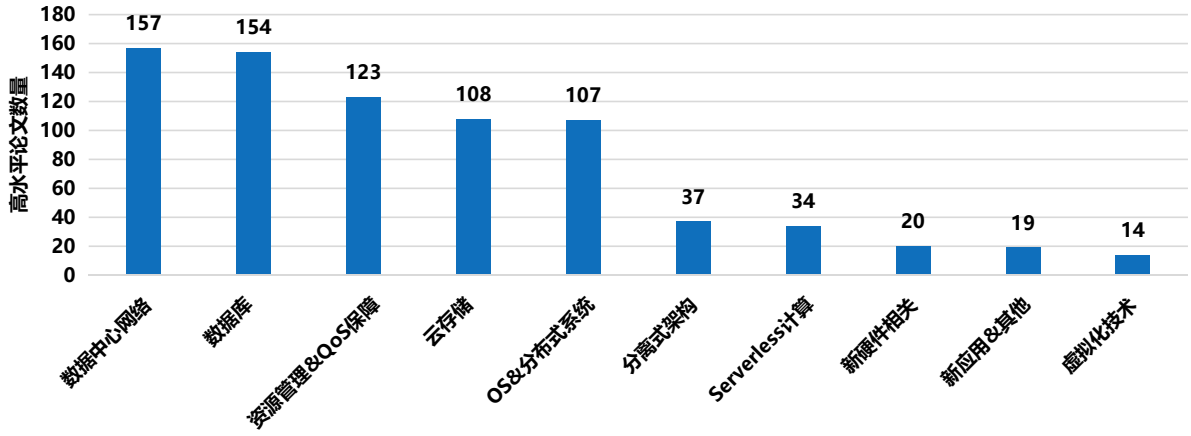


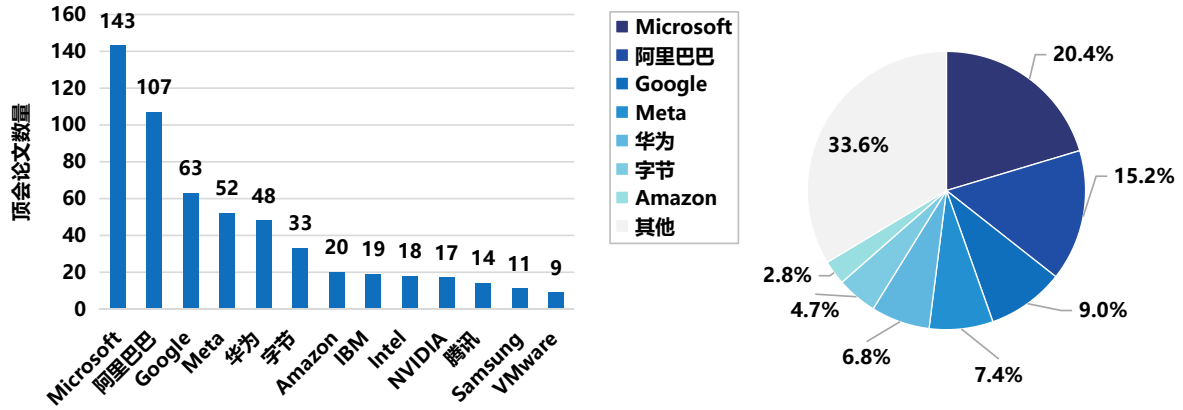
图 1.6: 近三年企业参与的云计算热点研究领域文章发表数量

通过调研近 3 年和云计算产业相关的 14 个顶级学术会议（NSDI, ASPLOS，SC，SOSP，VLDB 等）收录的 3,000 余篇高水平论文（以 CCF-A 类为主），本文从中筛选出近 700 篇云计算领域有企业参与的已发表文章，采用统计的方式进行全面的研究热点剖析。

现有学术研究聚焦通用计算云和 AI 智能云两大主体，涵盖包括数据中心基础架构、云操作系统、任务调度与编排框架、中间件、应用性能调优等在内的 30 余个具体的研究点。通过进一步的筛选与合并，本文将上述涉及的所有研究单元整理为 10 个基础研究方向，分别为数据中心网络、云存储、数据库、数据中心资源管理与云服务 QoS 保障、OS 与分布式系统、虚拟化技术、分离式数据中心架构与资源池化、

Serverless 计算技术、新硬件相关应用加速技术以及新兴业务场景（见图 1.6）。其中与 “AI + 云” 相关的文章按照研究主体被分为 AI for Cloud 和 Cloud for AI 两类。为保持清晰，这两类文章未单独列出，而是按照研究内容和涉及的领域归并到了 10 个基础研究方向内（注：同一篇文章可能涉及多个方向）。

近三年来，数据中心网络、云存储、数据库、分布式系统以及 QoS 保障下的资源管理是主要学术热点。在参与统计的近 700 余篇学术研究工作中，与这五项相关的研究工作数量占据了整体的 90% 以上，其中又以数据中心网络和数据库最为突出，两者在学术论文数量分别为 157 和 154 篇，均超过了整体的 20%。同时，在数据中心内部，存算分离的分离式基础架构设计（例如 CXL 内存池化）、Serverless 计算也具有相当程度的热度，两者总和也占据了全部研究工作的十分之一。此外，还有一些研究工作关注新硬件的设计与应用（例如近数计算，存内计算等硬件），新兴业务场景（例如 IoT，低空经济等）和虚拟化技术。



(a) 近三年头部企业研究数量分布（从高到低排序）　　(b) 近三年头部企业研究数量占总体比例分布

图 1.7: 近三年头部企业在研究成果的影响力分布

企业在学术研究领域话语权逐年攀升，大约三分之一的近三年学术成果背后都有企业参与，且大部分以国外厂商为主。在参与统计的近 700 篇学术论文中，参与的企业数量超过 100 个。如图 1.7(a)所示[3]，本章仅将发文数量靠前的 13 家企业进行了展示，这些企业的发文量占据了所有企业总和的 80% 以上。具体地，包括一些拥有大量公有云计算业务的厂商例如 Microsoft，阿里巴巴，Google，华为，Amazon 等，也包括以互联网业务和私有云数据中心建设为主体的 Meta，字节跳动，腾讯等企业，此外还包括一些大型的服务器提供商、硬件厂商和虚拟化厂商例如 IBM，Intel，Nvidia，Samsung 和 VMware。其中，上述这些企业又以国外为主，国内厂商仅有 4 家在列（占三分之一）（图 1.7(b)）。

在头部的 13 家企业中，学术研究成果的分布又呈现出明显的 “分布倾斜” 现象。其中，Microsoft 和阿里巴巴的发文量分别达到总量的 20.4% 和 15.2%，以超过 100 篇的研究成果稳居第一梯队。值得注意的是，这两家企业的学术产出主要来自其专门的研究机构（如 Microsoft research、达摩院和蚂蚁研究院），这也凸显了其对科研创新的重视以及近年来持续不断的科研投入所带来的显著成效。第二梯队由 Google、Meta 和华为构成，三家企业的发文量合计占比超过 20%，这与它们在云计算基础设施和前沿技术探索方面的深厚积累密切相关。第三梯队则由字节跳动、Amazon、IBM，这里不再逐一介绍，但有两家企业值得注意：新兴科技公司字节跳动凭借其对技术创新的热忱，近年来在学术界逐渐崭露头角；全球云计算市场的领导者 Amazon 早期在云计算领域发表了大量具有影响力的学术论文 [7]，近年来将人工智能技术确立为下一阶段的增长重点，尽管系统类论文发表数量虽有所降低，但其依旧保持着卓越的技术影响力，其现网产品和解决方案仍被大量引用。

为进一步了解各大头部企业在云计算领域的研究投入，本文统计了各家企业在各研究领域的发文数

[3]数据说明：统计近三年各个企业发表的文章中，归属于前 7 个热点研究方向的占比（同一篇文章可能涵盖多个方向）。这 7 个头部企业共参与发表了 400 篇左右的文章，大约占所有企业的 2/3。

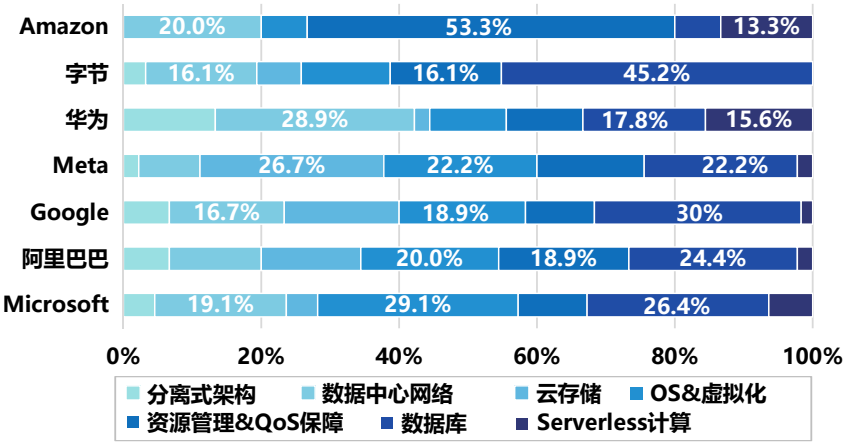


图 1.8: 近三年主要企业在云计算各热点研究领域的投入侧重

量在其总发文量中的占比。为了提高可读性，本节根据研究热点所在的层级和技术相似度，将 OS、分布式系统与虚拟化 3 个方向的研究成果进一步合并，而新硬件和新兴领域其他杂项由于较低的相关性或缺乏短期盈利能力，暂不考虑在内。

发文量排名前 7 的头部企业几乎在各个热点研究领域均有布局和投入。如图 1.8所示，头部企业从经典的数据中心网络，存储，数据库，虚拟化以及能效管理等领域，到新兴的 Serverless 计算业务以及以 CXL 内存池化等为代表的分离式架构都有研究涉猎。例如，数据中心网络向来是云厂商所关注和重点布局的方向，尤其是近年来 RDMA、智能网卡、DPU 等新技术又持续不断地增强了数据中心的网络能力。在云存储以及 OS 与虚拟化方面，阿里巴巴，Google 和 Meta 均有接近或超过 15% 的自身研究投入占比。而在数据库方面，除了 Amazon，其余 6 家企业均在自身研究布局中有接近 20% 的投入，这也反映了当前数据量飞速增长的时代下对数据查询对于业务能力提升的重要性，特别是以 AIGC，大模型训练等为代表的业务场景更是增加了对大规模数据检索在稳定性，可靠性和响应性能方面的要求。

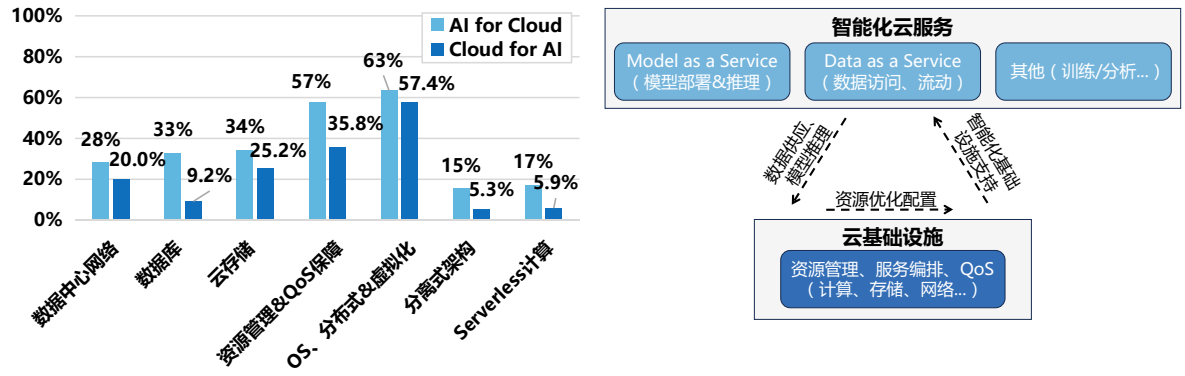
通过进一步的分析，本文将各个研究领域成果所涉及的技术点关键词进行了统计。如图 1.9所示。例如在数据库和存储领域，数据库基于分离式架构 [8]、多主架构 [9, 10] 和智能服务 [11] 提供了云原生 Serverless 可扩展的数据库服务，并面向 AI 场景基于向量检索 [12] 提供非结构化数据的加速检索服务。云存储方面，为支撑新兴的大规模智算和海量存储需求，利用全闪/混闪架构和新型的高速、高密度存储介质（如 ZNS/Open-channel 闪存、叠瓦式存储等）[13, 14, 15]，辅以软件层面优化（重删压缩、负载均衡、元数据扩展等）[16, 17, 18, 19]，实现更大规模、更高性能、更低成本的存储服务。在 Serverless 计算领域，主要通过镜像压缩、快照、代码裁剪等技术实现函数实例的冷启动加速 [20, 21, 22, 23, 24, 25, 26, 27]，而沙箱共享，复用等机制则广泛用于函数的工作流编排优化 [28, 29, 30, 31, 32]，用于降低函数运行时资源占用和通信开销。在采用 RDMA [33]、CXL [34] 等技术实现的分离式内存池化架构中，针对云负载的冷热分层动态 Profiling [35, 36]，页表地址翻译 [37, 38] 等关键技术被反复提及。表 1.2具体列出了全球头部云厂商在重点关注的 7 大热门研究领域中的主要代表性成果，研究分类以及所发表的国际会议名称。

图 1.9: 研究热点词云

表 1.2: 头部企业重点关注的七大热门研究领域

研究点	研究方向概述	会议	研究主要关注点与代表性工作
数据中心网络	网络是连接数据中心内计算、存储等资源的关键核心组件，头部云厂商长期对此重点投入研究。	SIGCOMM NSDI EuroSys	• **网络架构：** 阿里、Google 等持续探索创新网络架构，优化网络密集型应用的调度策略和运行效率 [39, 40]； • **流量调度：** 腾讯云等通过零排队的拥塞控制和可编程调度以提升用户体验 [41, 42]； • **网络协议：** 阿里、Meta 等利用 DCTCP 和 QUIC 等新型协议以优化网络传输 [43, 44]。
数据库	云数据库厂商主要基于三层池化架构探索 Serverless 化、多主可拓展的云原生数据库架构，同时在湖仓一体、数据库智能化、向量检索、数智融合等方面重点研究。	SIGMOD PVLDB ICDE CIDR EDBT	• **池化 & 多主：** 阿里云基于三层池化架构构建 Serverless 化的云原生数据库 [8]，阿里云 [10] 和华为云 [9] 都在探索存算分离平台下支持多写多读的分布式数据库架构； • **湖仓一体：** Databricks 在 Spark 基础上提出了 Delta Lake [45] 的技术，让数据湖生态支持事务和范式约束能力，进一步引领湖仓一体 [46] 技术体系； • **向量检索：** Zilliz 推出云原生的向量数据库 Miivus [12] 加速 AI 场景相似性检索的效率，Zilliz 目前在向量数据库赛道第一； • **数智融合：** 华为 GaussDB [11] 积极布局数据库库内智能计算的能力。
云存储	作为云计算的核心组成部分，云上数据存储朝着高可扩展性、低成本、高性能、稳定可靠、易用安全的方向加速变革。	FAST SC EuroSys	• **底层硬件：** 阿里、IBM 等云厂商利用高速闪存和高密度存储介质在提升存储系统性能的同时持续降低成本 [15, 47]； • **存储平台：** 华为、百度、Whamcloud 等企业针对数据去重、缓存协议、元数据管理等关键技术上持续进行智能化与动态化的探索 [18, 48, 49]； • **使用场景：** 华为等企业针对大模型场景，研究新型分布式 KVCache 存储系统 [50]。
资源管理与 QoS 保障	数据中心研究主要集中在改善资源利用率、优化和保障云服务的 QoS 等方面。近年来，诸如 GPU 应用加速、异构算力管理，绿色低碳也是新兴研究热点。	ASPLOS OSDI EuroSys SoCC NSDI	• **云原生 & 自动化运维 & 节能：** 阿里云结合静态策略和运行时调整技术构建低成本、快速响应的日志存储系统 [51]；字节跳动针对内部的大规模 Spark 集群进行负载分析，优化资源分配 [52]；Meta 利用机器学习算法分析业务特征，优化数据中心能耗管理 [53]； • **应用加速 & 异构资源管理:** Intel 利用智能网卡减少数据中心跨节点的数据 IO 传输和内存占用开销 [54]；Microsoft 面向多阶段 GPU 推理服务的 QoS 保障资源管理 [55]； • **SLO 保障 & 负载聚合及扩缩容:** 阿里云针对大规模集群的负载聚合策略研究 [56]；Google 构建应用感知的自动扩缩容机制，以改善微服务的 QoS 与资源分配 [57]。
OS 与分布式系统	OS 与分布式系统的研究目前广泛聚焦于内存管理、大语言模型服务、分布式通信及云原生技术，以提升系统性能、响应速度和资源利用效率。	OSDI SOSP ATC ASPLOS EuroSys	• **大规模云数据中心内存资源管理：** Google、Meta 在其数据中心集群内存分级技术，降低数据密集型负载的存储成本，同时提升数据中心内存资源利用率 [35, 36, 58]； • **分布式应用通信加速：** 阿里云、天翼云分别提出采用 CXL 共享内存实现分布式服务间 RPC 加速 [59] 和 Socket 通信加速 [60]； • **大语言模型服务优化：** 阿里通过在多个模型推理服务实例之间迁移任务来获得更好的负载均衡 [61]; Microsoft 在 LLM 服务请求的预填充阶段引入分块，实现调度加速 [62]。
分离式架构	传统计算与存储的分离架构逐渐出现资源利用不均、弹性粒度不足等问题。分离式架构将“内存池”进行独立资源管理优化，以提升资源利用率，解决资源匹配和分配不均问题。	OSDI SOSP ASPLOS EuroSys NSDI	• **CXL 内存池系统：** Microsoft 和 Intel 利用 CXL 高速互联总线技术进行内存池化场景下的多租资源分配，以提升内存资源使用率，并减少内存性能劣化 [34, 63]； • **RDMA 远程键值内存池系统:** 华为云解决现有内存分离架构中键值存储系统的索引问题，提升该架构下存储系统性能 [64, 65]； • **内存池高可用：** 阿里和 Google 分别研究现有 RDMA、CXL 内存池架构中的高可用问题，减少分离式架构带来的爆炸半径扩大影响 [66, 67]； • **分离式语言运行时：** 华为提出在分离式架构下支撑数据密集型系统的分布式运行时，该运行时可使得用户不必感知分离式场景下的数据布局以及底层硬件的状态 [68]。
服务器无感知计算	Serverless 计算是新兴的云计算编程范式，研究热点目前主要集中在优化冷启动时延、加速函数间通信、提高业务场景适配性以及优化函数运行时性能及安全性等方面。	OSDI ASPLOS EuroSys ATC SC	• **函数编排调度优化：** 阿里云性能感知的函数资源调度 [69]，华为云基于机器学习方法预测虚拟机内多实例混部下的性能干扰，提高函数部署密度； • **冷启动优化:** Amazon 基于镜像分块和分层缓存实现按需加载的容器冷启动优化 [70]； • **应用适配与移植:** IBM 基于 Serverless 构建弹性存储服务，降查询成本 [71]；Meta 在边缘利用 Serverless 实现流处理应用的部署和服务成本优化 [72]；华为云利用普通函数和非对称函数结合的方式实现函数推理加速 [73]； • **Trace 分析:** 华为分析商用 Serverless 集群的负载特征，为业界提供研究方向 [74]。

1.2.2 智能技术与云计算相结合



(a) 近三年企业参与研究成果与 AI 结合比重　　(b) 智能技术与云计算的关系示意

图 1.10: 智能技术与云计算相辅相成

对云计算领域近三年的研究分析表明，智能技术与云计算的结合度显著提升，当前研究以智能技术赋能云计算为主，而云支撑智能技术的研究仍在不断深化。根据 AI 所扮演的角色，图 1.10(a)中展示了各个研究领域中 AI 参与的研究工作占比。可以看到，利用 AI 算法优化资源分配，进行智能化决策和应用性能加速已经得到了相对广泛的应用，在整体研究中占比接近三分之一，在部分领域甚至占比超过了 50%（该部分详细内容请见本文第三章）。而在面向 AI 的系统或平台优化设计方面，现有的研究工作占比则相对较低（平均低于 20%）。在一些较为新兴的研究领域中，例如分离式架构和 Serverless 计算的研究内容与 AI 的结合度更低，在未来还有很大的探索空间。

前沿研究指出，智能化的前提是数据先行（Data First）[75]，数据是智能化的核心驱动力，而云则是支撑数据高效流转的关键底座。随着智能化浪潮的兴起，云计算面临着前所未有的挑战与机遇，不仅需要提供强大的计算能力和存储资源，还需通过数据流通、服务交付等功能，成为智能化应用落地的核心支柱。尤其是在近年来，基于云计算的“数据即服务（Data as a Service，DaaS）”和“模型即服务（Model as a Service，MaaS）”等创新服务模式，进一步推动了智能技术与云计算的深度融合，为智能化发展开辟了更多可能性。图 1.10(b)展示了智能化云服务与云基础设施间相辅相成，互相成就的关系。具体来说，在智能云时代，智能技术与云计算的结合主要体现在以下两个方面：

（1）AI 驱动下一代云从上层服务到底层设施的智能化变革

AI 技术在多领域推动云服务智能化革新，让传统云服务更简单易用。在云游戏、云视频和云教育等领域，通过大模型生成文本和图像的技术，实现了更低延迟的实时服务和更智能化的交互反馈，给用户带来了前所未有的沉浸式体验。而对于数据库等传统上需要专业技能操作的云服务，利用 Text-to-SQL [76] 技术，用户只需简单的交互即可实现高效查询与自动优化，使得云服务更加“触手可及”[77, 78]。

AI 技术为数据中心资源管理提供新解法，实现数据中心多维降本增效。由于云的负载多样性和规模庞大，数据中心内的海量资源管理是个长期存在的难题。过去，云服务提供商主要依赖于启发式算法和专家规则，而 AI 技术的引入则提供了新的解决方案。目前，如多层感知机、梯度提升机、决策树等机器学习算法已经在用户负载预测、数据布局优化和异构资源管理方面取得了显著成效，成功提高了数据中心的资源利用率并降低了能源消耗和硬件成本 [79, 80]。

AI 技术在云基础设施方面已初步应用，展现出构建智能基础设施的巨大潜力。尽管底层硬件管理难度较高，但 AI 技术在云上的计算、存储和网络等方面已得到了初步探索，诸如学习型索引 [81] 和学习型缓存等智能算法已经在数据中心上得到了部分应用。AI 技术在实现更高效的内存管理、更智能的缓存协议、更简洁的数据格式和更稳定的互联网络等方面已展现了巨大的潜力 [82, 83]。

（2）云技术助推大模型多维突破创新

云服务厂商加速构建超大规模智算中心，助力基础模型持续创新。为支持人工智能研究中的“Scaling laws”，云厂商需要提供强大的计算资源并进行持续的技术创新和优化，以适应大规模 AI 模型的需求。在构建具备 TB 级网络吞吐、百万级存储 IO 及互联万卡集群的基础设施时，云厂商在多维混合并行、大模型检查点支持与训练能耗感知等方面深入开展技术创新，朝着构建高可扩展、高速度、低能耗的下一代智算中心持续演进，以保障大规模模型训练的高效性与稳定性 [84, 85]。

云基础设施聚焦联合海量异构硬件能力，实现软硬件协同优化，释放云上智算潜力。在各类智算硬件和新型加速器涌现的同时，如何充分发挥硬件潜力以助力云上训练，也是云厂商关心的问题。针对模型训练与推理场景，从数据中心异构 GPU 管理、AI 加速器与智能网卡应用、分布式 KVCache 管理、分布式通信压缩到算子编译框架设计等展开一系列软硬件协同优化，旨在精准提升硬件利用率，充分释放云上智算潜力 [50, 86]。

云新型架构拓展人工智能应用场景边界，构建云边端一体架构，促进多领域融合落地。在万物互联的背景下，云厂商成为推动智能技术应用场景广泛覆盖与垂直发展的关键力量。目前，云厂商在提供云侧丰富算力资源的基础上，还持续挖掘边缘算力和端侧设备的潜力，如利用联邦学习技术将数据预处理及推理任务下放边缘与端侧。目前，实现数据在云边端的高效流转与协同处理，构建多元、无感知且安全隐私的云边端一体计算架构，是推动 AI 技术与智能交通、工业制造、远程医疗等多领域深度融合广泛落地的重要走向 [87, 88]。

1.3 下一代电信云的展望和发展建议

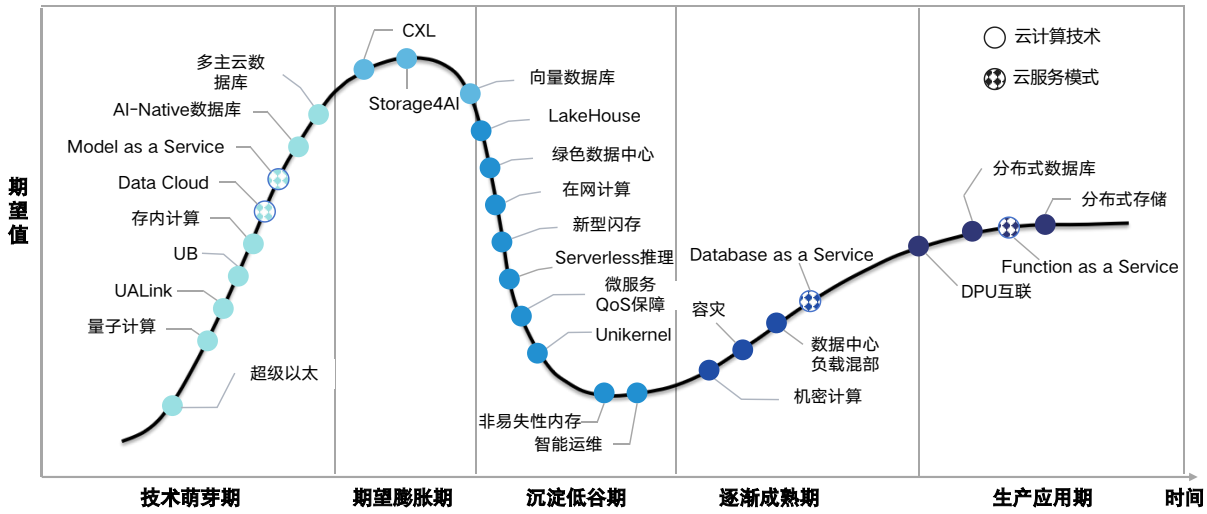


图 1.11: 云计算研究图谱技术成熟度曲线 2024

技术发展的规律性研究一直是学术界关注的重点。其中，Gartner 公司提出的技术成熟度曲线 (Hype Cycle) 作为分析新兴技术发展轨迹的重要工具，在全球范围内获得广泛认可。该曲线通过“技术萌芽期 → 期望膨胀期 → 沉淀低谷期 → 逐渐成熟期 → 生产应用期”五个阶段，形象地刻画了新技术从出现到最终成熟的完整过程。在这些阶段中，技术的期望值是定性和相对的。在技术萌芽期和期望膨胀期，期望值是相对较高的，通常表现为过度的乐观预期，而在沉淀低谷期，期望值则显著下降，显示出对技术的失望。这些期望值是通过市场反应、媒体报道和投资者关注等因素进行感知和推测，因而呈现出定性的变化趋势。基于 Gartner 成熟度曲线的分析方法，本节也构建了云计算技术领域的成熟度曲线（如图1.11所示）。与传统 Gartner 曲线不同，本文重点关注云计算生态系统中的近 30 项关键技术，从而帮助读者对当前云计算技术的发展现状与演进趋势形成相对清晰的认识，为有关领域研究与投资提供决策依据。

1.3.1 未来云计算技术趋势与服务模式展望

自适应、自学习、自调优的智能化技术将助力构建意图驱动的云服务模式，实现黑盒化的云上应用降本增效。云服务的复杂部署、使用与管理是限制用户业务上云的一大障碍。但通过大模型的自然语言理解技术，用户意图可高效解析，消除上云的心智和管理负担。此外，智能资源管理与黑盒调优技术，还有望在无用户干预的情况下实现云服务的自动降本增效。目前，实现智能化云服务模式是云厂商进一步开拓市场的重要机遇。

以 Function-as-a-Service (FaaS) 为代表的 XaaS 将成为未来云服务模式的主流。近年来新兴的 Serverless 计算极大地简化了云应用开发的编程方式，其高度弹性的细粒度资源供给方式和按需付费模式也为传统的云计算服务模型（例如 IaaS 和 PaaS）带来变革，目前已在大量业务场景中得到了广泛应用。未来，诸如 MaaS, Database-as-a-Service (DBaaS) 在内的云产品 Serverless 化将变得更加开放和多样化，不同云服务模型间的边界也越来越小，云用户对于更加高效灵活的云资源使用模式和高度透明的 QoS 保障能力的迫切需求也将推动 XaaS 在云计算市场中占据重要位置。

新一代数据中心的架构持续演进，例如基于 DPU 的高效灵活架构、资源池化的分离式计算架构等。分离式计算架构通过实现计算、存储、网络等资源的解耦与按需重组，不仅能够支持各类资源的独立扩展与按需分配，还能凭借高速互联总线技术实现大规模异构资源的互联互通，优化数据密集型等工作负载性能，减少资源闲置。除以此外，分离式数据中心架构还能为大模型训练等新兴应用场景提供更高效的基础设施支持，可更好适应未来数据中心对敏捷部署、智能调度和性能优化等多样化需求。

1.3.2 云计算未来发展建议

推动数据先行，建立数据采集和数据管理的体系化方案，为研究、开发和运营（RDO）铺设桥梁。高质量、规模化的数据是开展智能技术研究的前提，也是智能技术应用落地的必要条件。基础研究需要使用数据，而数据采集通常在应用运营环节，数据管理则可以看作系统开发的一部分。集团层面体系化的数据采集和数据管理方案将成为连接研究、开发、运营的桥梁。数据在研究、开发、运营各个环节的无缝流转将催化智能技术的快速迭代，最大化发挥智能技术在云计算中的潜力。

促进智能技术与云技术深度融合，构建内生智能和智算原生的下一代云服务。在当前的智能时代，云服务不仅需要引入智能技术进一步变革，也需要为智能技术做好基础支撑：一方面，推进智能技术（如意图理解、负载预测、资源调优等）在数据中心的广泛落地与赋能，实现自治管理、资源优化的内生智能云服务；另一方面，针对智算应用的新兴需求（如 KVcache 管理、检查点存储、GPU 直通等），实现智算原生的云服务，助力智能技术进一步发展。

未来持续增加对 PaaS 层的投入，在通用计算领域和垂直领域产品 Serverless 化同时发力。在当前国内 IaaS 市场已近饱和的形势下，PaaS 层的产品能力将更能体现云服务商在云资源供给，云服务编排能力等方面的技术成熟度，也是突破增长瓶颈进一步提高资源收益率的良好契机。借助当前火热的大模型业务需求和云服务 Serverless 化趋势，通过构建以 MaaS 为代表的面向垂直领域深度优化的 PaaS 产品，同时与通用 Serverless 计算产品相结合，实现云服务商从以“售卖资源”为主体到以“售卖服务能力和解决方案”为主体的转变。

积极探索底层架构的技术创新，引领下一代云计算关键技术的发展与变革。例如探索基于 DPU 的数据中心架构、支持内存池化的分离式架构等。分离式架构的创新实践主要体现在三个方面：一是建设基于新型高速总线技术（如 CXL，UB 等）的分离式内存池，解决资源池高可用、内存拉远导致的性能劣化等关键问题；二是推进现有数据库、大数据等平台软件实现内存池化的改造适配，加速分布式应用的性能优化；三是探索向用户提供内存即服务等新型云产品服务模式。这些探索不仅能为用户提供更灵活的资源供给方式、更优的性能体验和更低的使用成本，还将为云计算基础设施的演进开辟新的技术路径。

第二章

围绕云计算的云网融合研究

以 2020 年 11 月发布的《云网融合 2030 技术白皮书》为标志 [89]，中国电信一直致力于在研究、开发、运营的各个层面上协同推进云网融合战略的发展和落地，云计算研究院作为中国电信内部的学术研究机构，主要承接围绕云计算的云网融合研究。本章以第一章的研究探讨为基础，同时基于对《云网融合 2030 技术白皮书》中愿景架构图等内容的理解，探讨云网融合相关的研究话题。

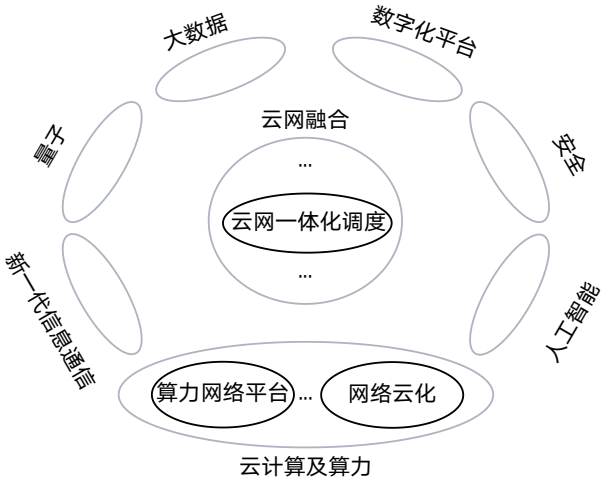


图 2.1: 本章主要讨论云网融合的三项技术

云网融合的研究重要且宏大，如图 2.1所示，七大战新领域紧密围绕在云网融合核心战略周边，各自承接相应的研究工作。云计算研究院聚焦在云计算及算力领域承接的云网融合研究，同时也在力所能及的范围内对云网融合核心研究做一定贡献。本章主要围绕三项技术展开讨论：在云网融合核心研究中，基于云计算研究院团队过往的研究积累，主要探讨一个需要长期研究投入的技术难题，即如何实现最优的云网一体化调度，以此作为云计算研究院参与云网融合核心研究的一点贡献。在云计算及算力领域，本文探讨两项与云计算密切相关的重要技术：算力网络平台和网络云化。

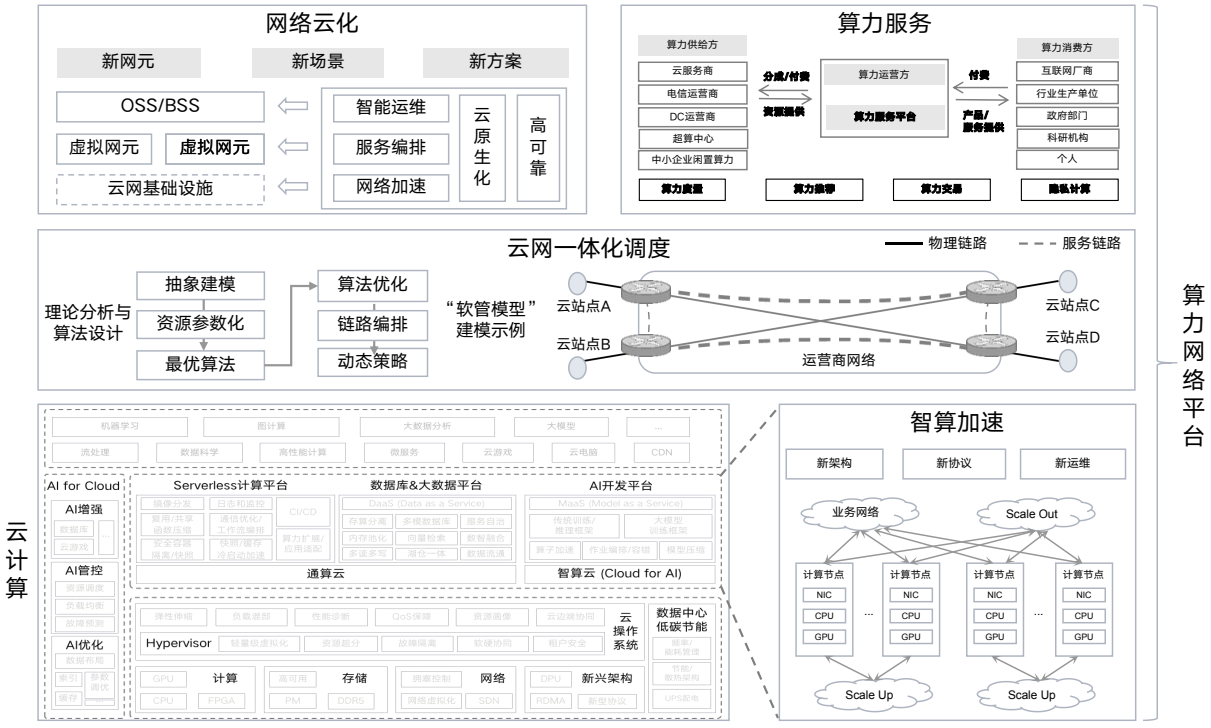


图 2.2: 研究图谱：云计算和云网融合三项技术的架构关系（云网一体化调度、算力网络平台和网络云化）

本章内容的组织方式与第一章类似，2.1节介绍云计算与云网融合三项技术组成的研究图谱，以及相关的产业和技术分析，2.2节围绕三项技术讨论研究热点和研究难题，2.3节提出展望和建议。

