中国电信
CHINA TELECOM

云计算研究白皮书

2025年

中国电信云计算研究院

2025年12月31日

前言

在成立第二年的 2025 年底，中国电信云计算研究院按计划第二次发布年度云计算研究白皮书。白皮书主体结构按照云计算研究院 **“三个面向一个围绕” 的四大研究方向**组织，即：一、面向下一代云计算的研究，体现云计算的专业定位；二、面向云网融合的研究，体现对中国电信战略的承接；三、围绕智能算法的研究，体现云计算研究院综合基础理论、核心算法与技术创新的特点；四、面向新兴技术的研究，体现前沿研究的属性。每章开篇以**研究图谱 2025 版**的形式呈现本章的内容范畴和各个研究点之间的组织关系。每章第一节都包含**趋势分析**和**方向聚焦**，首先利用行业数据、技术进展、政策导向、代表案例等分析本章研究方向的整体趋势，然后基于趋势分析结论，结合云计算研究院常年学术研讨的沉淀，聚焦到两个或者三个最值得关注的热点子方向。全文四章共沉淀**十个热点子方向**，分布在十个章节中论述，其中每一个章节都详细讨论该热点方向包含的主要研究点和代表性研究工作。每章最后一节则延续去年的方式，首先借用 **Gartner 技术成熟度曲线（Hype Cycle）**，对本章研究方向所涉及的技术点作发展现状的逐一判断，然后在本章研究方向范围内提出**趋势展望**，提供**发展建议**。此外，本年度白皮书增加了第五章，首先阐述云计算研究院 2025 年提出的研究愿景 – 智能泛在云，最后对本年度白皮书作整体总结。

本年度白皮书延续一贯的内容风格，以国际国内的最新行业趋势为导向，以详尽的产业数据分析和全面的学术界进展梳理为主要论述依据，共**引用 IDC、Gartner、Mckinsey 以及信通院等机构的国际国内行业报告和各类技术白皮书 60 余篇，引用高水平论文近 700 篇**。十个热点方向的详细论述中也介绍了**云计算研究院 2025 年度研究成果中的 16 项**，各项成果均已在高水平会议/期刊发表，或者已被接收录用。下面简要介绍各章的特色内容。

第一章，面向下一代云计算的研究，首先分析了 2024 年国际国内的市场数据，解读了 **IaaS、PaaS、SaaS 的占比及变化趋势**，然后讨论了云计算开源和标准方面的进展，之后对行业软硬件技术热点做追踪，并利用近三年高水平论文统计梳理了技术热点，涉及 **14 个 CCF 收录的顶级学术会议的 2800 余篇论文**。统计结果显示，企业参与论文依然占据 1/3，新的变化是更多企业新面孔开始出现，例如中国电信等传统运营商、智谱等创业期企业。下一代云计算讨论了三个热点方向，分别是：（1）分离式数据中心架构和关键技术（2）面向 AI 场景的 PaaS 数据平台层技术（3）智能化云运维、可信安全与能效优化。

第二章，面向云网融合的研究，本年度围绕的重点是**中国电信 2025 战略升级 – 云改数转智惠，一方面分析解读，另一方面探讨在理论研究和技术创新上的承接方向**。首先借用云计算研究院参与撰写，于年底发布的《云网融合 2035 技术白皮书》，阐述战略升级的核心驱动力、**云网融合的科学理论内涵**以及由供给-运营-服务的三层体系构成的核心愿景架构。然后在三层架构上分别识别出前沿技术趋势，分别是云网一体化调度依然是核心、网络基础设施 DC 向 AIDC 全面转型和云边端能力加速分化与融合，并由此引出本章的三个热点方向：（4）云网一体化调度（5）面向智算的云网基础设施（6）云边端协同。

第三章，围绕智能算法的研究，首先分析了云计算与云网融合相关的各类智能算法的发展趋势和应用场景，包括**运筹优化、深度学习、强化学习、大模型**和 **AI 智能体**，然后借用本章第一个热点方向：（7）算法赋能云计算，详细论述了运筹优化、深度学习和强化学习在云计算和云网融合中的应用。本章第二个热点方向围绕 2025 年最火热的话题开展介绍和论述：（8）**AI Agent 和 Agentic AI**。

第四章，面向新兴技术的研究，首先是新兴产业与新兴技术的前瞻分析，涵盖政策、技术和国内外云厂商的代表案例，涉及的新兴技术包括**工业互联网、视联网、智慧金融、低空经济、6G** 和**量子计算**。然后借用本章第一个热点方向：（9）新兴技术及应用，详细论述了新兴计算、6G 和低空智能以及这些新兴技术对云网的新需求。本章第二个热点方向围绕另一个热点话题开展介绍和论述：（10）**AI 安全**。

第五章，智能泛在云，介绍了云计算研究院基于云计算技术趋势和中国电信战略所提出的**研究愿景**，即，立足于泛在融合的云网基础设施，依托于云计算系统和 AI 算法深度融合的未来云计算新范式。本章介绍了**智能泛在云的背景与特征、技术挑战与创新机会、定位与展望**。

目录

第 一 章

面向下一代云计算的研究

目前，世界各国正在加速推动云计算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，以 DeepSeek 为代表的人工智能大模型应用取得突破性进展，众多 AI+ 应用成为云增长的新引擎，加速推动全球云计算产业向智能化方向发展。本章将从上述云计算产业的新变化入手，探讨全球云计算技术的发展现状以及前沿技术演进趋势，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 2025：云计算产业和技术分析

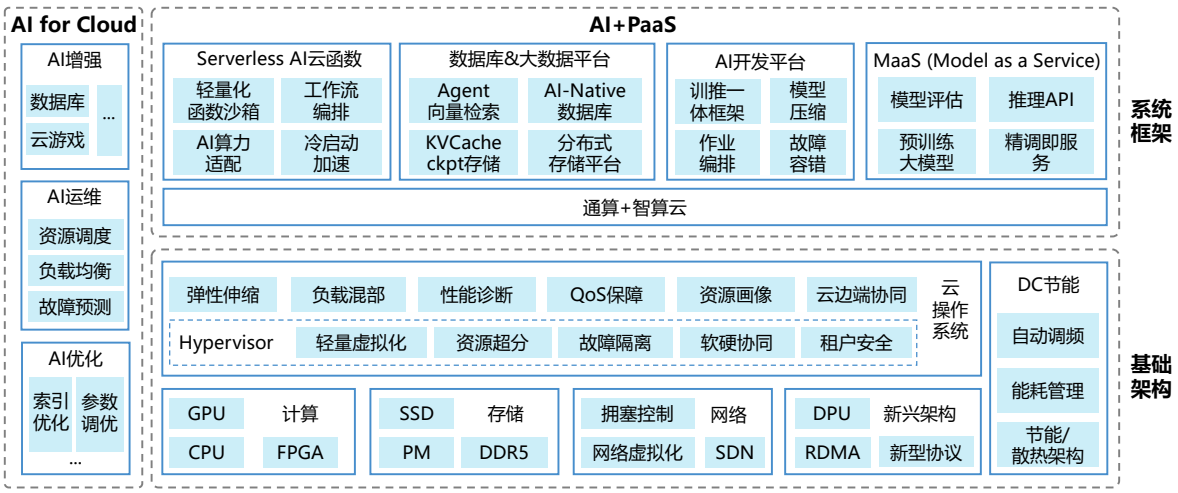


图 1.1: 云计算研究图谱（由云计算研究院总结形成）

传统的云计算服务模型主要由基础设施即服务 IaaS、平台即服务 PaaS 和软件即服务 SaaS 三大核心层面构成。随着人工智能技术的不断革新，AI 正从实验室走向千行百业，从工作场景深入生活场景。社会对算力的需求呈现出前所未有的普惠化、场景化与生态化特征。在此背景下，云计算服务模式正在加速向“AI+”深度转型，推动形成以 AI IaaS、AI PaaS、MaaS 和 AI SaaS 为代表的全产业链服务体系，构筑人工智能时代的新质生产力范式，图 1.1列举了当前阶段云计算领域的技术研究图谱。

在基础架构层，AI IaaS 成为支撑大模型时代的核心底座。基于 CXL（Compute Express Link）的内存池化架构显著提升异构算力资源的调度灵活性与利用率 [1, 2]；DPU 与 RDMA 等新兴架构技术强化了数据传输效率与系统控制能力 [3, 4]；GPU/FPGA/ASIC 等专用芯片与存算分离技术、高带宽存储介质深度融合，构建面向 AI 训练与推理的高性能智算云平台。同时，云操作系统通过 AI 驱动的资源调度、能耗管理与故障预测，实现数据中心的高效、低碳运行，践行绿色可持续发展。

在系统框架层，AI PaaS 正在重塑开发者体验。Serverless 计算平台结合冷启动加速、函数压缩与工作流编排，支持 AI 应用的极致弹性与快速迭代；面向 AI 开发的一站式平台集成训练框架、推理优化、向

量检索与模型压缩能力，降低开发门槛；数据库与大数据平台向“湖仓一体”、“实时分析 +AI 内嵌”演进，支撑复杂的数据科学任务。CI/CD 流程也扩展至 MLOps（Machine Learning Operations）范畴，实现模型交付的自动化与可追溯。与此同时，模型即服务 MaaS（Model as a Service）正成为连接模型能力与行业应用的关键枢纽 [5]。通过提供预训练大模型托管、精调接口、推理 API 及模型市场，MaaS 使企业无需从零训练即可获取先进 AI 能力，极大加速 AI 落地进程。

AI for Cloud 广泛应用于基础架构层和系统框架层两个层级，通过 AI 算法优化资源调度、网络拥塞控制、能耗管理与安全防护，云计算系统自身也变得更加智能、稳定与高效。这种“以 AI 优化云，以云承载 AI”的双向赋能机制，正在推动云计算进入一个自我进化、持续增效的新阶段。总的来说，在 AI 智能时代，云计算已不再仅是资源供给平台，而是演变为集 AI 算力供给、AI 能力构建、模型服务化与智能应用输出于一体的全栈 AI 服务平台。这一变革更催生出全新的商业模式与产业生态。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发展趋势。

1.1.1 趋势分析

过去一年里，AI 技术在云计算的各个方面加速渗透，正成为重塑云计算产业格局的核心驱动力，从算力需求、服务模式到应用场景，全方位推动着行业的创新与变革。接下来，本节将从全球市场格局、关键技术演进和行业开源标准三个维度，深入剖析云计算产业的最新变化趋势。

1.1.1.1 AI 技术驱动全球云计算市场持续变革

2024 年，Gartner 数据显示全球云计算市场继续保持稳健增长，规模达 6929 亿美元，同比增速为 20.3%；同时预计到 2030 年，全球云计算市场规模将接近 2 万亿美元 [6]。综合《云计算研究白皮书（2024）》 [7] 对 2021—2023 年全球云计算市场的系统研判，以及 Gartner 对未来五年云计算领域增长率的预测，可以发现当前云计算市场正经历关键转型：从高速扩张期步入结构优化期，增速虽有所放缓但更趋稳定。在此背景下，AI 技术的爆发式发展正成为打破原有市场平衡、重塑三大云服务版图的关键变量。

(a) 2024 年全球 IaaS，PaaS 和 SaaS 市场份额比例

(b) 近三年全球 IaaS，PaaS 和 SaaS 市场份额增速

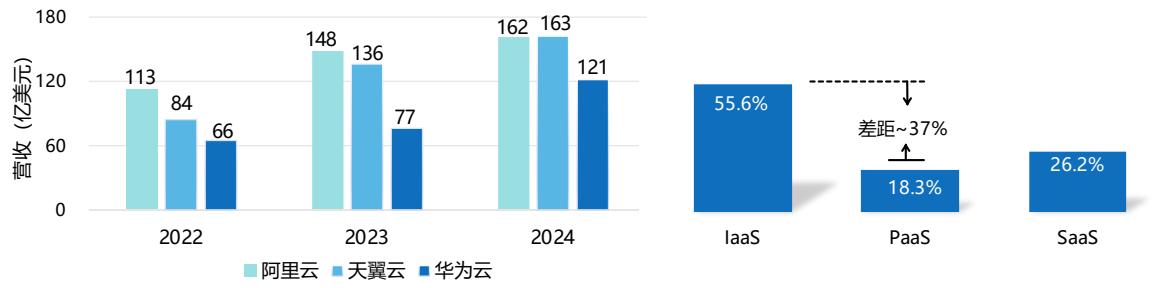
图 1.2: 全球 IaaS，PaaS 和 SaaS 市场份额分析

2024 年，全球云计算市场的增长情况展现出“稳中有变”的发展格局，AI 浪潮的持续深化和 AI Infra 的火热为 IaaS 市场注入了新的活力。如图 1.2所示，Gartner 等机构的数据 [6, 8] 指出 2023-2024 年 PaaS 与 SaaS 延续增长，增速分别达 19.2% 和 17.4%。最引人注目的是，AI 应用的火热带来了 IaaS 市场的强劲反弹：增长率从 2022-2023 年的 13.4% 跃升至 22.5%，成为三大云服务类型中增速最高的领域。一方面，Meta、字节跳动、天翼云等大型科技公司持续加码 AI Infra 建设，大幅扩充智算中心规模 [9]；另一方面，以 DeepSeek 为代表的 AI 独角兽则通过开源模式降低技术门槛 [10]，激发了更广泛的市场参与热情，推动中小企业和开发者对 IaaS 资源的需求快速增长。这种“头部企业重资产投入 + 开源生态普惠创新”的双轮驱动模式，正在重塑全球云计算市场格局，标志着产业进入 AI 深度融合的新增长周期。

2024 年，全球 PaaS 市场份额与 IaaS 市场份额基本持平，但累计份额已打破了过去 SaaS 长期占据主导地位的局面。具体而言，其市场份额分别达到约 1718 亿美元和 1707 亿美元，同时，IaaS 和 PaaS 的

市场比例分别提升至 28.6% 和 28.4%，虽然单一份额仍低于 SaaS 的 43%，但两者合计已超过 SaaS，显示出其在云服务市场中的重要地位，打破了过去 SaaS 长期占据主导地位的局面。这一结构性转变主要得益于生成式 AI 和大模型等新兴技术的推动，企业和开发者对底层算力和平台级服务的需求持续增长，使 IaaS 和 PaaS 成为驱动市场扩张的核心动力。生成式 AI 和大模型等新技术推动了企业和开发者对于底层算力资源和平台级开发环境的需求，促使他们选择高效灵活的云服务解决方案。尤其是 PaaS 平台，凭借一站式模型开发与部署能力，成为众多中小企业和开发者构建 AI 应用的首选。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持续攀升，功能也将进一步强化与拓展。与此同时，IaaS 市场的基础设施升级和规模扩展将继续为整个云服务生态提供坚实支撑。

全球云计算 SaaS 市场增速逐年放缓，这一趋势既源于行业发展的阶段性瓶颈，更与 AI 技术引发的行业变革密切相关。从市场基础来看，饱和态势已形成明显增长阻力。当前全球 99% 的企业已引入至少一种 SaaS 应用，美国大型企业的 SaaS 普及率更是高达 91%，办公协同、客户管理等核心场景的刚需客户增量锐减，剩余潜在市场的开拓不仅需要更高营销成本，还需适配小众化需求，整体获客效率持续走低。与此同时，IaaS 与 PaaS 的高速增长进一步分流资源。IaaS 依托 AI 训练等高算力需求保持强势增长，PaaS 因深度融合 AI 与大数据保持 20% 的增速，企业将更多预算投向这些支撑定制化 AI 服务的底层设施，进一步挤压了传统 SaaS 市场。而 AI 技术的渗透则从根本上重塑了 SaaS 的发展逻辑：生成式 AI 推动 SaaS 产品从传统流程自动化工具升级为智能决策伙伴，通过实时数据洞察、业务风险预判等能力重构服务价值；这种变革不仅催生了“按效果分成”等新型盈利模式，让传统按账号订阅的收费逻辑面临挑战，更因 GPT 等技术展现出的“需求即生成应用”能力，对标准化 SaaS 产品形成替代压力。



(a) 云厂商云计算业务营收（亿美元）　　(b) 云计算三大市场份额占比

图 1.3: 2024 年国内云计算厂商营收与市场份额占比

我国的云计算市场规模仍然保持着较高的增长态势，目前市场规模已达到 8288 亿元，同比增长 34.4%，增速远超全球平均水平。在国内云计算市场，随着产品服务竞争加剧和行业需求持续多样化，主要云服务厂商的云业务营收整体呈增长态势，但云厂商间的格局正出现新变化。如图1.3(a)所示，天翼云在 2024 年的营收首次超越阿里达到 163 亿美元，增速达 19.8%；阿里云营收 162 亿美元，增长 9.4%，增速放缓；华为云营收 121 亿美元，增速 22.4% [11]。造成这一格局变化有以下两个关键因素，一方面是国家对运营商云的政策支持力度持续加大，推动天翼云等电信系云厂商进行算力布局、在以政务云为代表的领域快速扩张；另一方面，行业用户对云服务的需求更加多元，各家云厂商深化混合云、行业云的布局，并通过降低价格与优化服务来争夺客户，加速了传统 IT 基础设施向云平台的迁移。因此，在价格、服务与政策等多重因素叠加下，国内云市场的渗透率进一步提升，导致云厂商排名与梯队结构的重新洗牌。

以智能体（AI Agent）、自动化模型服务、多模态生成等为代表的智算类业务驱动云计算市场从传统算力需求向“智算需求”转移，成为国内 IaaS 与 PaaS 的核心增量来源。2024 年国内云计算市场份额中，IaaS、PaaS 和 SaaS 市场份额占比分别为 55.6%、18.3%、26.2%（如图 1.3(b)所示）。根据 Gartner 预测，未来两年至少有 15% 日常工作决策将由智能体自主完成，33% 的企业软件应用将包含智能体。相比于 2024 年未产生智能体的阶段，智能体在运营商、制造业、金融服务等领域的规模化应用，直接拉动了企业对高

性能 GPU 以及大规模算力集群的需求，智能计算成为 IaaS 增长最快的份额。同时，智能体训练、推理的持续迭代需要自动化的数据管理、模型开发以及部署测试，促使企业进一步依赖云厂商提供的 AI PaaS、MLOps、向量数据库、MaaS 等平台级能力，从而也推动了 PaaS 业务收入显著提升。从技术发展趋势上看，智能体驱动下的算力需求、工具链需求和行业场景需求的三重叠加，将在未来成为支撑国内 IaaS 与 PaaS 高增长的主引擎。

1.1.1.2 软硬件创新驱动云基础设施持续进化

过去一年，云计算在硬件架构革新、软件系统智能化升级以及开源生态共建方面取得显著突破，形成以“硬件突破—软件革新”协同发展的技术发展格局。本部分聚焦行业实践，从硬件基础设施、软件平台能力两个方向梳理年度标志性事件与技术跃迁路径。