2.1 研究图谱及其产生：云网融合产业和技术分析

为便于直观理解云网融合的三项技术与第一章云计算研究的关系，图 2.2以架构图的形式把图 1.1和上述三项技术做了关联。其中，云网一体化调度作为云网融合的一个关键技术难题放在中间位置；算力网络平台则涵盖了上方的算力服务，中间的调度，以及云计算研究本身也涵盖的智算加速（见第 1.2.2小节）；网络云化则作为云计算的一个业务需求放在上方。

本节的后续内容围绕云网融合三项技术，分别介绍云网融合的基本概念和发展现状、国内外行业标准，以及国内外产业进展。

2.1.1 基本概念和发展现状

本小节介绍云网融合的基本概念和发展现状，基于本章开篇对内容范畴的讨论，本小节围绕云网融合的三项技术，分别介绍每项技术的基本概念和发展现状。

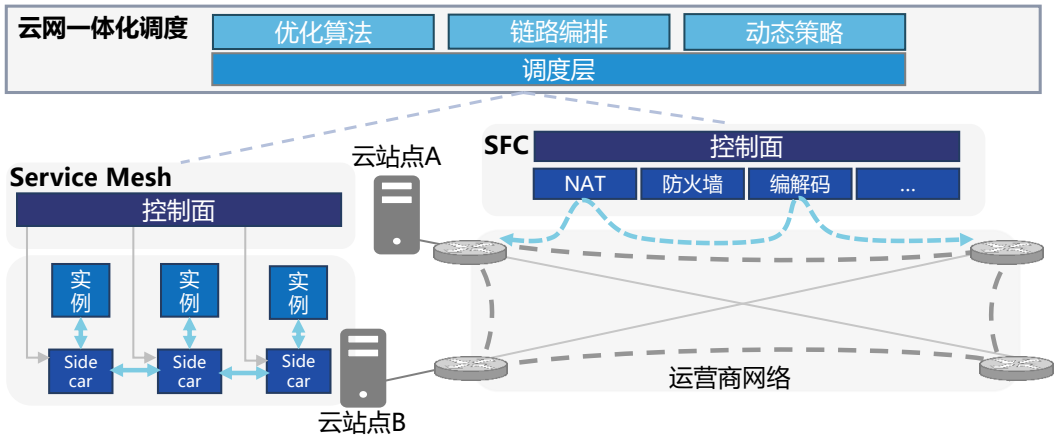


图 2.3: 云网一体化调度应用示例

云网一体化调度的理论难点是最优调度与计算复杂度的弹性平衡。云网一体化调度作为云网融合的核心能力之一，可以将云计算资源与网络资源进行高效整合和统一调度，以实现云网资源对上层业务的一体化支撑。在进行基础理论研究时，可以通过“软管模型（hose-based model）”[90] 对调度问题进行抽象简化，该模型为每个站点赋予一个“软管”，表示总带宽需求。通过动态调整流量，避免单独规划带宽，实现灵活调度。然而，求解此模型通常为 NP-Hard 问题，导致在大规模系统中获取精确解几乎不可行，通常只能得到近似解。云网一体化调度技术关键理论难题是依据实际业务动态需求，在调度最优解与计算复杂度间寻求弹性平衡。调度算法的理论分析通常包括抽象建模、资源参数化、最优算法、算法优化、弹性链路编排和动态策略六个步骤，以实现云网资源的跨域协同调度。由于全局最优调度方案的复杂性，常采用次优解和近似算法，以降低计算复杂度并在合理时间内提供接近最优的解决方案。根据现有研究文献 [91, 92]，某些特定问题的近似比率可达到 *2-*近似来实现在多项式时间内找到一个接近最优解的解，在实际应用中，近似结果的误差范围可能会因具体问题而异。另一类解决方案为基于 AI 算法，尽管这类算法可以通过对复杂模式的识别来加速寻求最优解 [93, 94]，该类算法方案在应用时缺乏可解释性，在遇到异常情况时，使得问题排查和修正变得更加困难，影响运营效率。此外，云网一体化调度还面临系统层面挑战：如图 2.3展示了云网一体化调度应用部署的一种示例；与传统的服务功能链（Service

Function Chain, SFC）和服务网格（Service Mesh）框架相比，云网一体化调度应用部署更注重跨域联动协同。关键挑战包括动态性和复杂性管理、负载均衡策略制定、数据一致性等。

算力网络是云网融合关键技术路径，智算引领算力网络焕发活力。云网融合的愿景目标是通过实施虚拟化、云化和服务化，形成一体化的融合技术架构，最终实现简洁、敏捷、开放、融合、安全、智能的新型信息基础设施的资源供给。算力网络作为这一愿景的关键技术路径，支撑了云网融合从云内、云间到入云的各个阶段，不断推进和深化云网融合的服务能力。算力网络结合网络信息和用户业务需求，提供计算、存储、网络等资源的分发、关联、交易与调配能力，实现全网整体算力资源的优化配置和使用。因数字化转型与经济发展、算力资源分布不均与供求失衡、新兴应用对算力的爆发性需求三大因素驱动，算力网络自提出以来一直热度不减。中国电信在 ITU-T Y.2501 中提出了算力网络的概念与架构，涵盖的关键技术主要包括算力度量、算力感知、算力路由、算力交易、算力推荐及隐私计算等。AI 大模型引领算力网络焕发活力（Network for AI），AI 大模型 scaling law 带来的爆发式算力增长需求使得算力网络成为提升算效的重要解决方案。然而 AI 大模型训练/推理等工作需要在大量的计算单位中传递海量数据，对算力网络提出了新需求：具备超大规模、超低时延、超大带宽、超高可靠等关键特征。当前产业界围绕智算集群的技术攻关方向主要还是以集群内部为主。当前集群内，围绕着 GPU 存在三大互连，分别是业务网络互连、横向扩展（Scale Out）网络互连、横向扩展（Scale Up）网络互连，如图2.4所示，它们分别承载了不同的职责：跨业务、集群内、超级点 GPU 之间连通性。根据智算对网络的新要求，其涵盖的技术点聚焦在新兴网络拓扑、高性能无损网络技术、集合通信算法优化、分布式协同训练与推理、故障感知与恢复等。

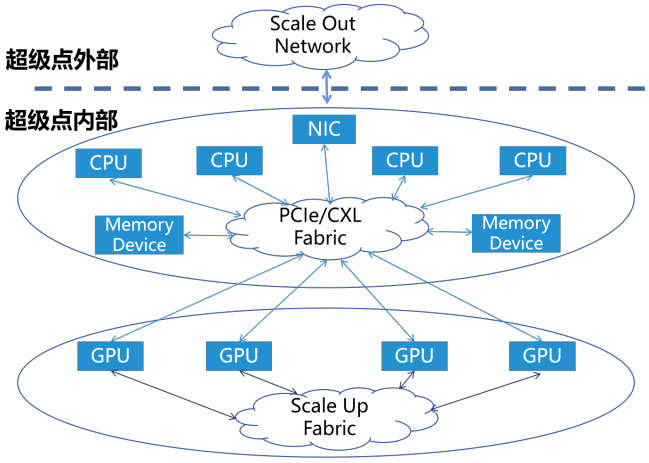


图 2.4: 智算集群网络架构

网络云化是网络的发展方向，需要云能够满足网络的特殊需求。随着新技术和应用的不断创新，网络从传统封闭刚性向更加开放灵活的方向发展。网络功能从以硬件为主体的架构向以软件为主体的架构演进，旨在实现弹性资源分配、敏捷灵活组网、自动智能运行等目标。网络功能虚拟化（NFV）技术的发展实现了软硬件的解耦，重新定义了通信网络的开发、部署、运营模式。当前阶段，如何利用云计算技术的先进性、云资源的弹性能力进一步为网络赋能，已经成为研究的热点，也是云网融合的关键问题之一。网络云化主要关注网络功能云化和云基础设施承载两部分。在网络功能云化方面，通过引入微服务、Serverless 等服务化架构和编程范式，使网络功能更加开放解耦，在应用层实现灵活可扩展；通过引入持续集成/持续部署（CI/CD）等敏捷服务开发范式，缩短服务的开发周期。在云基础设施承载方面，网络功能因其复杂性和定制化需求，对云提出了更高要求。包括：增强网络能力，实现低时延大容量网络处理；提供更加可靠的基础设施能力，满足网络功能高可靠运行要求；实现跨域多专业网络功能的统一编排与管理，满足网络功能多样化部署需求；建立自主、可控、可信的网络云安全防护体系，提供与传统物理网元等效、甚至更高的安全运行环境等。

2.1.2 国内外行业标准

相比目前云计算中行业标准的相对辅助性的作用（见第 1.1.2节），云网融合更强调标准的引领作用。本小节依然围绕云网融合的三项技术，分别介绍每项技术在国内外行业标准方面的进展。

云网一体化调度快速发展，聚焦资源管理调度框架。在技术标准方面，各标准组织（如 ETSI、ITU-T）

正在推动云网一体化的标准化进程，以实现不同系统之间的互操作性，例如 ETSI GR IPE 002 讨论了在 IPv6 系统下对于云网一体化调度的要求，以及国内由信通院主导的 ITU-T Y.3538 分布式云全局管理框架、ITU-T Y.ccts-frame 等。海外业界目前主要关注统一安全访问服务边缘 (SASE)、广域网即服务 (WANaaS)、多云网络等方面，Microsoft 推出了 SASE 产品，Google 提出 Cross-Cloud 框架，Amazon 与 Fortinet 合作引入了 SASE 能力。国内业界目前主要关注云网资源的统一调度管理。关于网络资源调度的标准工作较为广泛，例如 ETSI 确立的 MANO 网络功能调度参考模型，为网络资源管理调度方案设立了较为统一的标准。由于计算行业的特点使然，关于云计算资源调度的方案目前以开源框架为主导，以 Kubernetes 为代表的开源容器编排平台成为主流方案之一，在此基础上，OCI、CNCF 等技术组织通过对 API 和规范的制定来推进云资源调度的标准化工作，确保不同工具和平台间的互操作性，例如 runtime-spec 容器运行时规范以及 kube-scheduler 框架规范。

算力网络标准逐步走向体系化，智算网络标准待进一步完善。当前算力网络标准化主要在国内 CCSA、国际 ITU 及 IETF 开展工作，主要参与方为国内三大运营商。算力网络标准体系根据 Y.2501 架构开展，如图2.5所示。其中标准涉及的关键技术主要可以分为需求架构类、关键技术类、平台系统类、设备类及协议标准类。关键技术集中方向为算力度量、算力感知、算力路由、算力交易、确定性承载与算网编排调度。智算网络相关标准仍处于起步阶段，智算网络标准体系有待进一步完善。当前国内的智算网络标准化工作主要集中在 CCSA，主要参与方为三大运营商以及厂商、互联网公司及信通院。主要包含总体技术要求、基于 RoCE 协议的无损以太网、存算一体/异构算力、安全、承载智算业务的广域网能力要求及设备/平台互联互通。国际上智算网络标准化工作主要在 ITU 及 IETF 开展，主要参与方为国内三大运营商、信通院、华为等。中国联通、中国电信、信通院、紫金山实验室围绕 NGNe 在 SG13 启动智算立项，研究分布式智算中心在 NGNe 中的网络增强需求和能力及对广域无损网络的控制器提出功能要求，以增强控制器在路径计算、流量调度、流量控制、拥塞控制等方面的能力。在 IETF 中，中国移动牵头成立算力路由 (CATS) 工作组，中国联通/华为、中国移动/新华三提出相关文稿研究广域网中实现无损技术的用例和需求及基于 RoCEv2 的集合通信卸载。

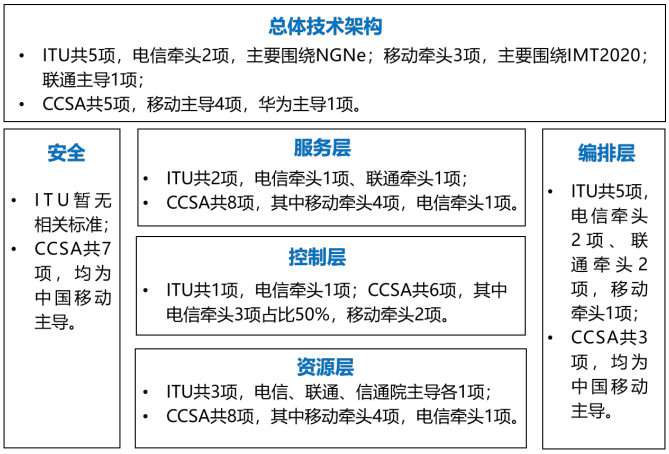


图 2.5: 算力网络标准体系

标准是网络云化最重要的推动力量，助力网络云化的规模化应用。ETSI NFV ISG 是网络云化的发源地，后续 3GPP、5GPPP、IETF、TM、BBF、O-RAN 等标准组织和联盟也开展了相关的研究。NFV 的发展经历了多个阶段，如图2.6所示，每个阶段面向不同的挑战和需求进行演进。在 Release 1 中，通过定义行业普遍接受的术语、架构框架以及适用于 NFV 系统的高级需求，为 NFV 的发展奠定了基础。随着 Release 2 的到来，现场试验和互操作部署的出现使得解决互操作性问题变得迫切，NFV 的网络功能分解带来了多供应商互操作的额外挑战，要求虚拟化网络功能（VNFs）能够独立于供应商进行打包，与独立开发的管理和编排系统进行互操作。Release 3 随着 NFV 部署从现场试验转向大规模部署，增强了对部分特性（如多站点服务管理、软件更新/升级、故障排除等）进行规范的需求。Release 4 以“云化和简化”为目标，基于利用先进的云计算和网络管理技术来简化 NFV 部署，重点集成容器管理和自主网络技术。Release 5 以“整合和生态系统”为口号，进一步解决运营问题，并考虑生态系统中其他组织开发的新用例或技术，如 O-RAN 联盟。最后，Release 6 重点关注接口、模型等架构和基础设施，包括架构的演进和简化，新基础设施，新的虚拟化形式，以及时延等关键问题。ETSI NFV 标准已助力各国运营商构建跨层和多厂商互操作的超大规模电信云基础设施。例如，中国电信基于 ETSI NFV 标准架构构建云基础设施，包括数百个数

据中心的分布式电信云。此外，网络中的虚拟化占比在不断增高，公有云厂商宣布在其电信网络管理服务解决方案中支持 ETSI NFV 标准。

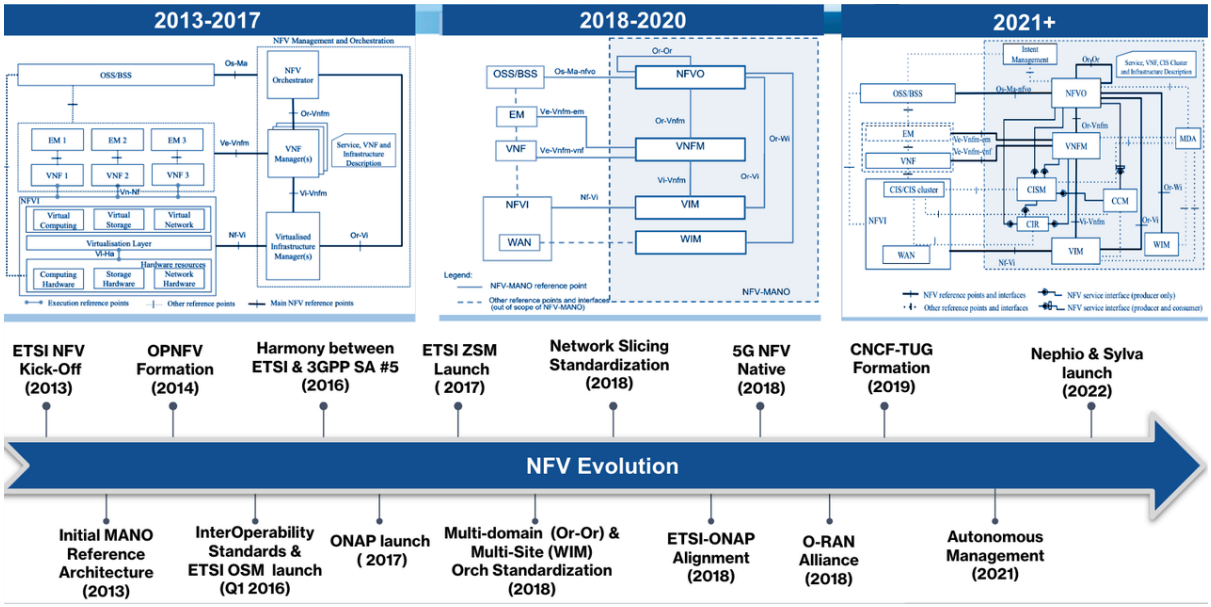


图 2.6: NFV 标准发展历程

2.1.3 国内外产业进展

本小节介绍云网融合的国内外产业进展，依然基于本章开篇对于内容范畴的讨论，围绕云网融合的三项技术分别介绍各项技术的国内外产业进展，并在表2.1给出了云网融合当前产业相关产品。

云网一体化调度推动云网服务一点供给。云网一体化调度作为云网融合的理论基础，国内三大运营商都在持续推进云网一体化服务。中国电信在“云网融合”的战略规划下持续加强天地云网一体化布局，中国电信天翼云推出了“息壤”算力服务平台，同时，中国电信还推出了“昆仑”云网能力开放平台，将多因子全局最优调度列为核心技术之一 [95]。中国移动提出了“一朵云、一张网、一体化服务”的云网一体化策略。中国联通构建了面向政企客户的线上云网一体自服务平台云联网系统。在工业界的实际部署中，云计算资源调度与网络资源调度仍处在两个独立的系统中，分属于云厂商和网络运营商，两边供应商分别调度各自系统内的资源，难以统一起来。在云计算资源调度方面，云厂商基于云产品情况会开发各自的资源调度平台，如国内厂商的阿里云资源编排服务、天翼云息壤算力服务平台，海外云厂商的 AWS Lambda、Azure Logic Apps 等，目前这些平台的开发会参考开源组织旗下的有影响力的开源项目，比如 Kubernetes、Istio、Linkerd 等资源调度框架。在网络资源调度方面，软件定义网络（SDN）及 NFV 等技术推动了网络的开放及可编程性，催生出了基于 ETSI MANO 参考模型的开源方案 OSMANO，以 ONF 为代表的开源组织及针对网络服务调度的开源网络控制器（如：ONOS、OpenDaylight），和 Microsoft 主导的针对网络交换设备资源调度的开源 SONiC 方案。在这样的背景下，云网一体化调度的另一主要挑战来源于云计算与网络资源调度系统在前期的各自演进，给调度优化方案在系统内整合统一后的实际性能带来了极大的挑战。

算力网络以打造算力服务平台实现产业化布局。随着算力时代的到来，中国三大运营商已制定战略规划，积极构建算力网络和平台产品，如中国电信的“云网融合”战略和息壤平台，中国移动的“算力网络”规划，以及中国联通在 CUBE-Net 架构下的算力网络发展。这些行动显示了国内运营商整合算力资源、提供高效服务的决心。同时，国内外企业如并行科技和 io.net cloud 也在推动分布式算力网络的建设，使算力网络成为全球竞争的新焦点。在全球智算网络领域，科技巨头正通过技术创新来满足日益增长的 AI 模型训练和计算需求。国内领军企业如阿里云推出的 HPN7.0 智算集群网络架构，以其全自研 51.2T

表 2.1: 云网融合行业产品

技术方向	企业名称	企业属性	产品	产品描述
云网一体化资源调度	中国电信	国内运营商	“昆仑”云网能力开放平台	实现云网服务一点供给
	中国移动	国内运营商	智能云网编排平台	实现业务统一受理、开通自动化、端到端运维
	中国联通	国内运营商	云联网系统	面向政企客户的线上云网一体自服务平台
	Azure	国外云厂商	统一安全访问服务边缘 (SASE)	实现安全性和网络连接的融合
算力及智算服务	中国电信	国内运营商	息壤一体化智算服务平台	纳管算力 27E，打造智算生态系统
	中国移动	国内运营商	算网星图算力并网服务平台	通算 8.2E、智算 26E、三方算力 3.4E
	中国联通	国内运营商	算网一体化怀来编排调度平台	运营 1400PLops 算力，规划 5 万架标准机柜能力
	并行科技	国内科技创业公司	超算云服务平台	总计算力超 1000P, 存储资源超 800PB
	鹏城实验室	国家实验室	鹏城云脑 II	E 级高性能人工智能算力平台
	io.net cloud	国外科技创业公司	去中心化计算网络	规模 2 万 + 个 GPU，65+ 集群
网络云化	中国电信	国内运营商	天翼网络云	承载虚拟化网元和网络业务平台
	中国移动	国内运营商	AUTO 行云	敏捷网络，自动驾驶，网络智能化
	中国联通	国内运营商	WoMANO	提供网络功能虚拟化管理和编排
	AWS	国外云厂商	AWS TNB、AWS Wavelength	定制化增强云服务以适配 ETSI-NFV 架构
	Azure	国外云厂商	Azure for Operators	实现电信级底座能力增强
	阿里云	国内云厂商	超轻量 5G 核心网设备	小型化 All In One 解决方案
	腾讯云	国内云厂商	Azure for Operators	包括 5G 专网在内的云原生系列产品

交换机和先进的网络技术，已成为行业的技术标杆。腾讯云的智能高性能网络 IHN 以其大规模和超高速特性，展现了网络控制系统和端侧控制系统在精准监控和调度方面的实力，有效消除网络拥塞。百度的 AI-Pool 网络方案通过优化节点间通信，提升了智算节点的互访效率。字节跳动则通过构建大规模训练集群和 MegaScale 系统，强化了其在大语言模型训练领域的能力。国际方面，Meta 通过定制化的数据中心网络和 RoCEv2 通信机制，提升了 AI 训练网络的效率。Google 在其 TPU 网络中采用的 OCS 技术，通过拓扑重构增强了性能和可用性。Microsoft 利用英伟达的 IB 方案构建智算集群，进一步扩展了其在智能计算领域的影响力。这些进展不仅展示了全球科技巨头在智算网络领域的积极布局，也反映了该领域在技术创新和应用实践方面的快速发展。

运营商、云商和设备厂商积极参与网络云化。NFV 规模推广前期，主要采用一体化设备和二层解耦的方案，部署 vBRAS、vIMS 等传统电信应用，这类应用中网络功能处于一个相对封闭的状态，设备厂商占据较大的主导权。考虑到稳定性和成本，运营商没有动力采用突破性技术，目前演进缓慢。随着云计算技术的不断引入，网络功能逐渐走向开放，5GC、vRAN 称为行业实践的主要抓手，运营商、云商和设备厂商都在积极参与其中，运营商逐渐占据主导地位。在基础设施建设方面，多数运营商选择自建网络私有云来部署网络功能及网络业务，如国内三大运营商和美国的 Verizon 等，他们使用自研或第三方云平台。部分运营商近年来开始尝试将 5GC 部署到公有云上，例如 AT&T、Dish、Telefonica 和 Swisscom。私有云模式因其安全可靠和高可控性受到青睐，而公有云尽管其成熟度相对较低，则以其灵活性和较低的短期成本吸引运营商。国内外云服务商借助于 5G 2B 市场参与进来。国内外设备厂商逐步打开网络功能，使之能够部署在电信私有云平台和云服务商公有云平台。通过与运营商和云服务商共同构建解决方案，利用云服务商的技术优势，设备厂商积极推动网络功能的创新和商业化进程。随着网络功能不断开放，以及对公有云先进技术、资源能力、成本优势的期望，将 NFV 基础设施扩展到公有云或者混合云成为行业焦点。总体而言，产业界当前主要关注网络云化规模落地过程中面临的工程技术挑战。

2.2 研究洞察：当前云网融合的研究热点和难题

随着用户业务种类和规模的高速增长，云网融合的趋势日益显著。业界期望通过云网融合来提升业务服务质量，进而催生新场景和新应用。尽管云网融合是大势所趋，但由于云计算和网络系统过去几十

年的相对独立发展，融合过程中仍面临一些关键挑战，需要进一步研究和解决。本节通过广泛调研云网融合领域的学术研究，围绕本章开篇讨论的三项技术，梳理了各项技术的热点问题和主要难题，同时概述了学术研究的相应进展，并进行了列举讨论（如下所示）。

研究热点和难题

1. **一体化调度算法复杂度**：如何在大规模场景中平衡服务的实时性需求与资源调度的最优解？
2. **调度算法优化部署及动态策略**：如何优化算法部署及调度策略以满足低时延服务的动态需求？
3. **算力网络架构与应用**：如何面向分布式异构算力资源设计一体化广域管理架构，实现面向多源业务需求的自适应服务？
4. **算力服务效率与性能提升**：如何实现算力网络中异构算力能力高效评估、任务调度与资源匹配，满足用户 QoS 并提升平台效用？
5. **智算加速与分布式智算协同**：如何面向 AI 模型大规模分布式训练及推理提升算效比及性能？
6. **网络功能云原生化**: 如何在云计算编程范式下构建灵活可扩展的虚拟网络功能？
7. **面向电信场景的云基础设施**: 如何构建高效、可靠、安全的云基础设施能力满足电信级要求？

2.2.1 热点研究问题的剖析

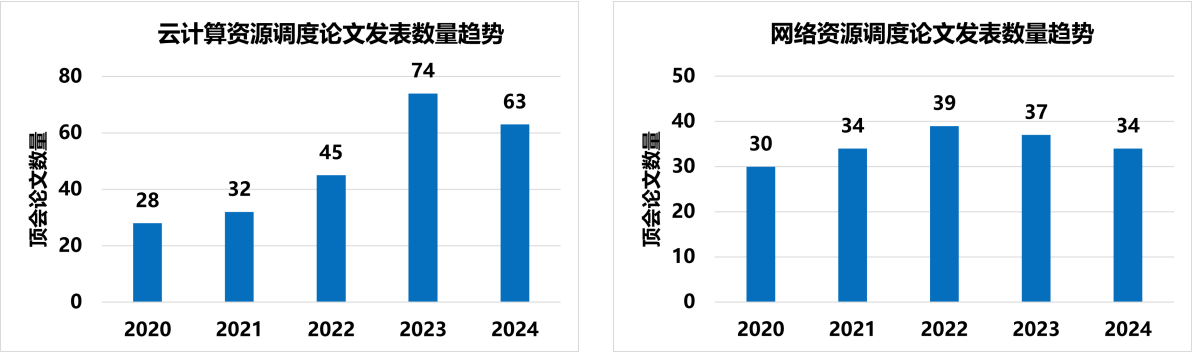
本小节围绕本章开篇讨论的云网融合的三项技术，分别针对每项技术的热点研究问题进行剖析。

研究关注由网络资源调度问题向云网一体化调度问题演进。随着云计算及算力资源接入网络，对于网络资源调度的研究也逐渐过渡到对云计算及网络资源的一体化调度研究。2019 至 2024 年对于云计算资源调度的研究论文呈现上升趋势（如图2.8），其中由于技术架构的演进，研究课题也从针对微服务架构转向新兴的 Serverless 架构。同时，近年来对于网络资源调度的研究成果趋稳。在这些针对网络资源调度的论文成果中，研究课题由传统的网络资源调度逐渐转向如何适应云/算资源接入后的调度扩展。

图 2.7: 研究热点词云

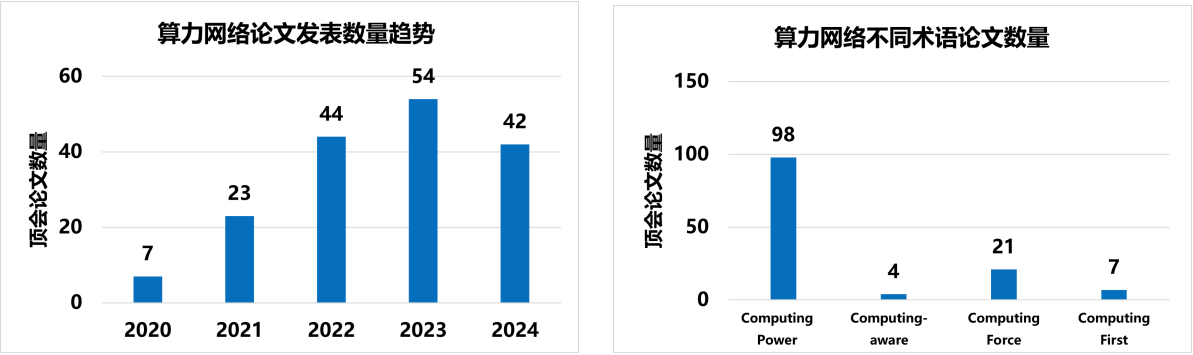
一体化调度的研究热点目前主要聚焦于大规模服务应用中服务实时性需求与资源调度效率的动态平衡。这涉及到在快速响应服务需求的同时，确保资源调度能迅速找到最优解，以优化成本效益和性能。研究热点问题涵盖有对服务实时性与资源调度效率的关系评估，对动态资源分配算法的设计以适应变化的需求和环境。一体化调度算法的动态自适应还涉及系统层面的问题，即：构建响应云网资源负载变化的调度策略框架，开发自适应算法实时监控资源使用并动态调整策略，以保持服务稳定性和效率。此外，一体化调度算法部署优化面临跨调度主体、跨层、跨域、跨云边端的挑战，侧重于全局资源协调优化。研究对云网一体化算法部署的优化，可以减少服务延迟，满足时延敏感的应用场景需求。

算力网络作为应对应用算力需求的下一代网络解决方案，正获得业界越来越高的认可，其研究关注度也在持续上升（如图2.9所示）。自 2019 年算力网络概念提出以来，相关技术论文数量逐年增长（2024 年部分论文尚未检索）。目前算力网络处于研究初期，未来将长期受到关注。在我国，因算力与业务分布不均，算力网络展现出显著发展潜力，最早获得国内关注，并衍生出如 Computing-aware Network（CAN）、Computing Force/First Network（CFN）、Cpmputing Power Network（CPN）等不同命名。Cpmputing Power Network 已成为普遍术语，主要研究者包括电信运营商及高校，而 Computing-aware Network 和 Computing Force Network 主要由中国移动及其合作单位推动，Computing First Network 则以军事科学院为主。



(a) 云计算资源调度论文发表数量趋势 (b) 网络资源调度论文发表数量趋势

图 2.8: 云网一体化调度相关论文趋势分析



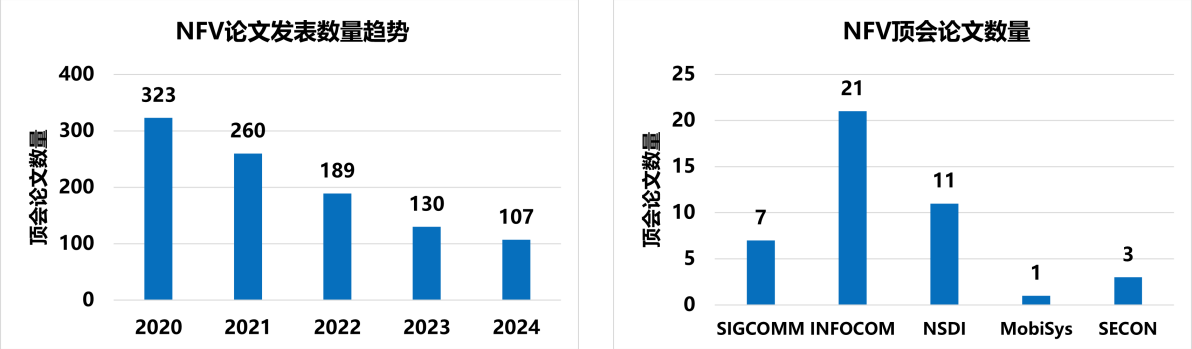
(a) 算力网络论文发表数量趋势 (b) 算力网络不同术语论文数量

图 2.9: 算力网络论文趋势分析

当前针对算力网络技术的研究热点主要围绕算力度量、算力交易、任务调度与资源分配等关键技术实现算力服务效率与性能的提升，以及面向新场景与新应用算网架构的优化设计与展望。主要研究团队是国内院士为代表的团队对算力网络带来的技术体制变革及框架提出了指导。高校、企业及研究机构团队则是在具体的技术点上深入研究，包括 AI 分布式训练及推理架构与算力网络的结合，任务的调度与分配，算力的交易等，因算力网络具有的拓扑属性及时空资源属性，目前调度和路由的研究方法通常是将强化学习，图神经网络等 AI 算法引入，以适应其中动态复杂的多维资源特性。此外考虑到算法的收敛性、高效性与可解释性，一些研究专注于经典的近似算法和贪心算法的研究。值得关注的是，目前算力网络的技术思路已经受到了美国相关研究机构的关注，美国亚利桑那大学对算力网络面向科学计算的前景和技术点进行了论述和分析。目前，研究领域文章侧重算法及理论研究的主要以高校及研究机构为主，企业方涉及较少，未来针对产业及产品化场景中实际应用问题的方法研究仍具有较大空间。

随着十多年的发展，NFV 相关技术逐步走向成熟。如图2.10所示，学术界的研究热度正在逐年下降。分析近 3 年重点会议学术论文，学术界的研究热点转向新的网络功能、新的应用场景以及新的技术方案。研究的主要力量分布在中国和美国的高校（卡内基梅隆、约翰霍普金斯、加州大学、清华大学、北京大学、浙江大学、南京大学等）、云商（阿里巴巴、Microsoft）和设备商（英特尔、华为）。**相对于产业界，学术界的研究内容更加超前，同时关注产业界在规模化落地过程中遇到的云原生化、管理编排等问题**。

在新的网络功能方面，研究方向逐步从 5G 核心网的虚拟化到 5G 接入网的虚拟化，解决 vRAN 如何使用 Kubernetes 部署、故障转移、时延优化、硬件加速、资源共享等问题；在新的应用场景方面，6G、卫星通信、无线通信、车载云等开始引入云化网络；在的新技术方面，使用云计算和人工智能的前沿方法，解决网络云化的核心问题，包括：使用大模型、强化学习、启发式算法实现资源预测、故障检测、编排调度、意图识别等；通过 FPGA、GPU、SmartNiC，以及 eBPF 内核旁路、并行优化技术等实现软硬件加速；利用硬件加速框架、异构 VNF 框架、可扩展应用框架等框架优化，进一步提升资源和应用管理效率；基



(a) NFV 论文发表数量趋势 (b) 2022-2024 年 NFV 顶会论文数量

图 2.10: NFV 论文趋势分析

于软件故障隔离实现 NFVI 资源在多租户环境下的安全；有状态网络功能的分解以及声明式 API 的应用。如表2.2所示，本文从网络功能云原生化和面向电信场景的云基础设施对这些研究方向进行了梳理。

2.2.2 智能技术与云网融合相结合

本小节讨论智能技术与云网融合的结合，类似于智能技术与云计算的结合（见第 1.2.2小节），分为以下两个方面展开讨论。

（1）智能技术对云网融合产生新需求

如在本节 2.1 中所述，AI 负载特别是生成式 AI 对网络提出了新的需求。智算网络的研究热点工作集中在理解 AI 工作负载的独特需求、设计新的网络架构以支持这些需求、探索新的协议以提高网络通信效率，以及实现快速故障定位与恢复。在 2024 年顶会上，面向智算的网络优化的可谓是首次全面开花。在 SIGCOMM 2024 上，阿里提出 HPN 智算网络架构，腾讯提出自研的集合通信库 TCCL 与流量的联合优化，Meta 提出了在大规模集群上基于 RDMA 的训练。在 NSDI 2024 上，字节提出了在万卡上进行训练的核心技术。ATC2024 上，三星则提出了其在异构 GPU 上自动化训练的方案。此外，还有香港科技大学等高校专注于面向系统的分布式训练架构的研究，以提升训练效率。

（2）智能技术赋予云网融合新的机遇

以大模型为代表的人工智能技术在网络云化中发挥着越来越重要的作用，主要体现在三个方面：任务及资源的管理与调度、性能优化、智能运维及运营。在任务与资源管理调度方面，通过精准感知资源和流量状态，运用强化学习和深度学习等 AI 算法，实现任务调度和资源分配的智能化，并根据任务和资源的实时动态变化，自动调整策略，在确保服务质量的同时，最大化资源利用效率。在性能优化方面，借助大模型和深度学习的预测与分析能力，AI 技术能够深入挖掘网络功能的性能瓶颈和潜在优化点。在智能运维与运营方面，通过构建网络大模型和意图网络，AI 技术为自动化和智能化的运维运营提供了强大支持：利用深度学习和强化学习技术进行跨层多维数据分析，AI 能够实现异常检测、故障定位和根因分析，从而提高网络的稳定性和可靠性。因 scaling law 的存在及存算技术升级与变革，未来面向 AI 的网络技术仍然是产学研界关注的热点。