全球云计算硬件基础设施正加速向高性能、异构化与资源池化方向演进。例如华为发布的 CloudMatrix 384 超节点架构成为年度最具影响力的硬件创新之一。该架构采用全对等互联与全栈协同设计，集成了自研鲲鹏 CPU、Ascend 910C NPU 及高速统一总线 UB（Unified Bus）网络，构建了总算力达 300PFLOPs 的超大规模 AI 云底座。NVIDIA H200 GPU 已在 Amazon、Google Cloud 和 Microsoft Azure 大规模部署，搭载 HBM3e 显存，带宽达 4.8TB/s，配合 GB200 Superchip 与 NVLink Switch 系统，在千卡集群中实现通信延迟下降近 40%，显著提升大模型训练效率。AMD Instinct MI300X 作为重要替代选择，凭借 192GB HBM3 内存和 CDNA 3 架构，在 Meta、Microsoft 等平台落地应用，支持 AI 与 HPC 融合负载。Amazon 推出自研 Trainium2 芯片与 Graviton4 CPU 组合，构建端到端可控的 EC2 UltraClusters，支撑千亿参数模型训练，并探索基于 CXL 3.1 的内存扩展架构以缓解 GPU 显存瓶颈。Google 在其 TPU v5p 集群中引入液冷封装与动态调频技术，提升能效比，同时试点 GDDR7+CXL 混合内存方案，拓展通用内存容量。整体来看，云厂商正通过芯片自研、高速互联与内存服务化，推动硬件架构从“封闭堆叠”向“开放协同”转型。

在软件层面，云计算正迈向以智能调度、自主运维与语义感知为核心的下一代操作系统阶段。阿里云在灵骏智算集群中集成智能运维引擎，利用时序预测模型对 GPU 利用率、温度、显存占用等指标进行分钟级异常预警，结合历史故障图谱实现根因定位，SLA 违规率降低超 35%。Google Cloud 的 Autopilot [12] 系统能够结合机器学习分析容器历史负载，自动推荐最优资源配置；Monarch 系统日均处理超百万条监控流，支持跨区域性能诊断与容量规划。Microsoft Azure 将因果推断与知识图谱应用于告警聚合，将数千条原始事件归并为可操作的故障单元，缩短平均修复时间达 40% 以上；Microsoft 的 Azure Machine Learning 平台采用弹性训练调度器，动态增减分布式训练节点，在保障收敛性的前提下降低 30% 以上计算成本。Amazon 通过 DevOps Guru 实现基于无监督学习的异常检测，可识别 Lambda 函数冷启动激增、RDS 慢查询等典型问题，并提供修复建议；Karpenter 弹性节点控制器可在秒级内响应 Pod 调度需求，大幅提升 EKS 集群资源利用率。

1.1.1.3 云计算行业开源组织与行业标准布局

全球云计算开源生态进入爆发期，开源项目成为技术标准制定与产业话语权争夺的主要战场。国际上，Linux 基金会主导的 Open Acceleration Framework 整合 CUDA、ROCm 与 CANN 生态，推动 AI 加速器接口标准化，打破厂商锁定困境。在国内，开源力量同样迅猛崛起。华为推出的 Open YuanRong 项目，聚焦 AI 推理框架开源，兼容 PyTorch 与 MindSpore 模型，支持异构硬件自动优化，上线三个月即被 30 余家云服务商集成。天翼云则发布 TeleDB——一款面向云原生的分布式数据库开源项目，支持多模态数据处理与强一致性事务，在电信级高并发场景中表现优异，已被多家省级政务云采用。2025 年中国云计算开源项目不仅在数量上快速增长，更在关键技术自主可控与生态协同方面展现出强大生命力，正逐步改变全球云计算技术格局。

标准制定正从“辅助支撑”角色转变为引导技术路线演进的核心力量，成为产业协同创新的枢纽。2025 年云计算领域在标准化建设方面取得关键进展，全球主要标准组织与产业联盟加速推进技术规范制

定，推动异构算力协同、多云互操作、绿色低碳等共性能力的统一化与规模化落地。在国际方面，分布式管理任务组 DMTF（Distributed Management Task Force）持续推动 Redfish 标准演进，增强对现代数据中心基础设施的建模能力，支持 GPU、FPGA 等加速器资源的发现与管理，为未来 AI 工作负载调度和资源拓扑暴露奠定数据模型基础。同时，Internet 工程任务组 IETF 通过 SCIM（System for Cross-domain Identity Management, RFC 7643）、NETCONF/YANG 模型（RFC 6241）、I2NSF 安全策略框架等标准，正在构建身份、配置与策略语义层面的跨域协同机制，为解决“多云孤岛”问题提供技术路径。在国内，中国电子技术标准化研究院发布的《云计算异构计算资源池化技术要求》（草案/团体标准）提出了基于虚拟化与解耦架构的资源池参考模型，探索异构硬件互联技术在算力资源整合中的应用前景，华为、阿里、中科曙光等企业参与了该标准的技术研讨与验证试点 [13]。同时，开放原子开源基金会推动 OpenHarmony、OpenEuler 等项目与国家标准对接，实现“开源—标准—产业化”闭环。

2025 年云计算标准工作呈现出“技术引领、场景驱动、全球协作、产研联动”的鲜明特征。标准不再滞后于技术发展，而是前瞻性地定义接口、协议与评估体系，为技术创新提供稳定预期与规模化路径。为明确标准化与开源在行业中的关键地位和作用，以下将从多云协同、产品兼容、合规监管和生态创新等方面，具体阐述标准化与开源在云计算中起到的实际作用。

在多云协同方面，企业为避免供应商锁定普遍采用多家云服务，但不同厂商架构和接口差异较大，容易形成“信息孤岛”。统一标准与主流开源协议有助于规范数据交换与接口对接，实现多云间的资源共享和协同管理，降低迁移成本。在产品兼容方面，云存储、数据库及安全组件若缺乏统一接口，企业跨云迁移与系统集成将面临较高技术门槛。通过标准化约束接口规范与性能指标，并结合开源软件的开放开发与测试机制，可显著提升云产品的互操作性与可靠性。在合规监管方面，随着《数据安全法》《网络安全法》等法规的实施，云服务必须满足安全与合规标准才能合法运营。行业标准为监管评估提供了统一依据，开源安全工具和合规方案则帮助企业更便捷地落地合规要求、降低运营风险。在生态创新方面，统一标准降低了技术集成与合作的门槛，促进第三方服务在云平台上的繁荣生长。开源通过开放源代码吸引全球开发者参与，推动技术快速迭代；谁能主导关键开源项目，谁就更容易在产业生态中掌握技术主导权和话语权。

1.1.2 方向聚焦

为全面把握全球云计算技术发展方向，本节从云领域学术界与工业界关于技术研究前沿探索出发，系统梳理并深入分析 2025 年度云计算领域的关键进展与趋势演进。

通过整合产业一线的创新成果与学术前沿科研动向，为政策制定者、技术研发人员提供权威、前瞻的洞察参考。本节持续跟踪调研了近 3 年和云计算产业相关的 14 个顶级学术会议（ASPLOS，SC，SOSP，VLDB 等）收录的 2,800 余篇高水平论文（以 CCF-A 类为主），从中筛选出近 860 篇云计算领域有企业参与的已发表文章。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、AI 与系统、任务调度与编排框架、中间件、AI 加速器、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，本节将上述涉及的所有研究整理为 10 个基础研究方向，分别为 MLSys、数据库、DC 与服务管理、文件与存储系统、加速器硬件、OS 与分布式系统、分离式数据中心、AI for Cloud、虚拟化技术以及量子计算，以构建一个从硬件到软件、从基础设施到智能优化较为完整的云计算热点研究方向洞察（见图 1.4）。

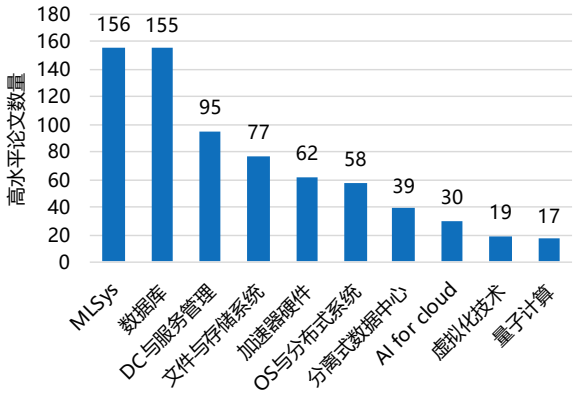
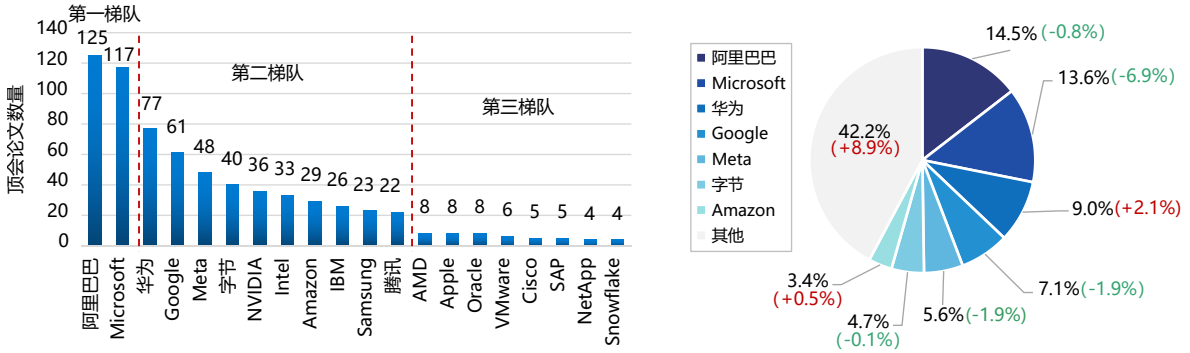


图 1.4: 近三年企业参与的云计算热点研究领域文章发表数量

相比 2024 年的统计结果，受益于大模型的快速崛起，近三年的 MLSys 方向的发文数量和硬件加速器方向论文呈现稳步增长，成为了云计算研究体系中最受关注、增长最快、企业参与最集中的方向。而 DC 与服务管理和文件存储相关领域的发文数量呈现少许下降趋势。量子计算技术的发文数量逐年攀升，是未来的潜在热点领域。



(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比及相较去年变化

图 1.5: 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活跃度保持不变，大约 31% 的近三年学术成果有企业参与，且大部分以国外厂商为主。以华为和阿里巴巴为代表的国内厂商学术影响力增加显著。如图 1.5(a)所示。本节将发文数量较多的企业划分为三个梯队，相比 2024 年统计结果，阿里云近三年发文量超越 Microsoft 成为第一，华为超越 Google、Meta 进入前三，位居第二梯队首位。而 Samsung 进入第二梯队并超越腾讯，AMD，Apple，Oracle，Cisco，SAP，NetApp 和 Snowflake 成为第三梯队新晋成员，VMware 则保持第三梯队不变。

相比 2024 年白皮书统计结果，近三年企业参与发表的论文数量的“分布倾斜”现象有所缓解。从图 1.5(a)可以看到，贡献超过 80% 研究成果的头部企业数量从去年统计结果的 13 个增加到 20 个，其中新增了不少新面孔，这也代表着有更多的企业初步在云计算和系统领域学术研究上崭露头角。图1.5(b)进一步展示了阿里，Microsoft，华为，Google，Meta，Amazon 这些拥有云计算营收业务的头部企业研究成果占总体发文数量的比值（红色代表相比 2024 年白皮书统计结果有增长，绿色则代表有所下降），未展示的企业则全部归类于其他类别。可以看到，除华为和 Amazon 发文量占比有轻微增长外，其他头部企业研究成果发表论文数量均有所下降，例如阿里云论文发表数量占整体比重降低 0.8%，Google 发文量占比下降 1.9%，而 Microsoft 发文量占比下降幅度最大接近 7%，其他企业则上升了近 9%。

创新驱动力在系统领域持续扩散，企业新面孔的加速涌现，为学术研究和产业发展注入了新鲜血液。相比 2024 年白皮书统计结果，在今年的系统领域顶级学术会议上，有七十余家单位如商汤科技、智谱科技、超威半导体、百度、联想，以及电信、移动等首次亮相，涵盖了 AI 初创、互联网公司、传统 IT 与半导体巨头、通信运营商等多元行业。它们在体系结构、操作系统、数据库、云计算等多个顶会发表了聚焦大模型推理、分布式调度、云原生等前沿主题的论文，展现出学科交叉和新兴应用的广泛需求。企业直接参与论文署名愈发普遍，研究紧贴实际生产和业务需求，产学研融合不断加深。同时，国际机构的积极参与进一步推动了全球学术交流。顶会对新团队和新思想的包容性持续提升，越来越多新兴单位能够在学术舞台崭露头角，行业创新生态更加开放与活跃。

综上，2025 年云计算领域在软硬件协同、AI 原生架构和分布式系统等方面取得突破，学术与产业的深度融合加速了新技术的落地。云计算正围绕大模型驱动的算力升级、数据平台智能化以及云服务高可用与安全能效三大主题进行演进。具体来看，分离式数据中心架构与关键技术通过异构算力、内存池化及新型互连技术，显著提升了资源利用率和远程内存访问效率，为大模型场景下的高效调度与弹性扩缩容提供支撑；面向 AI 场景的 PaaS 数据平台层技术聚焦于数据库和存储系统的创新，推动高性能、可扩展、智能化数据平台的构建，以满足 AI 业务对海量数据管理与加速的需求；智能化云运维、可信安全与

能效优化方面，通过 AI 运维和安全机制的不断完善，实现了云服务的高可用、可信与绿色发展。由此，下文将围绕上述三大热点方向进行系统梳理与深入分析。

1.2 热点方向一：分离式数据中心架构与关键技术

在云计算和大数据时代，数据中心面临着资源利用率、弹性扩展和高效运维等多方面的挑战。传统的数据中心架构已难以满足日益增长的业务需求和技术演进。分离式数据中心架构作为一种创新模式，通过资源池化与功能解耦，实现了更高的灵活性与可扩展性，为云服务和新型应用场景提供了坚实基础。本节将围绕分离式数据中心架构展开，重点介绍其关键技术及发展趋势。具体而言，后续三个小节将分别探讨弹性可扩展的云数据中心资源优化、面向资源池化的分离式数据中心架构，以及支持分离式数据中心架构的软件栈等关键问题。为更好地理解相关技术路径与研究进展，表 1.1重点遴选了部分具有代表性的关键研究成果。

表 1.1: 头部企业重点关注的分离式数据中心架构与关键技术研究领域

研究点	研究方向概述	主要会议	研究主要关注点与代表性工作
弹性可扩展的云数据中心资源优化	传统云数据中心架构日益暴露出资源搁浅和弹性粒度不足等问题。随着规模扩大，提升资源利用率、采用分层存储和动态收割技术进行协同调度，已成为云服务商应对资源闲置与弹性需求的核心关注。	OSDI SOSP ASPLOS EuroSys NSDI	• **内存分级、资源动态收割与复用：** Meta、Google 和 Microsoft 广泛研究了云基础设施中关于使用 SSD、NVM、CXL 内存等异构存储介质进行动态卸载能力 [14, 15, 16]，中国电信云计算研究院进一步探索了该场景下的多租资源公平分配，以减少共置负载性能劣化 [17]；Microsoft 针对数据中心闲置资源开展关于智能动态资源收割与复用的一系列研究工作 [18, 19, 20]，实现了在云平台中动态、安全地回收和复用虚拟机暂时闲置的计算和内存资源，显著提升了资源利用率和服务器效能。 • **智算资源细粒度管理：** NVIDIA 提出了一系列 GPU 资源共享技术。Kimi、DeepSeek 以及华为等深入研究了智算云基础设施中不同应用中不同阶段的资源需求，通过算力、显存、DRAM 等不同资源的细粒度管控，提升计算与存储效率，实现整体智算资源的效率提升 [10, 21, 22, 23]。
面向资源池化的分离式数据中心架构	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	• **RDMA、NVLink 高速互联内存池：** Google 通过细粒度、高效的远程内存分配机制，解决了池化架构中“分配开销与内存浪费”的兼顾挑战 [24]；还基于现有 RDMA 内存池架构中的高可用问题，研究如何减少分离式架构带来的爆炸半径扩大影响 [25]；阿里通过将集群中的 GPU 中显存统一管理，对推理服务过程中产生的数据统一放置，实现多 GPU 资源之间的资源共享，减少 GPU 资源的“碎片空间”，提升了显存的使用率 [26]。 • **CXL 共享内存池：** 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 [1, 27]；阿里云利用 CXL 交换机，实现云数据库的内存池化和数据共享 [28]。
支持分离式数据中心架构的软件栈	支持分离式架构的软件栈主要集中在简化编程复杂性、提升远程资源访问效率、优化资源池化与调度策略，以及增强系统可扩展性与高可用性等方面，为大规模异构资源的统一管理与高效利用提供支撑。	OSDI SOSP NSDI ATC HotOS	• **分离式操作系统:** 华为提出 FlacOS 操作系统，通过在内存互联的机架级架构中实现内核数据结构的共享和无锁同步机制，使得单一操作系统统一管理机架级资源，解决传统分离式资源管理带来的同步瓶颈和故障恢复挑战 [29]; 天翼云提出“聚合计算”产品理念，通过将池化算力资源按需聚合，为 HPC 等场景提供高效、灵活的算力服务 [30]。 • **分离式运行时:** 华为云在分离式数据中心运行时方面，提出通过分层接口和声明式 API 实现数据系统与硬件的解耦，提升了资源利用率和系统扩展性。随后，又通过挖掘多线程程序的异步性，进一步优化了分离式内存的访问性能和编程易用性 [31, 32]。 • **IR 运行时:** Google 和 OneFlow 针对 AI 场景下的多样的硬件资源集群，提出了支持算子粒度任务执行的运行时，通过统一抽象算子来支持不同 CPU、GPU、FPGA 等异构硬件，实现不同算力需求与资源之间的高效匹配 [33, 34]。

1.2.1 弹性可扩展的云数据中心资源优化

现行云计算资源分配粒度粗，资源利用率不高且成本居高不下，高效利用迫在眉睫。追求像用水用电一样灵活地使用资源是云计算发展的核心目标。然而，受限于硬件体系结构和操作系统抽象，且高速网络互联、内存、存储等模块发展速度出现严重不均衡现象，使得主机间资源难以高效共享。人工智能

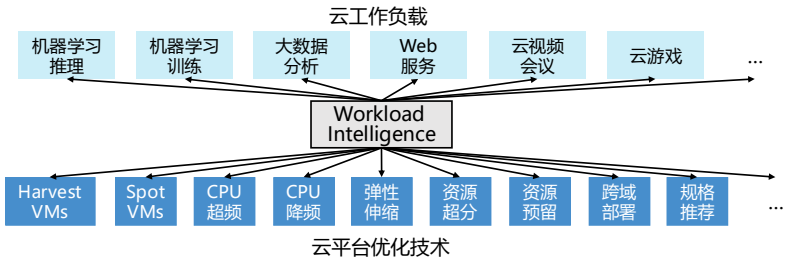


图 1.6: Workload Intelligence 概览 [35]