2.3 云网融合研究的展望和发展建议

借鉴 Gartner 成熟度曲线，本节构建了云网融合技术领域的成熟度曲线（如图2.11所示）。可以看到，一体化调度和算力网络相关技术处于技术萌芽期、期望膨胀期和沉淀低谷期这三个阶段。网络云化发展相对成熟，部分产品和技术已经在生产中应用，目前主要是在新技术和新场景驱动下演进。

表 2.2: 云网融合研究领域热点

研究点	研究方向概述	会议及期刊	研究主要关注点与代表性工作
云网一体化调度算法	对不同算法设计对调度模型复杂性进行优化，缩短求解时间和计算表现。	SIGCOMM NSDI INFOCOM ICDCS	• **分布式经典优化算法：** 面向服务 QoS 需求，韩国研究团队联合 VMWare 提出基于优先级的调度算法 [96]；天津大学团队主导提出基于不准确信息的流控算法 [97]； • **基于图的算法：** 通过图的形式构建服务间的复杂依赖关系，北京理工大学团队联合 Microsoft 提出基于图的跨域调度算法 [98]； • **基于 AI 的算法：** 通过对服务间依赖关系模式的高效识别，提升调度编排性能，华为研究团队提出 DNN 算法 [93]。
云网一体化调度架构与应用	优化云网一体化算法的部署，以确保对时延敏感的服务能实现快速响应。	SIGCOMM NSDI ATC	• **并行优化：** 美国哈佛大学联合 Microsoft 提出对调度算法的并行优化从而显著加速求解计算过程 [99]； • **实时监测：** 美国 UIUC 团队联合 IBM 提出分布式非侵入的实时监测框架来优化调度性能 [100]； • **动态自适应：** 阿里联合浙江大学及清华大学针对实际云计算业务需求，对调度框架进行解耦实现资源利用率动态优化 [50]。
算力网络架构与应用	面向分布式广域异构算力资源实现一体化服务能力的架构研究，针对下一代网络技术的演进及现有网络技术局限性，以院士为代表的团队在架构方面进行了大量研究。	INFOCOM IEEE NETWORK SCIS	• **智慧标识网络体系：** 北京交通大学张宏科院士团队研究标识网络至算力网络的演进，包含原理、体系与技术，通过打造标识体系与智慧映射体系，消除传统网络三重绑定限制 [101]； • **确定性算力网络体系:** 紫金山实验室刘韵洁院士研究基于 IP 网络的确定性保障技术下的确定性算力网络架构 [102]； • **通感算一体网络:** 北京邮电大学张平院士团队通过边缘智能下的通感算及智能空口研究面向 6G 的算力网络体系架构 [103, 104]； • **面向科学计算的算力网络:** 美国亚利桑那大学关注算力网络面向科学计算场景的技术应用和架构 [105]。
算力服务效率性能	算力服务区别于传统的云服务，存在多方异构算力资源及不同服务要求，因此针对算力服务尚存挑战，研究热点聚焦在异构算力度量与评估，任务调度、算力交易等方面。	INFOCOM LCN IWQoS IEEE TSC	• **算力度量与评价：** 嵩山实验室、紫金山实验室对算力服务能力及面向服务的评估进行研究 [106, 107]； • **算力交易：** 浪潮、济南超算中心、天津大学等团队对算力交易平台框架及机制展开研究与实践，实现平台利益最大化，包括博弈论及拍卖算法 [108, 109, 110]； • **任务调度与资源分配：** 鹏城实验室对算力网络上的任务调度进行了体系化的研究 [111]，阿里、天津大学、南京大学对算力网络中 AI 任务的推理及分割框架进行了优化研究 [112, 113, 114]。
网络功能云原生化	多数基于服务化架构的网络功能都是有状态的，需在大规模场景中支持实时配置，在实现弹性的同时保障状态的一致性、操作的有效性成为挑战。	INFOCOM SIGCOMM NetSoft	• **有状态网络功能的动态扩展：** Microsoft、Uber 通过分解网络状态、消除或最小化网络状态来解决网络功能在多个核心或服务器上的扩展问题 [115, 116]； • **声明式配置：** 通过声明式 API 实现网络功能的管理配置，并结合意图简化网络管理复杂性，提高配置可读性和可维护性 [117]； • **SFC 的高效部署:** 网络服务功能链可以灵活的构建复杂的网络服务，华为通过分布式代理协同处理的方法解决 SFC 工作负载的动态调度与部署 [118]；美国研究团队基于流量的变化实现 SFC 工作负载的弹性调度 [119]。
面向电信场景的云基础设施	网络系统相较 IT 系统，在实时性、安全性、可靠性等方面有更高、更严苛的要求，需要云基础设施增强相应的能力以满足网络系统的要求，近期研究热点主要聚焦在性能加速、编排管理等方面。	SIGCOMM INFOCOM NSDI MobiSys SECON	• **网络功能编排调度优化：** Meta 提出了面向 eBPF 网络功能的编排 [120]；阿里巴巴提出了针对 vRAN 的编排调度模型 [121]；西安交通大学提供了在异构框架下整合网络功能的研究 [122]；。 • **云原生基础设施承载：** 5G 核心网、vRAN、6G、SFC 在 Openshift、Kubernetes 云原生基础设施上的部署问题 [123, 124, 125]； • **网络加速：** 加州大学研究控制面性能加速 [126]；阿里巴巴、Microsoft 等基于异构硬件、内核优化等研究数据面加速 [40, 127, 128, 129]。 • **运维与安全：** Microsoft 在百毫秒内实现故障转移和零停机的软件升级 [130]；阿里巴巴提出了一种软件故障隔离的安全 NFVI，能够实现更好的性能 [131]。

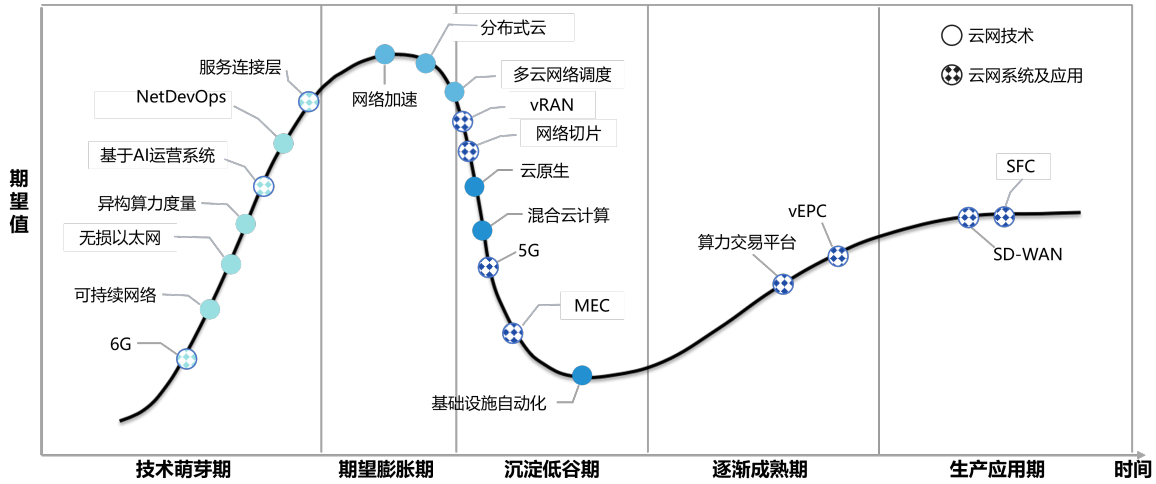


图 2.11: 云网融合研究图谱技术成熟度曲线 2024

2.3.1 云网融合的未来研究方向和关键技术展望

优化一体化调度计算复杂度，在大规模分布式场景下提升调度性能与计算效率。在资源规模化扩展的背景下，未来将更加重视计算复杂度的管理。随着数据量的激增和计算需求的复杂化，需要对算法做进一步优化以适应这一变化。这不仅涉及到对算法的基础理论研究，也涉及到更高效的系统框架，如并行计算、分布式系统设计等，以实现资源的最优利用。在资源规模化的同时，保持甚至提升计算效率，以确保在面对大规模服务调度需求时，云计算系统能够快速、准确地调度云网资源来处理需求。

实现内生性能保障、优化经济模型与数据流通机制，并提升服务的扩展性与弹性，是算力网络平台研究的重要方向，从而能够全面支撑高效、智能的计算服务。通过内生性能保障机制，算力网络实现资源的协同调度与融合部署，提供全场景、跨层跨域的确定性服务能力，确保时延、安全及可靠性，从而为计算提供可预期的服务保障。在经济模型和数据流通方面，随着算力市场的逐步成熟，研究将集中于资源定价、成本分配和市场机制优化，同时构建隐私保护和数据安全的流通机制，借助数据确权、追溯技术及智能合约提升流通效率，防止数据滥用。此外，算力网络平台的扩展性与弹性研究将着眼于新型架构与容错机制，以支持不断增长的计算需求，并提升平台面对故障时的自愈和弹性能力，从而构建更加灵活和可靠的算力服务体系。

深入优化软件、异构硬件以及软硬融合技术，全面提升网络处理能力。在软件层面，通过内核旁路、并行优化、新型虚拟化等技术，减少资源及性能损失，达到与物理硬件相近的处理能力。在硬件层面，着力于打破通用处理器的性能瓶颈，引入智能网卡、DPU、FPGA 等为高速网络专门设计的加速卡和定制化计算解决方案；利用可编程白盒设备提供的灵活性和可扩展性，满足不断变化的网络需求。同时，软硬融合通过任务卸载、计算资源的高效协同，进一步优化工作负载的性能，满足未来网络对高带宽、低延迟及灵活调度的多样化需求。

网络云原生化是网络云化发展的必要阶段，标志着网络架构向更加灵活、更加高效的云服务模式转变。随着容器、微服务和容器集群管理等技术的引入，传统的 NFV 架构得到了革命性的改变。这些技术不仅提高了资源利用率，还加快了服务部署速度，增强了系统的可伸缩性和可维护性。展望未来，网络云原生化将进一步结合轻量虚拟化、Serverless 架构、声明式 API 等云原生技术，简化配置管理，降低运营成本，增强系统的可伸缩性和可维护性，提高资源利用率和部署效率。同时，高性能 Service Mesh 将优化微服务间的通信，提升服务的可观测性和安全性。

2.3.2 云网融合的发展建议

关注算法理论研究，实现大规模资源调度的极致弹性与智能规划。深入研究算法理论，能够为资源调度提供科学的理论支撑，使调度更加精准、可解释。云计算在不同行业场景落地时，业务需求是多样

且多变的。对算法理论的研究，可以使调度算法在不同场景下呈现更强的弹性，根据业务要求的调度响应时间，灵活适配满足计算效率的调度，从而在业务量突然发生变化时，迅速做出响应，避免资源浪费或不足的情况出现。此外，借助算法理论研究，还可以对大规模异构资源进行全面、深入的分析和规划，从而结合业务的特点和需求，预测未来的资源需求趋势，提前进行资源布局和调整，实现资源的智能规划，提高资源利用率，降低运营成本，保障服务质量及响应速度。

推动算力网络逐步走向泛在智能协同，促进技术成熟和应用范围扩展。近期，随着大模型催生的集群算力进一步扩展，研究将重点围绕 Scale-out 和 Scale-up 的协同与融合发展。同时，确定性网络、远距离高性能网络及传输技术的快速进步，使智算分布式跨域逐渐成为算力网络的重要应用场景。这些发展有助于加速算力网络技术的全面成熟。在中远期，随着泛在互联、具身智能等新型应用需求的不断涌现，算力网络将向泛在一体化方向迈进，通过云边端协同智能训推一体技术与移动算力协同技术的突破，进一步提升算力的灵活性和高效性。这些技术的融合发展，不仅将推动算力网络适应复杂多变的智能化需求，还将开辟更广泛的应用前景，为未来智慧社会建设提供重要的支撑。

增强云基础设施满足网络的定制化需求，实现网络功能跨云跨域流动。NFVI 正逐步从电信私有云延伸至公有云和混合云，从数据中心扩展至客户设备乃至深边缘场景，以提供更加灵活、贴近用户的计算与网络能力。通过增强云基础设施，虚拟网络功能能够灵活部署到任意用户需要的地点（无论是中心节点还是边缘节点），从而满足差异化、低延迟的应用需求。云基础设施能力全方位拓展，构建敏捷高效的网络，云网能力实现进一步融合，从而更好地满足物联网、自动驾驶、低空经济等新兴应用对低时延和高可靠的严苛要求。

聚焦新场景，局域组网和专网正成为云服务商拓展网络能力的重要方向。在低空互联网、卫星无线网、车联网等新兴应用场景中，网络功能展现出小容量、短时性和移动性等特点。通过分布式云的跨域能力，这些场景将迎来新的发展机遇，进一步推动网络技术与服务模式的创新。此外，6G 网络以原生在云上构建为目标，将云计算和网络技术深度融合，为未来网络提供更加灵活、高效和可扩展的服务能力。6G 不仅能满足超可靠低延迟的要求，还能够支撑沉浸式通信等多样化应用，为虚拟现实、全息视频和智慧工业等领域提供坚实的基础。通过聚焦新场景，云计算和通信技术的结合将开辟更多的可能性，为未来智能化社会提供强大动力。

第 三 章

智能算法赋能的研究

云计算深刻改变了信息服务的开发、部署、运维和计费方式。依托互联网，它创建了一个强大的云环境，使得用户可以随时随地访问和管理关键资源。这种模式不仅提高了工作效率，还通过灵活的工作流程和按需付费的定价机制显著降低了企业运营成本。因此，云计算的普及和应用已经跨越行业界限，覆盖了医疗、金融和社交网络等多个领域。Gartner 预测，到 2028 年，云计算将成为企业保持竞争力和生存发展的核心要素。另一方面，国务院印发实施了《新一代人工智能发展规划》，明确将人工智能定位为国家未来重要的发展战略，并预期在 2030 年建设成为世界主要人工智能创新中心。这一决策标志着智能算法作为当代科技创新的核心，根植于数十年的技术发展和理论研究，已成为推动社会进步和产业转型的关键力量。它不再仅存于想象之中，而是实实在在地影响着经济结构和人类生活的各个方面。

云计算与智能算法的深度融合正在重塑未来技术格局。一方面，云计算是释放智能算法潜力的关键动力。Nvidia CEO 黄仁勋指出，智能算法的计算性能可能遵循超摩尔定律的加速增长趋势。因此，云计算为智能算法应用的部署和扩展提供了坚实的平台，使得企业能够在不进行专用硬件投资的情况下，实施智能算法驱动的解决方案。另一方面，智能算法带来的自动化和快速决策能力，使得管理超大规模云系统变得更加高效。传统上，云基础设施的管理需要大量的手动操作和专业知识，而智能算法的引入带来了革命性的变化，为实现自我管理和自我优化的云环境铺平了道路。

本章将概述与云计算和云网融合关联的关键智能算法，如图 3.1 所示。重点聚焦智能算法如何赋能云计算智能化进程，深入探讨当前研究的最新进展、面临的开放性问题及未来的攻关方向。

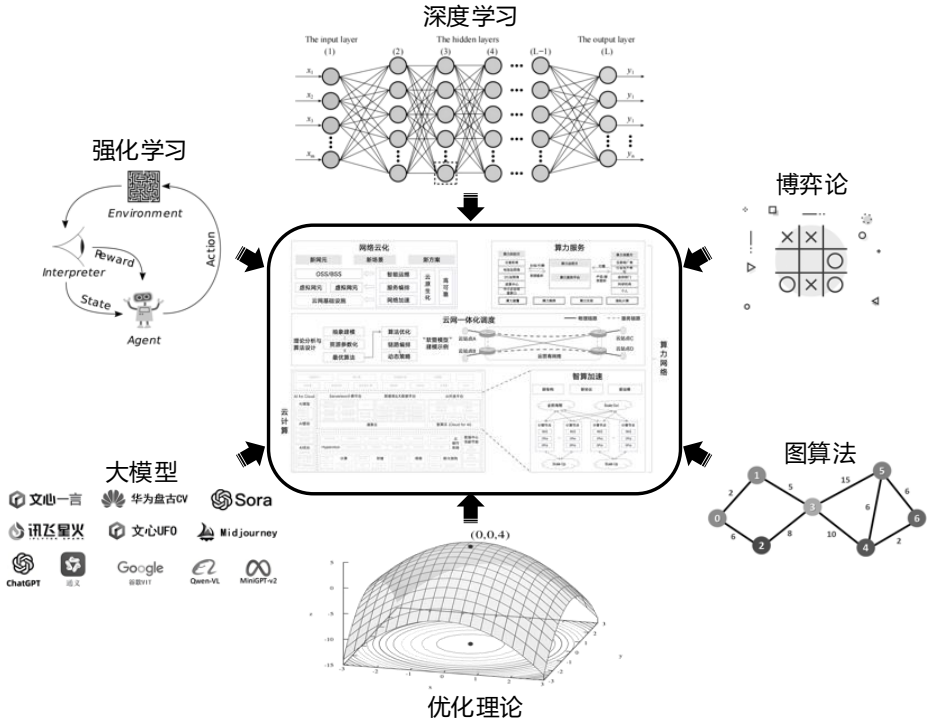


图 3.1: 研究图谱：赋能云计算和云网融合的智能算法

3.1 研究图谱及其产生：赋能云计算和云网融合的智能算法

本节将追溯智能算法数十年的发展历程，涵盖从早期的优化算法、图算法和博弈论等基础理论，到具有深远影响的里程碑技术，如深度学习和强化学习，再到当下炙手可热的大模型，全面审视这些关键智能算法的演进路径及未来趋势。具体而言，每小节将从各个技术的发展历程、核心思想和经典算法出发，进一步延伸至当前备受关注的研究热点和实际应用范式，旨在帮助读者全面了解智能算法从理论到实践的全貌。此外，表 3.1 分类整理了这些智能算法的类型、目标、代表性方法及其核心特点，为其与云计算的深度融合奠定全面而清晰的认知基础。

3.1.1 优化理论及其应用

优化方法（Optimization Methods）是在约束条件下对目标函数进行极值求解的技术与理论体系，广泛用于科学研究与工程实践中。根据问题的不同性质，优化可分为凸优化、非凸优化、线性规划、整数规划、组合优化等类别。此外，启发式优化算法（如遗传算法、模拟退火）为解决高维复杂优化问题提供了灵活有效的工具 [135, 159]。优化方法的核心目标是寻找最优解，其理论基础源于数学中的变分学和最优控制理论。在凸优化问题中，目标函数和约束条件的凸性确保了全局最优解的唯一性，这类问题通常可以通过梯度下降、牛顿法等经典算法高效求解 [132]。而非凸优化问题存在局部最优解，在一些场景中可以通过松弛、近似等方法转化为凸优化或易于处理的形式 [133]。在实际应用中，往往采用随机梯度下降方法提升求解效率。在离散优化领域，线性规划和整数规划为资源调度和路径规划提供了理论框架 [160]，而组合优化则通过图论和排列组合等方法解决复杂的离散选择问题 [134]。除此之外，遗传算法、模拟退火等元启发式算法借助生物学和物理学的启示，能够在复杂环境中找到近似最优解，在一些问题中得到了广泛的应用。近年来，优化理论的研究方向逐步转向大规模分布式优化、实时在线优化以及与机器学习结合的领域。例如，在深度学习的训练过程中，优化方法被用于寻找网络参数的最优解；在动态系统优化中，在线优化通过不断更新模型实现对环境变化的实时响应。

随着云计算和网络技术的快速发展，优化理论及方法成为提升资源利用效率和服务质量的重要工具。云计算环境中涉及的资源调度、负载均衡、网络路由优化以及能耗管理等问题，均具有高维、动态和非线性等特点，而优化理论为这些问题的解决提供了系统化的理论支撑和算法支持。在负载及资源状态实时变化的云计算平台中，实时任务优化是最大化系统效率的关键，在线优化方法可以在动态环境中不断更新策略，适应变化的需求和约束条件，应对环境的不确定性和复杂性。在涉及云边端众多计算节点的物联网场景中，分布式优化算法可以通过多节点协同来提升计算效率，满足复杂的计算需求。在不确定性的网络环境中，鲁棒优化在保证系统稳定性方面发挥重要作用，安全优化则提升了系统对故障的应对能力。在具有大量节点及复杂拓扑结构的网络场景中，启发式优化方法可以高效的搜索高维、非线性的求解空间，找到满足条件的解。优化理论和相关算法的应用不仅是提升云计算效率的关键，还为云平台的智能化和自动化提供了技术基础。在未来，优化方法将继续为推动云计算的性能提升与成本降低贡献重要力量，同时为下一代智能化云服务的实现提供理论与实践支持。

3.1.2 图算法及其应用

图计算领域的历史悠久，其起源可追溯至 18 世纪数学家欧拉提出的“七桥问题”。图是一种关键的数据结构，由节点 V（代表个体）和边 E（代表个体间的联系）组成，通常以 G=(V, E) 的形式表示 [161]。这种抽象的图数据结构广泛应用于通信网络、社交网络、推荐系统、金融风控等多个领域 [162]。不同领域的图数据结构如图 3.2 在数据日益复杂的现代社会，图数据模型凭借其在表征复杂关联性和提供可解释性计算上的优势，成为研究和应用的焦点。近年来，图数据的规模呈现指数级增长，节点数量已达到数十亿级，边的数量更是高达数万亿。这种规模扩张不仅反映了数据复杂性的提升，也凸显了对高效图计算算法需求的迫切性。作为图数据处理的核心，图算法提供了解析和优化图结构数据的基础工具。其中，遍历算法作为底层基础，支撑了众多高级算法的开发与应用。典型的图算法包括最短路径算法、图划分

表 3.1: 智能算法技术图谱

算法类型	算法概述	代表性技术	技术特点
优化理论	通常描述为在一定约束条件下最大化目标函数的问题，往往涉及问题的松弛及转化、求解方法设计、收敛性分析等，核心在于根据问题结构设计高效算法，平衡求解速度与质量。	凸 (非凸) 优化 组合优化 启发式优化 随机优化	• **凸 (非凸) 优化：** 凸优化中最优解唯一，算法高效稳定，非凸优化复杂度高，常进行凸松弛转化或借助随机搜索寻找近似解 [132, 133]； • **组合优化：** 优化变量取值离散、复杂度高，常利用图论、整数规划等工具求近似解，应用非常广泛 [134]； • **启发式优化:** 灵活且通用，模拟自然现象来寻找复杂问题的近似解，无需严格数学模型，适合高维、非线性、多峰值问题 [135]； • **随机优化:** 引入随机策略探索解空间，可以应对非线性、非凸或高维复杂问题，能跳出局部最优，具有全局搜索能力 [136]。
图算法	图算法基于图论，用于高效处理分析图结构数据，通过优化节点和边的关系，解决最短路径、网络流等问题。超图作为图的扩展形式，增强了算法在复杂场景中的表达能力，拓宽了图算法的建模能力。	图划分 图匹配 路径发现 密集子图识别	• **图划分算法：** 旨在将图分割为若干子图优化节点间的分布关系。其历史可以追溯至上世纪 70 年代的并行计算研究 [137]； • **图匹配算法：** 旨在挖掘子图间相似性的算法。目标是通过比较节点、边或子图的结构、属性或拓扑关系，找到图之间的最佳匹配 [138]； • **路径发现算法：** 旨在基于图遍历算法探索节点间路径，该算法从某个节点开始遍历，进而在条件限定的情况下用来识别最优路径 [139]； • **密集子图识别算法：** 旨在挖掘图中一组连接紧密的节点。早期算法是严格挖掘完全子图，而后续的 k-Core、谱聚类和 Modularity 优化等算法放宽了密集性定义 [140]。
博弈论	研究多个参与者相互博弈过程中的策略选择及结果，涉及策略空间建模、均衡解和稳定性分析，重点在于理解参与者的相互影响。	纳什均衡 合作博弈 演化博弈	• **纳什均衡：** 一种所有参与者均无法通过单方面改变策略获益的策略组合，策略具有稳定性，其存在和唯一性由严格条件决定 [141]； • **合作博弈:** 分析参与者通过协作达成共赢的博弈，涉及联盟形成和收益分配问题，如资源共享和联合决策 [142]； • **演化博弈：** 研究策略随时间演化的博弈，借助复制动态等工具分析策略分布变化，动态性强，着眼于长期稳定状态 [143]。
深度学习	基于多层神经网络，通过对数据进行逐层非线性变换，从低层次特征提取到高层次抽象表示，实现特征自动提取和复杂模式的学习。	卷积神经网络 循环神经网络 图神经网络 生成对抗网络 Transformer	• **卷积网络：** 通过局部感知与权值共享，提取多层次特征 [144]； • **循环网络：** 连接历史状态，学习序列数据的时序特性 [145]； • **图网络:** 为图结构数据设计，建模复杂拓扑，支持图级任务优化 [146]； • **生成对抗网络：** 生成器与判别器对抗优化，模拟复杂场景建模 [147]； • **Transformer：** 基于多头注意力机制，高效建模全局依赖关系，处理大规模复杂任务 [148]。
强化学习	通过智能体与环境交互，基于奖励信号学习最优策略的机器学习方法，核心在于通过序列决策优化策略以最大化长期累计奖励；	Q 学习 SARSA 算法 深度 Q 网络 近端策略优化 异步 A3C	• **Q 学习：** 通过时间差分更新动作价值，无需环境模型 [149]； • **SARSA 算法：** 基于策略依赖更新，考虑实际执行反馈 [150]； • **深度 Q 网络：** 结合深度网络逼近 Q 函数，解决高维空间问题 [151]； • **近端策略优化：** 通过限制策略更新幅度稳定训练，平衡样本效率与策略优化，广泛应用于工业级强化学习场景 [152]； • **异步 A3C：** 异步并行更新策略与值函数，提高训练效率 [153]。
大模型	通常是基于 Transformer 的架构、具有大规模参数和计算能力的生成模型，通过预训练-微调的深度学习方法，以实现自然语言的理解与生成。	上下文学习 人类反馈的 RL 检索增强生成 专家混合模型 低秩自适应	• **上下文学习：** 将下游任务的输入输出作为 prompt 引导模型给出预测结果，实现在推理时通过提示中的少量样本学习新任务 [154]； • **人类反馈的 RL：** 利用人类评估指导模型优化其行为和决策 [155]； • **检索增强生成：** 从外部检索信息，提升模型准确性与适应性 [156]； • **专家混合模型：** 整个模型由多个专家（子模型）组成，在推理时动态选择部分专家参与计算 [157]； • **低秩自适应：** 引入低秩矩阵，实现大模型的高效微调 [158]。

算法、密集子图识别、图匹配算法、k-覆盖算法、图聚类算法、图传播算法、链路预测算法和 PageRank 等。这些算法紧密结合实际需求，推动了多个领域的发展。随着数据复杂性的持续提升和关联模式的不断演化，传统图计算模型和算法在处理多维度、多关系复杂场景时已显现出诸多局限性。作为图的自然推广形式，超图 [163] 能够通过超边表示多个节点之间的高阶交互关系，从而在数据建模上显现出显著优势。超图建模不仅能够全面捕捉复杂系统中多节点协同作用的特性，还可以更加精准地表征真实世界中

通信网络　　交通网络　　社交网络

图 3.2: 不同领域的图数据的典型示例

非二元关系的多维度关联性。超图算法的高阶性和灵活性使其具有巨大研究和应用潜力。

随着云计算技术的快速发展，图算法在解决数据库管理、分布式计算以及资源调度与编排等关键技术问题中发挥了至关重要的作用。云计算环境中的数据库管理、分布式计算和资源调度与编排等问题通常具有高度的相互依赖性，图算法通过构建图模型，为这些问题的解决提供了强有力的理论和算法支持。例如，最短路径算法广泛应用于路由协议和内容分发网络中，通过优化传输路径显著降低时延；图划分算法在分布式计算的任务分配、云资源的负载均衡以及通信网络的拓扑优化中具有重要作用，通过划分网络子图有效提升任务执行效率和资源利用率；图聚类算法则在数据中心负载优化、通信网络社区检测及微服务拆分等场景中，显著优化了资源分配效率和网络结构。此外，图匹配算法在任务分配、虚拟机与服务器匹配等领域，通过最大化匹配效率有效提升资源利用率并降低调度成本；链路预测算法则在动态网络拓扑设计和路由优化中显著提高了系统的扩展性。PageRank 和图传播算法通过优化节点重要性排序和信息传播路径，为故障定位与网络攻击检测提供了关键支持。总体而言，图算法在云计算领域的广泛应用不仅推动了云资源的高效配置，还为智能化、自动化的网络管理奠定了坚实基础，为云计算的未来发展指明了方向。值得注意的是，在最新的学术研究与行业应用中，图算法正与机器学习、深度学习等前沿技术深度融合，为进一步优化云网系统的性能和效率提供了强有力的工具支持。这一趋势表明，图算法将在云计算的未来发展中扮演更加重要的角色。

3.1.3 博弈论及其应用

博弈论（Game Theory）研究理性决策者之间的互动博弈行为，目标是分析参与者在特定规则和激励机制下如何选择策略以实现利益最大化，包含纳什均衡、演化博弈论、机制设计等核心概念 [164]，为计算机科学中的相关问题提供了强有力的分析工具。博弈论相关研究主要包括规范性方法和描述性方法。规范性方法以构建精确数学模型为目标，强调对博弈参与者行为的逻辑推理和最优策略分析；而描述性方法更关注现实世界中人类或智能体的行为特性。近年来，博弈论与计算技术结合日益紧密，催生了算法博弈论（Algorithmic Game Theory）的研究分支 [165]，重点探讨如何设计高效算法来求解复杂博弈问题。算法博弈论研究方向包括多主体博弈中的分布式计算、在线博弈、拍卖机制设计等，在处理大规模多主体问题时，博弈论往往会结合机器学习与优化技术，通过启发式算法或强化学习解决实际问题。

随着云计算技术的广泛应用，博弈论在解决云计算资源调度、系统性能提升以及算力网络平台方面展现了巨大的潜力。云计算环境中，资源分配和任务调度问题具有异质性和多方竞争的特征，博弈论的策略分析与均衡求解为这些复杂问题提供了重要解决手段。在多租户云环境中，博弈论模型可用于分析用户间的竞争与合作关系，设计公平且高效的资源分配机制。在面向能效的云平台优化领域，基于重复博弈的机制可以激励数据中心在不同时间段动态调整计算负载，实现节能目标，降低整个云平台的运行成本。在云网融合的算力网络应用中，博弈论为算力分配和资源共享提供了理论基础，对算力网络中计算设备的动态竞价和实时任务分配能够保证参与方的计算需求得到满足，资源方的效益得到保证。博弈论有助于建立灵活的市场化机制，支持云服务的动态定价与资源供需调节，从而提升云计算系统的灵活性与智能化水平。总体而言，博弈论的策略分析与机制设计能力为下一代云计算和网络架构的发展提供了重要支持，其研究和应用成果推动云计算向更高效、更绿色、更智能的方向不断迈进。

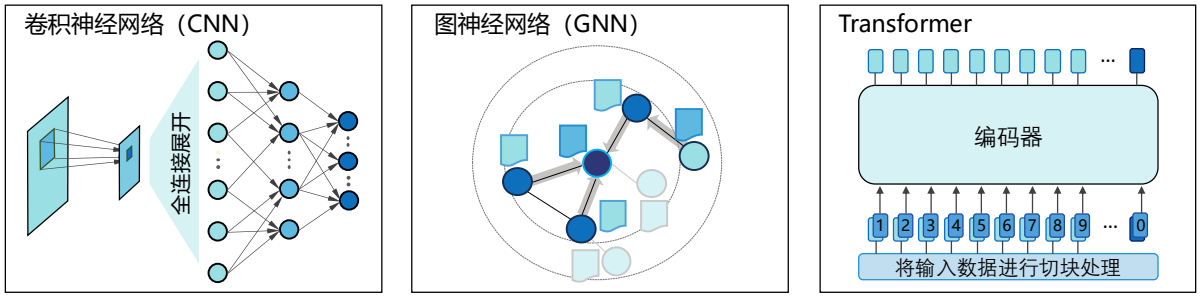


图 3.3: 深度学习经典算法的网络结构示意图

3.1.4 深度学习及其应用

深度学习（Deep Learning）作为智能算法的一个重要分支，近年来取得了广泛的关注和显著的进展。**深度学习的核心思想是通过模拟人脑神经网络的结构和工作机制，以多层非线性变换逐步提取数据中的高层抽象特征，从而在复杂任务中实现自适应的学习 [166]**。深度学习领域包含多种经典算法，例如，卷积神经网络（Convolutional Neural Networks, CNNs）、循环神经网络（Recurrent Neural Networks, RNNs）、生成对抗网络（Generative Adversarial Networks, GANs）、图神经网络（Graph Neural Networks, GNNs）及 Transformer [167]。网络结构如图 3.3 所示。CNNs 通过局部连接与权值共享的结构显著降低参数复杂度，强化模型对空间信息的捕获能力，其多层次特征提取机制在图像分类、目标检测等任务中表现卓越。RNNs 及其改进版本（如 LSTM、GRU）通过门控机制有效缓解梯度消失问题，擅长捕获序列数据中的时间依赖关系，成为时间序列预测与自然语言处理的核心技术。GANs 通过生成器和判别器的动态博弈训练机制，在图像生成、云计算领域展现出广泛应用，例如数据增强、流量生成与预测、网络安全威胁检测，以及资源分配优化等。GNNs 则是深度学习处理图结构数据的重要工具，通过聚合节点邻居信息，学习节点和图的表征，被广泛应用于社交网络分析、推荐系统和分子结构预测等领域。Transformer 则通过自注意力机制实现全局特征建模，并通过多头注意力机制捕获丰富的特征关系，显著提升了自然语言处理和视觉任务中的建模能力。这些结构各具特点，共同奠定了深度学习的技术基石。

然而，深度学习的高复杂性和数据驱动的本质使其“黑箱”特性成为一大挑战，特别是在高风险领域中，模型决策的透明性和可信性愈发重要。“可解释机器/深度学习”（XAI）旨在揭示模型决策背后的逻辑与依据，从而提升透明性、用户信任以及模型的实际应用潜力 [168]。研究方法主要分为两类：一是模型本身的内在可解释性，设计透明的模型结构，如决策规则嵌入或注意力机制；二是模型输出的后验解释性，利用如 Grad-CAM [169] 和 SHAP [170] 等特征归因方法，解释模型对输入数据的响应。这些方法提升了模型的透明性，优化了模型的性能和鲁棒性。

近年来，云计算为深度学习的发展提供了强大的算力支撑，同时深度学习也在推动云计算的快速变革。一方面，大规模分布式深度学习框架的持续优化显著提升了云平台的模型训练效率和扩展能力，支持更复杂、更大规模的模型开发。另一方面，深度学习在云计算领域的应用日益深入，包括利用神经网络优化资源调度与任务分配，以及提升系统性能等。此外，云—边—端协同架构的兴起使得深度学习模型能够在云端高效训练，并在边缘设备快速部署，从而满足低延迟、高实时性的应用需求。这展现出深度学习与云计算深度融合的趋势，为技术创新和行业应用开辟了广阔的前景 [171]。

3.1.5 强化学习及其应用

强化学习（Reinforcement Learning, RL）是一种通过智能体与环境的交互，在试错中通过奖励信号学习最优策略的机器学习方法，其核心在于通过序列决策最大化长期累计奖励。强化学习算法主要分为基于模型的方法和无模型的方法：基于模型的方法依赖环境动力学的准确建模，通过规划技术（如动

态规划）寻找最优策略，适用于环境完全可知的场景，但在高维或非线性问题中计算成本高昂；无模型方法无需环境模型，直接通过试验经验更新策略，主要包括蒙特卡罗方法和时间差分学习，前者基于整条情节的累积奖励进行更新，适合离散任务，后者则通过逐步估计进行即时更新，结合了动态规划的递归思想，具有更高的实时性和灵活性 [172]。其中，经典算法如 Q 学习和 SARSA 在处理离散状态和动作空间问题上表现出色，但在高维或连续空间中表现有限。为此，深度强化学习（DRL）通过深度神经网络实现值函数和策略函数的非线性逼近，突破了传统强化学习在高维状态和动作空间中的瓶颈，并在游戏（如 AlphaGo）、机器人控制和智能驾驶等复杂领域展现了卓越性能。整体而言，强化学习通过理论与实践的不断融合，成为解决复杂、不确定序列决策问题的关键技术。