的崛起加剧了高投入与低资源利用率之间的矛盾。Microsoft、Google 等云厂商宣布上百亿美元建设 AI 专用数据中心，以 GPU 为核心配置成为了云厂商提供的主流基础资源。然而，主流云服务商采用大颗粒度“切割” 物理服务器资源（如 CPU、内存、GPU），以虚拟机或容器实例售卖，进一步限制了租户的资源灵活性，导致资源搁浅和成本浪费。Microsoft Azure 等公开数据表明，数据中心内存搁浅比例高达 6 - 30%，GPU 算力使用效率仅仅在 30% - 50%，而内存成本占物理服务器总成本的 37-50% [27]。除此之外，随着大数据业务和内存密集型应用的持续增长，云业务普遍采用大容量内存缓存以及算力独占的方式以提升性能，负载长期占据大量内存与计算资源，资源分配率居高不下，但实际高频访问的数据仅占很小比例。为此，学术界与工业界在近些年逐步开始探索通过部署低成本存储介质（如 SSD、NVM 等）以及基于高速互联技术（CXL、RDMA 等）的资源池化技术尝试解决。

针对资源搁浅与成本浪费的挑战，业界已积极探索远内存、自动资源配置等多种技术路径，以提升资源利用率和降低成本。Meta、Google 和 Microsoft 等公司，广泛研究了如何利用利用 SSD、NVM、CXL 内存等异构存储介质进行动态数据卸载 [14, 15, 16]。比如，中国电信云计算研究院针对多租户环境下不同负载间的内存资源竞争问题，提出了 QoS 感知的分级内存管理框架 Vulcan。该框架设计了基于负载特征的用户态内存页面迁移机制，显著提升了多应用场景下的灵活性与适应性。通过工作负载敏感性的快速内存容量动态公平分配策略，有效避免了传统热度驱动分配导致的 “冷页困境”，保障了延迟敏感型与吞吐量型负载的性能隔离与公平性。实验结果表明，Vulcan 在云服务多租户内存资源管理领域展现出显著优势 [17]。除此以外，Microsoft 在数据中心闲置/搁浅资源的收割方面持续创新，先后推出用以高效收割闲置 CPU、内存资源 [18, 20, 19] 的 Harvest 系列 VM，实现多资源联合调度与细粒度分配，显著提升了资源利用率和业务保障，有效推动了云基础设施的智能化和降本增效。然而，目前这些优化方式多为平台单向、透明管理，没有租户的直接参与。虽便于部署但受限于狭窄的资源分配接口，云平台难以直接洞察租户实际需求，导致实际响应滞后且效率仍有优化空间。随着云应用复杂性提升，协作式资源管理逐步成为趋势。以 Microsoft 在 SC ’25 大会提出的 WI（Workload Intelligence） [35] 框架为代表，新型协作机制支持租户主动表达业务特性（如可用性、延迟容忍度等），平台则智能匹配并启用多种优化机制（如自动扩缩容、Spot/Harvest VM、超频/降频、区域迁移等），显著提升资源利用率和经济性，图 1.6展示了 WI 的概览。研究表明，WI 框架可为租户平均节省约 48.8% 的资源成本，并提升绿色低碳水平。

针对智算中心的巨大投入与 GPU 资源使用效率的低下，业界与学术界积极探索不同租户间资源划分与调度，任务资源度量与抢占等多种技术，以此提升资源的使用效率。NVIDIA 在硬件架构、驱动层和软件栈上为智算中心的多场景混合运行构建了完善的资源共享机制。CUDA 的 Context、Stream 与 Hyper-Q 在软件与运行时层面提供了基础的并行与软隔离能力；MPS（Multi-Process Service）在进程级别上将多进程请求合并到同一 GPU 上下文中，提升了多任务并发度。针对多租户环境的强隔离场景, NVIDIA 又提出了 MIG（Multi-Instance GPU）技术，使得云原生环境下能够实现划分物理 GPU。在此基础上，学术界也围绕不同服务的资源使用模式开展了一系列工作。SpotServe [36] 基于多实例负载变化，提出使用可回收实例来降低服务成本并提升资源效率。FlexLLM [23] 则从单实例内部的多应用共存出发，利用推理和微调在计算和显存需求上的互补，实现两类资源的同时提升，为不同服务的资源共享提出新方案。MuxServe [37] 更进一步从单个应用内部特征分析，识别推理过程中密集和内存密集阶段的差异，通过跨请求进行模型的请求阶段组合，为模型的服务模式提供了新思路。

1.2.2 面向资源池化的分离式数据中心架构

分离式资源池化数据中心架构，通过算力、存储、网络三大资源的解耦与池化，有效应对了生命周期失配、资源利用不均、弹性调度不足和协同效率瓶颈等四大挑战。随着数据中心规模的持续扩展和业务形态的日益复杂，传统数据中心架构正面临多重挑战。首先，数据保存周期远长于服务器硬件的更新周期，导致数据迁移与运维成本显著增加，存储与算力资源的生命周期严重失配。其次，资源利用在时空维度上呈现显著不均衡，部分计算节点或存储设备长期处于低负载状态，而高峰期资源紧张，整体利用率难以提升。第三，云原生应用不断涌现，对计算与存储资源的弹性分配提出了更高要求，传统架构难以满足其动态扩缩和敏捷调度的诉求。最后，数据中心在算力、存储和网络资源之间的协同效率面临瓶颈，资源孤岛和跨域性能损耗制约了整体服务能力。针对上述困境，分离式数据中心架构应运而生，正如图 1.7所示，其核心理念是通过资源池化实现算力、存储与网络的解耦与独立调度。一方面，多元化业务场景驱动算力异构化发展，异构计算资源池可根据任务类型按需分配 CPU、GPU、FPGA 等多种算力，实现高效资源利用。另一方面，低时延网络技术的发展为内存与磁盘的分离及池化提供了技术基础，网络层的优化有效降低了数据访问延迟，支撑资源池间的高效协作。此外，新型应用不断推动高效共享存储的发展，存储资源池不仅提升了数据访问的弹性和可靠性，还为多租户环境下的数据隔离和共享提供了保障。总体而言，面向资源池化的分离式数据中心架构通过算力、存储、网络三大资源的解耦与池化，显著提升了资源利用率与服务弹性，增强了对新兴业务场景的适应能力，为下一代数据中心的发展奠定了坚实基础。

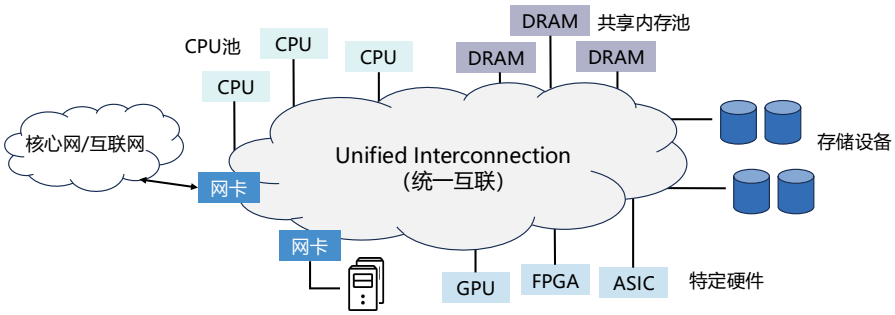


图 1.7: 分离式数据中心架构示意

分离式数据中心通过高速互联实现算力、内存和存储资源的池化与解耦，推动了 RDMA 和 CXL 等技术的应用，同时也带来了内存高可用性和大模型显存池化等新的挑战与机遇。在分离式数据中心架构中（如图 1.7），算力、内存和存储等各类资源通过高速网络实现解耦与互联，每个服务器节点通常专用于某一类功能，如计算、内存或持久化存储，从而构建出多样化的资源池。应用程序可以灵活地从这些资源池中获取所需资源，实现高度弹性的扩展能力。目前，计算与持久化存储的分离已在业界广泛落地，内存资源池化则成为新的研究热点。分离式内存技术（Disaggregated Memory）通过将远程服务器的未使用内存或共享内存池纳入统一管理，打破了单机内存容量的限制，提升了资源利用率。随着高性能互联技术（如 RDMA、CXL 和 UB）的发展，内存池化可以在机架内或机架间灵活扩展，实现集群级别的资源调度与弹性分配。然而，远程内存访问带来的性能开销、故障影响范围扩大以及资源管理复杂性等挑战，仍需进一步技术突破。当前，RDMA 和 CXL 等技术正推动内存池系统的创新，相关容错与成本优化机制也在持续探索。工业界如国际上 Meta 和 Microsoft Azure 和国内阿里已提出原型方案，但成熟的分离式内存系统仍处于发展阶段。在此基础上，CXL 与 RDMA 为内存池系统的构建提供了关键技术支撑，内存高可用性成为系统落地的核心挑战，此外，面向大模型的显存池化也成为分离式架构下的新兴研究方向。

基于 RDMA 等技术的高速互联内存池系统有效缓解了资源搁浅问题，但距离工业落地仍有挑战。基于 RDMA 以及 NVLink 的远端内存访问技术，将多台服务器的内存与显存整合为统一池，实现数据透明迁移，有效扩展本地内存资源，提升大规模机器学习等负载的性能。其中，在内存管理中，Fastswap [38] 结合