近年来，强化学习的研究聚焦于解决其在复杂环境中适应性与效率方面的不足，以下是一些备受关注的研究热点。（1）多智能体强化学习（Multi-Agent RL）[173]：研究多个智能体在共享环境中的协作和竞争，解决单智能体无法应对的复杂任务，通过共享奖励和通信协议等增强智能体间的协调和效率，如 MADDPG 方法（Multi-Agent Deep Deterministic Policy Gradient）等方法能够有效处理多个智能体在共享环境中的复杂交互。（2）离线强化学习（Offline RL）[174]：利用固定数据集进行训练，适用于数据收集成本高或存在安全风险的场景，通过基于模型和无模型的方法减少数据需求，如 MOReL（Model-Based Offline Reinforcement Learning）通过建立环境模型来模拟并生成虚拟经验，在有限数据情况下表现出色。（3）人类反馈强化学习（RLHF）[155]：通过人类反馈而非预设奖励函数训练智能体，使其更好地对齐人类目标，采用自然语言、比较等反馈形式，如直接偏好优化（Direct Preference Optimization, DPO）方法直接优化模型以符合人类偏好。（4）层级强化学习（Hierarchical RL）[175]：通过将稀疏奖励问题分解为层级子任务以提升探索效率，低层策略负责动作，高层策略制定目标，如选取子目标的半马尔可夫决策过程（Semi-Markov Decision Process，SMDP）提升智能体在复杂任务中的表现。

强化学习（RL）在云计算领域的应用正逐步走向体系化和智能化，成为解决复杂优化问题的重要手段。一方面，云计算环境中的资源调度、负载均衡、能耗优化等问题具有高维度、非线性和动态变化等特性，传统方法难以适应。而强化学习通过构建动态交互模型，使智能体能够在任务调度和资源分配过程中自主探索优化策略。例如，深度强化学习结合云平台的大规模计算能力，能够在多任务环境中高效解决资源分配优化问题，提高系统吞吐量与资源利用率。另一方面，强化学习推动了云—边—端协同计算的深度融合，为实时性和低延迟应用提供了全新解决方案。在边缘计算场景中，RL 模型可用于动态优化数据卸载策略和边缘节点协作机制，平衡计算负载和传输延迟。而在云端，强化学习支持智能化的基础设施管理，如动态扩容与节能调度，提升服务质量与成本效益。整体来看，强化学习在云计算中的广泛应用不仅显著提高了系统性能，还为下一代智能化云服务架构的创新提供了重要支持。

3.1.6 大模型技术及其应用

大语言模型（LLM）是基于神经网络的大规模预训练模型，语言模型发展经历了四个阶段：统计语言模型、神经语言模型、预训练语言模型及大语言模型。自 2017 年 Google 提出 Transformer 架构以来，大模型经历了快速发展，从 BERT、GPT 等基础模型到更复杂的指令调优模型，参数规模达到数十亿至数千亿级别。在此过程中，尺度定律（Scaling Law）揭示了模型规模、数据量和性能之间的关系，推动了大模型在规模化与效率方面的优化。图 3.4展示了从 2017 年发展至今具有代表性的大模型。国际上，Google、OpenAI、Meta 等公司不断推出具有更强理解和生成能力与更多模态的模型，如 Gemini、GPT-4o、LLaMA 等。与此同时，中国的百度、华为、阿里巴巴、清华大学等团队也在大模型领域取得了重要突破，推出了如 ERNIE、PanGu、Qwen、ChatGLM 等具有多模态融合和产业化应用能力的模型。到 2024 年，国内大模型在多模态理解、垂域行业赋能方面实现了全面提升，特别是阿里云的通义千问、字节跳动的豆包，以及初创团队推出的月之暗面 Kimi 和阶跃星辰 Step-2 等，电信也推出了星辰大模型 TeleChat [176], 推动了国内大模型赋能千行百业和核心竞争力的全面跃升。

相比于预训练语言模型，LLM 不仅模型规模更大，而且语言理解与生成能力更强，尤其是出现了小

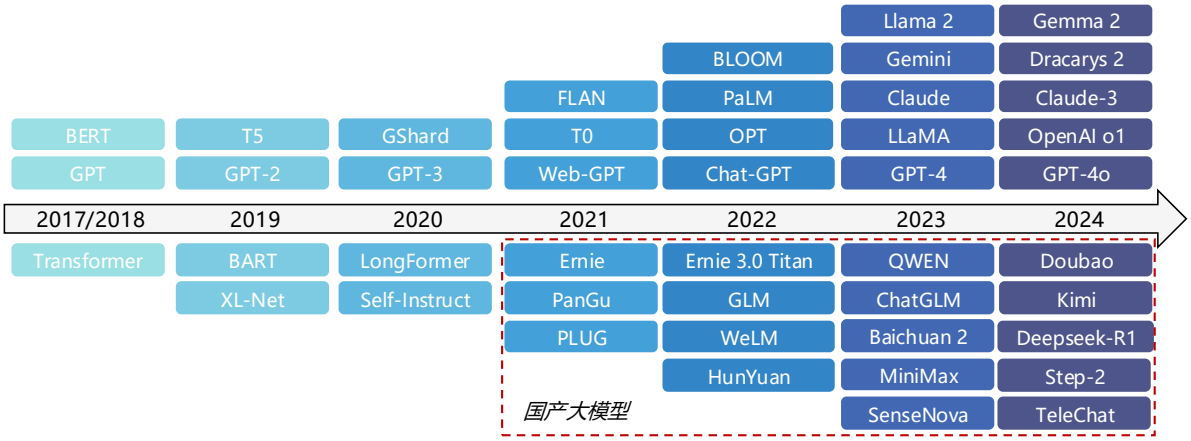


图 3.4: 代表性大模型发展历程。

模型不具备的“涌现能力”，具体而言，LLM 能够通过以下方式实现创新：(1) 上下文学习 [154]，允许模型在推理时通过提示中的少量样本快速适应新任务；(2) 指令跟随 [177]，支持模型在缺乏明确示例的情况下，根据抽象任务指令完成多样化需求；(3) 多步推理 [178]，通过“思维链”提示将复杂问题分解为多个中间推理步骤，显著提高解决复杂任务的能力。此外，LLM 可以通过集成检索外部知识和工具，在信息交互与执行任务的效率上进一步优化，同时通过引入反馈机制（如 RLHF 算法）持续改进模型的生成质量和用户适配能力。这些特性使得 LLM 能够在跨领域任务中表现出前所未有的灵活性与智能性，从而推动了智能应用的边界。

LLM 的构建涵盖从数据预处理到模型优化的多个关键阶段，每个阶段通过特定技术手段共同提升模型的质量与效率，最终实现对多样化用户需求的高效响应与生成能力 [179]，技术包括：(1) 数据清理 [180]：包括数据过滤（去除噪声、处理异常、解决数据不平衡、文本预处理）和去重，目标是提高数据质量，确保模型在干净且高质量的数据上进行训练。(2) 分词：将原始文本转换为模型能够理解的符号，常用的方法包括字节对编码、词片编码和句子片编码。(3) 位置编码：在模型中加入序列信息，使得模型能够理解输入词汇的相对位置，包括绝对位置编码、相对位置编码、旋转位置编码 [181] 和相对位置偏差。(4) 模型架构：确定模型的基础结构，影响模型的表达能力；包括只使用编码器 [182]、只使用解码器、以及编码器-解码器等不同的架构类型。(5) 模型预训练：使用大规模无标签数据进行预训练，目标是学习通用的语言表示，常见的方法包括掩码语言模型、因果语言建模、下一句预测以及混合专家 [157] 等。(6) 微调和指令调优：在特定任务或领域数据上对预训练模型进行微调，包括监督微调、通用微调、多轮指令微调以及指令跟随。(7) 对齐：使得模型生成符合用户需求的输出，包括监督学习、人类反馈强化学习、直接偏好优化以及 Kahneman-Tversky 优化 [183] 等。(8) 解码策略：确定生成文本的方式，常见的方法有贪婪搜索、束搜索、Top-k 采样和 Top-p 采样。(9) 高效训练/推理、适应/压缩：降低模型的计算和存储成本，代表性方法包括零冗余优化器、加权键值接收、低秩适配 [158]、知识蒸馏和量化等。

LLM 的可解释性研究在自然语言处理领域至关重要，直接关系到模型的透明性、可信度与伦理性应用。目前的研究方法主要分为两大类：(1) 局部分析：聚焦于模型对特定输入的预测机制，常见方法包括特征归因解释 [184] 和 Transformer 块分解 [185]。前者通过为输入词元分配相关性评分来量化其对模型预测的贡献，后者则深入研究 Transformer 模型中注意力机制与前馈网络的交互过程。(2) 全局分析：侧重于揭示模型中编码的语言知识与行为能力，代表方法包括探测模型的隐藏表示 [186] 与机制解释性 [187]。探测方法通过在模型激活值上训练分类器，揭示模型中学到的语义与句法信息，而机制解释性则通过计算子图发现（Circuit discovery）、因果追踪与词汇透视等方法，解构深度网络的推理机制与内部结构。未来，研究人员将继续探索更具扩展性与通用性的解释性技术，推动大型语言模型在多样化应用中的透明性与可控性发展。

云计算环境常涉及大规模、分布式的系统架构，任务和数据具有高度的动态性、复杂性和异构性。LLM 能够从结构化（如配置文件、监控数据）与非结构化数据（如日志、错误报告、用户反馈）中提取有效信

表 3.2: 研究热度分析：各类智能算法在各个云计算和云网融合研究热点上的应用热度

智能算法 \ 研究热点		数据管理	负载预测/均衡	参数调优	调度/编排	故障诊断
优化理论	凸 (非凸) 优化		★★		★★★	
	组合优化			★	★★	
	元启发优化	★	★★★	★★	★★★★	
图算法	图划分	★★	★		★★★★★	★
	图匹配	★★★			★★★	
	路径发现	★★★			★★★	★★★★
	密集子图识别	★★★★★			★★★	★★★
博弈论		★			★★★	
深度学习	卷积神经网络	★	★★★		★★	★★★
	循环神经网络		★★★★★	★★★	★	★★★
	图神经网络		★★★★		★★★★★	★★★★★
	生成对抗网络	★★	★		★★	★★★
强化学习	在线学习		★	★★★	★★★★	
	深度强化学习	★★★★	★★★	★★★★	★★★★★	★★
大模型		★		★★★★		★★★★★

息，并结合跨平台的数据源进行深度分析。通过其强大的语言理解与生成能力，LLM 能够有效处理云计算系统中的多维度信息，支持跨数据源的整合与分析，自动化地进行决策和优化。在如故障诊断、资源管理、配置优化、安全防护等场景中，LLM 都能够提供解决方案，提升云计算系统的智能化、自动化和高效性，使其成为推动现代云计算系统发展的新兴重要技术。

3.2 研究洞察：智能算法驱动的云计算和云网融合研究热点和难题

本节通过调研 NeurIPS、AAAI、NSDI、ASPLOS 等 20 个智能算法和云计算相关会议近 5 年发表的以智能算法赋能云计算智能化为主题的 200 余篇论文，系统梳理了该领域中研究者们重点关注的研究方向。具体而言，遵循从基础设施即服务（IaaS）到软件即服务（SaaS）这一自底而上的业务逻辑，深入分析云计算中的智能化需求，最终对现有文献进行全面梳理，重点涵盖数据管理、工作负载预测与均衡、参数调优、调度与编排、故障诊断等核心问题（如下所示）。针对这些问题，每小节将首先介绍其定义、目标与挑战，并概述最具代表性的智能算法解决方案。此外，表 3.2 统计整理了各类智能算法在云计算和云网融合研究热点上的应用热度，直观展现不同算法在当下前沿技术融合进程中的活跃程度差异，旨在帮助读者理解技术趋势，并为后续研究提供方向性指导。

研究热点和难题

1. **数据管理**：如何通过高效手段实现海量数据的存储、检索、缓存及优化利用？
2. **工作负载预测与均衡**：如何通过精准的负载预测与资源分配，优化资源利用率，高效应对突发负载，最大化云计算平台性能？
3. **参数调优**：如何在数据库系统配置及任务参数优化方面提升资源利用率和系统效率？
4. **调度与编排**：在多任务、多资源环境下，如何合理分配计算资源并高效调度任务执行？
5. **故障诊断**：如何快速定位故障根因及其影响范围，确保系统的稳定性和高可用性？
6. **其他研究热点**：在云计算与大规模分布式系统快速发展的背景下，除了上述关键问题，还需关注诸如程序设计、软件安全、网络设计与配置等领域的重要研究问题。

3.2.1 数据管理中的智能算法研究

数据管理旨在通过技术手段优化分布式系统中的海量数据存储与利用，是云计算领域的重要研究方向。**其主要目标是利用高效的技术手段，实现分布式系统中海量数据的存储、检索、缓存和优化利用**。在技术层面，数据管理依托分布式存储系统，如分布式文件系统、对象存储和数据库系统，提供高可靠性和高弹性的存储解决方案。此外，数据库检索技术通过索引构建、查询优化和分布式查询执行，能够加速复杂数据的访问和分析，提升系统性能。缓存技术则通过存储热数据及中间计算结果，显著提高系统响应性能，降低访问延迟，进一步优化数据流转过程。除这些核心技术外，数据一致性协议、事务处理、分片与重分布、数据压缩与去重等技术也是数据管理的重要组成部分，旨在进一步提高系统的性能、可靠性以及成本效率。随着云计算向边缘计算、混合云和多云架构方向演进，数据管理的关注点逐渐转向动态数据迁移、跨区域数据共享、隐私与安全保护等前沿领域，这为分布式数据处理的高效性与智能化提供了新的思路与解决方案。

在数据管理建模中，图算法作为一种关键工具，能够显著提升数据存储效率与数据库查询性能。在数据存储方面，为了最小化数据存储的资源能耗，相关研究 [188] 提出了一种基于图覆盖模型的能效存储策略，该策略通过选择最小边覆盖，找出满足数据可用性要求的最小数据节点集合，从而关闭不必要的服务器以节省能耗。此外，为了降低数据访问延迟，一种基于图划分的数据存储算法 [189] 也被提出，该算法综合优化图划分和数据复制策略，将数据高效分布存储于不同的数据中心，有效缩短数据访问路径。在图数据库查询方面，使用场景主要分为实时查询和离线数据分析。在实时查询中，常用的算法包括图遍历搜索算法和路径发现算法（如最短路径算法、最小生成树算法等 [190]，并支持高级查询需求，例如频繁子图挖掘 [161]、子图匹配查询 [191] 和社区搜索 [192] 等。为满足复杂网络分析需求，大多数图数据库已经集成了这些常见算法，以便更高效地从复杂图数据中挖掘价值信息。而在离线数据分析中，图算法则根据不同的目标提供多种方法，例如路径查找算法 [193] 用于最短路径搜索，中心性分析用于识别关键节点，链路预测 [194] 用于推测节点间潜在关系，以及社区发现算法 [195] 用于揭示网络的分组结构。这些算法通过从复杂网络中提取有用信息，为研究人员更好地理解数据结构特性提供了有力支持。

相比于传统索引方法，学习型索引在数据管理领域展现了新的潜力，尤其在高效数据查找与操作性能优化方面具有显著优势。相关研究团队 [81] 针对学习型索引的性能表现，构建了一个测试平台，对其在、键查找、插入、并发操作和批量加载等关键组件中的表现进行了系统性比较。在模型设计上，该团队优先选择非线性模型以提升预测精度；在插入策略上，结合 Delta 缓冲区与结构化调整，有效降低模型重训练的开销；在并发操作中，通过改进锁机制大幅提升操作效率；而在批量加载环节，则采用基于成本模型的节点划分方法优化了加载性能。实验结果表明，学习型索引在简单数据分布和读密集场景中表现尤为突出，不仅实现了更高的查找效率，还为未来学习型索引的设计与优化提供了宝贵参考。这一研究进一步拓展了数据管理领域的技术边界，与图算法在数据存储和数据库查询优化中的应用相辅相成，共同推动了数据管理的智能化发展。

3.2.2 工作负载预测与均衡中的智能算法研究

工作负载预测与均衡通过精确负载预测和有效资源分配来优化资源利用率，**旨在高效应对突发性负载，最大限度提高云计算平台的经济性、性能和服务质量**。然而，实现这一目标面临多个关键挑战。包括：（1）工作负载具有很强的动态性和不可预测性，尤其用户行为的多样化和突发性的请求峰值导致资源需求经常剧烈波动。（2）云环境高度复杂，由多层次的资源（如计算、存储和网络）组成，各资源间的相互依赖增加了预测与均衡难度。（3）为了保证云平台服务的连续性与高效性，资源调度需要实时响应，而实现实时的高效决策极富挑战性。传统的工作负载预测和均衡方法通常基于静态阈值、预定义规则或者简单的历史平均值来进行预测和决策。这些方法在面对高度动态、复杂和不确定的工作负载时效率低下，难以处理突发事件。而 **AI 解决方案则通过数据驱动的方式，动态地学习和适应复杂的工作负载模式**，降低运营成本，减少对人工干预的依赖，最终推动云计算平台向更加自主、智能和高效的方向发

展。本节将简要介绍针对该问题的经典 AI 解决方案，展示该方向的主流技术思路。

图结构能够有效刻画节点间的关联关系，捕捉数据中的复杂模式和动态变化，尤其适用于描述工作负载的时间依赖性和空间关系。与传统方法相比，图建模在处理高维度、非线性以及时序变化问题时具有更强的能力，因此在工作负载预测中展现出明显优势。例如，相关研究团队 [196] 发现，尽管工作负载的资源使用模式在短期内较为稳定，但在长期范围内会发生显著变化。基于此，他们首次提出了基于图神经网络的进化学习算法 EvoGWP 来预测长期动态变化。该方法通过自动提取形状元（shapelets）显式识别工作负载的资源使用模式，并同时考虑时间和空间因素进行预测。阿里巴巴、腾讯和 Google 数据集上的实验结果表明，EvoGWP 相较现有方法，其预测准确度最高提升了 58.6%，同时模型收敛速度更快。

优化算法，尤其是强化学习和启发式算法，因其在应对动态环境中的不确定性和复杂性方面的优势，已成为解决工作负载均衡问题的有效工具。具体地，研究团队针对边缘计算中节点计算能力和成本的高动态性与不确定性，提出了基于预测的动态任务分配算法 [197]，利用指数平滑法（EMA）和 ARIMA 模型预测节点计算能力和成本，并结合历史数据学习最优分配策略（PA-OPT 算法）。此外，团队还设计了基于强化学习的在线任务分配算法，实现了实时任务优化和工作负载均衡。实验结果表明，PA-OPT 算法在计算能力可预测的环境中接近离线最优解，而强化学习在低预测性环境中表现突出，有效提高了工作负载均衡，并降低了任务完成时间和系统成本。此外，相关研究团队为优化负载均衡并降低运营成本，提出了结合在线与离线的启发式算法 [198]，通过任务依赖图模型进行虚拟机放置。而针对现有方法未能保证任务执行顺序、明确截止期限以及高执行成本的问题，研究者将任务最早完成时间预测与基于蚁群优化的元启发式方法结合 [199]，通过最小成本和截止期限对任务排序，计算最优成本和实际完成时间，并将任务分配至成本最低、完成时间最短的虚拟机。接着，根据完成时间的阈值确定未充分利用的虚拟机，利用人工蚂蚁在虚拟机间转移负载，实现负载均衡。

3.2.3 参数调优中的智能算法研究

参数调优是优化系统性能、提升资源利用率的关键环节，在数据库系统配置、任务参数配置优化等方面有着广泛的应用。然而，由于云计算相关应用涉及计算、存储、网络等多方面要素，参数设置非常复杂，依靠人力及专业经验进行参数调优，往往只能找到次优配置。另一方面，云计算场景中任务负载及资源状态不断变化，具有高度动态性，任务参数或数据库服务配置都需要及时调整以适应变化，人力调优成本高且易出错，无法满足这种需求。**为了应对参数空间复杂以及系统动态变化的挑战，智能化参数调优在云计算中的应用是近年来的研究热点**。以强化学习、在线学习、大语言模型为代表的技术可以较好的适应动态变化的环境，通过与环境交互并进行增量学习，在系统运行过程中逐步优化参数设置，并有效的降低参数搜索空间的复杂性，为解决参数调优问题提供了新的思路。

利用在线学习技术评估任务参数对最终性能的影响，并动态的调整参数设置，能够将参数调优整合到系统本身的运行过程中，省去离线性能测试的开支，是参数调优相关研究的代表性方法。针对计算集群中神经网络模型训练任务的参数调优场景，相关研究团队从优化云计算平台能源效率的角度出发，设计了基于在线学习技术的调优框架 Zeus [200]。这一框架以在线的方式对神经网络训练任务进行性能分析，将探索–利用相结合，能够深入分析能源效率和模型训练性能之间的权衡，为训练任务寻找最优的作业级和 GPU 级配置。在实际云计算平台上的实验表明，这一工作避免了昂贵的离线测试，能够适应数据的动态变化，将不同类型任务的能源效率提高了 15.3% 到 75.8%，极大的降低了平台运营成本。

利用大语言模型进行数据库系统性能调优是目前的研究热点。相关研究团队研究团队先后提出了 DB-BERT [201] 和 λ-Tune [202]。DB-BERT 基于 BERT 模型微调权重，使用数百份关于数据库调优的文本文档作为输入，将自然语言提示转化为推荐设置，并通过强化学习指导调优设置的选择。λ-Tune 是一个基于大语言模型的数据库系统自动调优框架，这一框架通过生成完整的输入文档来描述调优上下文，生成调优配置脚本，可以生成多种候选配置，并采用系统化的策略选择最佳配置。国内阿里巴巴团队提出的 DB-GPT [203] 包括检索增强生成（RAG）知识系统、自适应学习机制，以及一个服务导向的多模型框架

(SMMF)，并配备了强大的数据驱动代理。DB-GPT 的核心创新在于其私有 LLM 技术，经过针对领域特定语料库的微调，以确保用户隐私并保障数据安全。这三项工作展示了大语言模型在数据库领域的广泛应用，特别是在自动调优、SQL 查询生成和数据库系统的自然语言交互方面，推动了技术创新和用户体验的提升。

3.2.4 调度与编排中的智能算法研究

调度与编排是计算资源管理中的核心问题，旨在有效地利用有限的资源，满足不同任务的需求，并优化系统整体性能。调度与编排的关键在于如何在多任务、多资源环境下，合理地分配计算资源并安排任务的执行。在云计算场景中，资源的分配和任务调度受到任务负载、资源种类（如 CPU、内存、存储）、网络带宽等因素的影响。此外，随着任务负载和资源需求的不断变化，调度和编排需要具备一定的动态适应能力。因此，这一问题不仅仅是单一的资源分配问题，而是涉及多个维度和复杂优化目标的多目标决策问题。网络及云计算的很多场景均涉及到调度及编排问题，在云计算平台上，资源调度和任务编排是云服务提供商运营的关键，常见的应用场景包括虚拟机和容器的分配、弹性扩展、负载均衡等；在云边端协同和物联网场景中，任务调度和资源编排的复杂性进一步增加，因为资源通常是分布式和异构的，常见的应用包括边缘计算任务的卸载、联邦学习任务的分配；在分布式数据库场景中，任务调度和资源编排主要涉及查询请求在不同机器间的分配、存储的管理，典型应用场景包括查询优化、负载均衡、数据迁移与副本管理。

在网络及云计算相关应用中，调度与编排问题面临着以下三方面的挑战。(1) 资源和负载动态变化：云计算、网络以及物联网环境下，任务的负载和资源的可用性常常是动态变化的。任务负载的变化可能是由于用户需求的波动、系统性能的变化，或者是因网络带宽、设备故障等外部因素引起的。因此，调度系统需要具备强大的动态适应能力，能够实时监测系统状态并根据变化进行资源重新调配。(2) 任务间存在依赖和协同关系：云计算平台上的许多任务具有前后依赖关系，需确保调度时序正确；任务也可能需要多种资源（计算、存储和网络），需协调分配。(3) 优化目标多样化：由于云计算平台涉及租户和平台两方，且对外提供 IaaS、SaaS、PaaS 层面的多种服务，进行编排调度时需要在多种指标间权衡，包括资源利用率、延迟、能耗、吞吐量等。此外，在调度中需保证多租户间的公平性，同时兼顾整体性能。针对这些技术挑战，业界及学术界开展了丰富的实践和研究，利用包括优化算法、博弈论、强化学习在内的多种工具，在算法设计及系统实现上持续创新，为云计算和网络的发展提供了关键的支撑。接下来，我们对近些年出现的针对调度及编排的典型工作进行更详细的介绍。

基于图算法的调度与编排研究往往利用图结构描述任务间的通信和依赖关系，以网络成本、能源效率、负载均衡和响应延迟为优化目标。针对降低网络成本和通信延迟的目标，可利用图划分算法或图聚类算法将虚拟机划分群组，将同群组虚拟机集中放置，以实现最小化网络成本 [204, 205]。针对优化设备能效的目标，首先通过图划分技术将虚拟机分组，以减少跨分区的通信流量，然后结合 Bin Packing（BP）算法，进一步优化分组结果以减少交换机使用数量和能源消耗 [206]。

基于博弈论的算法在资源调度与编排方面，通常用于解决与经济相关的问题，如定价、资源组合策略或经济方面的优化。在多云协调良好的联邦云中，基于合作博弈论的方法可以最大化买方利润或卖方社会福利。当完整的信息可用时，研究者主要采用集中式算法和动态规划，当信息不完全时，则采用分散式算法。在无合作的多云环境中，研究者采用非合作博弈论分析和设计了多方竞争资源采购的定价方案，并证明了在一些特定情况下，只存在一个纳什均衡，验证了 StackelBerg 平衡点的存在性和唯一性。

图神经网络和强化学习技术能够较好的捕获负载在时空上的相关性，适应动态变化的系统，在协调分布式数据中心并降低运营成本中得到了广泛的应用。研究者将多数据中心任务调度问题和资源扩展问题进行联合优化，提出了基于图神经网络和强化学习的双时间尺度优化框架 [207]。短时间尺度的任务调度可以快速缓解计算任务的突发到达，而长时间尺度的资源扩展可以很好地适应工作负载的长期变化。实验结果表明，该算法能够在保持合理成本的同时，减少任务完成时间和任务超时率。

3.2.5 故障诊断中的智能算法研究

在云计算领域，**故障诊断的核心目标是确保系统的稳定性与高可用性，通过迅速定位故障根源及其影响范围来防止故障蔓延**。故障诊断通常从异常检测开始，首先通过监控系统中的关键指标和日志，识别出偏离正常行为的异常情况，例如资源使用超出阈值、服务延迟增加等。这些异常信息为后续的故障诊断提供了初步的警报和症状描述，帮助及时发现潜在问题。一旦发现异常，故障诊断进入深入分析阶段，进一步利用监控数据和日志，通过事件追踪技术识别故障组件，进而找出根本原因和故障的影响范围。为了实现这一目标，现代故障诊断采用了多种先进技术，如聚类算法、图算法、大语言模型等。这些技术帮助系统在云计算环境中进行高效的故障定位和恢复，确保及时应对复杂、动态的故障场景。

在大规模分布式系统中，图算法凭借对节点之间关联性的揭示，能够有效识别故障的根本原因，并借助图聚类和链路预测等技术实现精准定位与主动预防。图聚类算法能快速定位并隔离具有相似故障特征的节点，而链路预测算法则有助于识别潜在的故障链路，支持故障的主动防范。相关研究团队 [208] 提出了基于图神经网络的微服务异常检测方法 DeepTraLog，使用统一的图表示来描述调用轨迹结构，并将日志事件嵌入其中，获得了高精度的异常检测结果。由于微服务架构中服务之间依赖关系复杂，根因分析尤为困难。对此，有研究团队将轨迹数据构建为服务依赖图结构，借助图相似性和图匹配算法，将系统图与以往异常图进行比对，从而实现有效的根因分析 [209]。

除上述技术外，深度学习与大语言模型也为故障诊断注入了新的活力。相关研究团队 [210] 提出了一种基于时间卷积网络（TCN）和自编码器（AE）的无监督异常检测方法，能够在传感系统故障导致的数据异常中，通过重构误差识别异常，并结合自适应阈值算法，有效提升检测精度，克服传统固定阈值方法导致的误判与漏判。在微服务架构中，大语言模型能够自动识别故障源并提供修复建议，通过分析微服务之间的交互和依赖关系，实现高效的故障诊断。此外，LLM 在云原生管理平台中的应用也逐步完善，如 Kubernetes 中的 GenKubeSec 和 K8sGPT 工具，已成功实现配置检测、故障修复以及自然语言生成诊断报告，极大地简化了故障诊断流程并提高了修复效率。

3.2.6 其他研究热点

在云计算和大规模分布式系统的快速发展过程中，除了前述关键问题外，还有许多研究热点亟待深入关注，例如软件开发、软件测试、数据库交互、网络设计与配置等。这些领域直接影响系统的稳定性、安全性以及高效的设计、开发与管理。随着云环境中设备的异构性和网络复杂性日益增长，如何有效提升程序设计与测试的效率，保障软件的安全性，以及实现网络配置和管理的智能化，已成为学术界与工业界广泛探讨的重要课题。

程序设计。在云计算环境中，程序设计面临诸多挑战，尤其是软件开发、测试及数据库交互方面。LLM 的应用成为解决这些问题的重要方向，通过智能化工具显著提升云服务的开发、部署与维护效率。在软件开发中，LLM 简化了需求收集、系统设计及编程支持流程，例如 GitHub Copilot 通过代码自动补全和问题解决，大幅缩短开发时间。在软件测试领域，LLM 提升了代码测试覆盖率和漏洞检测能力，例如 Meta 的 TestGen-LLM 优化了 Instagram 的测试流程，使更多代码能快速进入生产环境 [211]。在数据库交互中，LLM 驱动的 Text-to-SQL 技术显著降低了非技术用户的数据查询门槛，例如 DIN-SQL [212] 和 DAIL-SQL [213] 增强了查询效率及分析能力。

网络设计与配置。在云计算和大规模分布式系统中，异构设备的多样性和网络性能需求增加，使得网络设计与配置复杂化。LLM 辅助网络设计与配置成为新兴研究方向。在网络设计中，LLM 通过分析设备性能指标和历史模式优化设备选择与网络规划，提升效率与弹性。在网络配置中，LLM 简化了复杂的设备配置流程，推动自配置网络的发展，减少人为错误并提高系统稳定性。例如，ChatNet [214] 结合多模块和工具集成实现高效的网络规划解决方案，GeNet [215] 通过多模态交互优化拓扑设计和设备配置。这些技术推动了网络设计与管理的智能化发展，为云计算的稳定与高效运行奠定了基础。

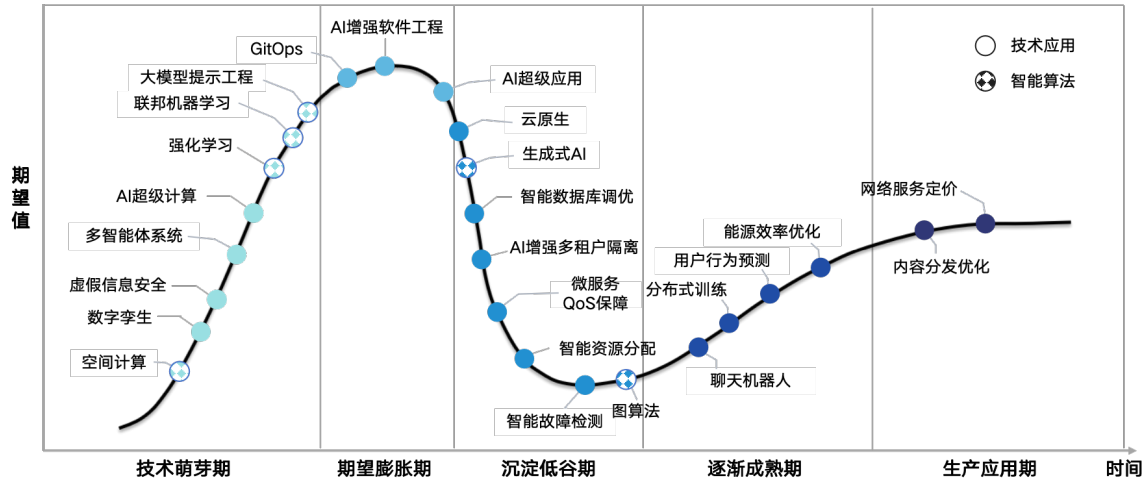


图 3.5: 智能算法研究图谱技术成熟度曲线 2024

3.3 智能算法研究的展望和发展建议

基于技术成熟度曲线的分析方法 [216]（如图 3.5 所示），智能算法领域正沿着这一曲线持续演进，不断推动云计算的智能化转型。本节将聚焦大模型与深度学习、图算法以及优化技术三大关键方向，通过分析其未来研究方向和关键技术，探讨智能算法在赋能云计算生态系统中的作用，并提出针对性的发展建议，为智能算法的研究与应用提供有力支持。

3.3.1 智能算法的未来研究方向和关键技术展望

大模型与深度学习正推动云计算智能化与应用创新，核心方向涵盖隐私优化、自动化运维、增强软件工程及内容生成，为未来云平台升级提供关键支撑。云计算平台的弹性与高并发特性，结合大模型及深度学习的预测与推理能力，可进一步推动云计算智能化的变革升级。关键研究方向与技术包括：（1）联邦机器学习的隐私保护与资源优化。在分布式数据环境下研究联邦学习算法，解决数据隐私保护与跨节点计算资源优化的问题，实现云边协同的智能模型训练。（2）智能运维实现自愈与优化。应用深度学习与大模型技术实现运维自动化，聚焦故障诊断、根因分析、预测性维护等场景，提升大规模云平台的稳定性与运维效率。AI 增强软件工程促进开发效率。通过大模型技术支持自动代码生成、测试用例生成和代码优化，提高云计算相关软件开发的效率和质量，减少开发周期。（3）生成式 AI 推动云端服务创新。研究生成式 AI 在文本、图像、视频等生成任务中的应用，推动智能内容生成和创新型云服务的落地。

图算法通过对依赖关系的建模，能够为复杂系统提供可靠、高效的解决方案，在优化云计算系统的性能、提升资源利用效率以及实现智能化管理方面展现出广阔的应用前景。通过对复杂问题进行建模，可推动云计算系统的智能化运行。关键研究方向与技术包括：超图通过超边有效建模多个节点之间的高阶关联，为云计算中的多维复杂数据分析提供了新的思路。超图算法将在数据中心的资源管理、网络性能优化以及复杂任务分解等方面发挥更大作用，推动高阶关系建模的深入研究与应用。图划分技术在分布式计算、资源管理中至关重要。未来的研究将聚焦于开发基于社区结构、节点重要性或通信特性的智能图划分算法，减少跨分区通信开销，优化负载均衡，提升分布式任务执行的效率和稳定性。云计算环境中网络拓扑和资源状态随时间动态变化。通过研究动态图算法，可以实现对动态网络结构的实时更新与分析，从而优化资源调度、流量管理以及任务分配，提高系统响应速度和灵活性。结合流计算框架，动态图分析能够显著提升云计算系统在高频变化场景下的适应能力。将图神经网络（GNN）与传统图算法相结合，是未来智能云管理的重要方向。GNN 可以通过学习节点和边的特征，挖掘潜在关联，实现智能流量预测、异常检测、任务调度等功能，为云计算系统提供更智能的决策支持。

智能优化及序列决策相关技术具有稳定可靠的特点和较好的可解释性，能够为复杂系统提供透明、高

效的解决方案，在资源管理、网络性能提升和用户服务中发挥了关键作用。通过对复杂问题进行建模和决策，可推动云计算系统的智能化运行。关键研究方向与技术包括：在线学习与强化学习优化动态决策。结合在线学习和强化学习技术，设计动态资源分配与负载均衡策略。通过实时反馈机制优化算法性能，适应用户需求变化，提升系统的灵活性与稳定性。分布式优化助力大规模资源调度。研究云计算平台中多节点资源分配与调度问题，结合分布式计算与通信优化，支持大规模任务的高效分配与协同执行，实现跨区域云资源的全局最优管理。在线凸优化支持实时定价与需求预测。基于在线凸优化技术，开发实时定价模型与用户行为预测算法，动态调整云服务价格策略与资源配置，提升收益与用户满意度，实现服务与需求的精准匹配。

3.3.2 智能算法的发展建议

大模型与深度学习驱动云计算智能化升级。随着云计算平台的持续演进，智能化与自动化资源管理、个性化交互体验和开发流程优化正成为提升云服务效率和用户满意度的关键方向。在资源管理方面，通过大模型和深度学习的负载预测、异常检测和实时分析，云资源可以在高并发场景下精准调度，避免资源浪费或不足。例如，基于 LSTM 的时间序列预测模型可提前识别访问流量高峰，而自动编码器则可及时检测硬件故障或网络瓶颈，实现自动化修复。个性化服务方面，基于用户行为和偏好建模的深度学习算法推动了 SaaS 平台的界面优化和功能推荐，结合多模态智能客服技术，显著提升了用户交互的精准度与效率。开发流程方面，利用自然语言接口生成架构设计和 API 定义，结合深度学习进行性能分析和自动化测试，正推动云计算开发流程的智能化转型。未来，跨云和跨数据中心的智能化、自动化资源调度、服务定制和开发工具将进一步提升系统的稳定性、灵活性和创新性，有效降低运维和开发成本。

图算法推动研究热点与实际场景需求的深度融合。企业在针对复杂问题进行建模时，应结合自身业务特点，精准聚焦云计算领域的核心应用场景的真实问题。同时，将这些实际需求与当前研究热点深度结合，不仅能够加速技术成果的落地转化，还能显著提升企业的技术竞争力，推动行业的持续创新与高质量发展。建议企业通过建立联合实验室或研发合作项目，将实际问题与学术研究相结合，以更高效地探索前沿技术的应用潜力。作为当前研究的热点之一，图算法以其在建模复杂关联关系中的突出优势，能够为资源调度、负载均衡、网络优化等实际场景需求提供高效的解决方案。因此，图算法也成为企业技术创新的理想切入点，助力技术理论与实际需求的深度对接，为推动行业智能化发展奠定了坚实基础。

基于优化理论与智能算法提升云计算系统的效率、稳定性和可解释性。随着云计算和网络系统向规模更大、协同更复杂、场景更多元的方向发展，资源调度、网络优化和系统设计面临着动态性增强和不确定性加剧的挑战。提升系统的决策效率、增强运行的鲁棒性并确保优化过程的透明性，是支撑未来发展的核心需求。在资源分配中，凸优化与随机优化能够高效应对动态需求，结合在线学习技术可实现透明化与实时性；在网络优化中，强化学习结合组合优化可为动态路径调整提供可靠性保障；在复杂系统设计中，在线学习与强化学习通过持续优化和反馈机制，可助力数据中心的能耗管理与服务质量提升。建议持续深化优化理论与智能算法在实际场景中的研究与实践，为未来云计算与网络系统的可持续发展提供技术支持。

第 四 章

面向新兴技术的研究

在全球科技革命与产业变革加速演进的当下，新兴技术产业已成为推动经济高质量发展的关键引擎。我国高度重视新兴技术产业发展，出台了一系列政策战略文件加以引导和支持，为产业蓬勃发展创造了良好环境。与此同时，中国电信作为通信领域的领军企业，积极响应国家战略，在新兴技术产业布局中展现出强大的引领和推动作用。云计算作为新兴技术产业的核心基础设施和关键支撑，在我国的政策战略布局中占据重要地位。例如，《“十四五”数字经济发展规划》明确提出要推动云计算等新兴技术在各领域的深度应用与融合创新，为数字经济发展提供有力支撑。在这一政策指引下，中国电信于 2020 年启动“云改数转”战略，以云计算为核心，大力推动企业自身数字化转型，并助力千行百业上云，实现全社会的数字化变革。近年来，中国电信在新兴技术产业领域持续发力，2023 年提出全面布局云计算及算力、大数据、人工智能、安全、量子、数字平台、新一代信息通信等七大战略新兴业务。通过持续加大技术创新和资源投入，中国电信不仅在国内云计算市场取得了显著进展，如天翼云形成了全球“9+30+X+N”的云资源布局，实现了“集中化 + 区域化 + 属地化 + 边缘化”的云网基础设施，具备了超过 113T 的带宽能力，还在国际舞台上展现出强大的竞争力，开启了海外业务拓展的新篇章。

在云计算技术的强劲赋能与中国电信等企业的积极推动下，新兴技术产业呈现出多元融合、蓬勃发展的态势。云计算所提供的强大的计算能力、灵活的资源调配以及可靠的数据存储与管理功能，为工业互联网、空天地海一体化信息网络、智慧交通、政企数字化转型、医疗信息化、教育智能化以及金融科技化等多领域的创新应用搭建了坚实的舞台，催生出一系列极具潜力与变革性的新兴业态。本章将深入剖析上述云计算相关重点领域的发展现状、面临的挑战、关键技术和研究热点，并对未来发展提出建议。

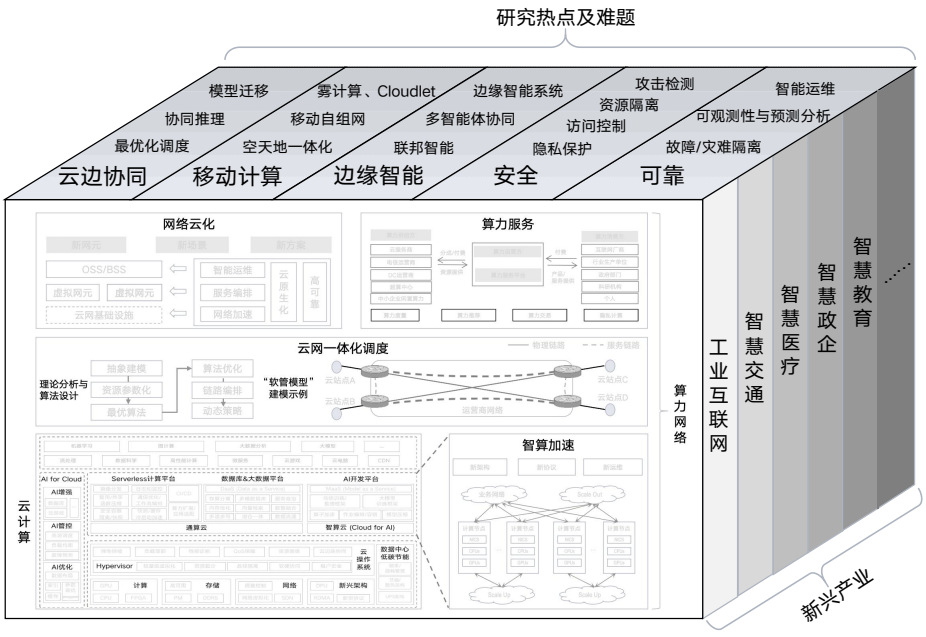


图 4.1: 面向新兴技术的研究图谱：云计算与云网融合在新兴产业的应用与关键技术

4.1 研究图谱及其产生：面向新兴技术的云计算与云网融合研究

当前，云计算和云网融合技术正广泛应用于工业互联网、智慧交通、智慧医疗、智能政务和智慧教育等新兴领域。这些应用又涵盖了云边协同、移动计算、边缘智能、安全以及可靠性等关键技术的研究和实践。为便于直观理解云计算与云网融合在新兴产业的这些应用与关键技术，本节在第一章和第二章研究架构之上，给出面向新兴技术的云计算与云网融合研究图谱，如图 4.1所示。本节的后续内容，将分别介绍云计算和云网融合相关的新兴技术产业分析、云计算和云网融合面临的挑战，以及国内外云厂商支撑新兴技术产业的案例。

4.1.1 产业分析：云计算和云网融合相关的新兴技术产业

工业互联网是指新一代信息通信技术与工业经济深度融合的新型基础设施、应用模式和工业生态。它通过人、机、物的全面互联，实现全要素、全产业链、全价值链的全面连接，将工业生产过程中的各类数据进行采集、传输、存储和分析处理，从而驱动工业生产方式和企业形态的根本性变革，提高工业经济的质量和效益。工业互联网借助云计算的虚拟化、分布式计算、弹性扩展等特性，能够快速响应工业生产中的各种需求变化，实现工业资源的优化配置和协同共享。在全球工业互联网发展进程中，政策引导的重要性不言而喻。国际上，美国政府提出“工业互联网参考架构”等概念，美国国家标准与技术研究院（NIST）制定标准并以税收优惠激励企业创新；德国“工业 4.0”战略通过政策与资金扶持企业创新，且在国际标准制定方面积极作为，提升其全球话语权，各方政策共同推动工业互联网在全球范围内的发展与变革。我国极为重视工业互联网发展，相继出台了诸如《工业互联网创新发展行动计划（2021-2023 年）》等一系列政策，明确了提升工业互联网平台核心能力、推动工业设备和业务系统上云上平台以及培育新模式新业态等目标，并通过资金扶持、税收优惠等措施大力促进企业创新应用。

智慧交通是指利用新一代信息通信技术对交通系统进行智能化管理、优化和服务，从而提高交通效率、减少拥堵、保障交通安全、降低环境污染，最终实现交通系统的智能化、自动化和可持续发展。其中典型代表是车联网和低空经济。车联网是指以车辆为主体，通过信息和通信技术，实现车内、车与车、车与人、车与路、车与云的互联互通、信息共享。近年来，我国政府相关部门已出台了一系列与车联网相关的政策，鼓励发展智能网联汽车、自动驾驶、智能车载系统等领域。低空经济是指在低空空域范围内，以无人机和通用航空器为载体，结合信息技术与现代服务业，实现资源整合与经济增长的新型经济模式。我国的低空经济起步较慢，经历了严格管控阶段，直到 2021 年，《国家综合立体交通网规划纲要》将低空经济作为重点打造的交通形态之一，我国低空经济才进入快速培育期。2023 年，中央经济工作会议明确提出低空经济是战略性新兴产业之一。2024 年，低空经济被写入《国务院政府工作报告》。

智慧医疗是指一种将现代信息技术与传统医疗服务深度融合的新型医疗模式，其核心是通过大数据、云计算、人工智能、物联网等技术手段，优化医疗资源配置、提升医疗服务效率和质量，从而实现“精准诊疗、便捷服务和高效管理”。与传统医疗不同，智慧医疗更注重患者体验，通过技术赋能，让患者能够享受到更个性化、更智能化的医疗服务。智慧医疗的具体应用领域包括远程医疗、智能诊断、健康管理、药物研发、智能设备监测等。欧盟“数字欧洲计划”（2021-2027）投入巨额资金发展包括智慧医疗在内的数字化基础设施，同时鼓励成员国合作建立跨境医疗数据共享平台。我国近年来高度重视智慧医疗的发展，并将其纳入“健康中国”战略。早在 2016 年就提出《“健康中国 2030”规划纲要》，明确提出要推动“互联网 + 健康医疗”模式发展，支持智慧医疗平台建设。2018 年《关于促进“互联网 + 医疗健康”发展的意见》提出推进互联网医院建设、推广远程医疗应用、实现线上线下一体化医疗服务。《“十四五”全民健康信息化规划》明确指出，深化“互联网 + 医疗健康”服务体系，完善健康医疗大数据资源要素体系。此外，我国还在多个试点城市建设智慧医疗示范区，例如上海的长三角智慧健康一体化示范区和杭州的智慧医疗产业园，这些措施为全国推广智慧医疗提供了宝贵经验。未来，智慧医疗需要在技术创新和政策引导下不断完善，实现更大规模的应用和推广，为全球健康事业作出贡献。

智慧政企是指将新一代信息技术应用于政企管理与服务的新模式，旨在提升政府和企业的数字化治理能力和服务水平。智慧政企通过整合资源、优化流程，实现精准决策、实时响应与高效协作，推动公共管理和企业运营从传统模式向智能化、数据驱动的模式转变。智慧政企的应用领域广泛，主要包括智慧政务、智慧企业管理、政企协同、智慧城市建设等。2019 年美国政府发布的《联邦数据战略》，推动政府数据开放与共享，支持智慧政务与企业数据服务的发展。2023 年，我国中共中央、国务院印发的《数字中国建设整体布局规划》中指出要加强数字政府建设，推动政府数字化转型，促进信息系统网络互联互通、数据按需共享、业务高效协同，提升政务数字化服务水平。《"十四五" 推进国家政务信息化规划》提出，到 2025 年，我国政务信息化建设总体迈入以数据赋能、协同治理、智慧决策、优质服务为主要特征的融慧治理新阶段。智慧政企是数字化转型背景下的一项重要发展方向，融合技术创新与管理创新，能够显著提升政府治理效率和企业竞争力。

智慧教育是指利用新一代信息通信技术，构建数字化、智能化的教育生态系统，以提升教学质量、优化资源分配和个性化学习体验为核心目标的教育模式。智慧教育不仅关注传统课堂教学的数字化转型，还强调通过技术赋能实现 "因材施教"，帮助学生自主学习、教师精准教学、学校高效管理，从而推动教育公平与质量提升。2023 年，美国教育部发布《人工智能与教学的未来》，报告指出美国教育部致力于支持利用人工智能技术改善教与学，并支持整个教育系统的创新。欧盟制定并颁布《数字教育行动计划 (2021-2027)》提出要发展高效的数字教育生态系统，支持各成员国教育和培训系统的数字化转型。2023 年，教育部、国家发展改革委、财政部联合发布《关于实施新时代基础教育扩优提质行动计划的意见》，意见指出将提升国家中小学智慧教育平台建设应用水平，加大在智慧课堂、智慧作业、个性化学习等方面的功能，促进优质教育资源广泛共享。智慧教育是教育领域数字化转型的重要方向，其通过技术与教育的深度融合，推动教学方式和学习模式的创新，实现资源均衡与教育质量的全面提升。

这些新兴产业正处于快速发展阶段，通过技术创新驱动行业变革。作为数字化转型的关键支撑，云计算以其强大的数据存储、计算和分析能力，赋能各行业实现资源整合、智能决策和高效协同，为推动产业转型升级提供了重要技术基础和发展动力。

4.1.2 云计算和云网融合面临的挑战

新兴产业蓬勃兴起，持续推动着云服务在存储、计算以及网络方面的变革，在为数据处理和资源管理开拓新路径的同时，也引发了诸多挑战。表 4.1列出了新兴技术在推动云计算与云网融合的发展进程中，在存储、计算及网络领域，围绕协同性、移动性、智能性、安全性和可靠性所面临的新挑战。妥善应对这些挑战，是构建高效、安全、可靠云环境的关键，对于云技术产业的稳健发展意义重大。接下来，本小节将基于五个性能维度详细展开介绍。

表 4.1: 云服务随着新兴产业发展遇到的新挑战

	存储	计算	网络
协同性	数据异构	资源差异	海量传输
移动性	多源同步	能源受限	有限覆盖
智能性	管理复杂	聚合困难	通信频繁
安全性	数据易泄露	资源边界模糊	环境复杂
可靠性	数据易丢失	跨域数据整合	高动态环境

端云协同面临着海量异构数据、差异化的设备资源、大规模数据传输带来的挑战。(1) 在存储方面，终端设备产生海量的监测数据，包括数值型数据、图像、音频等多模态数据，而边缘设备和云上则包含大量结构化数据和非结构化数据。不同类型的数据在存储需求、存储格式和访问方式上都有很大区别，给端云协同存储系统的设计和管理带来了复杂性。(2) 在计算方面，云边端各级节点的计算资源和能力存在巨大差异，从高性能的中心云服务器到低性能的终端物联网设备都参与到数据的处理和存储中。针对不同能力的云边端设备，难以使用统一的标准进行任务卸载和资源管理。(3) 在网络方面，随着大规模边缘设备的接入，海量数据需要从终端设备或边缘节点传输至云端，以及从云端反

馈处理结果至终端或边缘，这使得云网络传输的数据量大幅增加，极易导致网络带宽拥塞。如何针对海量数据传输，提高网络吞吐量，降低数据传输延迟是目前存在的挑战。

动态复杂的移动环境下，多源数据的实时同步、移动设备有限的计算与网络资源对云服务提出了新的要求。(1) 在存储方面，多个移动设备需要与云端存储的数据保持一致。然而，在移动网络环境下，由于网络延迟、设备性能等因素的影响，数据的同步可能会出现问题，导致不同设备上的数据不一致。如何针对多源移动设备与云端存储的数据在移动互联网上维持同步是目前存在的挑战。(2) 在计算方面，移动边缘设备通常依靠电池供电，能源有限使得其计算单元（如 CPU、GPU 等）不能长时间以高性能模式运行。例如，在一些物联网传感器设备中，由于能源限制，其处理器可能会降频以节省电量。这就导致复杂的计算任务难以在边缘设备上高效完成。(3) 在网络方面，随着我国低空经济和海洋发展战略的推进实施，云计算的服务区域需要从地面扩展到空域、海域等无人区，为多维空间中的用户与传感设施提供立体多维、覆盖全时、全域、全空间的云网服务。然而，现有通信与算力网络严重依赖地面基础设施，难以在低空、远海、荒漠等区域提供高质量服务。

云服务的智能化面临着数据管理复杂、模型聚合困难、通信开销巨大等一系列新的难题。(1) 在存储方面，智能化依赖的大规模数据通常分布在多个边缘设备、本地服务器和云端存储节点上，其分布与一致性管理非常复杂。这种分布式数据的复杂分布增加了数据管理的难度。云存储系统需要能够有效地管理这种分布式的数据，确保数据的一致性和完整性。(2) 在计算方面，在联邦智能中，模型通常是在分布式环境下进行训练和更新的。云计算平台需要协调各个节点上的模型训练过程，并将各个节点的模型参数进行聚合和更新。考虑到不同节点的模型质量、数据分布差异等因素，如何确保聚合后的模型能够准确地反映全局数据的特征，这增加了模型管理的复杂性。(3) 在网络方面，在联邦智能中，数据需要在分布式的节点和云计算平台之间频繁传输，这会导致大量的通信开销。例如，在联邦学习的参数聚合过程中，大量的模型参数需要在本地设备和云端服务器之间来回传输，可能会造成网络拥塞，尤其在网络带宽有限的情况下，会严重影响联邦智能的运行效率。

云服务的安全性方面正面临着敏感数据泄露、云资源边界模糊、网络环境复杂多样等严峻挑战。(1) 在存储方面，随着数据量的增长，大规模数据存储在云端，面临的最大问题就是安全性和隐私性。例如，在低空经济中，不同类型的航空器产生的数据格式和安全级别各异，需要更加精细的数据分类和加密策略。而且云端数据是动态变化的，数据会不断地被添加、修改、删除和共享。传统的加密技术和访问控制机制可能难以应对如此大规模、复杂、动态的数据安全管理需求。(2) 在计算方面，云计算安全主要关注云计算环境中的资源（如虚拟机、容器、计算集群等）的安全性，确保这些计算资源在运行时资源隔离，以防止未经授权的访问、恶意攻击或服务中断等问题。DaaS 和 MaaS 的出现使得资源隔离的边界变得模糊。以 DaaS 为例，数据作为一种服务被多个用户共享使用，这些数据可能包含敏感信息。不同用户对数据的访问权限、数据的加密级别、数据的合规性要求等都不同，这就需要更精细的资源隔离技术来确保数据的安全性和隐私性。(3) 在网络方面，云服务中的网络安全涉及保护云计算环境中的所有网络组件—如数据传输、网络连接、API 接口等免受攻击、滥用、数据泄露和其他安全威胁。新兴技术的网络环境复杂多样，可能包括卫星通信、5G 网络、物联网等多种通信方式的融合。云服务提供商需要确保在这种复杂网络环境下的通信安全，防止网络攻击。

随着新兴产业不断发展，数据易丢失、跨域数据整合的复杂性以及高动态环境等问题日益凸显，对云服务的可靠性提出了更高的要求。(1) 在存储方面，随着新兴产业与云计算的紧密结合，云存储的数据量呈爆炸式增长。但由于数据量巨大且产生速度极快，云存储系统面临着巨大压力。例如在低空经济领域，无人机飞行数据如飞行轨迹、姿态信息等需要实时存储以便后续分析飞行安全性和优化飞行计划，一旦云存储系统遭遇硬件故障、软件错误或恶意攻击，数据丢失可能使无人机的运营安全评估失去依据，无法准确判断飞行风险。因此对云存储的可靠性提出了极为严苛的要求，传统存储模式难以确保数据在复杂环境下的完整性与安全性。(2) 在计算方面，新兴产业的云服务应用往往涉及跨域数据的处理，例如在空天地海一体化场景中，卫星网络、地面网络和海洋网络各自产生的数据在格式、精度、时效性等方

面存在显著差异。数据的异构性与海量性使得传统云计算架构难以高效处理，在确保计算准确性与时效性方面面临巨大挑战，需要全新的计算策略与资源调配机制来保障计算的可靠性。（3）在网络方面，以智慧交通为例，车辆高速移动且大量接入云网络，网络拓扑快速变化，传统故障恢复机制难以迅速响应。一旦发生故障，如网络拥塞或基站故障，在保障数据完整性与业务连续性上，现有的故障 / 容灾管理体系难以满足要求。网络故障与任务动态变化相互交织，对故障的实时监测、精准定位以及快速恢复提出更高标准，需要构建更智能、灵活且具前瞻性的故障恢复机制，以有效应对复杂多变的高动态环境，确保云网络的可靠性与稳定性，进而保障新兴产业业务的正常运转与持续发展。

4.1.3 国内外云厂商案例

在当今数字化浪潮的席卷下，国内外各大云厂商如 Amazon、Microsoft、Google、NVIDIA、阿里巴巴、华为云、腾讯云等积极投身新兴产业，与众多企业携手合作，催生出一系列极具创新性与影响力的应用案例，在全球范围内掀起了一场技术赋能产业变革的热潮。

在工业互联网领域，国外 Microsoft Azure 与上海振华重工合作，利用 IoT 和 AI 技术构建设备物联网，实现预测性维护与远程监测运营，创建新门户提供物流监控服务；国内华为云与宝钢合作，利用 5G 网络的大带宽、低时延、高可靠等特性，实现了工业领域的视频监控回传、远程控制、数据采集与预测性维护等应用，打造了 5G 智慧工厂；腾讯云旗下的腾讯 WeMake 工业互联网平台连续三年入选国家级双跨平台，已服务 42 万家制造企业，覆盖 26 个行业，为三一重工、工业富联等龙头企业提供数字底座。

在空天地海领域，国外 Google 与澳大利亚联邦科学与工业研究组织、塔斯马尼亚大学海洋与南极研究所等机构联合，借助 Google 云的 AI 平台 Vertex AI 在超过 7000 平方公里的卫星图像中对海藻森林进行高速识别、定位和分析；Microsoft 启动 AzureSpace 项目，为卫星通信和遥感数据处理企业提供强大的云计算能力和存储支持，为海洋监测、气象预报、地理测绘等空天地海相关应用提供服务；国内华为云充分发挥在云计算、大数据等领域的技术优势，协助文昌市打造空天地海一体化平台，推动遥感云、航天数字经济等产业发展，为海洋科研及航天企业分别提供数据处理与技术支持。

在智慧交通领域，国外 NVIDIA 借助 Jetson AGX Orin 和 Omniverse 平台助力 Kodifly 构建交通基础设施的数字孪生与实时三维分析，凭借强大的 GPU 计算能力处理大量实时交通数据，其 Omniverse 平台为数字孪生模型创建与管理营造良好环境，加速数据处理与模型构建进程，削减时间和成本，增强整体解决方案的性能与可靠性；Amazon 获得美国联邦航空管理局的批准，在美国亚利桑那州菲尼克斯西部的托利森市测试其新型小型送货无人机，该无人机重量更轻、噪音更小，可以运送超过 50000 种货物，有望为物流配送领域带来革新；国内华为携手云南移动，在洱海完成西部首个 5.5G 通感一体低空试点，实现了实时显示无人机测量速度、角度、位置精度等信息，还具备轨迹跟踪、黑飞入侵、电子围栏告警等关键功能，为低空经济发展注入新动能。

在智慧医疗领域，阿里云于浙江丽水开展 “医疗 AI 多癌早筛公益项目”，借助 “平扫 CT+AI” 技术助力癌症早筛；华为云与上海润达医疗携手打造医疗 AI 大模型；Google 云推出护士交接数字助理 HCA Healthcare Katie 及 Vertex AI Search；Amazon 云科技与大米和小米合作推出特需儿童康复 AI 解决方案并发布 AWS Health Scribe；Microsoft Azure 联合医生集团构建智能医疗云平台并开发疾病智能预测系统等。在智慧政企方面，阿里云中标中核核电核工业数据中心云平台等项目；腾讯云拿下上海智慧健康松江一体化云平台项目并助力数据库业务拓展；华为云与安顺经开区管委会等合作搭建 “安顺国电南自-华为数字生态云”；百度智能云在交通等领域收获邯郸市峰峰矿区智慧交通等项目；AWS 与埃森哲合作服务医疗保健等行业。在智慧教育领域，Google 推出一系列教育应用和工具，如 Google Classroom，通过云计算技术为教师和学生提供在线教学和学习平台，支持课程管理、作业布置与批改、在线交流等功能，促进了教育的数字化和个性化发展；华为云为高校提供混合云支持科研大数据分析平台建设。

未来，在新兴产业布局时应充分利用自身在网络技术、云服务、大数据和 AI 领域的优势，深化与各行业龙头的合作，推动产业数字化转型。通过强化网络基础设施建设，提升云服务能力和数据处理技术，

支持更多创新应用的落地，满足不同行业对数据处理和存储的需求。同时，积极探索新的服务模式和应用场景，以用户需求为导向，提供更加个性化和智能化的服务，推动技术创新和产业升级。

4.2 研究洞察：面向新兴技术的研究热点和难题

随着信息技术的高速发展，尤其是 5G、物联网时代的到来，入云的设备数量正在指数级增长，由此产生了大量数据，需要一种新的云计算模式来满足对实时性、移动性、数据安全性的要求。本节通过调研近几年发表的新兴领域与云计算相关的学术论文，梳理了国内外研究学者们重点关注的研究方向，包括云边协同、移动计算、边缘智能、安全性和可靠性，及其代表性技术，并整理如表4.2所示。

研究热点和难题

1. **云边协同**：如何通过任务卸载和资源调度，进一步提升其运行效率与智能水平?
2. **移动计算**：如何在基础设施受限的移动场景下更好地提供云计算服务?
3. **边缘智能**：如何优化智能算法在边缘设备上的部署，提供就近入云的高质量服务?
4. **安全性**：如何保障大规模入云数据的隐私安全与智能防护?
5. **可靠性**：如何预防、诊断云网系统的故障，增强云计算服务的可靠性?

4.2.1 云边协同研究

面向海量的物联网设备与移动终端的大规模数据入云需求，传统的云计算模型面临着高响应延迟和带宽限制的挑战，而这些新兴移动计算任务往往具有时延敏感的特性 [277]。在此情况下，边缘计算的出现为解决该问题提供了新的思路：将数据在靠近用户的边缘设备上进行处理，从而提供更低的延迟，并且优化带宽的利用率 [278]。边缘计算具有响应速度快，但计算资源相对有限的特点，与虽有强大的计算能力但难以满足时延敏感的任务的云计算形成了互补。云边协同能够结合边缘计算的低时延、本地处理优势与云计算的强大计算存储能力，让数据在边缘端进行预处理与筛选，将重要数据或需要深度分析的数据传输至云端进一步处理，实现资源的高效配置与利用，满足不同类型应用场景对于计算性能、响应速度、数据安全等多方面的综合需求，推动物联网、智能交通、工业互联网等众多领域的创新发展。在云边协同的体系架构中，如何进一步提升其运行效率与智能水平成为了关键的研究方向。最优化调度旨在根据边缘设备与云端的实时资源状况、任务的紧急程度及数据特性等多方面因素，制定出最为合理的任务分配与资源调配策略，确保整个云边协同系统能够流畅且高效地运行 [217]。此外，近年来涌现的人工智能技术与机器学习方法也成为了云边协同下的研究热点：如何联合云端和本地的计算资源实现深度学习模型的协同推理，以及如何将云端训练好的强大模型迁移至资源有限的边缘设备上提供实时服务 [221, 225]。

针对资源分配与任务卸载机制，设计最优化调度算法，优化系统的整体吞吐量、时延、以及整体能耗。为了向数量激增的边缘设备提供大规模、低时延的云计算服务，近年来海内外研究机构纷纷在最优化调度方面开展研究：如何面向资源差异较大的边缘节点与云计算中心，设计资源调度算法和任务卸载机制，优化任务的吞吐量、时延、以及整体能耗。最近的研究工作如下：（1）最大化吞吐量。针对海量计算任务的随机到达与有限的无线信道资源之间的矛盾，研究人员提出了渐进式的任务调度方法，利用李雅普诺夫（Lypunov）函数来实现最优调度，在最大化移动边缘网络吞吐量的同时兼顾分配机制的公平性 [218]。（2）最小化时延。针对不同任务的异质性特征，最近的研究工作提出基于博弈论的任务卸载决策优化算法，减少批量任务的处理时间 [219]。（3）最小化能量消耗。针对协同计算中的能耗问题和实时性要求，研究人员提出了基于当前任务执行成本的动态计算卸载与资源调度算法，在满足任务时延需求的前提下，降低整体能耗 [220]。

联合端云设备实现高效的深度学习模型推理，满足低时延、高可靠的服务需求。近年来，深度学习模型在终端设备上发挥着重要的作用，支撑着例如人脸识别、异常动作检测、火灾预警等应用。然而，当前无人机等终端设备的计算资源有限，为了在终端设备上实现快速实时的任务响应，近年来国内外的研

表 4.2: 新兴技术研究热点

研究方向	方向概述	代表性技术
云边协同	随着物联网和移动终端设备的数量剧增，传统的数据中心节点面临着传输带宽有限以及响应时延高的挑战，而边缘节点虽然响应迅速但是计算资源有限。云边协同将两者的优势结合，在边缘设备进行数据预处理后，将后续任务卸载至云端快速处理，实现低时延的协同计算。	• **最优化调度：**针对边缘节点和云计算中心的计算资源差异，设计资源调度算法和任务卸载机制，从而实现任务执行吞吐量的最大化、时延以及整体能耗的最小化等目标 [217, 218, 219, 220]； • **协同推理：**通过将深度学习模型分割后分别部署在边缘侧和云端，实现推理任务的部分卸载，从而联合利用端云设备的可用资源，实现低延迟的协同推理 [221, 222, 223, 224]； • **模型迁移:** 将部署在云端的深度学习模型通过模型压缩、知识蒸馏等技术快速迁移至边缘侧，适应其有限的计算资源 [225, 226, 227, 228]。
移动计算	随着智能手机、平板电脑等移动终端的广泛普及，用户对于在移动设备上获得无处不在的云计算服务的需求正急剧增长。借助移动无线网络技术，使得用户能够在任何地点实时、可靠连接到云端资源的移动计算技术正成为近年来的研究热点。	• **空天地海一体化：**在缺乏地面基础设施的地区，将卫星侧基础设施作为地面的关键补充，从而构建一体化的全空间立体基础设施，提供全域无缝覆盖的服务 [229, 230, 231, 232]； • **移动自组网：**通过将无人机等移动设备作为网络节点，实现无基础设施的无线自组网络，满足复杂场景下的通信需求 [233, 234, 235, 236]； • **雾计算：**面向低时延数据传输的需求，将计算、存储和网络服务拓展至网络边缘，使其更靠近用户与数据侧，从而有效降低时延 [237, 238]； • **Cloudlet：**即小型化的云数据中心，能够部署在更靠近于移动用户的网络边缘，从而缓解云计算远程访问带来的局限性 [239, 240]。
边缘智能	为了缓解大规模边缘设备所产生的海量数据给云计算中心带来的巨大压力，在靠近数据源的边缘设备（如智能网关、边缘服务器等）上部署智能算法，实现数据的就近预处理与初步特征提取，减少数据传输量，降低中心云的负担，并且提高智能算法的整体响应速度。	• **联邦智能：**在多个边缘设备上分布式地收集数据并进行本地处理，使本地计算出的特征与参数参与到模型训练中，在充分利用边缘设备资源的同时确保数据隐私安全 [241, 242, 243, 244]； • **多智能体协同:**针对多个分散的边缘设备，将其建模为多智能体进行整体协调部署与调度，从而解决复杂任务并实现全局最优 [245, 246, 247, 248]； • **边缘智能系统：**针对边缘设备内存有限、设备异构、算力不足等问题，设计面向边缘设备的定制化智能系统，充分利用有限的资源实现深度学习的模型推理等任务 [249, 250, 251, 252]。
安全性	随着业务的多样化和复杂化，云服务需要提供更加灵活的计算和存储能力，同时要保证敏感数据不被恶意访问或泄露，并持续不断地优化虚拟化平台的安全性，进行细粒度的访问控制。传统的静态规则检测和防护手段显得力不从心，云服务需要更多智能化、自动化的安全技术来应对动态、复杂的威胁，利用人工智能技术辅助云网安全防护成为近年来的研究热点。	• **隐私保护与访问控制：**通过加密技术保护数据隐私，同时采用访问控制技术控制用户对数据的访问权限，防止未经授权的访问和数据滥用，提高了云平台的安全性和效率 [253, 254, 255]； • **资源隔离:**通过虚拟化、TEE 等技术确保共享资源（如计算、存储、网络等）被严格划分，使得每个租户或实例仅能访问其分配的资源 [256, 257, 258]； • **网络攻击与防御：**针对窃取数据、破坏系统或操控资源的网络攻击，不断发展实时检测与响应的防御技术，以增强网络防御能力 [259, 260, 261]； • **人工智能辅助云网安全：**通过机器学习和深度学习算法，人工智能能够从海量日志和网络流量中快速识别异常行为和潜在威胁，构建更智能、高效和动态的云网安全体系 [262, 263, 264, 265, 266]。
可靠性	随着新兴行业的数据剧增，其可靠性面临诸多新挑战。通过分析海量系统数据，提前预判潜在故障风险，研究分布式系统的容错机制，通过冗余设计和动态资源调配，确保在部分组件失效时系统仍能稳定运行，为云计算在复杂多变的新兴技术应用中提供可靠保障，满足不同行业的严格需求。	• **故障/灾难隔离：**面对如硬件漏洞、软件缺陷、外部攻击等引发的故障 / 灾难时，防止其在系统中扩散，保障服务连续性和数据安全性，加强应急响应 [267, 268, 269]； • **可观测性和可预测性分析：**针对云计算和微服务系统状态难以及时精准把控、性能波动难以提前预估等难题，对系统运行数据进行深度采集、分析与解读，实现对系统潜在问题的洞察和性能预测 [270, 271, 272]； • **智能运维：**传统运维方式在面对新兴场景下复杂多变的云计算环境及海量数据时，存在的故障发现与修复滞后、资源调配不灵活、运维效率低下且人力成本高等问题，利用智能方法实现对系统实时监测、故障自动预测与精准定位、资源智能优化调配及运维流程自动化 [273, 274, 275, 276]。

究者提出了协同推理技术，即将神经网络模型进行分割，并分别部署在终端和云端，将推理任务部分卸载至云端，从而兼顾低延迟和强大计算能力的需求。近年来的研究热点主要关注网络动态波动、随机丢包带来的挑战，并且致力于优化批量协同任务的并行效率。（1）动态卸载。针对网络的动态波动对特征

传输带来的挑战，研究人员提出了动态卸载的方法，渐进式地向云端传输特征的同时在本地继续执行推理任务，从而在网络状态不佳的情况下更多地依赖本地进行推理 [222]。（2）差错容忍。针对网络随机丢包对特征传输造成的干扰，近期的研究工作设计了差错容忍的协同推理方法，在发送的特征数据上进行随机交织编码和不等差错保护，从而提高对随机丢包的抵抗性 [223]。（3）批量重构。针对模型分割方式的差异，研究人员在云端设计了批量重构方法，将大量差异化分割的模型批量对齐，以提高云侧大规模并行推理的效率 [224]。

将深度学习模型通过压缩、蒸馏等方式迁移至端侧设备，提供就近的低时延服务。将云端强大的深度学习模型迁移至边缘设备上部署，能够提供就近的实时服务，同时也避免了在边缘设备上重复进行大规模模型训练，节省了边缘的计算资源和能源消耗。近年来，许多研究工作关注着模型压缩、知识蒸馏、迁移学习等技术，为云端深度学习模型的快速化迁移和本地部署提供支持。（1）模型压缩。为了压缩神经网络模型，去除冗余的模型参数，研究人员设计了基于判别感知的模型剪枝方法，利用注意力机制保留神经网络中最具判别力的通道，自适应地进行模型压缩，同时保持较好的模型性能 [226]。（2）知识蒸馏。为了在资源受限的边缘设备上部署深度学习算法，研究者利用云端服务器模型作为教师模型，在边缘设备上监督轻量级的学生模型的训练，降低边缘模型的由于压缩带来的性能损失 [227]。（3）迁移学习。为了将云端强大的神经模型的能力迁移至边缘设备上，研究团队设计了面向工业物联网系统的迁移学习框架，减少了边缘模型的训练时间，同时提高了模型准确性 [228]。

4.2.2 移动计算研究

随着智能手机、平板电脑等移动终端的广泛普及，用户对于在移动设备上获取云计算服务的需求急剧增长：借助移动网络技术，用户能够在任何地点即时连接到云端资源，要求云计算服务能够支持无处不在的可访问性。然而移动网络严重依赖地面网络基础设施，如何在移动场景下更好地提供云计算服务实现移动计算成了近年来的重要研究热点 [279]。为了更好地应对移动计算的复杂需求与挑战，一系列关键技术应运而生。其中，空天地海一体化网络技术构建了全方位、多层次的通信网络架构，确保移动计算在不同地理环境下都能无缝对接云计算资源 [229]；移动自组网技术能够在自然灾害救援等缺少网络基础设施的特殊场景下满足大量移动设备的接入需求 [233]；此外，雾计算与 Cloudlet 技术在移动计算与云计算的协同中发挥着关键作用。雾计算将计算、存储和网络服务推向网络边缘，更靠近移动用户 [237]；Cloudlet 则作为一种小型的、部署在靠近移动设备端的云资源，能够快速响应移动设备的请求，有望推动移动计算在云计算领域的深度应用与创新拓展 [239]。