远内存感知调度，提高整体吞吐量。然而，这类系统在多应用并发场景下容易发生性能干扰，Canvas [39] 通过交换路径隔离，为每个应用分配独立的交换分区和带宽，实现自适应优化，显著减少了性能波动。此外，部分方案如 AIFM [40] 和 Carbink [25] 将远存管理显式暴露给开发者，要求应用自行管理远程内存，虽然提升了灵活性，但增加了开发复杂度。在显存管理中，Aegaeon [41] 中根据不同模型请求的实时负载，动态决定模型在 GPU 的资源占比，并采用 Token 级细粒度调度实现灵活的资源分配。通过低开销的 KVCache 管理与上下文切换，使多模型共享显存成为可能。进一步的，eLLM [42] 将模型推理过程的所有模型权重、激活与 KVCache 在统一的虚拟显存池中进行管理，并解耦虚拟地址与物理显存的映射构建可扩展的显存抽象。Infinite-LLM [43] 通过自适应、分布式注意力机制，将 KVCache 拆分为细粒度单元并跨节点动态放置，实现无感从集群空闲实例分配内存，实现全局范围的灵活、高效内存池化。更进一步的，为了更高效的使用显存空间，Mooncake [21] 在以 GPU 为核心的池化分级存储中提出了基于预测的早期拒绝策略与启发式热点迁移机制。通过缓存副本平衡跨实例间数据的复用热度，从而提升缓存复用效率。尽管资源逻辑池化已经取得显著进展，但保障池化后的可靠性仍是不可或缺的关键一环。为了提升容错能力并降低存储开销，Google 的 Carbink 系统 [25] 采用纠删码（Erasure Coding）替代传统复制机制，将本地驱逐的数据编码后分散存储于多个远程节点，同时结合远程内存压缩和可卸载奇偶校验计算，实现了高效冗余与快速恢复，显著降低了故障带来的影响。

基于原生内存语义的 CXL 共享内存池系统催生了一系列架构创新。近年来，远程内存管理技术不断演进，基于 CXL 技术的远存管理则带来了新的突破。CXL 打破物理服务器的内存边界，实现池化和跨主机动态分配，极大提升了资源利用率。Microsoft Azure 的 Pond 方案采用机器学习预测负载时延敏感性，将不敏感的虚拟机优先分配池化内存，并通过 QoS 监控和自动回滚机制缓解“内存搁浅”问题 [27]。学术和产业界也在积极探索 CXL 远存管理的新模式。Tigon [44] 系统针对分布式数据库场景，利用 CXL 内存实现跨主机原子操作，显著降低同步延迟，提升事务处理性能。PolarCXLMem 聚焦云原生数据库，通过 CXL 交换机实现内存池化，并创新 PolarRecv 机制支持数据库瞬时恢复和缓冲池热身，同时提出新型一致性协议，提升多节点数据共享效率。实验证明，基于 CXL 的管理方案不仅提升了内存资源的灵活性和利用率，还显著改善了数据库等关键应用的性能。尽管如此，CXL 远存管理同样面临着“爆炸半径”问题。阿里云提出通过基于引用计数的分布式内存管理机制，即使部分节点故障或进程崩溃，也能保障远程内存资源的安全释放和回收，有效避免内存泄漏和双重释放，提升了系统的弹性和可靠性 [45]。需要指出的是，当前主流的研究更多是在提升远程内存池系统对外围故障的应对能力。例如，纠删码和分布式管理可以降低单点失效带来的数据丢失风险，热迁移和一致性协议则有助于快速恢复业务和保障多节点协同。但这些机制本质上仍是围绕数据和资源管理展开，尚未从根本上解决内存池底层硬件或核心服务发生故障时所带来的爆炸半径问题。如何提升内存池自身的容错和隔离能力，仍是未来远程内存池系统落地部署的重要挑战。

1.2.3 支持分离式数据中心架构的软件栈

尽管分离式数据中心架构通过将计算、存储、内存等关键资源进行解耦与池化，为云服务带来了更高的灵活性与可扩展性，但这种架构也带来了编程复杂性提升、远程资源访问效率降低、资源调度与管理难度增加等新挑战。为此，国际国内均开始围绕分离式操作系统、分离式运行时及 IR（Intermediate Representation）运行时的软件栈进行创新设计，图 1.8展示了三者关系。

高速互联技术的演进，正在催生以完全资源分离和分布式部署为核心特征的分离式操作系统新架构。传统操作系统（如 Linux、Windows 等）通常是单机、单内核设计，即所有资源（CPU、内存、存储、网络等）的管理和调度都由一个内核负责，资源被严格限制在一台物理机。分离式操作系统将操作系统的各个功能模块（如内存管理、存储管理、网络管理等）拆分出来，分别运行在不同的专用硬件或服务器上，通过高速网络互联，实现资源的“池化”和“按需分配”。也就是说，分离式操作系统把操作系统的服务变成了“分布式服务”。2018 年，美国普渡大学提出 LegoOS [46]，首次提出了 splitkernel 架构，其将传统操作系统功能拆分为多个分布式监控器，各自运行在独立硬件组件上，通过网络消息协同完成资源

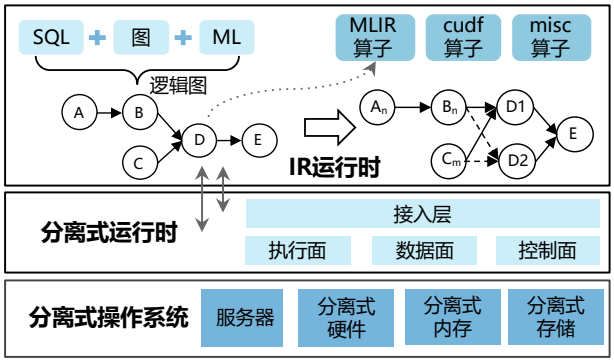


图 1.8: 分离式数据中心架构的软件栈

管理与故障处理。这种架构突破了操作系统依赖单一物理服务器的限制，实现了 CPU、内存、存储等资源的彻底解耦，使得资源分配不再受限于物理边界，极大提升了数据中心的资源利用率和弹性，该研究成果获得了当年大会 Best Paper Award。2021 年 VMware 提出的 NrOS [47] 进一步关注多核、多节点环境下操作系统内核的可扩展性和正确性。NrOS 通过高效的内核状态复制与共享机制，简化了内核同步，提高了系统的扩展性和可靠性，为分离式和分布式硬件环境下的操作系统开发提供了新的思路。与此同时，分别来自英国爱丁堡大学和上海交通大学提出的 Aggregate VM [48] 和 GiantVM [49] 等系统则从虚拟化层面推动了分离式资源管理的落地。其通过分布式 Hypervisor，将来自不同物理主机的碎片化资源临时聚合为一个虚拟机实例，支持 vCPU、内存和 I/O 设备的动态迁移和调度，提升了资源利用率。

分离式软件运行时在提升远程内存可用性和资源弹性的同时，也引入了访问延迟和编程复杂度等开销，需要在性能收益与系统复杂性之间权衡。分离式软件运行时是一种专为分离式内存架构设计的运行时系统，它的核心目标是让应用能够高效地利用分布在不同服务器上的远程内存资源，从而突破单机物理内存的限制，实现资源的弹性扩展和高效利用。在传统的数据中心架构中，计算和内存资源被固定地绑定在同一台服务器上，导致资源利用率不均衡，部分服务器内存闲置而部分服务器因内存不足而性能受限。分离式运行时通过网络将各服务器的内存池化，使得应用在本地内存不够时可以直接访问远程内存，避免了频繁的磁盘换入换出带来的性能瓶颈。然而，远程内存访问带来的微秒级高延迟成为新的技术挑战。传统做法是通过多线程同步编程模型来隐藏远程访问延迟，但这种模式下频繁的线程切换不仅带来调度开销，还会破坏数据局部性，造成缓存失效和更多的 CPU 资源浪费。华为云提出一种新型分离式运行时框架 Beehive [32]，其基于协程的异步执行模型，允许每个线程在遇到远程内存访问时无需阻塞，并将代码自动拆分为多个可异步调度的小单元，通过高效的协程调度机制实现远程访问的高并发和低开销，以最大程度地保持数据局部性。Beehive 进一步借助 Rust 语言的类型系统自动将传统同步代码转换为异步代码，极大简化了开发者的编程负担。通过这些创新，分离式运行时不仅让应用能够像使用本地内存一样灵活高效地使用远程内存，也显著提升了资源利用率和整体性能，为云数据中心的弹性计算和大规模数据处理提供了坚实的基础。

支持不同异构硬件平台之间无缝适配与高效执行，提升 AI 系统可扩展性与兼容性的多级中间表示运行时框架。 CPU、GPU、NPU、FPGA 以及各类专用加速器的不断涌现，AI 系统正面临着通用场景中前所未有的异构性挑战。直接通过软件框架适配不同硬件，开发成本高并且难以在不同平台间保持性能的一致性。因此，需要构建一个能够承接计算图并无感适配多类型后端的中间表示。MLIR（Multi-Level Intermediate Representation）[50] 提出了多级中间表示的设计，阐述了多级中间表示如何支持跨域优化与可扩展编译器基础设施。ONNX-MLIR [51] 将 ONNX 模型映射到 MLIR 中，利用 MLIR 与 LLVM 的协同作用，将统一的 ONNX 模型编译到不同硬件架构上，提升了跨平台的兼容性和性能。通过多层 IR 优化编译方法，解决了 FPGA 等可重构硬件的开发难度和优化复杂性，提高了系统实现、调试和扩展的效率 [52]。Google 的开源项目 [53] 通过结合硬件特性和运算图优化，有效地提升了 TensorFlow 在各种硬件平台上的执行效率。Intel 的开源项目 [54] 将 MILR 扩展到其硬件，使 ML 通用编译器基础设施能够无缝兼容底层

硬件，是 MLIR 在工业硬件厂商端适配的典型实践。

1.3 热点方向二：面向 AI 场景的 PaaS 数据平台层技术

以大模型为代表的 AI 技术的飞速发展尤其是 DeepSeek 等模型的正式开源，驱动国内外各家云厂商的平台层技术投入重心逐步向 AI PaaS 倾斜。举例来说，各大厂商正积极布局 Serverless 化的大模型推理服务；此外，AI 场景不仅需要对复杂数据进行大量的实时处理，也对海量数据的高性能、低成本存储提出了更高要求，进而催生了诸多新的技术挑战。图 1.9展示了 Serverless 计算、数据库服务、存储技术三者在智能时代的平台层技术发展重心，表 1.2总结了近年具有代表性的关键成果案例。