通过空天地海一体化技术提供全域无缝覆盖的通信与云计算服务，拓展传统服务的覆盖范围。随着我国低空经济和海洋发展战略的推进实施，云计算的服务区域需要从地面服务扩展到空域、海域等无人区，为多维空间中的用户与传感设施提供立体多维、覆盖全时、全域、全空间的云网服务。然而，现有通信与算力网络严重依赖地面基础设施，难以在低空、远海、荒漠等区域提供高质量服务。为了提供全域无缝覆盖服务，卫星侧的计算与通信基础设施需作为地面的补充，与传统地面设施共同构建一体化的全空间立体基础设施，即空天地海一体化。由于传输距离较远、卫星移动较快等问题，空天地海一体化网络面临着拓扑复杂、传输时延大、部署成本高等问题。针对上述问题，最近的研究热点如下：（1）流量自适应卸载。针对一体化网络中日益增长的通信流量需求，来自日本的研究团队利用深度强化学习中的 Q-learning 算法训练各个节点，使其能够根据本地历史信息以及邻近节点信息智能化选择流量卸载策略，从而缓解了网络拥塞 [230]。（2）动态卫星路由。针对低轨卫星之间复杂多变的路由，研究人员提出一种基于图神经网络的分布式动态路由算法，将路由问题建模为部分可观察的马尔可夫决策过程，使得每个卫星仅与相邻一跳的邻居节点共享信息，随后利用图注意网络得到多跳信息的隐藏特征，从而进行动态路由决策 [231]。（3）星地协调互联。针对现有星地互联方案网络波动大、时延高的问题，研究团队利用不同地面站之间可见卫星分布的相似特性，设计了地面站间分布式部署的星地协调互联算法，协调多地面站之间星地链路的建立，从而最小化传输延迟，同时保持稳定的路由和高网络可达性 [232]。

在缺乏基础设施的极端环境下，利用无人机等移动节点构建移动自组网，提供应急的通信与计算服

务。随着移动计算和无线通信技术的飞速发展，云计算的覆盖范围得到拓展。移动自组网作为一种没有基础设施且自组的无线网络，其发展满足了战场、防灾等场合的需求。移动自组网的网络节点由无人机等组成可以任意移动，网络的拓扑结构动态变化，这要求网络能够适应这种动态变化并保持可靠传输。为了应对这种挑战，最近的研究热点如下：（1）可扩展式组网。为了提高移动自组网的可拓展性，可以在无人机之间随机集中和按需聚类，以减少无人机和地面之间的通信需求，从而实现更好的可扩展性 [234]。（2）移动网元管理。为了管理大规模无人机等组成的网元节点以形成良好的移动自组网，来自加拿大的研究团队提出了一种基于能量感知的无人机群和移动性预测方案，支持无人机路线和路径规划、移动性预测和多跳通信管理，从而实现更高的传输效率 [235]。（3）延迟容忍网络。为了在拓扑变化频繁的情况下提高移动自组网对于网元动态移动的鲁棒性，延迟容忍网络（DTN）中的信息传播技术是值得关注的。美国的研究团队提出了一种基于轨迹的分布式容忍网络路由算法以解决车辆的移动随机性对于网络的影响，通过预测车辆移动轨迹在车辆到达前调度数据包，从而提高路由性能 [236]。

利用雾计算与 Cloudlet 技术在更接近用户的位置布置边缘节点，提供大规模、低时延的云计算服务。移动设备产生海量数据需实时处理与低延迟响应，云计算虽有强大处理能力，但因数据传输距离远，会导致高延迟、网络拥塞及隐私安全隐患。在此背景下，雾计算与 Cloudlet 受到了关注。雾计算将计算、存储和网络服务拓展至网络边缘，更靠近数据源与用户，能有效弥补云计算不足；Cloudlet 是一种位于网络边缘的小型云计算数据中心，它将云计算的能力延伸到网络的边缘，更接近用户和终端设备，从而为用户提供低延迟、高带宽的计算和存储服务。当前围绕这两种新兴技术的研究热点如下：（1）模糊卸载。由于雾节点数量众多，将海量物联网应用任务卸载至雾节点具有较大的决策搜索空间，针对这一问题，研究团队设计了一种模糊卸载策略，利用多目标分布估计算法来从各种应用程序中学习和优化该策略，从而缩小搜索空间，节省系统资源 [238]。（2）部署优化。为了使得 Cloudlet 提供最小化延迟、实现负载平衡、最小化成本与能源损耗，近年来很多研究关注 Cloudlet 的部署优化问题。例如美国的研究团队通过设计双因子近似算法解决异构 Cloudlet 的部署问题，以保证有限的延迟和放置成本，同时将任务完全映射到适当的 Cloudlet [240]。

4.2.3 边缘智能研究

随着近年来边缘设备计算能力的不断提高，将智能算法部署在边缘设备提供就近服务的边缘智能技术得到了国内外研究人员的重视。边缘智能能够在靠近数据源的边缘设备（如智能网关、边缘服务器等）上能够实现数据的预处理与初步特征提取，减少了数据传输量，降低中心云的负担，并且提高了整体系统的响应速度。然而，边缘智能在发展过程中也面临诸多挑战，这促使一系列创新技术与协同策略的产生。其中，联邦智能成为解决边缘智能数据隐私与协同训练难题的关键 [241]；多智能体协同则进一步提升了大规模边缘智能系统的协同能力 [245]；边缘智能系统则使得能力有限的边缘设备能够更高效地执行深度学习任务，为边缘智能的更广泛应用提供了基础 [249]。

通过联邦智能技术联合多个边缘设备实现数据收集、模型训练及推理，在保障用户的数据隐私安全的同时，充分利用边缘设备的资源。针对海量边缘设备产生的数据，为了利用海量数据训练边缘智能同时保护数据隐私，联邦智能受到了研究人员的重视：在多个边缘设备上分布式地收集数据、在本地处理后分布式地参与到模型训练中，充分利用边缘设备的资源，同时确保数据隐私的安全。近期的研究热点如下：（1）分布式数据收集。针对边缘设备产生的大量数据，如何从多个数据源节点收集数据用于训练是联邦智能的主要挑战之一。针对大量并发设备参与的数据收集过程，研究人员设计了协作式地图数据收集框架，支持大规模用户并行收集地图数据 [242]。（2）分布式训练。不同边缘设备差异化的计算速度，神经网络的分布式训练速度会存在差异，导致模型难以聚合。对此，研究人员通过分析异构设备上模型训练算法的收敛目标不一致的问题，设计了标准化的梯度平均方法，实现训练过程的快速收敛 [243]。（3）分布式推理。由于资源受限的边缘设备难以支持 Transformer 等深度学习模型较高的计算成本，来自加拿大的研究团队设计了面向通信和计算资源有限设备的分布式推理框架，在多个边缘设备上均衡负载，提高推理速度 [244]。

将分散的边缘设备视作多智能体进行协同部署与控制，从而优化智能系统的整体性能。现代应用场景往往涉及到复杂的任务，单一的边缘设备无法独立完成。以智能交通系统为例，要实现整个城市交通的优化，不仅需要路口摄像头监测交通流量，还需要车辆自身的智能感知以及可能的无人机辅助进行路况勘查等，需要多种设备协同工作才能达成诸如交通拥堵缓解、事故快速响应等复杂目标。多智能体协同技术将边缘智能环境中大量分散的边缘设备，视作多智能体进行协同部署和调度，从而实现全局最优。近期的研究热点如下：(1) 群体控制。针对每个智能体难以获得全局信息的局限性，研究团队设计了深度循环图神经网络，通过图卷积实现智能体之间的信息传递，从而在动态拓扑下控制多智能体 [246]。(2) 差异化调度。针对不同无人机设备的差异性，近期的研究提出了基于多智能体模仿学习的无人机部署方法，为地面用户提供差异化的通信服务，同时最大化运营者的效益 [247]。(3) 可扩展式协同。针对多智能体场景中智能体数量的动态增长，研究人员设计了可扩展的多智能体协作 SLAM 框架，通过服务器-客户端同步机制、优先级感知的任务调度器有效解决智能体数量不断增长所导致的数据爆炸问题 [248]。

针对异构边缘设备在内存、算力等方面的瓶颈，设计专门的边缘智能系统，实现智能算法在边缘设备上的高效部署与性能优化。受到边缘设备有限计算资源的制约，在边缘设备上部署深度学习等智能算法存在着以下挑战：内存有限，使得现有智能模型难以直接部署在端侧；设备异构，不同的边缘环境往往需要量身定制的模型架构，单一模型难以适应不同设备的异构环境；算力不足，边缘设备的 GPU 能力往往非常有限，仅能提供与 CPU 相同量级的算力服务。针对上述挑战，最近的工作设计了以下方法：(1) 内存受限下的高效推理。研究人员通过设计一种高效且自适应的内存管理框架，优化神经网络在边缘设备上推理任务的内容占用机制，从而适应内存有限的约束 [250]。(2) 异构设备中的自行应推理。研究人员设计了一种深度学习模型的弹性化方法，在边缘环境上自适应地进行神经网络结构优化，从而适应不同异构设备的环境 [251]。(3) 边缘设备上的推理加速。研究人员发现边缘设备的 CPU 和 GPU 具有相当的算力，但是 GPU 与 CPU 之间的数据共享开销会导致两者共同使用时的性能下降，对此提出了基于混合维度划分和计算链优化的数据共享方法，从而实现更高效的 GPU、CPU 协同计算 [252]。

4.2.4 安全性研究

工业互联网、智慧交通、智慧医疗、智慧政企等新兴产业的发展，带来了新的技术和安全问题。这些行业的数字化和智能化转型过程中，数据的集中化与开放性使得隐私保护和数据安全面临更大挑战。例如，智慧交通系统收集了大量的车辆和个人信息，如车辆行驶轨迹、车主身份信息、出行习惯等。这些数据的泄露可能会导致车主的隐私被侵犯，同时也可能被用于恶意商业目的或犯罪活动 [280]。同时，云上租户数量增加，将要求云服务商提供更安全的计算服务，确保不同租户之间的资源隔离，防止数据泄露或越权访问。还需要提升虚拟化平台的安全性，以防止虚拟机逃逸、恶意代码注入等攻击。此外，复杂的系统架构和多样化的终端设备增加了网络攻击的潜在入口，传统的安全防护手段难以完全覆盖。

不同行业的业务场景和需求各异，对网络安全技术提出了更加定制化和动态化的要求。因此，在推动新兴产业发展的同时，需要加大对技术标准的完善、云网安全技术的创新，传统的安全检测方式往往依赖于规则匹配和人工配置，但面对云计算环境中海量、复杂和动态的网络数据，这些方法的效率和准确性受到极大限制。人工智能技术，尤其是机器学习和深度学习的引入，为安全检测注入了智能化能力，使其能够高效、精准地识别潜在威胁，为构建更安全的数字化环境奠定坚实基础。

通过加密和访问控制技术对云端数据进行安全防护，防止敏感数据泄露或被恶意访问。随着数据量的增长，数据的关联性和多样性增加，使得匿名化后的信息更容易被重新识别。同时，对于云端存储的海量数据，使用传统加密算法进行加密可能导致存储系统的性能大幅下降。近年来国内外学者纷纷投入到云数据安全的研究中，提出了许多创新性的方法和技术，以应对日益复杂的安全威胁，研究热点如下：(1) 隐私保护。用户在自身数据的收集和使用上可选项很少，为了解决云-物联网应用程序中的数据隐私问题，有研究提出新的隐私保护架构和数据共享模型，该架构使用基于类别的数据访问模型，涵盖了从物联网设备的数据收集到云服务共享整个数据生命周期，可以使用户对其数据进行细粒度的控制 [253]。(2) 密文检索。敏感数据被上传到云服务器进行存储之前往往需要加密，在海量的数据中，如何对加密数据进行

有效和准确地检索是一个难题。有研究提出一种基于同态加密的安全高效的相似性检索方案，设计了双云服务器协同检索模型和消息认证方案，能够保证访问模式的隐私安全和传输数据的完整性。此外，有研究提出一种改进方案，使用 Simhash 算法生成查询和特征向量，减少存储开销 [254]。(3) 访问控制。多个用户通过云边缘共享数据时面临着数据隐私泄露的问题，因此，不能在发送者和接收者之间任意共享。针对这个问题，面向电子医疗云上多个医疗机构之间共享电子病历的场景，有研究提出一种跨域内积方案，可以防止单个发送者机构发送密文，同时保护数据和接收者的隐私 [255]。

探索容器等云资源隔离技术存在的脆弱点，并通过可信执行环境构筑安全云。云服务资源隔离是构建安全可信云环境的关键，随着技术的进步和需求的变化，原来的虚拟化隔离策略需要动态调整，否则，容易被攻击者通过漏洞进行恶意攻击。为了应对这一挑战，研究者们从攻防两个角度出发，一方面，针对现有的虚拟化隔离技术，通过探索新的攻击手段来发现其脆弱点。另一方面引入更先进的隔离机制，例如基于硬件的可信执行环境（TEE），通过硬件层面提供强制隔离和保护，从根本上防止虚拟机逃逸和侧信道攻击。研究热点如下：(1) 容器隔离。在公有云中，将容器放置在轻量级的虚拟机内运行可以充分利用虚拟机的安全性和容器的高效性。研究人员发现了一种可用于破坏基于轻量级虚拟机容器隔离的新攻击，称为操作转发攻击。攻击者可以通过操作转发来利用主机内核的漏洞并耗尽主机资源。在多个公有云上的实验结果显示，这种攻击可以降低受害容器的 IO 和 CPU 性能，甚至导致主机崩溃 [256]。(2) 机密虚拟机。云平台上使用基于硬件的 TEE 提供机密虚拟机用于托管安全敏感代码和数据，这其中有一些不受信任的虚拟机管理程序控制着多个资源管理和配置任务。研究者提出一种新的攻击，向虚拟机管理程序中注入恶意的非定时中断来破坏机密虚拟机的机密性和完整性 [257]。(3) TEE。公有云平台利用 TEE 技术提供机密计算服务。然而，受 TEE 保护的应用程序仍会受到回滚或分叉攻击，这些攻击会导致应用程序的状态回滚到过时的版本或者分叉为多个版本。针对这个问题，有研究提出了一个安全实用的分布式 TEE 系统，该系统使用区块链进行初始化，以较小的交互开销奠定去中心化信任基础，同时利用分布式系统的高性能状态提供连续性保护 [258]。

分析新型网络攻击技术，发展更加鲁棒、全面、灵活的安全防御技术。随着新技术的发展，云服务在网络层面的攻击和防御进入了更加复杂和动态的博弈阶段。攻击者利用伴随新技术产生的漏洞发起更具针对性的攻击，使得拒绝服务（DOS）攻击、域名服务器（DNS）攻击等针对云环境的攻击方式变得更加隐蔽和高效。近年来，针对云服务的网络攻击和防御研究热点如下：(1) DOS。有学者提出了一种新的攻击，可以导致 Serverless 计算平台和外部内容服务器中间拒绝服务。Serverless 计算平台在所有属于不同用户的 Serverless 功能之间共享同一组出口 IP，以访问外部内容服务器。因此，该平台上的恶意用户可以故意行为不当，导致这些出口 IP 被内容服务器阻止，从而导致整个平台的拒绝服务 [259]。(2) DNS。域名服务器的安全性和稳健性对于互联网的总体运行至关重要，因此域名所有者需要部署多个候选域名服务器以实现流量负载平衡，一旦负载平衡机制被破坏，攻击者就可以操纵大量合法 DNS 请求到指定的候选 DNS。有研究团队报告了一类 DNS 漏洞，并提出了一种新的攻击，允许攻击者以低成本秘密破坏权威域名服务器的 DNS 负载均衡 [260]。(3) 网络侧信道。网络侧信道通过数据包时间和大小泄露机密，这使得公有云的 IaaS 层中任何租户都能够间接观察到受害者的流量形状。最近，研究人员提出了一个端到端消除公有云 IaaS 网络侧信道泄露的系统，该系统使流量形状在设计上与机密解耦 [261]。

探索 AI 在云安全防御中的应用，构建自动化响应与防御模型，以实现更精准和高效的防御。随着人工智能技术的发展，其强大的数据处理能力、智能化的学习和适应能力、对复杂模式的理解以及动态环境中的实时响应能力，都将为云网安全检测注入了前所未有的能力。因此，使用人工智能技术辅助云网安全检测的研究越来越多，研究热点如下：(1) 图算法辅助云网态势感知。从入侵者的视角可以对网络环境和漏洞信息进行攻击图建模，根据攻击图类型的不同，顶点可以表示主机、服务、漏洞、权限等网络安全相关要素，也可以表示账户被攻击者破解、权限被攻击者获取等网络安全状态，边用于表示攻击者攻击行为的先后顺序。在入侵告警关联方面，一种基于队列图的攻击图匹配结构框架被提出 [262]，该结构采用广度优先搜索算法来遍历告警节点之间的关联，查找攻击路径，为告警数据的实时处理提供了理论基础。在风险评估和网络加固方面，需要着重分析漏洞之间被利用次数及依赖关系，通过图节点

重要性排序算法辨别需要优先修复的脆弱性节点 [263]。在攻击路径分析方面，有学者提出基于图神经网络的可迁移模型 SPGNN-API，通过识别最短路径来检测网络攻击路径，并主动调整网络防火墙规则和零信任策略来切断关键攻击路径。（2）LLMs 辅助安全。LLMs 通过预训练于大规模多模态数据（文本、图像、视频、代码等），具备处理多语言和多媒体内容的能力，可以理解多种数据形式并进行解释和总结，在有害内容检测、网络钓鱼防护、安全日志解释以及内容审核和优先级管理方面展现出重要作用。例如，ChatSpamDetector [264] 和 Phishpedia [265] 都是研究者提出的基于 LLMs 检测网络钓鱼电子邮件的系统，结合文本理解和图像分析可显著提升检测效果。HuntGPT [266] 被提出用于做可解释的入侵检测，旨在以易于解释的格式提供检测到的威胁，强调用户理解并提供流畅的交互体验。

4.2.5 可靠性研究

随着新兴行业对云计算服务质量要求的不断提高，对可靠性提出了诸多新要求与挑战。一方面，云计算环境的复杂性与动态性大幅增加，多种新兴技术的集成使得系统架构更为庞大，硬件、软件及网络组件之间的交互错综复杂，任何一个环节出现故障都可能引发连锁反应，影响整体可靠性 [281]。另一方面，数据量呈爆炸式增长且对数据安全性与完整性要求极高，一旦发生数据丢失或损坏，后果不堪设想。此外，新兴技术应用场景的多样性也要求云计算能够快速适应不同业务需求的可靠性变化，在高并发、低延迟等特殊要求下仍能稳定运行，这无疑给云计算的可靠性保障带来了前所未有的压力 [282]。最近，这些问题得到了国内外研究人员的重视，一系列关键技术应运而生。在故障/灾难隔离方面，着力于构建全栈隔离机制与灾难恢复策略，保障系统安全稳定；通过可观测性和可预测性分析，聚焦于构建指标体系与预测模型，提前把握系统趋势；智能运维则借助人工智能与机器学习，实现自动化运维与精准故障诊断，从多维度为云计算的可靠性筑牢根基，推动云计算在新兴技术浪潮下持续稳健发展。

有效的故障/灾难隔离可防止故障在系统中扩散，保障服务的连续性和数据的安全性。故障 / 灾难可能源于硬件漏洞、软件缺陷或外部攻击等，一旦发生，若缺乏有效隔离机制，将导致服务中断、数据泄露等严重后果。实现完全可靠的隔离面临诸多挑战，如硬件和软件的复杂性、不断演变的攻击手段以及性能与安全性之间的平衡等，以上挑战受到国内外众多研究人员的关注。近期主要研究热点如下：（1）安全隔离架构。GPU 云已成为一种流行计算平台，众多云架构也被提出，通过强制实施强安全策略隔离不可信的虚拟机管理程序与客户虚拟机（VM），以确保客户应用程序的安全执行环境。因此有研究人员针对攻击目标选择和故障注入精度在时间和位置上面临的主要挑战，提出了敏感目标搜索算法和遗传故障注入参数搜索算法。揭示了共享 GPU 云安全隔离中存在的潜在严重威胁，并提出了一种硬件不可信的安全隔离架构作为应对措施 [267]。（2）灾难恢复层。为了实现业务连续性，有研究人员提出一种灾难恢复层（Disaster Recovery Layer，DRL），其基于自主组件和 OpenStack 模块扩展，具备可扩展架构，通过分布式灾难检测机制识别灾难并告警，利用 BGP 任播在两个测试数据中心间重定向流量，经实验验证该技术能有效保护虚拟机和存储卷，在故障时服务中断小且开销低，为保障数据中心业务连续性提供了有效方案 [268]。（3）容错虚拟网络功能 (VNFs) 放置。网络功能虚拟化（NFV）使网络功能从专用硬件转移到 VM 中的软件实现，增强了灵活性与经济性。然而，VNFs 易受多种故障影响，如软件配置错误、VM 故障和软件故障等，确保其故障容错和可靠性成为关键挑战。工作 [269] 针对有限资源下的故障容错 VNF 放置问题提出基于联合资源可用性的启发式算法；针对无带宽或计算资源约束的问题，分别提出具有特定近似比的近似算法，确保在性能和资源利用间取得平衡。

可观测性和可预测性分析对微服务应用至关重要，前者有助于快速定位和解决故障，后者关乎服务质量、资源管理和风险应对。然而目前实现微服务的可观测性依赖特定工具且成本相关、缺乏系统方法支持；提高可预测性可能会影响资源利用率，需要在多目标间权衡。针对以上问题，近期主要研究热点工作如下：（1）软件设计权衡。研究人员提出一种系统的方法，以达到持续可评估的可观测性设计决策。重点关注云原生微服务应用的故障可观测性，并将其转化为可测试和可量化的属性，使用一个流行的开源微服务应用程序演示了所提出的方法，并展示了不同可观测性设计决策所涉及的权衡 [270]。（2）多云观测性。智慧城市的高效运作依赖于区域间的沟通与信息共享，尤其在交通管理、应急响应等方面。因此，智慧城市需要云原生技术来整合和分析交通传感器和公共交通车辆等多源数据。为此，研究人员提

出一种多云观测性方法来聚合不同地区的数据。该解决方案旨在提供一套完整的可观测性套件，能够跨多云的层收集数据并集成现有的开源项目 [271]。（3）负载平衡策略。在现代分布式计算环境中，计算能力的提供是一个非常重要的问题。云计算虽为开发者提供看似无限的能力，但数据中心的复杂供应和容量短缺风险仍需通过优雅降级技术来应对，这涉及牺牲用户体验以换取可预测性，并可能影响基础设施级别的其他决策，如负载平衡。一种负载平衡策略被提出 [272]，在必要时处理容量不足和平稳降级。该方法基于一种可靠的控制理论方法，并证明了其在应急管理和用户体验方面能实现更高的性能。

通过智能故障诊断与资源优化，有力保障系统的稳定高效运行，提升运维效率与系统可靠性、可用性。随着 IT 系统规模和复杂度剧增，传统运维方式难以满足需求，智能运维借助人工智能和大数据分析技术，从海量数据中挖掘模式，通过实时监测系统状态，实现精准异常检测和故障定位，被广泛应用于云服务、数据中心管理、网络运维等领域 [273, 274]。近年来，该领域涌现出诸多研究热点：（1）异常检测。随着移动感知技术的普及，大量时间序列数据在各领域产生，推动了众多实际应用的发展。为解决分散数据与集中算法间的差距，并应对隐私问题，研究者提出了联邦异常检测框架 PeFAD [283]，首次将预训练语言模型（PLM）作为客户端模型核心，通过高效联邦训练模块减少通信成本，仅需微调少量参数。同时采用异常驱动掩码选择策略和知识蒸馏，解决了数据异构性问题。（2）根因分析。为保证云服务的可靠性和可用性，需要对云事件进行高效的根因分析（RCA），但传统的 RCA 方法依赖于对日志和轨迹等数据源的人工调查，费时费力易出错，具有挑战性 [275]。最近，有学者提出了一种自动化的端到端云事件根因分析解决方案 [284]，该方案集成了一个大型语言模型。该模型能够根据警报类型将传入事件匹配到相应的处理程序，聚合关键的运行时诊断信息，预测事件的根本原因类别，并提供解释性说明。这一创新展示了大型语言模型在根因分析领域的潜力。（3）告警优化。针对当前网络系统中由于检测方法产生的大量错误报警现象，有学者提出一种无监督方法 pVoxel，将与告警关联的流量特征向量视为流量特征空间中的一个点，利用点云分析捕获点之间的拓扑特征来对告警进行分类。可为现有的基于机器学习的检测系统识别误报，而不需要任何关于警报的先验知识 [285]。

4.3 面向新兴技术的展望和发展建议

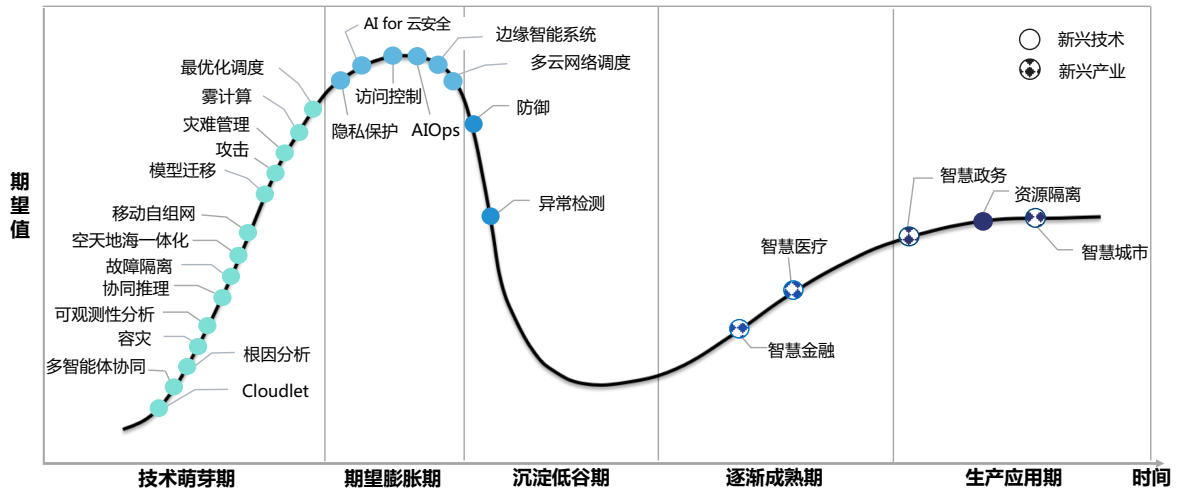


图 4.2: 新兴技术研究图谱技术成熟度曲线 2024

本节同样构建了新兴领域关键技术的 Gartner 成熟度曲线，如图4.2所示。通过此分析框架，可以清晰地识别出不同新兴关键技术所处的发展阶段，为新兴产业的研究与投资提供决策依据。展望未来，云计算在协同性、移动性、智能性、安全性、可靠性等方面正面临着前所未有的发展机遇。对新兴技术的展望与发展建议如下：

4.3.1 新兴技术的未来研究方向和关键技术展望

随着边缘终端计算能力的快速发展，将智能算法部署在边缘从而为用户提供低延迟、高可靠的智能服务成为可能。边缘智能有望在未来广泛应用于工业、交通、医疗等各个重要新兴产业中，在靠近数据源或用户的边缘设备上执行复杂的智能任务，如实时图像识别、数据分析与决策等，有效缓解了云计算中心的压力，并减少了数据传输过程中的带宽消耗和隐私泄露风险。而云边协同则进一步拓展了这种模式的潜力。通过云计算与边缘计算的紧密协作，二者优势互补。云计算凭借其强大的计算资源和海量存储，可为边缘设备提供模型更新、数据备份以及复杂任务的二次处理。边缘设备则利用其本地快速响应能力，对实时性要求高的任务进行即时处理，并将关键数据上传至云端进行深度分析与长期存储。在未来，边缘智能与云边协同将在更多领域展现其价值，如智能交通系统中实现车辆的实时精准导航与交通流量优化；工业生产里达成设备的智能故障诊断与生产流程的高效管控；医疗领域助力远程医疗设备的精准诊断和治疗决策，为各行业的智能化升级提供关键支撑，推动整个社会更加智能、高效的发展。

随着 5G、6G、以及卫星通信等新兴通信技术的发展，移动计算进一步与云计算深度融合，提供广域覆盖的低时延高可靠云服务。特别伴随着近年来低轨卫星技术的快速发展，更多功能强大、轨道布局合理的通信卫星被发射升空，极大拓展了云计算服务的覆盖范围，实现对全球各个角落的无缝覆盖，将低空、海洋、荒漠等区域纳入云服务体系，让空天地海的信息流通顺畅无阻。而近年来无人机组网技术的日益成熟，则使得移动自组网等应急通信技术能够在灾害等极端情况下，提供不依赖基础设施的应急云服务。在云计算基础设施方面，随着各大云厂商也在全球广泛建设移动边缘云节点，雾计算与 Cloudlet 也不断演进，实现多边缘节点的联合部署，更智能地分配任务、更高效地管理资源，真正实现低延迟、高性价比的服务，为物联网时代的海量数据处理和实时响应需求提供坚实保障，推动移动计算与云计算在智能交通、智能家居、工业自动化等领域的深度融合与创新应用，让人们的生活更加便捷、高效和智能。

随着云服务环境的复杂性增加，零信任架构将成为未来安全体系的核心。随着新技术的发展，零信任安全架构正逐步从理论走向实践，并在深化和扩展中不断演进。未来，零信任的研究将围绕动态信任评估与实时授权、微分段策略优化以及无边界安全实践展开，通过行为分析和多因素信任指标动态调整权限，实现从单点授权到全生命周期持续验证的转变。同时，零信任架构将在多云和跨租户环境中发挥重要作用，重点解决跨云身份与访问管理、统一策略编排以及威胁情报共享问题。结合人工智能，零信任将进一步提升威胁检测的智能化水平，通过预测性分析和自适应响应机制增强安全性。在场景化应用方面，不同行业将根据自身需求对零信任架构进行定制化设计，如智慧医疗中的数据隐私保护、工业互联网中的设备安全管理，以及金融行业中的实时交易验证。此外，研究还将探索标准化与互操作性、自动化部署和可视化管理，以降低架构复杂性并提升可操作性。尽管零信任的实施面临高复杂性和性能开销等挑战，但其通过动态化、智能化和多层次的安全策略，能够有效应对云计算环境中的复杂威胁，逐步成为构建安全可信网络体系的核心支柱。

边缘计算与分布式安全研究将为云服务的安全性、可靠性以及高效性注入新的内涵和保障机制。首先，分布式安全架构将得到优化，研究将探索如何在异构环境中实现统一、安全策略和动态调整，提升威胁检测与响应效率。其次，轻量化的安全协议与加密技术将成为重点，研究将专注于设计高效的安全协议以适应边缘设备的资源限制。零信任架构将在边缘计算中得到广泛应用。此外，AI 和机器学习驱动的威胁检测与防御机制将在边缘计算中得到应用，提升实时分析和响应能力。可信计算技术将确保边缘设备的安全性，防止恶意篡改和数据泄露，而跨域安全协作与多方数据保护也将成为未来研究的重点，确保数据在不同平台间的安全传输与隐私保护。最后，边缘计算的标准化将推动跨平台、跨厂商的安全互操作性，提升分布式防御体系的协同能力。

随着新兴产业的蓬勃发展, Serverless 计算架构、多云架构和边缘计算架构等新兴架构的研究显得至关重要。在这些新兴产业中，Serverless 计算的函数级可靠性保障不容忽视。例如在工业生产的实时控制场景以及金融高频交易场景下，函数冷启动的高效性和并发执行的正确性直接影响着业务的连续性和准确性，而服务集成的优劣则关系到整个业务生态的协同性。对于多云架构，在智能交通的多区域数据融

合以及医疗的跨机构数据共享等应用中，跨云服务的无缝切换、数据一致性保障以及故障隔离恢复能力，是确保系统稳定运行的关键。同时，AIOps 的融入为这些新兴架构的可靠性保障带来了新的机遇和挑战。借助 AIOps，能够对海量的运维数据进行智能分析，提前预测潜在的可靠性风险，例如在教育在线学习平台的服务器负载预测、金融交易系统的网络故障预警等方面。然而，AIOps 也面临着数据质量参差不齐、模型可解释性不足以及与现有架构融合的复杂性等问题。总之，攻克这些难题将有力推动新兴云计算架构在各产业的应用，为其数字化变革筑牢根基，提升用户信任，繁荣产业生态。

4.3.2 新兴技术的发展建议

关注无人区等特殊环境下的服务难题，拓展传统云计算的服务范围。针对低空经济等新兴业务需求的涌现，云服务商应当拓展云计算服务的范围，在低空、海洋、荒漠等无人区提供高可靠的云服务，助力新兴业务的快速发展。通过加强空天地海一体化的基础设施建设，云服务商可以借助卫星侧基础设施，在缺乏地面基础设施的地区构建一体化的全空间立体服务区域，提供全域无缝覆盖的服务。此外，通过将无人机、船舶等移动设备作为网络节点构建移动自组网，能够在无基础设施的极端复杂环境下实现高可靠的云服务。随着这些特殊环境下云计算服务的逐步完善，云服务商不仅能够开拓全新的业务蓝海，更将推动相关新兴产业实现跨越性发展，为全球经济发展注入新的活力。

深度挖掘边缘设备的计算潜力，通过智能化技术推动云边协同发展。一方面，云服务商应当优化云边资源的协同调度，设计更加灵活的云边协同调度方法，针对不同场景的计算任务和多样化的计算资源实现灵活的动态调度，降低系统的整体延迟并提升计算效率。考虑到优化目标繁多带来的高复杂性，强化学习等 AI 技术有望成为破局的关键利器，通过构建智能模型，学习不同任务的优先级、资源需求以及网络状况的实时波动，精准实现资源配置，实现高效协同。另一方面，随着边缘设备能力的逐渐提升，云服务商应当重视智能化算法在靠近数据源的边缘设备上的部署，提供低延迟、高可靠的服务。考虑到边缘设备的资源局限性，模型压缩、迁移学习等方案有望快速实现轻量级的边缘智能模型，而联邦智能、多智能体协同等技术有望进一步强化多个边缘设备之间的协同性，实现边缘设备集群效能的最大化释放。

增强基础安全能力，提供针对新兴技术的定制化安全解决方案。一方面，云服务商需要强化基础安全能力，构建端到端的数据加密和零信任架构，覆盖数据的传输、存储和处理环节。要加强数据隐私合规，针对不同国家和地区的法规要求（如欧盟的 GDPR、中国的《数据安全法》），云服务商应提供多样化的合规工具和服务，帮助客户高效满足本地及国际法律要求。另一方面，随着攻击方式的不断演变，云服务商需持续更新安全技术。针对云边协同、边缘智能等新兴技术领域的安全需求提供定制化的安全解决方案，重点涵盖数据传输、边缘节点和设备管理的全方位保护。通过加密通信协议和轻量化的端到端加密技术，确保云边数据传输的安全性；在边缘节点部署零信任架构和安全网关，防范本地攻击和未授权访问；引入基于 AI 的实时威胁检测系统，动态识别边缘设备的异常行为。

提升智能化运维和管理水平，增强云基础设施可靠性。一方面，云服务商要引入机器学习与大数据分析等技术构建智能运维工具，全方位、实时监控系统性能指标，如 CPU 使用率、内存占用等，及时发现故障并精准预测潜在问题，通过深度挖掘历史故障与运行数据，建立预测模型，提前防范硬件故障、软件漏洞及网络拥塞等状况，保障系统稳定，降低停机损失。另一方面，全力发展智能化管理系统，依据实时业务负载智能调配计算、存储与网络资源，如电商平台促销时自动增加资源，活动后回收，避免资源浪费，帮助客户降本增效。再者，着重强化运维可视化能力，打造直观工具，以图表形式呈现系统运行与资源使用详情，让客户对服务器负载、存储进度、网络流量等一目了然，快速决策，增强信任，为新兴技术在云服务领域的发展筑牢根基，满足业务与市场需求。

参考文献

[1] Gartner. Gartner Says Worldwide IaaS Public Cloud Services Revenue Grew 16.2% in 2023. https://www.gartner.com/en/newsroom/press-releases/, 2024.

[2] 国际数据公司 (IDC). 2023 年全球公共云服务收入.

[3] 国际数据公司 (IDC). 中国公有云服务市场 (2023 上半年) 跟踪.

[4] 国际数据公司 (IDC). 中国公有云服务市场 (2023 下半年) 跟踪.

[5] Zhenkun Yang, Chuanhui Yang, Fusheng Han, Mingqiang Zhuang, Bing Yang, Zhifeng Yang, Xiaojun Cheng, Yuzhong Zhao, Wenhui Shi, Huafeng Xi, et al. Oceanbase: a 707 million tpmc distributed relational database system. *Proceedings of the VLDB Endowment*, 15(12):3385–3397, 2022.

[6] Wei Cao, Feifei Li, Gui Huang, Jianghang Lou, Jianwei Zhao, Dengcheng He, Mengshi Sun, Yingqiang Zhang, Sheng Wang, Xueqiang Wu, et al. Polardb-x: An elastic distributed relational database for cloud-native applications. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 2859–2872. IEEE, 2022.

[7] Amazon. Publications - Amazon Science. https://www.amazon.science/publications, 2022.

[8] Wei Cao, Yingqiang Zhang, Xinjun Yang, Feifei Li, Sheng Wang, Qingda Hu, Xuntao Cheng, Zongzhi Chen, Zhenjun Liu, Jing Fang, et al. Polardb serverless: A cloud native database for disaggregated data centers. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2477–2489, 2021.

[9] Alex Depoutovitch, Chong Chen, Per-Ake Larson, Jack Ng, Shu Lin, Guanzhu Xiong, Paul Lee, Emad Boctor, Samiao Ren, Lengdong Wu, et al. Taurus mm: Bringing multi-master to the cloud. *Proceedings of the VLDB Endowment*, 16(12):3488–3500, 2023.

[10] Xinjun Yang, Yingqiang Zhang, Hao Chen, Feifei Li, Bo Wang, Jing Fang, Chuan Sun, and Yuhui Wang. Polardb-mp: A multi-primary cloud-native database via disaggregated shared memory. In *Companion of the 2024 International Conference on Management of Data*, pages 295–308, 2024.

[11] Guoliang Li, Xuanhe Zhou, Ji Sun, Xiang Yu, Yue Han, Lianyuan Jin, Wenbo Li, Tianqing Wang, and Shifu Li. opengauss: An autonomous database system. *Proceedings of the VLDB Endowment*, 14(12):3028–3042, 2021.

[12] Rentong Guo, Xiaofan Luan, Long Xiang, Xiao Yan, Xiaomeng Yi, Jigao Luo, Qianya Cheng, Weizhi Xu, Jiarui Luo, Frank Liu, et al. Manu: a cloud native vector database management system. *arXiv preprint arXiv:2206.13843*, 2022.

[13] Tzu-Wei Yang, Seth Pollen, Mustafa Uysal, Arif Merchant, and Homer Wolfmeister. Cachesack: Admission optimization for google datacenter flash caches. In Jiri Schindler and Noa Zilberman, editors, *Proceedings of the 2022 USENIX Annual Technical Conference, USENIX ATC 2022, Carlsbad, CA, USA, July 11-13, 2022*, pages 1021–1036. USENIX Association, 2022.

[14] Shushu Yi, Shaocong Sun, Li Peng, Yingbo Sun, Ming-Chang Yang, Zhichao Cao, Qiao Li, Myoungsoo Jung, Ke Zhou, and Jie Zhang. Biza: Design of self-governing block-interface zns afa for endurance and

performance. In *Proceedings of the ACM SIGOPS 30th Symposium on Operating Systems Principles*, pages 313–329, 2024.

[15] Su Zhou, Erci Xu, Hao Wu, Yu Du, Jiacheng Cui, Wanyu Fu, Chang Liu, Yingni Wang, Wenbo Wang, Shouqu Sun, et al. SMRSTORE: A storage engine for cloud object storage on HM-SMR drives. In *21st USENIX Conference on File and Storage Technologies (FAST 23)*, pages 395–408, 2023.

[16] Roei Kisous, Ariel Kolikant, Abhinav Duggal, Sarai Sheinvald, and Gala Yadgar. The what, the from, and the to: The migration games in deduplicated systems. *ACM Transactions on Storage*, 18(4):1–29, 2022.

[17] Saurabh Kadekodi, Shashwat Silas, David Clausen, and Arif Merchant. Practical design considerations for wide locally recoverable codes (lrcs). *ACM Transactions on Storage*, 19(4):1–26, 2023.

[18] Yiduo Wang, Yufei Wu, Cheng Li, Pengfei Zheng, Biao Cao, Yan Sun, Fei Zhou, Yinlong Xu, Yao Wang, and Guangjun Xie. Cfs: Scaling metadata service for distributed file system via pruned scope of critical sections. In *Proceedings of the Eighteenth European Conference on Computer Systems*, pages 331–346, 2023.

[19] Yiduo Wang, Cheng Li, Xinyang Shao, Youxu Chen, Feng Yan, and Yinlong Xu. Lunule: an agile and judicious metadata load balancer for CephFS. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC 21)*, pages 1–16. IEEE, 2021.

[20] Dong Du, Tianyi Yu, Yubin Xia, Binyu Zang, Guanglu Yan, Chenggang Qin, Qixuan Wu, and Haibo Chen. Catalyzer: Sub-millisecond startup for serverless computing with initialization-less booting. In James R. Larus, Luis Ceze, and Karin Strauss, editors, *ASPLOS '20: Architectural Support for Programming Languages and Operating Systems, Lausanne, Switzerland, March 16-20, 2020*, pages 467–481. ACM, 2020.

[21] Alexandru Agache, Marc Brooker, Alexandra Iordache, Anthony Liguori, Rolf Neugebauer, Phil Piwonka, and Diana-Maria Popa. Firecracker: Lightweight virtualization for serverless applications. In Ranjita Bhagwan and George Porter, editors, *17th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2020, Santa Clara, CA, USA, February 25-27, 2020*, pages 419–434. USENIX Association, 2020.

[22] Alessandro Randazzo and Ilenia Tinnirello. Kata containers: An emerging architecture for enabling MEC services in fast and secure way. In Mohammad A. Alsmirat and Yaser Jararweh, editors, *Sixth International Conference on Internet of Things: Systems, Management and Security, IOTSMS 2019, Granada, Spain, October 22-25, 2019*, pages 209–214. IEEE, 2019.

[23] Zijun Li, Jiagan Cheng, Quan Chen, Eryu Guan, Zizheng Bian, Yi Tao, Bin Zha, Qiang Wang, Weidong Han, and Minyi Guo. Rund: A lightweight secure container runtime for high-density deployment and high-concurrency startup in serverless computing. In Jiri Schindler and Noa Zilberman, editors, *Proceedings of the 2022 USENIX Annual Technical Conference, USENIX ATC 2022, Carlsbad, CA, USA, July 11-13, 2022*, pages 53–68. USENIX Association, 2022.

[24] Nikita Lazarev, Varun Gohil, James Tsai, Andy Anderson, Bhushan Chitlur, Zhiru Zhang, and Christina Delimitrou. Sabre: Hardware-accelerated snapshot compression for serverless microvms. In Ada Gavrilovska and Douglas B. Terry, editors, *18th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2024, Santa Clara, CA, USA, July 10-12, 2024*, pages 1–18. USENIX Association, 2024.

[25] Edward Oakes, Leon Yang, Dennis Zhou, Kevin Houck, Tyler Harter, Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau. SOCK: rapid task provisioning with serverless-optimized containers. In Haryadi S. Gunawi and Benjamin Reed, editors, *2018 USENIX Annual Technical Conference, USENIX ATC 2018, Boston, MA, USA, July 11-13, 2018*, pages 57–70. USENIX Association, 2018.

[26] Mohammad Shahrad, Rodrigo Fonseca, Iñigo Goiri, Gohar Chaudhry, Paul Batum, Jason Cooke, Eduardo Laureano, Colby Tresness, Mark Russinovich, and Ricardo Bianchini. Serverless in the wild: Characterizing and optimizing the serverless workload at a large cloud provider. In Ada Gavrilovska and Erez Zadok, editors, *2020 USENIX Annual Technical Conference, USENIX ATC 2020, July 15-17, 2020*, pages 205–218. USENIX Association, 2020.

[27] Yanan Yang, Laiping Zhao, Yiming Li, Shihao Wu, Yuechan Hao, Yuchi Ma, and Keqiu Li. Flame: A centralized cache controller for serverless computing. In Tor M. Aamodt, Michael M. Swift, and Natalie D. Enright Jerger, editors, *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 4, ASPLOS 2023, Vancouver, BC, Canada, March 25-29, 2023*, pages 153–168. ACM, 2023.

[28] Guowei Liu, Laiping Zhao, Yiming Li, Zhaolin Duan, Sheng Chen, Yitao Hu, Zhiyuan Su, and Wenyu Qu. FUYAO: dpu-enabled direct data transfer for serverless computing. In Rajiv Gupta, Nael B. Abu-Ghazaleh, Madan Musuvathi, and Dan Tsafrir, editors, *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3, ASPLOS 2024, La Jolla, CA, USA, 27 April 2024- 1 May 2024*, pages 431–447. ACM, 2024.

[29] Xingda Wei, Fangming Lu, Tianxia Wang, Jinyu Gu, Yuhan Yang, Rong Chen, and Haibo Chen. No provisioned concurrency: Fast rdma-codesigned remote fork for serverless computing. In Roxana Geambasu and Ed Nightingale, editors, *17th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2023, Boston, MA, USA, July 10-12, 2023*, pages 497–517. USENIX Association, 2023.

[30] Istemi Ekin Akkus, Ruichuan Chen, Ivica Rimac, Manuel Stein, Klaus Satzke, Andre Beck, Paarijaat Aditya, and Volker Hilt. SAND: towards high-performance serverless computing. In Haryadi S. Gunawi and Benjamin Reed, editors, *2018 USENIX Annual Technical Conference, USENIX ATC 2018, Boston, MA, USA, July 11-13, 2018*, pages 923–935. USENIX Association, 2018.

[31] Swaroop Kotni, Ajay Nayak, Vinod Ganapathy, and Arkaprava Basu. Faastlane: Accelerating function-as-a-service workflows. In Irina Calciu and Geoff Kuenning, editors, *2021 USENIX Annual Technical Conference, USENIX ATC 2021, July 14-16, 2021*, pages 805–820. USENIX Association, 2021.

[32] Zinuo Cai, Hao Wang, Tao Song, Yang Hua, Ruhui Ma, and Haibing Guan. Chrion: Optimizing recurrent neural network inference by collaboratively utilizing cpus and gpus. *CoRR*, abs/2307.11339, 2023.

[33] Chenxi Wang, Yifan Qiao, Haoran Ma, Shi Liu, Wenguang Chen, Ravi Netravali, Miryung Kim, and Guoqing Harry Xu. Canvas: Isolated and adaptive swapping for Multi-Applications on remote memory. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 161–179, Boston, MA, April 2023. USENIX Association.

[34] Huaicheng Li, Daniel S Berger, Lisa Hsu, Daniel Ernst, Pantea Zardoshti, Stanko Novakovic, Monish Shah, Samir Rajadnya, Scott Lee, Ishwar Agarwal, et al. Pond: Cxl-based memory pooling systems for cloud platforms. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 574–587, 2023.

[35] Padmapriya Duraisamy, Wei Xu, Scott Hare, Ravi Rajwar, David E. Culler, Zhiyi Xu, Jianing Fan, Christopher Kennelly, Bill McCloskey, Danijela Mijailovic, Brian Morris, Chiranjit Mukherjee, Jingliang Ren, Greg Thelen, Paul Turner, Carlos Villavieja, Parthasarathy Ranganathan, and Amin Vahdat. Towards an adaptable systems architecture for memory tiering at warehouse-scale. In Tor M. Aamodt, Natalie D. Enright Jerger, and Michael M. Swift, editors, *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3, ASPLOS 2023, Vancouver, BC,*

Canada, March 25-29, 2023, pages 727–741. ACM, 2023.

[36] Johannes Weiner, Niket Agarwal, Dan Schatzberg, Leon Yang, Hao Wang, Blaise Sanouillet, Bikash Sharma, Tejun Heo, Mayank Jain, Chunqiang Tang, and Dimitrios Skarlatos. TMO: transparent memory offloading in datacenters. In Babak Falsafi, Michael Ferdman, Shan Lu, and Thomas F. Wenisch, editors, *ASPLOS '22: 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Lausanne, Switzerland, 28 February 2022 - 4 March 2022*, pages 609–621. ACM, 2022.

[37] Reto Achermann, Ashish Panwar, Abhishek Bhattacharjee, Timothy Roscoe, and Jayneel Gandhi. Mitosis: Transparently self-replicating page-tables for large-memory machines. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '20, page 283–300, New York, NY, USA, 2020. Association for Computing Machinery.

[38] Hongliang Qu and Zhibin Yu. Wasp: Workload-aware self-replicating page-tables for numa servers. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS '24, page 1233–1249, New York, NY, USA, 2024. Association for Computing Machinery.

[39] Yunhong Xu, Keqiang He, Rui Wang, Minlan Yu, Nick Duffield, Hassan Wassel, Shidong Zhang, Leon Poutievski, Junlan Zhou, and Amin Vahdat. Hashing design in modern networks: Challenges and mitigation techniques. In *2022 USENIX Annual Technical Conference (USENIX ATC 22)*, pages 805–818, 2022.

[40] Xing Li, Xiaochong Jiang, Ye Yang, Lilong Chen, Yi Wang, Chao Wang, Chao Xu, Yilong Lv, Bowen Yang, Taotao Wu, et al. Triton: A flexible hardware offloading architecture for accelerating apsara vswitch in alibaba cloud. In *Proceedings of the ACM SIGCOMM 2024 Conference*, pages 750–763, 2024.

[41] Feixue Han, Qing Li, Peng Zhang, Gareth Tyson, Yong Jiang, Mingwei Xu, Yulong Lan, and ZhiCheng Li. ETC: An elastic transmission control using End-to-End available bandwidth perception. In *2024 USENIX Annual Technical Conference (USENIX ATC 24)*, pages 265–284, 2024.

[42] Shibo Wang, Shusen Yang, Xiao Kong, Chenglei Wu, Longwei Jiang, Chenren Xu, Cong Zhao, Xuesong Yang, Jianjun Xiao, Xin Liu, et al. Pudica: Toward Near-Zero queuing delay in congestion control for cloud gaming. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, pages 113–129, 2024.

[43] Jiaxing Zhang, Furong Yang, Ting Liu, Qinghua Wu, Wu Zhao, Yuanbo Zhang, Wentao Chen, Yanmei Liu, Hongyu Guo, Yunfei Ma, et al. TECC: Towards efficient QUIC tunneling via collaborative transmission control. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, pages 253–266, 2024.

[44] Abhishek Dhamija, Balasubramanian Madhavan, Hechao Li, Jie Meng, Shrikrishna Khare, Madhavi Rao, Lawrence Brakmo, Neil Spring, Prashanth Kannan, Srikanth Sundaresan, et al. A large-scale deployment of DCTCP. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, pages 239–252, 2024.

[45] Michael Armbrust, Tathagata Das, Liwen Sun, Burak Yavuz, Shixiong Zhu, Mukul Murthy, Joseph Torres, Herman van Hovell, Adrian Ionescu, Alicja Łuszczak, et al. Delta lake: high-performance acid table storage over cloud object stores. *Proceedings of the VLDB Endowment*, 13(12):3411–3424, 2020.

[46] Michael Armbrust, Ali Ghodsi, Reynold Xin, and Matei Zaharia. Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In *Proceedings of CIDR*, volume 8, page 28, 2021.

[47] Benjamin Reidys, Yuqi Xue, Daixuan Li, Bharat Sukhwani, Wen-Mei Hwu, Deming Chen, Sameh Asaad, and Jian Huang. Rackblox: A software-defined rack-scale storage system with network-storage co-design. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 182–199, 2023.

[48] Hongming Huang, Peng Wang, Qiang Su, Hong Xu, Chun Jason Xue, and André Brinkmann. Palantir: Hierarchical similarity detection for post-deduplication delta compression. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 830–845, 2024.

[49] Yingjin Qian, Wen Cheng, Lingfang Zeng, Marc-André Vef, Oleg Drokin, Andreas Dilger, Shuichi Ihara, Wusheng Zhang, Yang Wang, and André Brinkmann. Metawbc: Posix-compliant metadata write-back caching for distributed file systems. In *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–20. IEEE, 2022.

[50] Bin Gao, Zhuomin He, Puru Sharma, Qingxuan Kang, Djordje Jevdjic, Junbo Deng, Xingkun Yang, Zhou Yu, and Pengfei Zuo. Cost-Efficient large language model serving for multi-turn conversations with CachedAttention. In *2024 USENIX Annual Technical Conference (USENIX ATC 24)*, pages 111–126, 2024.

[51] Junyu Wei, Guangyan Zhang, Junchao Chen, Yang Wang, Weimin Zheng, Tingtao Sun, Jiesheng Wu, and Jiangwei Jiang. Loggrep: Fast and cheap cloud log storage by exploiting both static and runtime patterns. In Giuseppe Antonio Di Luna, Leonardo Querzoni, Alexandra Fedorova, and Dushyanth Narayanan, editors, *Proceedings of the Eighteenth European Conference on Computer Systems, EuroSys 2023, Rome, Italy, May 8-12, 2023*, pages 452–468. ACM, 2023.

[52] Yixin Wu, Xiuqi Huang, Wei Zhongjia, Hang Cheng, Chaohui Xin, Zuzhi Chen, Binbin Chen, Yufei Wu, Hao Wang, Tieying Zhang, Rui Shi, Xiaofeng Gao, Yuming Liang, Pengwei Zhao, and Guihai Chen. Towards resource efficiency: Practical insights into large-scale spark workloads at bytedance. *Proc. VLDB Endow.*, 17(12):3759–3771, 2024.

[53] Shuang Chen, Angela Jin, Christina Delimitrou, and José F. Martínez. Retail: Opting for learning simplicity to enable qos-aware power management in the cloud. In *IEEE International Symposium on High-Performance Computer Architecture, HPCA 2022, Seoul, South Korea, April 2-6, 2022*, pages 155–168. IEEE, 2022.

[54] Houxiang Ji, Mark Mansi, Yan Sun, Yifan Yuan, Jinghan Huang, Reese Kuper, Michael M. Swift, and Nam Sung Kim. STYX: exploiting smartnic capability to reduce datacenter memory tax. In Julia Lawall and Dan Williams, editors, *Proceedings of the 2023 USENIX Annual Technical Conference, USENIX ATC 2023, Boston, MA, USA, July 10-12, 2023*, pages 619–633. USENIX Association, 2023.

[55] Wei Zhang, Quan Chen, Kaihua Fu, Ningxin Zheng, Zhiyi Huang, Jingwen Leng, and Minyi Guo. Astraea: towards qos-aware and resource-efficient multi-stage GPU services. In Babak Falsafi, Michael Ferdman, Shan Lu, and Thomas F. Wenisch, editors, *ASPLOS '22: 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Lausanne, Switzerland, 28 February 2022 - 4 March 2022*, pages 570–582. ACM, 2022.

[56] Yongkang Zhang, Yinghao Yu, Wei Wang, Qiukai Chen, Jie Wu, Zuowei Zhang, Jiang Zhong, Tianchen Ding, Qizhen Weng, Lingyun Yang, Cheng Wang, Jian He, Guodong Yang, and Liping Zhang. Workload consolidation in alibaba clusters: the good, the bad, and the ugly. In Ada Gavrilovska, Deniz Altinbüken, and Carsten Binnig, editors, *Proceedings of the 13th Symposium on Cloud Computing, SoCC 2022, San Francisco, California, November 7-11, 2022*, pages 210–225. ACM, 2022.

[57] Vighnesh Sachidananda and Anirudh Sivaraman. Erlang: Application-aware autoscaling for cloud mi-

croservices. In *Proceedings of the Nineteenth European Conference on Computer Systems, EuroSys 2024, Athens, Greece, April 22-25, 2024*, pages 888–923. ACM, 2024.

[58] Hasan Al Maruf, Hao Wang, Abhishek Dhanotia, Johannes Weiner, Niket Agarwal, Pallab Bhattacharya, Chris Petersen, Mosharaf Chowdhury, Shobhit O. Kanaujia, and Prakash Chauhan. TPP: transparent page placement for cxl-enabled tiered-memory. In Tor M. Aamodt, Natalie D. Enright Jerger, and Michael M. Swift, editors, *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3, ASPLOS 2023, Vancouver, BC, Canada, March 25-29, 2023*, pages 742–755. ACM, 2023.

[59] Teng Ma, Zheng Liu, Chengkun Wei, Jialiang Huang, Youwei Zhuo, Haoyu Li, Ning Zhang, Yijin Guan, Dimin Niu, Mingxing Zhang, and Tao Ma. Hydrarpc: RPC in the CXL era. In Saurabh Bagchi and Yiying Zhang, editors, *Proceedings of the 2024 USENIX Annual Technical Conference, USENIX ATC 2024, Santa Clara, CA, USA, July 10-12, 2024*, pages 387–395. USENIX Association, 2024.

[60] Wenda Tang, Ying Han, Tianxiang Ai, Guanghui Li, Bin Yu, and Xin Yang. Yggdrasil: Reducing network i/o tax with (CXL-Based) distributed shared memory. In *Proceedings of the 53rd International Conference on Parallel Processing*, ICPP '24, page 597–606, New York, NY, USA, 2024. Association for Computing Machinery.

[61] Biao Sun, Ziming Huang, Hanyu Zhao, Wencong Xiao, Xinyi Zhang, Yong Li, and Wei Lin. Llumnix: Dynamic scheduling for large language model serving. In Ada Gavrilovska and Douglas B. Terry, editors, *18th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2024, Santa Clara, CA, USA, July 10-12, 2024*, pages 173–191. USENIX Association, 2024.

[62] Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, Alexey Tumanov, and Ramachandran Ramjee. Taming throughput-latency tradeoff in LLM inference with sarathi-serve. In Ada Gavrilovska and Douglas B. Terry, editors, *18th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2024, Santa Clara, CA, USA, July 10-12, 2024*, pages 117–134. USENIX Association, 2024.

[63] Yuhong Zhong, Daniel S. Berger, Carl A. Waldspurger, Ryan Wee, Ishwar Agarwal, Rajat Agarwal, Frank Hady, Karthik Kumar, Mark D. Hill, Mosharaf Chowdhury, and Asaf Cidon. Managing memory tiers with CXL in virtualized environments. In Ada Gavrilovska and Douglas B. Terry, editors, *18th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2024, Santa Clara, CA, USA, July 10-12, 2024*, pages 37–56. USENIX Association, 2024.

[64] Pengfei Li, Yu Hua, Pengfei Zuo, Zhangyu Chen, and Jiajie Sheng. ROLEX: A scalable RDMA-oriented learned Key-Value store for Disaggregated Memory Systems. In *21st USENIX Conference on File and Storage Technologies (FAST 23)*, pages 99–114, Santa Clara, CA, February 2023. USENIX Association.

[65] Jiacheng Shen, Pengfei Zuo, Xuchuan Luo, Tianyi Yang, Yuxin Su, Yangfan Zhou, and Michael R. Lyu. FUSEE: A fully memory-disaggregated key-value store. In Ashvin Goel and Dalit Naor, editors, *21st USENIX Conference on File and Storage Technologies, FAST 2023, Santa Clara, CA, USA, February 21-23, 2023*, pages 81–98. USENIX Association, 2023.

[66] Yang Zhou, Hassan M. G. Wassel, Sihang Liu, Jiaqi Gao, James Mickens, Minlan Yu, Chris Kennelly, Paul Turner, David E. Culler, Henry M. Levy, and Amin Vahdat. Carbink: Fault-tolerant far memory. In Marcos K. Aguilera and Hakim Weatherspoon, editors, *16th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2022, Carlsbad, CA, USA, July 11-13, 2022*, pages 55–71. USENIX Association, 2022.

[67] Mingxing Zhang, Teng Ma, Jinqi Hua, Zheng Liu, Kang Chen, Ning Ding, Fan Du, Jinlei Jiang, Tao Ma, and Yongwei Wu. Partial failure resilient memory management system for (cxl-based) distributed shared memory. In Jason Flinn, Margo I. Seltzer, Peter Druschel, Antoine Kaufmann, and Jonathan Mace, editors, *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 658–674. ACM, 2023.

[68] Cunchen Hu, Chenxi Wang, Sa Wang, Ninghui Sun, Yungang Bao, Jieru Zhao, Sanidhya Kashyap, Pengfei Zuo, Xusheng Chen, Liangliang Xu, Qin Zhang, Hao Feng, and Yizhou Shan. Skadi: Building a distributed runtime for data systems in disaggregated data centers. In *Proceedings of the 19th Workshop on Hot Topics in Operating Systems*, HOTOS '23, page 94–102, New York, NY, USA, 2023. Association for Computing Machinery.

[69] Huangshi Tian, Suyi Li, Ao Wang, Wei Wang, Tianlong Wu, and Haoran Yang. Owl: performance-aware scheduling for resource-efficient function-as-a-service cloud. In Ada Gavrilovska, Deniz Altinbüken, and Carsten Binnig, editors, *Proceedings of the 13th Symposium on Cloud Computing, SoCC 2022, San Francisco, California, November 7-11, 2022*, pages 78–93. ACM, 2022.

[70] Marc Brooker, Mike Danilov, Chris Greenwood, and Phil Piwonka. On-demand container loading in AWS lambda. In Julia Lawall and Dan Williams, editors, *Proceedings of the 2023 USENIX Annual Technical Conference, USENIX ATC 2023, Boston, MA, USA, July 10-12, 2023*, pages 315–328. USENIX Association, 2023.

[71] Jingyuan Zhang, Ao Wang, Xiaolong Ma, Benjamin Carver, Nicholas John Newman, Ali Anwar, Lukas Rupprecht, Vasily Tarasov, Dimitrios Skourtis, Feng Yan, and Yue Cheng. Infinistore: Elastic serverless cloud storage. *Proc. VLDB Endow.*, 16(7):1629–1642, 2023.

[72] Alireza Sahraei, Soteris Demetriou, Amirali Sobhgol, Haoran Zhang, Abhigna Nagaraja, Neeraj Pathak, Girish Joshi, Carla Souza, Bo Huang, Wyatt Cook, Andrii Golovei, Pradeep Venkat, Andrew Mcfague, Dimitrios Skarlatos, Vipul Patel, Ravinder Thind, Ernesto Gonzalez, Yun Jin, and Chunqiang Tang. Xfaas: Hyperscale and low cost serverless functions at meta. In Jason Flinn, Margo I. Seltzer, Peter Druschel, Antoine Kaufmann, and Jonathan Mace, editors, *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 231–246. ACM, 2023.

[73] Qiangyu Pei, Yongjie Yuan, Haichuan Hu, Qiong Chen, and Fangming Liu. Asyfunc: A high-performance and resource-efficient serverless inference system via asymmetric functions. In *Proceedings of the 2023 ACM Symposium on Cloud Computing, SoCC 2023, Santa Cruz, CA, USA, 30 October 2023 - 1 November 2023*, pages 324–340. ACM, 2023.

[74] Artjom Joosen, Ahmed Hassan, Martin Asenov, Rajkarn Singh, Luke Nicholas Darlow, Jianfeng Wang, and Adam Barker. How does it function?: Characterizing long-term trends in production serverless workloads. In *Proceedings of the 2023 ACM Symposium on Cloud Computing, SoCC 2023, Santa Cruz, CA, USA, 30 October 2023 - 1 November 2023*, pages 443–458. ACM, 2023.

[75] Yogendra Jain. Book review: Nitin seth. 2024. mastering the data paradox: The key to winning in the ai age. *Metamorphosis*, 23(1):90–91, 2024.

[76] Xinyu Liu, Shuyu Shen, Boyan Li, Peixian Ma, Runzhi Jiang, Yuxin Zhang, Ju Fan, Guoliang Li, Nan Tang, and Yuyu Luo. A survey of nl2sql with large language models: Where are we, and where are we going? *arXiv preprint arXiv:2408.05109*, 2024.

[77] Wei Zhang, Binghao Chen, Zhenhua Han, Quan Chen, Peng Cheng, Fan Yang, Ran Shu, Yuqing Yang, and Minyi Guo. PilotFish: Harvesting free cycles of cloud gaming with deep learning training. In *2022 USENIX*

Annual Technical Conference (USENIX ATC 22), pages 217–232, 2022.

[78] Yiwen Zhu, Yuanyuan Tian, Joyce Cahoon, Subru Krishnan, Ankita Agarwal, Rana Alotaibi, Jesús Camacho-Rodríguez, Bibin Chundatt, Andrew Chung, Niharika Dutta, et al. Towards building autonomous data services on azure. In *Companion of the 2023 International Conference on Management of Data*, pages 217–224, 2023.

[79] Kuo Zhang, Peijian Wang, Ning Gu, and Thu D Nguyen. Greendrl: managing green datacenters using deep reinforcement learning. In *Proceedings of the 13th Symposium on Cloud Computing*, pages 445–460, 2022.

[80] Gaurav Saxena, Mohammad Rahman, Naresh Chainani, Chunbin Lin, George Caragea, Fahim Chowdhury, Ryan Marcus, Tim Kraska, Ippokratis Pandis, and Balakrishnan Narayanaswamy. Auto-wlm: Machine learning enhanced workload management in amazon redshift. In *Companion of the 2023 International Conference on Management of Data*, pages 225–237, 2023.

[81] Zhaoyan Sun, Xuanhe Zhou, and Guoliang Li. Learned index: A comprehensive experimental evaluation. *Proceedings of the VLDB Endowment*, 16(8):1992–2004, 2023.

[82] Matteo Brucato, Tarique Siddiqui, Wentao Wu, Vivek Narasayya, and Surajit Chaudhuri. Wred: Workload reduction for scalable index tuning. *Proceedings of the ACM on Management of Data*, 2(1):1–26, 2024.

[83] Pengcheng Li, Yixin Guo, and Yongbin Gu. Predicting reuse interval for optimized web caching: an lstm-based machine learning approach. In *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15. IEEE, 2022.

[84] Haoran Qiu, Weichao Mao, Archit Patke, Shengkun Cui, Saurabh Jha, Chen Wang, Hubertus Franke, Zbigniew Kalbarczyk, Tamer Başar, and Ravishankar K Iyer. Power-aware deep learning model serving with mu-Serve. In *2024 USENIX Annual Technical Conference (USENIX ATC 24)*, pages 75–93, 2024.

[85] Zhuohan Li, Lianmin Zheng, Yinmin Zhong, Vincent Liu, Ying Sheng, Xin Jin, Yanping Huang, Zhifeng Chen, Hao Zhang, Joseph E Gonzalez, et al. AlpaServe: Statistical multiplexing with model parallelism for deep learning serving. In *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*, pages 663–679, 2023.

[86] Cedric Lichtenau, Alper Buyuktosunoglu, Ramon Bertran, Peter Figuli, Christian Jacobi, Nikolaos Papandreou, Haris Pozidis, Anthony Saporito, Andrew Sica, and Elpida Tzortzatos. Ai accelerator on ibm telum processor: Industrial product. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*, pages 1012–1028, 2022.

[87] Maximilian Lam, Jeff Johnson, Wenjie Xiong, Kiwan Maeng, Udit Gupta, Yang Li, Liangzhen Lai, Ilias Leontiadis, Minsoo Rhu, Hsien-Hsin S Lee, et al. Gpu-based private information retrieval for on-device machine learning inference. *arXiv preprint arXiv:2301.10904*, 2023.

[88] Hanpeng Hu, Junwei Su, Juntao Zhao, Yanghua Peng, Yibo Zhu, Haibin Lin, and Chuan Wu. Cdmpp: A device-model agnostic framework for latency prediction of tensor programs. In *Proceedings of the Nineteenth European Conference on Computer Systems*, pages 1054–1074, 2024.

[89] 中国电信集团公司. 云网融合 2030 技术白皮书. 2020.

[90] N. G. Duffield, Pawan Goyal, Albert Greenberg, Partho Mishra, K. K. Ramakrishnan, and Jacobus E. Van, der Merive. A flexible model for resource management in virtual private networks. In *Conference on Applications*, pages 95–108, 1999.

[91] Yu Ma, Weifa Liang, and Jie Wu. Online nfv-enabled multicasting in mobile edge cloud networks. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 821–830, 2019.

[92] Mania Abdi, Samuel Ginzburg, Xiayue Charles Lin, Jose Faleiro, Gohar Irfan Chaudhry, Inigo Goiri, Ricardo Bianchini, Daniel S Berger, and Rodrigo Fonseca. Palette load balancing: Locality hints for serverless functions. In *Proceedings of the Eighteenth European Conference on Computer Systems*, EuroSys '23, page 365–380, New York, NY, USA, 2023. Association for Computing Machinery.

[93] Cheng Gu, Xin Song, Ben Hok Ng, Qiao Xiang, Zehua Guo, and Geng Li. An ML-Accelerated Framework for Large-Scale Constrained Traffic Engineering. *Proceedings - International Conference on Distributed Computing Systems*, pages 47–58, 2024.

[94] Changgang Zheng, Haoyue Tang, Mingyuan Zang, Xinpeng Hong, Aosong Feng, Leandros Tassiulas, and Noa Zilberman. DINC: Toward Distributed In-Network Computing. In *Proceedings of ACM CoNEXT'23*, 2023.

[95] 中国电信. 中国电信发布云算网一体化调度产品.

[96] Hwijoon Lim, Juncheol Ye, Sangeetha Abdu Jyothi, and Dongsu Han. Accelerating Model Training in Multi-cluster Environments with Consumer-grade GPUs. *ACM SIGCOMM 2024 - Proceedings of the 2024 ACM SIGCOMM 2024 Conference*, pages 707–720, 2024.

[97] Wenxin Li, Xin He, Yuan Liu, Keqiu Li, Kai Chen, Zhao Ge, Zewei Guan, Heng Qi, Song Zhang, and Guyue Liu. Flow scheduling with imprecise knowledge. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, pages 95–111, 2024.

[98] Xiaofei Yue, Song Yang, Liehuang Zhu, Stojan Trajanovski, and Xiaoming Fu. Demeter: Fine-grained Function Orchestration for Geo-distributed Serverless Analytics. *Proceedings - IEEE INFOCOM*, pages 2498–2507, 2024.

[99] Zhiying Xu, Francis Y. Yan, Rachee Singh, Justin T. Chiu, Alexander M. Rush, and Minlan Yu. Teal: Learning-Accelerated Optimization of WAN Traffic Engineering. *SIGCOMM 2023 - Proceedings of the ACM SIGCOMM 2023 Conference*, pages 378–393, 2023.

[100] Sachin Ashok, Vipul Harsh, Brighten Godfrey, Radhika Mittal, Srinivasan Parthasarathy, and Larisa Shwartz. TraceWeaver: Distributed Request Tracing for Microservices Without Application Modification. *ACM SIGCOMM 2024 - Proceedings of the 2024 ACM SIGCOMM 2024 Conference*, pages 828–842, 2024.

[101] Hongke Zhang, Wei Quan, Han-chieh Chao, and Chunming Qiao. Smart identifier network: A collaborative architecture for the future internet. *IEEE network*, 30(3):46–51, 2016.

[102] Shuo Wang, Binwei Wu, Chen Zhang, Yudong Huang, Tao Huang, and Yunjie Liu. Large-scale deterministic ip networks on ceni. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6. IEEE, 2021.

[103] Zhiyong Feng, Zhiqing Wei, Xu Chen, Heng Yang, Qixun Zhang, and Ping Zhang. Joint communication, sensing, and computation enabled 6g intelligent machine system. *IEEE Network*, 35(6):34–42, 2021.

[104] Guangxu Zhu, Zhonghao Lyu, Xiang Jiao, Peixi Liu, Mingzhe Chen, Jie Xu, Shuguang Cui, and Ping Zhang. Pushing ai to wireless network edge: An overview on integrated sensing, communication, and computation towards 6g. *Science China Information Sciences*, 66(3):130301, 2023.

[105] Marilyne Apkarian. Prospects for the development of computing power network technology for scientific computing.

[106] Lin Wei, Jinyang Li, Yufei Gao, Lei Shi, Huijuan Lian, Guozhen Cheng, and Mengyang He. Resource matching algorithm based on multidimensional computing resource measurement in computing power network. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 2295–2300. IEEE, 2024.

[107] LI Yinan, TANG Qinqin, PENG Kailai, LIU Jia, XIE Renchao, and HUANG Tao. Research on measurement and modeling of service-centric computing power network. *Information and Communications Technology and Policy*, 49(5):21, 2023.

[108] Qunyang Lin, Luyang Liu, Hongyin Zhu, Haonan Tong, and Chuang Zhang. Efsc: an efficient, flexible and secure trading system for computing power network. In *2024 IEEE 49th Conference on Local Computer Networks (LCN)*, pages 1–7. IEEE, 2024.

[109] Xiaoxu Ren, Chao Qiu, Xiaofei Wang, Zhu Han, Ke Xu, Haipeng Yao, and Song Guo. Ai-bazaar: A cloud-edge computing power trading framework for ubiquitous ai services. *IEEE Transactions on Cloud Computing*, 11(3):2337–2348, 2022.