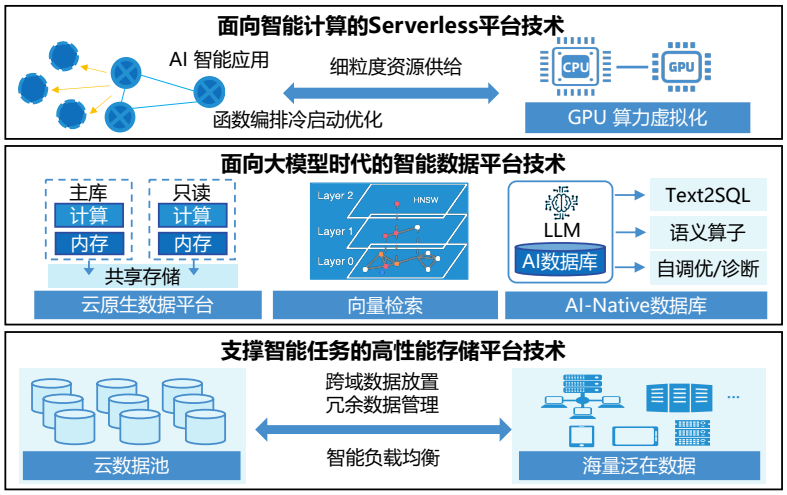


图 1.9: 面向 AI 场景的 PaaS 数据平台层技术概览

1.3.1 面向智能应用的 Serverless 计算平台技术

在数字经济加速渗透、以大语言模型和 AI Agent 等为代表的智能应用广泛落地的当下，云计算行业正处于向智能泛在云转型的关键时期。天翼云作为国家云基础设施建设与服务提供的主力军，既要满足海量的内部业务（如智能客服、智能运维、用户行为分析）与外部用户（如中小微企业 AI 建模、智慧城市边缘智能计算等）对 GPU 算力的多样化需求，又面临着传统云计算服务模式下 GPU 算力供给的多重瓶颈。基于当前流行的 Serverless 编程范式，各大云计算厂商纷纷推出了基于函数即服务 FaaS（Function-as-a-Service）编程模型的 AI 云函数产品，旨在向用户提供快速部署、高度弹性以及按需付费的智能应用开发平台服务。然而，AI 智能应用对 GPU 算力的需求呈现“多样化、碎片化和动态化”的特征，AI 模型较大的初始化加载时延也制约着服务弹性，使得 Serverless 平台设计和优化面临新的问题挑战。构建面向 AI 工作负载优化的 Serverless 运行时与资源编排体系，已成为主流云服务商重点关注的技术方向。

构建低成本、高弹性的 GPU 云函数沙箱，提供粗粒度算力分配能力。粗粒度的算力分配是当前云计算 GPU 资源供给模型存在的首要不足。传统物理 GPU 裸金属服务器或 GPU 云主机多以整机、整卡为单位分配，尽管这种方式常用于 AI 模型的训练，但在推理场景下，以“百 MB 级显存、分钟级算力”为需求的中小型应用往往占据业务主体。粗粒度的 GPU 算力供给方式不仅导致用户侧大量算力闲置浪费，也变相提高了算力计费成本，降低用户黏性；一些 FaaS 云厂商例如阿里云，Microsoft 推出了按需付费的 GPU 云函数 [55]，例如，Microsoft Azure 容器应用在某些地区为 Serverless 模型训练和推理提供了 NVIDIA A100 GPU，阿里云函数计算支持以 1GB 设备内存为单位为函数配置 NVIDIA V100 GPU [56]。然而，这些 FaaS 平台中的 GPU 函数分配粒度仍然较粗，无法精确匹配许多小模型的需求，导致了严重的资源浪费。尽管先前大量研究广泛采用 MPS [79] 技术来共享 GPU 设备从而改善利用率，但这些技术无法应用于需

表 1.2: 头部企业重点关注的数据平台层关键技术研究领域

研究点	研究方向概述	主要会议	研究主要关注点与代表性工作
面向智能应用的 Serverless 计算平台技术	AI 应用正在加速普及，针对传统云计算 GPU 资源供给模型存在的粗粒度分配、弹性能力不足以及运行成本高昂等问题，业界正在探索面向智能应用的 FaaS 平台技术来满足中小模型推理、边缘智能等“泛在化、动态化和碎片化”的 AI 算力需求。	OSDI SOSP ASPLOS EuroSys ATC	• **GPU 云函数沙箱：** 阿里云和 Microsoft 陆续推出了面向 AI 智能应用的 GPU 云函数服务，允许租户利用 FaaS 函数部署推理、训练等服务 [55, 56]。 • **函数冷启动问题：** 华为最新的数据中心 Trace 深入分析了 Serverless 计算平台内部的冷启动发生频率以及对函数性能的影响 [57]。而阿里云的最新研究成果则通过延迟调度请求以提高函数实例的复用率 [57]，或利用 Fork 机制加速实例启动过程 [58]，从而减少冷启动开销；中国电信云计算研究院同样聚焦函数冷启动问题提出了热点竞争感知的函数分区缓存技术以改善缓存效率。 • **资源利用率与性能优化：** 字节跳动和 CorkrocachDB 聚焦 Serverless 数据库提出了高并发扩容技术和多核心间节能方法，用于改善特定垂直领域应用的运行效率 [59, 60]。华为则面向 Serverless 大模型推理业务场景开展了大规模资源快速扩容技术研究 [61]。
面向大模型的智能数据平台技术	向量数据库为大模型的外部知识库的管理提供了极大的便捷；业界结合 AI 的推理能力，积极扩宽数据平台的能力边界。既可以内置 AI 增强数据库的交互形式，又可以加强对底层数据的理解能力。AI 基础设施的逐步普及也正在逐步改写数据库的架构设计。	SIGMOD PVLDB ICDE EDBT CIDR	• **向量检索：** Apple 公司采取倒排索引的技术路线提供了向量检索的服务 [51]。AlayaDB [62] 推出了基于向量检索的高效高质量长文本 LLM 推理的数据基础设施。 • **AI Inside 数据平台：** 阿里云百炼开源面向 Java 开发者的 NL2SQL 智能体框架 [63]。阿里云瑶池数据库团队推出的面向企业用户的数据分析智能体，可以根据自然语言描述进行需求分析，自动完成数据理解，并基于数据理解提出分析需求。Oracle 数据库支持自动索引创建和销毁的生命周期管理能力 [64]。 • **AI Infra 加速数据平台：** 阿里云瑶池数据库团队基于推出了基于 CXL 2.0 协议的 PolarCXLMem 多写数据库一体机 [65]。华为推出全球首个通用计算超节点 TaiShan 950 SuperPoD，并结合 GaussDB 推出替代 Exadata 一体机的技术方案。NVIDIA 持续推进各个数据库厂商集成 GPU 加速分析的合作。 • **氛围编程的数据库新诉求：** Microsoft 推出针对数据库高频列变更管理技术 [66]。
支撑智能任务的高性能存储平台技术	针对大模型训练与推理带来的海量存储及低延迟需求，传统存储架构面临语义、性能与成本的严峻挑战。业界正聚焦于通过软硬件协同、元数据优化和成本控制等，旨在构建支撑智能任务的下一代高性能存储底座平台。	OSDI SOSP FAST EuroSys SoCC	• **大模型训推中的存储优化：** 字节与阿里针对训练，利用增量与异步写入技术构建高效检查点存储降低阻塞 [67, 68]；月之暗面针对推理，通过以 KVCache 为中心的分层与重用机制缓解显存压力 [21]，Microsoft 则通过高效复用减少 RAG 场景下的推理开销 [69]； • **软硬协同的数据加速：** 华为为突破 I/O 瓶颈，一方面利用 GPU 直通存储技术，消除 CPU 数据拷贝开销 [70]，同时利用 DPU 卸载存储索引 [71]，Samsung 则通过新型 SSD 特性优化文件系统日志与数据放置效率 [72, 73]。 • **极致性能与成本压缩：** 百度为对象存储设计高效的层级元数据管理，在路径解析性能和扩展性之间取得较好权衡 [74]；IBM 与字节 [75, 76] 聚焦跨域调度与冷热分层以降低 TCO；华为与阿里云进一步优化了纠删码与流量偏斜下的资源效能 [77, 78]。

要虚拟机进行性能隔离的多租户 FaaS 环境。此外，还有一些研究提出了诸如 GPU 虚拟化和 MIG 的技术方案，但这些方案要么面临高运行开销，要么兼容性较差，且仅能在特定软件栈下工作。

为此，中国电信云计算研究院联合天翼云共同开展了面向 AI 智能应用场景的 Serverless GPU 函数沙箱研究，其核心设计聚焦 Serverless 函数的细粒度 GPU 算力切片及高度弹性的按需分配能力，以支持包括大中小型模型和 AI Agent 在内的海量智能应用的“泛在化”、“动态化”和“碎片化”加速器算力需求。该课题以 Kata-container 作为函数运行时载体，通过 I/O 直通的低开销 GPU 虚拟化技术，基于快速显存交换的虚拟 GPU 重映射技术以及调度延迟感知的 GPU 切片再分配技术，实现多租户函数对 GPU 设备的高效“时空动态”共享，从而显著提高 Serverless 平台内部的 GPU 资源利用率，大幅降低采购成本。

构建热点感知的高效函数缓存策略，降低 AI 应用冷启动开销。Serverless 的冷启动问题是影响 AI 云函数性能的一大挑战。Serverless 平台多基于无状态编程模型运行函数，尽管这种方式能让开发者无需关注环境与资源管理，常用于构建事件触发型应用，但在实际请求场景中，以“偶尔到达、需从零启动”为特征的函数调用往往占据业务主体。无状态特性带来的冷启动问题，不仅导致函数启动时间远高于执行时间，还成为制约 Serverless 计算性能的核心瓶颈。一些主流 FaaS 平台例如采用 TTL 策略的 OpenWhisk [80]、基于优先级策略的 FaasCache [81]，采用了函数缓存机制以缓解冷启动，然而这些缓存方案的效率仍存在明显不足：一方面，集群内部各个节点的本地缓存控制器缺乏全局工作负载视图与调度协同能力，导致节