[110] Yue Zhao, Jizhi Wang, Lingrui Kong, and Tongtong Sui. The study of multi-type computing power trading mechanism in computing power network based on blockchain and combinatorial double auction. In *Proceedings of the 2024 4th International Conference on Artificial Intelligence, Big Data and Algorithms*, pages 259–266, 2024.

[111] Shujiong Tang, Yue Yu, Hui Wang, Guiliang Wang, Wuhui Chen, Zenglin Xu, Song Guo, and Wen Gao. A survey on scheduling techniques in computing and network convergence. *IEEE Communications Surveys & Tutorials*, 26(1):160–195, 2024.

[112] Shuai Wang, Dan Li, Jiansong Zhang, and Wei Lin. Cefs: compute-efficient flow scheduling for iterative synchronous applications. In *Proceedings of the 16th International Conference on Emerging Networking EXperiments and Technologies*, CoNEXT '20, page 136–148, New York, NY, USA, 2020. Association for Computing Machinery.

[113] Xinjing Yuan, Lingjun Pu, Lei Jiao, Xiaofei Wang, Meijuan Yang, and Jingdong Xu. When computing power network meets distributed machine learning: An efficient federated split learning framework. In *2023 IEEE/ACM 31st International Symposium on Quality of Service (IWQoS)*, pages 1–10. IEEE, 2023.

[114] Mingtao Ji, Zhuzhong Qian, and Baoliu Ye. When cpn meets ai: Resource provisioning for inference query upon computing power network. In *2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 2261–2268. IEEE, 2023.

[115] Fabrício B Carvalho, Ronaldo A Ferreira, Ítalo Cunha, Marcos AM Vieira, and Murali K Ramanathan. Dyssect: Dynamic scaling of stateful network functions. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 1529–1538. IEEE, 2022.

[116] Deepak Bansal, Gerald DeGrace, Rishabh Tewari, Michal Zygmunt, James Grantham, Silvano Gai, Mario Baldi, Krishna Doddapaneni, Arun Selvarajan, Arunkumar Arumugam, et al. Disaggregating stateful network functions. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 1469–1487, 2023.

[117] Jacopo Massa, Stefano Forti, Federica Paganelli, Patrizio Dazzi, and Antonio Brogi. Declarative provisioning of virtual network function chains in intent-based networks. In *2023 IEEE 9th International Conference on Network Softwarization (NetSoft)*, pages 522–527. IEEE, 2023.

[118] Marcel Blöcher, Nils Nedderhut, Pavel Chuprikov, Ramin Khalili, Patrick Eugster, and Lin Wang. Train

once apply anywhere: Effective scheduling for network function chains running on fumes. In *IEEE INFOCOM 2024-IEEE Conference on Computer Communications*, pages 661–670. IEEE, 2024.

[119] Yingling Mao, Xiaojun Shang, and Yuanyuan Yang. Provably efficient algorithms for traffic-sensitive sfc placement and flow routing. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 950–959. IEEE, 2022.

[120] Theophilus A Benson, Prashanth Kannan, Prankur Gupta, Balasubramanian Madhavan, Kumar Saurabh Arora, Jie Meng, Martin Lau, Abhishek Dhamija, Rajiv Krishnamurthy, Srikanth Sundaresan, et al. Netedit: An orchestration platform for ebpf network functions at scale. In *Proceedings of the ACM SIGCOMM 2024 Conference*, pages 721–734, 2024.

[121] Jiamei Lv, Yi Gao, Zhi Ding, Yuxiang Lin, Xinyun You, Guang Yang, and Wei Dong. Providing ue-level qos support by joint scheduling and orchestration for 5g vran. In *IEEE INFOCOM 2024-IEEE Conference on Computer Communications*, pages 51–60. IEEE, 2024.

[122] Hao Li, Yihan Dang, Guangda Sun, Guyue Liu, Danfeng Shan, and Peng Zhang. Lemonnfv: Consolidating heterogeneous network functions at line speed. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 1451–1468, 2023.

[123] Stefano Maxenti, Salvatore D' Oro, Leonardo Bonati, Michele Polese, Antonio Capone, and Tommaso Melodia. Scalo-ran: Energy-aware network intelligence scaling in open ran. In *IEEE INFOCOM 2024-IEEE Conference on Computer Communications*, pages 891–900. IEEE, 2024.

[124] Tolga O Atalay, Sudip Maitra, Dragoslav Stojadinovic, Angelos Stavrou, and Haining Wang. Securing 5g openran with a scalable authorization framework for xapps. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2023.

[125] Jorge Baranda, Luca Vettori, Miquel Payaró, Josep Mangues-Bafalluy, Guillermo Gomez, and Sozos Karageorgiou. Multi-administrative domain service onboarding in a zsm-based orchestration architecture. In *2023 20th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 366–368. IEEE, 2023.

[126] Vivek Jain, Hao-Tse Chu, Shixiong Qi, Chia-An Lee, Hung-Cheng Chang, Cheng-Ying Hsieh, KK Ramakrishnan, and Jyh-Cheng Chen. L25gc: A low latency 5g core network based on high-performance nfv platforms. In *Proceedings of the ACM SIGCOMM 2022 Conference*, pages 143–157, 2022.

[127] Shuo Shi, Chao Zhang, Zongpu Zhang, Hubin Zhang, Xin Zeng, Weigang Li, Junyuan Wang, Xiantao Zhang, Yibin Shen, Jian Li, et al. vcrypto: a unified para-virtualization framework for heterogeneous cryptographic resources. In *IEEE INFOCOM 2024-IEEE Conference on Computer Communications*, pages 781–790. IEEE, 2024.

[128] Francisco Pereira, Fernando MV Ramos, and Luis Pedrosa. Automatic parallelization of software network functions. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, pages 1531–1550, 2024.

[129] Jiansong Zhang, Yongqiang Xiong, Ningyi Xu, Ran Shu, Bojie Li, Peng Cheng, Guo Chen, and Thomas Moscibroda. The feniks fpga operating system for cloud computing. In *Proceedings of the 8th Asia-Pacific Workshop on Systems*, pages 1–7, 2017.

[130] Nikita Lazarev, Tao Ji, Anuj Kalia, Daehyeok Kim, Ilias Marinos, Francis Y Yan, Christina Delimitrou, Zhiru Zhang, and Aditya Akella. Resilient baseband processing in virtualized rans with slingshot. In *Proceedings of the ACM SIGCOMM 2023 Conference*, pages 654–667, 2023.

[131] Soki Koizumi, Takao Kondo, and Fumio Teraoka. Poster: Secure nfv infrastructure based on software fault isolation considering multi-tenant environment. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*, pages 650–651, 2024.

[132] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[133] Panos M Pardalos, Antanas Žilinskas, Julius Žilinskas, et al. *Non-convex multi-objective optimization*. Springer, 2017.

[134] Laurence A Wolsey and George L Nemhauser. *Integer and combinatorial optimization*. John Wiley & Sons, 2014.

[135] Xin-She Yang. *Nature-inspired metaheuristic algorithms*. Luniver press, 2010.

[136] Johannes Schneider and Scott Kirkpatrick. *Stochastic optimization*. Springer Science & Business Media, 2006.

[137] Tomas Feder, Pavol Hell, Sulamita Klein, and Rajeev Motwani. Complexity of graph partition problems. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 464–472, 1999.

[138] Horst Bunke. Graph matching: Theoretical foundations, algorithms, and applications. In *Proc. Vision Interface*, volume 2000, pages 82–88, 2000.

[139] Günter Rote. Path problems in graphs. *Computational graph theory*, pages 155–189, 1990.

[140] Chuntao Jiang, Frans Coenen, and Michele Zito. A survey of frequent subgraph mining algorithms. *The Knowledge Engineering Review*, 28(1):75–105, 2013.

[141] David M Kreps. Nash equilibrium. In *Game theory*, pages 167–177. Springer, 1989.

[142] Rodica Branzei, Dinko Dimitrov, and Stef Tijs. *Models in cooperative game theory*, volume 556. Springer Science & Business Media, 2008.

[143] Jörgen W Weibull. *Evolutionary game theory*. MIT press, 1997.

[144] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[145] Jürgen Schmidhuber, Sepp Hochreiter, et al. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[146] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2022.

[147] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[148] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[149] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.

[150] Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering Cambridge, UK, 1994.

[151] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[152] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[153] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Tim Harley, Timothy P Lillicrap, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 1928–1937, 2016.

[154] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, and Zhifang Sui. A survey on in-context learning. In *EMNLP*, pages 1107–1128. Association for Computational Linguistics, 2024.

[155] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023.

[156] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[157] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[158] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[159] Philip E Gill, Walter Murray, and Margaret H Wright. *Practical optimization*. SIAM, 2019.

[160] Christodoulos A Floudas. *Nonlinear and mixed-integer optimization: fundamentals and applications*. Oxford University Press, 1995.

[161] Reinhard Diestel. *Graph theory*. Springer (print edition); Reinhard Diestel (eBooks), 2024.

[162] Jonathan L Gross, Jay Yellen, and Mark Anderson. *Graph theory and its applications*. Chapman and Hall/CRC, 2018.

[163] Alain Bretto. Hypergraph theory. *An introduction. Mathematical Engineering. Cham: Springer*, 1, 2013.

[164] John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior: 60th anniversary commemorative edition. In *Theory of games and economic behavior*. Princeton university press, 2007.

[165] Tim Roughgarden. Algorithmic game theory. *Communications of the ACM*, 53(7):78–86, 2010.

[166] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[167] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning*, volume 1. MIT press Cambridge, MA, USA, 2017.

[168] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[169] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[170] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red

Hook, NY, USA, 2017. Curran Associates Inc.

[171] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. Edge computing: Vision and challenges. *IEEE internet of things journal*, 3(5):637–646, 2016.

[172] Ashish Kumar Shakya, Gopinatha Pillai, and Sohom Chakrabarty. Reinforcement learning algorithms: A brief survey. *Expert Syst. Appl.*, 231(C), November 2023.

[173] Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, 55(2):895–943, 2022.

[174] Rafael Figueiredo Prudencio, Marcos ROA Maximo, and Esther Luna Colombini. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[175] Shubham Pateria, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(5):1–35, 2021.

[176] Zhongjiang He, Zihan Wang, Xinzhang Liu, Shixuan Liu, Yitong Yao, Yuyao Huang, Xuelong Li, Yongxiang Li, Zhonghao Che, Zhaoxi Zhang, et al. Telechat technical report. *arXiv preprint arXiv:2401.03804*, 2024.

[177] Renze Lou, Kai Zhang, and Wenpeng Yin. Large language model instruction following: A survey of progresses and challenges. *Computational Linguistics*, 50(3):1053–1095, September 2024.

[178] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics.

[179] Shervin Minaee, Tomás Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *CoRR*, abs/2402.06196, 2024.

[180] Nuo Chen, Ning Wu, Jianhui Chang, Linjun Shou, and Jia Li. ControlMath: Controllable data generation promotes math generalist models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12201–12217. Association for Computational Linguistics, 2024.

[181] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[182] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

[183] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

[184] Ali Modarressi, Mohsen Fayyaz, Ehsan Aghazadeh, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. Decompx: Explaining transformers decisions by propagating token decomposition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2649–2664, 2023.

[185] Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. Globenc: Quantifying global token attribution by incorporating the whole encoder layer in transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 258–271, 2022.

[186] Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. Copen: Probing conceptual knowledge in pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5015–5035, 2022.

[187] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024.

[188] Ting Yang, Haibo Pen, Wei Li, Dong Yuan, and Albert Y Zomaya. An energy-efficient storage strategy for cloud datacenters based on variable k-coverage of a hypergraph. *IEEE Transactions on Parallel and Distributed Systems*, 28(12):3344–3355, 2017.

[189] Jingya Zhou, Jianxi Fan, Juncheng Jia, Baolei Cheng, and Zhao Liu. Optimizing cost for geo-distributed storage systems in online social networks. *Journal of computational science*, 26:363–374, 2018.

[190] Liu YN and Fan BB. Survey on graph database development. *Computer Systems and Applications*, 31(8):1–16, 2022.

[191] Ye Yuan, Delong Ma, Zhenyu Wen, Zhiwei Zhang, and Guoren Wang. Subgraph matching over graph federation. *Proceedings of the VLDB Endowment*, 15(3):437–450, 2021.

[192] Shuheng Fang, Kangfei Zhao, Yu Rong, Zhixun Li, and Jeffrey Xu Yu. Inductive attributed community search: To learn communities across graphs. *Proceedings of the VLDB Endowment*, 17(10):2576–2589, 2024.

[193] Junhua Zhang, Wentao Li, Long Yuan, Lu Qin, Ying Zhang, and Lijun Chang. Shortest-path queries on complex networks: experiments, analyses, and improvement. *Proceedings of the VLDB Endowment*, 15(11):2640–2652, 2022.

[194] Ajay Kumar, Shashank Sheshar Singh, Kuldeep Singh, and Bhaskar Biswas. Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications*, 553:124289, 2020.

[195] Di Jin, Zhizhi Yu, Pengfei Jiao, Shirui Pan, Dongxiao He, Jia Wu, S Yu Philip, and Weixiong Zhang. A survey of community detection approaches: From statistical modeling to deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(2):1149–1170, 2021.

[196] Jialun Li, Jieqian Yao, Danyang Xiao, Diying Yang, and Weigang Wu. Evogwp: Predicting long-term changes in cloud workloads using deep graph-evolution learning. *IEEE Transactions on Parallel and Distributed Systems*, 2024.

[197] Xiaoheng Deng, Jun Li, Enlu Liu, and Honggang Zhang. Task allocation algorithm and optimization model on edge collaboration. *Journal of Systems Architecture*, 110:101778, 2020.

[198] Hadi Goudarzi and Massoud Pedram. Geographical load balancing for online service applications in distributed datacenters. In *2013 IEEE Sixth International Conference on Cloud Computing*, pages 351–358. IEEE, 2013.

[199] Amanpreet Kaur and Bikrampal Kaur. Load balancing optimization based on hybrid heuristic-metaheuristic techniques in cloud environment. *Journal of King Saud University-Computer and Information Sciences*, 34(3):813–824, 2022.

[200] Jie You, Jae-Won Chung, and Mosharaf Chowdhury. Zeus: Understanding and optimizing GPU energy consumption of DNN training. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 119–139, Boston, MA, April 2023. USENIX Association.

[201] Immanuel Trummer. Db-bert: a database tuning tool that" reads the manual". In *Proceedings of the 2022 international conference on management of data*, pages 190–203, 2022.

[202] Victor Giannankouris and Immanuel Trummer. {\lambda}-tune: Harnessing large language models for automated database system tuning. *arXiv preprint arXiv:2411.03500*, 2024.

[203] Siqiao Xue, Caigao Jiang, Wenhui Shi, Fangyin Cheng, Keting Chen, Hongjun Yang, Zhiping Zhang, Jianshan He, Hongyang Zhang, Ganglin Wei, et al. Db-gpt: Empowering database interactions with private large language models. *arXiv preprint arXiv:2312.17449*, 2023.

[204] Jiangtao Zhang, Hejiao Huang, and Xuan Wang. Resource provision algorithms in cloud computing: A survey. *Journal of network and computer applications*, 64:23–42, 2016.

[205] Ameni Hedhli and Haithem Mezni. A survey of service placement in cloud environments. *Journal of Grid Computing*, 19(3):23, 2021.

[206] Yefu Wang and Xiaorui Wang. Performance-controlled server consolidation for virtualized data centers with multi-tier applications. *Sustainable Computing: Informatics and Systems*, 4(1):52–65, 2014.

[207] Shuangwu Chen, Jiangming Li, Qifeng Yuan, Huasen He, Sen Li, and Jian Yang. Two-timescale joint optimization of task scheduling and resource scaling in multi-data center system based on multi-agent deep reinforcement learning. *IEEE Transactions on Parallel and Distributed Systems*, 2024.

[208] Chenxi Zhang, Xin Peng, Chaofeng Sha, Ke Zhang, Zhenqing Fu, Xiya Wu, Qingwei Lin, and Dongmei Zhang. Deeptralog: Trace-log combined microservice anomaly detection through graph-based deep learning. In *Proceedings of the 44th international conference on software engineering*, pages 623–634, 2022.

[209] Teng Zhong, Yinglei Teng, Shijun Ma, Jiaxuan Chen, and Sicong Yu. A microservices identification method based on spectral clustering for industrial legacy systems. In *2023 IEEE Globecom Workshops (GC Wkshps)*, pages 1331–1337. IEEE, 2023.

[210] Qiuhan Meng and Songye Zhu. Anomaly detection for construction vibration signals using unsupervised deep learning and cloud computing. *Advanced Engineering Informatics*, 55:101907, 2023.

[211] Nadia Alshahwan, Jubin Chheda, Anastasia Finogenova, Beliz Gokkaya, Mark Harman, Inna Harper, Alexandru Marginean, Shubho Sengupta, and Eddy Wang. Automated unit test improvement using large language models at meta. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, pages 185–196, 2024.

[212] Mohammadreza Pourreza and Davood Rafiei. Din-sql: Decomposed in-context learning of text-to-sql with self-correction. *Advances in Neural Information Processing Systems*, 36, 2024.

[213] Dawei Gao, Haibin Wang, Yaliang Li, Xiuyu Sun, Yichen Qian, Bolin Ding, and Jingren Zhou. Text-to-sql empowered by large language models: A benchmark evaluation. *Proceedings of the VLDB Endowment*, 17(5):1132–1145, 2024.

[214] Yudong Huang, Hongyang Du, Xinyuan Zhang, Dusit Niyato, Jiawen Kang, Zehui Xiong, Shuo Wang, and Tao Huang. Large language models for networking: Applications, enabling techniques, and challenges. *IEEE Network*, 2024.

[215] Beni Ifland, Elad Duani, Rubin Krief, Miro Ohana, Aviram Zilberman, Andres Murillo, Ofir Manor, Ortal Lavi, Hikichi Kenji, Asaf Shabtai, et al. Genet: A multimodal llm-based co-pilot for network topology and configuration. *arXiv preprint arXiv:2407.08249*, 2024.

[216] Gartner. Gartner 2024 hype cycle for emerging technologies highlights developer productivity, total ex-

perience, ai, and security, 2024. Accessed: 2024-12-27.

[217] Kuanishbay Sadatdiynov, Laizhong Cui, Lei Zhang, Joshua Zhexue Huang, Salman Salloum, and Mohammad Sultan Mahmud. A review of optimization methods for computation offloading in edge computing networks. *Digital Communications and Networks*, 9(2):450–461, 2023.

[218] Xinchen Lyu, Wei Ni, Hui Tian, Ren Ping Liu, Xin Wang, Georgios B Giannakis, and Arogyaswami Paulraj. Optimal schedule of mobile edge computing for internet of things using partial information. *IEEE Journal on Selected Areas in Communications*, 35(11):2606–2615, 2017.

[219] Miao Hu, Zixuan Xie, Di Wu, Yipeng Zhou, Xu Chen, and Liang Xiao. Heterogeneous edge offloading with incomplete information: A minority game approach. *IEEE Transactions on Parallel and Distributed Systems*, 31(9):2139–2154, 2020.

[220] Songtao Guo, Jiadi Liu, Yuanyuan Yang, Bin Xiao, and Zhetao Li. Energy-efficient dynamic computation offloading and cooperative task scheduling in mobile cloud computing. *IEEE Transactions on Mobile Computing*, 18(2):319–333, 2019.

[221] Jiasi Chen and Xukan Ran. Deep learning with edge computing: A review. *Proceedings of the IEEE*, 107(8):1655–1674, 2019.

[222] Jin Huang, Hui Guan, and Deepak Ganesan. Re-thinking computation offload for efficient inference on iot devices with duty-cycled radios. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pages 1–15, 2023.

[223] Song Wang and Xinyu Zhang. Neuromessenger: Towards error tolerant distributed machine learning over edge networks. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 2058–2067. IEEE, 2022.

[224] Jing Wu, Lin Wang, Qirui Jin, and Fangming Liu. Graft: Efficient inference serving for hybrid deep learning with slo guarantees via dnn re-alignment. *IEEE Transactions on Parallel and Distributed Systems*, 35(2):280–296, 2024.

[225] MG Sarwar Murshed, Christopher Murphy, Daqing Hou, Nazar Khan, Ganesh Ananthanarayanan, and Faraz Hussain. Machine learning at the network edge: A survey. *ACM Computing Surveys (CSUR)*, 54(8):1–37, 2021.

[226] Jing Liu, Bohan Zhuang, Zhuangwei Zhuang, Yong Guo, Junzhou Huang, Jinhui Zhu, and Mingkui Tan. Discrimination-aware network pruning for deep model compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4035–4051, 2022.

[227] Yoshitomo Matsubara, Ruihan Yang, Marco Levorato, and Stephan Mandt. Supervised compression for resource-constrained edge computing systems. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2685–2695, January 2022.

[228] Xing Liu, Wei Yu, Fan Liang, David Griffith, and Nada Golmie. Toward deep transfer learning in industrial internet of things. *IEEE Internet of Things Journal*, 8(15):12163–12175, 2021.

[229] Jiajia Liu, Yongpeng Shi, Zubair Md Fadlullah, and Nei Kato. Space-air-ground integrated network: A survey. *IEEE Communications Surveys & Tutorials*, 20(4):2714–2741, 2018.

[230] Fengxiao Tang, Hans Hofner, Nei Kato, Kazuma Kaneko, Yasutaka Yamashita, and Masatake Hangai. A deep reinforcement learning-based dynamic traffic offloading in space-air-ground integrated networks (sagin). *IEEE Journal on Selected Areas in Communications*, 40(1):276–289, 2021.

[231] Yongyi Ran, Yajie Ding, Shuangwu Chen, Jizhao Lei, and Jiangtao Luo. Fully-distributed dynamic packet routing for leo satellite networks: A gnn-enhanced multi-agent reinforcement learning approach. *IEEE Transactions on Vehicular Technology*, 2024.

[232] Yaoying Zhang, Qian Wu, Zeqi Lai, and Hewu Li. Enabling low-latency-capable satellite-ground topology for emerging leo satellite networks. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 1329–1338. IEEE, 2022.

[233] Marco Conti and Silvia Giordano. Mobile ad hoc networking: milestones, challenges, and new research directions. *IEEE Communications Magazine*, 52(1):85–96, 2014.

[234] Xi Chen, Gang Sun, Tao Wu, Ling Liu, Hongfang Yu, and Mohsen Guizani. Rance: A randomly centralized and on-demand clustering protocol for mobile ad hoc networks. *IEEE Internet of Things Journal*, 9(23):23639–23658, 2022.

[235] Iago Medeiros, Azzedine Boukerche, and Eduardo Cerqueira. Swarm-based and energy-aware unmanned aerial vehicle system for video delivery of mobile objects. *IEEE Transactions on Vehicular Technology*, 71(1):766–779, 2022.

[236] Li Yan, Haiying Shen, and Kang Chen. Mobit: Distributed and congestion-resilient trajectory-based routing for vehicular delay tolerant networks. *IEEE/ACM Transactions on Networking*, 26(3):1078–1091, 2018.

[237] Carlo Puliafito, Enzo Mingozzi, Francesco Longo, Antonio Puliafito, and Omer Rana. Fog computing for the internet of things: A survey. *ACM Transactions on Internet Technology (TOIT)*, 19(2):1–41, 2019.

[238] Chu ge Wu, Wei Li, Ling Wang, and Albert Y. Zomaya. An evolutionary fuzzy scheduler for multi-objective resource allocation in fog computing. *Future Generation Computer Systems*, 117:498–509, 2021.

[239] Zhengyuan Pang, Lifeng Sun, Zhi Wang, Erfang Tian, and Shiqiang Yang. A survey of cloudlet based mobile computing. In *2015 International conference on cloud computing and big data (CCBD)*, pages 268–275. IEEE, 2015.

[240] Dixit Bhatta and Lena Mashayekhy. A bifactor approximation algorithm for cloudlet placement in edge computing. *IEEE Transactions on Parallel and Distributed Systems*, 33(8):1787–1798, 2022.

[241] Dinh C Nguyen, Ming Ding, Pubudu N Pathirana, Aruna Seneviratne, Jun Li, and H Vincent Poor. Federated learning for internet of things: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 23(3):1622–1658, 2021.

[242] Xinran Zhang, Hanqi Zhu, Yifan Duan, Wuyang Zhang, Longfei Shangguan, Yu Zhang, Jianmin Ji, and Yanyong Zhang. Map++: Towards user-participatory visual slam systems with efficient map expansion and sharing. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 633–647, 2024.

[243] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

[244] Chenghao Hu and Baochun Li. When the edge meets transformers: Distributed inference with transformer models. In *2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS)*, pages 82–92. IEEE, 2024.

[245] Davide Calvaresi, Mauro Marinoni, Arnon Sturm, Michael Schumacher, and Giorgio Buttazzo. The challenge of real-time multi-agent systems for enabling iot and cps. In *Proceedings of the international confer-*

ence on web intelligence, pages 356–364, 2017.

[246] Zhenhui Ye, Ke Wang, Yining Chen, Xiaohong Jiang, and Guanghua Song. Multi-uav navigation for partially observable communication coverage by graph reinforcement learning. *IEEE Transactions on Mobile Computing*, 22(7):4056–4069, 2023.

[247] Xiaojie Wang, Zhaolong Ning, Song Guo, Miaowen Wen, Lei Guo, and H Vincent Poor. Dynamic uav deployment for differentiated services: A multi-agent imitation learning based approach. *IEEE Transactions on Mobile Computing*, 22(4):2131–2146, 2021.

[248] Jingao Xu, Hao Cao, Zheng Yang, Longfei Shangguan, Jialin Zhang, Xiaowu He, and Yunhao Liu. Swarmmap: Scaling up real-time collaborative visual slam at the edge. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*, pages 977–993, 2022.

[249] Xuanzhe Liu, Shangguang Wang, Yun Ma, Ying Zhang, Qiaozhu Mei, Yunxin Liu, and Gang Huang. Operating systems for resource-adaptive intelligent software: Challenges and opportunities. *ACM Transactions on Internet Technology (TOIT)*, 21(2):1–19, 2021.

[250] Xiangyu Li, Yuanchun Li, Yuanzhe Li, Ting Cao, and Yunxin Liu. Flexnn: Efficient and adaptive dnn inference on memory-constrained edge devices. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 709–723, 2024.

[251] Hao Wen, Yuanchun Li, Zunshuai Zhang, Shiqi Jiang, Xiaozhou Ye, Ye Ouyang, Yaqin Zhang, and Yunxin Liu. Adaptivenet: Post-deployment neural architecture adaptation for diverse edge environments. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pages 1–17, 2023.

[252] Fucheng Jia, Deyu Zhang, Ting Cao, Shiqi Jiang, Yunxin Liu, Ju Ren, and Yaoxue Zhang. Codl: efficient cpu-gpu co-execution for deep learning inference on mobile devices. In *MobiSys*, volume 22, pages 209–221, 2022.

[253] Maribel Fernández, Jenjira Jaimunk, and Bhavani Thuraisingham. A privacy-preserving architecture and data-sharing model for cloud-iot applications. *IEEE Transactions on Dependable and Secure Computing*, 20(4):3495–3507, 2022.

[254] Na Wang, Wen Zhou, Jingjing Wang, Yifan Guo, Junsong Fu, and Jianwei Liu. Secure and efficient similarity retrieval in cloud computing based on homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, 2024.

[255] Caiqun Shi, Qinlong Huang, Rui Jian, and Genghui Chi. Cross-domain inner-product access control encryption for secure emr flow in cloud edge. *IEEE Transactions on Information Forensics and Security*, 2024.

[256] Jietao Xiao, Nanzi Yang, Wenbo Shen, Jinku Li, Xin Guo, Zhiqiang Dong, Fei Xie, and Jianfeng Ma. Attacks are forwarded: Breaking the isolation of microvm-based containers through operation forwarding. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 7517–7534, 2023.

[257] Benedict Schlüter, Supraja Sridhara, Mark Kuhne, Andrin Bertschi, and Shweta Shinde. Heckler: Breaking confidential vms with malicious interrupts. In *USENIX Security*, 2024.

[258] Jianyu Niu, Wei Peng, Xiaokuan Zhang, and Yinqian Zhang. Narrator: Secure and practical state continuity for trusted execution in the cloud. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2385–2399, 2022.

[259] Junjie Xiong, Mingkui Wei, Zhuo Lu, and Yao Liu. Warmonger: inflicting denial-of-service via serverless

functions in the cloud. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 955–969, 2021.

[260] Fenglu Zhang, Baojun Liu, Eihal Alowaisheq, Jianjun Chen, Chaoyi Lu, Linjian Song, Yong Ma, Ying Liu, Haixin Duan, and Min Yang. Silence is not golden: Disrupting the load balancing of authoritative dns servers. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 296–310, 2023.

[261] Aastha Mehta, Mohamed Alzayat, Roberta De Viti, Björn B Brandenburg, Peter Druschel, and Deepak Garg. Pacer: Comprehensive network {Side-Channel} mitigation in the cloud. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2819–2838, 2022.

[262] 胡浩, 刘玉岭, 张玉臣, and 张红旗. 基于攻击图的网络安全度量研究综述. 网络与信息安全学报, 4(9):1–16, 2018.

[263] 叶子维, 郭渊博, 王宸东, and 琚安康. 攻击图技术应用研究综述. *Journal on Communications*, 2017.

[264] Takashi Koide, Naoki Fukushi, Hiroki Nakano, and Daiki Chiba. Chatspamdetector: Leveraging large language models for effective phishing email detection. *arXiv preprint arXiv:2402.18093*, 2024.

[265] Yun Lin, Ruofan Liu, Dinil Mon Divakaran, Jun Yang Ng, Qing Zhou Chan, Yiwen Lu, Yuxuan Si, Fan Zhang, and Jin Song Dong. Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 3793–3810, 2021.

[266] Tarek Ali and Panos Kostakos. Huntgpt: Integrating machine learning-based anomaly detection and explainable ai with large language models (llms). *arXiv preprint arXiv:2309.16021*, 2023.

[267] Rihui Sun, Pefei Qiu, Yongqiang Lyu, Dongsheng Wang, Jiang Dong, and Gang Qu. Lightning: Striking the secure isolation on GPU clouds with transient hardware faults. *CoRR*, abs/2112.03662, 2021.

[268] Luis Tomás, Panagiotis C. Kokkinos, Vasilios Anagnostopoulos, Oshrit Feder, Dimosthenis Kyriazis, Kalman Z. Meth, Emmanouel A. Varvarigos, and Theodora A. Varvarigou. Disaster recovery layer for distributed openstack deployments. *IEEE Trans. Cloud Comput.*, 8(1):112–123, 2020.

[269] Guochang Yuan, Zichuan Xu, Binxu Yang, Weifa Liang, Wei Koong Chai, Daphné Tuncer, Alex Galis, George Pavlou, and Guowei Wu. Fault tolerant placement of stateful vnfs and dynamic fault recovery in cloud networks. *Comput. Networks*, 166, 2020.

[270] Maria C. Borges, Joshua Bauer, Sebastian Werner, Michael Gebauer, and Stefan Tai. Informed and assessable observability design decisions in cloud-native microservice applications. In *21st IEEE International Conference on Software Architecture, ICSA 2024, Hyderabad, India, June 4-8, 2024*, pages 69–78. IEEE, 2024.

[271] Sofia Montebugnoli and Luca Foschini. A multicloud observability support based on elasticsearch for cloud-native smart cities services. In *IEEE Symposium on Computers and Communications, ISCC 2023, Gammarth, Tunisia, July 9-12, 2023*, pages 1–6. IEEE, 2023.

[272] Tommi Nylander, Marcus Thelander Andrén, Karl-Erik Årzén, and Martina Maggio. Cloud application predictability through integrated load-balancing and service time control. In *2018 IEEE International Conference on Autonomic Computing, ICAC 2018, Trento, Italy, September 3-7, 2018*, pages 51–60. IEEE Computer Society, 2018.

[273] Paolo Notaro, Jorge Cardoso, and Michael Gerndt. A systematic mapping study in aiops. In Hakim Hacid, Fatma Outay, Hye-young Paik, Amira Alloum, Marinella Petrocchi, Mohamed Reda Bouadjenek, Amin Beheshti, Xumin Liu, and Abderrahmane Maaradji, editors, *Service-Oriented Computing - ICSOC*

2020 Workshops - AIOps, CFTIC, STRAPS, AI-PA, AI-IOTS, and Satellite Events, Dubai, United Arab Emirates, December 14-17, 2020, Proceedings, volume 12632 of *Lecture Notes in Computer Science*, pages 110–123. Springer, 2020.

[274] Zhaoyang Yu, Shenglin Zhang, Mingze Sun, Yingke Li, Yankai Zhao, Xiaolei Hua, Lin Zhu, Xidao Wen, and Dan Pei. Supervised fine-tuning for unsupervised KPI anomaly detection for mobile web systems. In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee, editors, *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 2859–2869. ACM, 2024.

[275] Yicheng Pan, Yang Zhang, Tingzhu Bi, Linlin Han, Yu Zhang, Meng Ma, Xiangzhuang Shen, Xinrui Jiang, Feng Wang, Xian Liu, and Ping Wang. HEAL: performance troubleshooting deep inside data center hosts. In Michele Garetto, Andrea Marin, Florin Ciucu, Giulia Fanti, and Rhonda Righter, editors, *Abstracts of the 2024 ACM SIGMETRICS/IFIP PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS/PERFORMANCE 2024, Venice, Italy, June 10-14, 2024*, pages 41–42. ACM, 2024.

[276] Rui Wang, Xudong Mou, Renyu Yang, Kai Gao, Pin Liu, Chongwei Liu, Tianyu Wo, and Xudong Liu. Cutaddpaste: Time series anomaly detection by exploiting abnormal knowledge. In Ricardo Baeza-Yates and Francesco Bonchi, editors, *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 3176–3187. ACM, 2024.

[277] Linghe Kong, Jinlin Tan, Junqin Huang, Guihai Chen, Shuaitian Wang, Xi Jin, Peng Zeng, Muhammad Khan, and Sajal K Das. Edge-computing-driven internet of things: A survey. *ACM Computing Surveys*, 55(8):1–41, 2022.

[278] Wazir Zada Khan, Ejaz Ahmed, Saqib Hakak, Ibrar Yaqoob, and Arif Ahmed. Edge computing: A survey. *Future Generation Computer Systems*, 97:219–235, 2019.

[279] Nasir Abbas, Yan Zhang, Amir Taherkordi, and Tor Skeie. Mobile edge computing: A survey. *IEEE Internet of Things Journal*, 5(1):450–465, 2017.

[280] Efstathios Zavvos, Enrico H Gerding, Vahid Yazdanpanah, Carsten Maple, Sebastian Stein, et al. Privacy and trust in the internet of vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):10126–10141, 2021.

[281] Yifan Xiong, Yuting Jiang, Ziyue Yang, Lei Qu, Guoshuai Zhao, Shuguang Liu, Dong Zhong, Boris Pinzur, Jie Zhang, Yang Wang, Jithin Jose, Hossein Pourreza, Jeff Baxter, Kushal Datta, Prabhat Ram, Luke Melton, Joe Chau, Peng Cheng, Yongqiang Xiong, and Lidong Zhou. Superbench: Improving cloud AI infrastructure reliability with proactive validation. In Saurabh Bagchi and Yiying Zhang, editors, *Proceedings of the 2024 USENIX Annual Technical Conference, USENIX ATC 2024, Santa Clara, CA, USA, July 10-12, 2024*, pages 835–850. USENIX Association, 2024.

[282] Sa Meng, Liang Luo, Xiwei Qiu, and Yuanshun Dai. Service-oriented reliability modeling and autonomous optimization of reliability for public cloud computing systems. *IEEE Trans. Reliab.*, 71(2):527–538, 2022.

[283] Ronghui Xu, Hao Miao, Senzhang Wang, Philip S. Yu, and Jianxin Wang. Pefad: A parameter-efficient federated framework for time series anomaly detection. In Ricardo Baeza-Yates and Francesco Bonchi, editors, *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 3621–3632. ACM, 2024.

[284] Yinfang Chen, Huaibing Xie, Minghua Ma, Yu Kang, Xin Gao, Liu Shi, Yunjie Cao, Xuedong Gao, Hao Fan, Ming Wen, Jun Zeng, Supriyo Ghosh, Xuchao Zhang, Chaoyun Zhang, Qingwei Lin, Saravan Rajmohan,

Dongmei Zhang, and Tianyin Xu. Automatic root cause analysis via large language models for cloud incidents. In *Proceedings of the Nineteenth European Conference on Computer Systems, EuroSys 2024, Athens, Greece, April 22-25, 2024*, pages 674–688. ACM, 2024.

[285] Chuanpu Fu, Qi Li, Ke Xu, and Jianping Wu. Point cloud analysis for ml-based malicious traffic detection: Reducing majorities of false positive alarms. In Weizhi Meng, Christian Damsgaard Jensen, Cas Cremers, and Engin Kirda, editors, *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*, pages 1005–1019. ACM, 2023.