点间缓存资源使用不均衡出现较大的性能差异，例如一些节点内且超过 50% 缓存实例极少被调用，冷启动率波动达 38%。另一方面，每个节点内部的所有函数共享同一个缓存池，这种设计会天然地在热点函数之间引发无序竞争，加剧租户函数的延迟波动。这也导致现有的缓存策略并不能从根本上解决 Serverless 平台内部的缓存争用问题，使得其只能带来有限的缓存效率改善。

为解决该问题，中国电信云计算研究院提出了一种基于分区缓存池设计的函数缓存方案。核心创新在于将节点缓存资源划分为独立分区分配给热点函数，并根据热点函数的运行时性能动态调整分区大小和缓存资源配置，从而在缓解节点内热点函数间缓存争用的同时保持高的缓存命中率。结合用户性能反馈感知的分区调整算法，可在有效降低冷启动调用发生的同时解决传统单一缓存池的资源竞争问题。实验表面，该研究可以显著提升 FaaS 平台内部 20% - 40% 的缓存资源利用率，降低至少 50% 的 AI 云函数启动延迟。

1.3.2 面向大模型时代的智能数据平台技术

在大模型席卷全球的当下，人工智能正推动数据库领域从应用场景到底层架构的全面演进，不仅引爆了新的需求，更实现了大模型能力对数据库系统的深度赋能。在此背景下，向量数据库及氛围编程场景下的新型数据库快速崛起，同时 AI 基础设施也为数据库的架构创新与性能提升提供了核心动力。

向量数据库为大模型提供外部知识库的管理功能，承接大模型的长期记忆功能。向量数据库是一种对非结构化数据提供管理与检索的工具，常用于图像检索、推荐系统等任务。但是，随着大模型技术的爆发，向量数据库迅速成为数据库社区新一轮的研究热点。向量数据作为大模型的外部知识库，可以在发起大模型查询前通过检索外部知识增强输入上下文的信息量，从而提升大语言模型的输出结果质量。可以一定程度解决大模型的幻觉问题、知识过时和私域数据推理的问题。向量数据库目前比较有效的方法是基于领域图索引的方式进行近似查询，业内已经诞生了很多有效的图算法 [82, 83]。但是图索引的内存空间开销很大，对于十亿量级的向量数据，通常在单机内存很难做到管理。因此需要研究基于外部 SSD 的高效存储方案，业内比较有代表性的设计是由 Microsoft 提出的 DiskANN 索引 [84, 51]，但是其面临着页面读放大等磁盘读写效率不高的问题。中国电信云计算研究院针对 DiskANN 相关的读写优化技术进行了全面的梳理，对内存结构优化、磁盘结构优化和搜索加速技术总结了 8 类优化技术并对其进行全面评测，结合对 SSD 的访问效率进行分析，从中筛选出 4 种高效的优化方法组合，为向量数据库的产业应用提供实践指引。此外，我们还设计了向量索引针对跨查询间的页面复用方法，提出了一种针对图结构定制化的缓存技术，其相比于经典的 LRU 等缓存替换策略，更为适合于向量图索引检索查询过程的页面访问模式，可以让整体向量查询的性能相比 DiskANN 有数倍的查询效率提升。

大语言模型赋能的 AI Inside 大数据平台，推动数据查询从传统关键词匹配迈向以语义理解为核心的智能化查询新范式。当前数据智能体（Data Agent）目前仍然处于发展早期，它借助生成式大模型可以加强数据分析系统的语义理解能力 [85, 86]，大模型可以提供语义算子，增强传统的字面值匹配形式的算子。结构化查询语言（SQL）是各行业从数据库中获取信息的关键环节，这通常依赖用户具有一定的数据库知识和技能才能高效地构建查询。随着人工智能和自然语言处理技术的进步，Text-to-SQL 技术应用而生，研究者开始利用深度学习模型对文本数据进行训练，使得系统能够更准确地理解用户语言从而降低数据库使用门槛。LLMs 如 GPT-4 和 GLM-130B，凭借其强大的语言理解和生成能力在 Text-to-SQL 任务中展现出了巨大的潜力。这些模型通过预训练学习大量语言知识和结构信息，能够在少量样本甚至零样本（Zero-shot）的情况下生成准确的 SQL 查询 [87]。同时，借助 AI 的能力可以帮助数据平台进行自调优和自诊断。一方面，数据库引擎内部提出了很多自调优的适配算法，比如自适应的存储结构、自适应的优化器执行计划选择、基数估计、代价估计等技术。另一方面，对于数据库外部自动运维提出了索引推荐 [64]，慢查询根因分析 [88] 等自动化运维方法。

AI 基础设施的大规模普及，正在改写数据库的底层架构设计。面对 AI 基础设施的大规模普及，关系数据库必须进行根本性的架构变革。这种变革不是简单的功能叠加，而是要从底层重新设计数据库的架

构。主要体现在以下几个方面：首先，内存架构需要革新。传统的本地内存架构已经无法满足 AI 工作负载对内存容量和访问速度的需求。CXL 技术的出现为解决这一问题提供了新的思路，它允许 CPU 像访问本地内存一样访问远程内存，延迟可低至 200 - 500 纳秒。通过 CXL 高速互联通道，延迟可低至百纳秒级别，显著降低数据同步带来的延迟和带宽成本，并大幅提升跨节点数据一致性的处理效率。Tigon 数据库系统 [89] 基于 CXL 的分布式内存池解决方案，可实现和本地一样低延迟、高带宽的远程内存访问，延迟可低至百纳秒级，带宽吞吐达到数 TB/s，实现内存资源" 池化可共享、按需可调度"。阿里云 PolarDB 团队联合服务器团队率先构建了基于 CXL 2.0 协议的分布式内存池系统 PolarCXLMem [65]。华为云在 CXL 技术应用方面也有重要进展。华为率先把超节点技术引入通用计算领域，发布全球首个通用计算超节点 TaiShan 950 SuperPoD，结合 GaussDB 分布式数据库，能够彻底取代各种应用场景的大型机和小型机以及 Exadata 数据库一体机。其次，GPU 加速技术正在深刻改变数据库的计算层架构 [90]。NVIDIA 与各大数据库厂商的合作正在推动这一变革。Microsoft SQL Server 2025 在 GPU 加速方面取得了重要突破。Microsoft 与 NVIDIA 合作，将 NVIDIA Nemotron 开放模型和 NIM 与新的 Microsoft SQL Server 2025 相集成，以加速其中的 AI 应用。

氛围编程催生数据库场景新需求。AI 技术迅猛发展推动软件开发范式变革，OpenAI 联合创始人 Andrej Karpathy 于 2025 年初提出的“氛围编程（Vibe Coding）”理念，正引发全球开发者社区变革。这种新型编程模式下，开发者无需逐行编码，仅用自然语言描述功能目标，专用大模型便能生成对应代码，使工程师从“代码构建者”升级为“架构设计者”。这一范式革新不仅重塑了传统开发流程，更对数据库提出全新技术挑战，传统数据库的设计理念与架构已难以适配。氛围编程强调快速迭代、高频试错的特性，要求数据库具备高度自动化能力，可依据 AI 生成代码实现自动配置、优化与运维，彻底摆脱传统人工干预模式的效率瓶颈。在关键应用场景中，传统关系型数据库的短板尤为突出：部署初始化需数分钟至数小时的“慢速拉起”，与秒级响应需求相悖；无法支持弹性扩缩容以匹配开发各阶段资源波动；面对高频代码修改，缺乏高效的版本管理、快速增删列与快照功能，难以实现快速回滚及历史版本查询，这些都成为制约开发效率的核心痛点 [66]。在氛围编程主导的新开发范式下，数据库不再只是“被动存储组件”，而必须演进为能与大模型协同工作的“智能数据基础设施”。能否提供开箱即用、秒级拉起、自动调优并支持全链路可追溯的数据能力，将成为衡量下一代数据库竞争力的关键指标。

1.3.3 支撑智能任务的高性能存储平台技术

大模型应用的爆发式增长为存储系统带来了特有需求，尤其在训练阶段的检查点存储和推理阶段的 KVCache 管理方面，传统的存储平台正面临语义、可靠性与性能的严峻挑战。以大语言模型为代表的新型智能负载已成为当前存储系统需支撑的核心任务。然而，大模型通常拥有千亿甚至万亿级别的参数量，其训练和推理过程对计算、内存及存储资源的需求呈指数级增长。特别是在存储层面，如何高效管理推理过程中的 KVCache 和保障训练过程中的检查点（Checkpoint）存储，已成为制约大模型性能、资源利用率和系统可靠性的关键瓶颈。传统存储系统在面对大模型特有的高吞吐、低延迟、海量小文件以及极端并发访问模式时，往往难以满足其严苛要求，亟需针对性的存储优化方案。在大型语言模型的自回归推理过程中，KVCache 机制被广泛采用以避免重复计算，但其对 GPU 显存的巨大消耗和数据传输延迟构成了显著的性能瓶颈。业界正通过智能的 KVCache 管理、分层存储及数据压缩等技术积极应对。例如，月之暗面的 Mooncake 架构 [21] 采用了以 KVCache 为中心的多层存储池设计，旨在通过存储优化缓解显存压力；Microsoft 的 CacheBlend [69] 则设计了高效的 KVCache 重用机制，以减少其海量开销，共同目标是显著提升推理吞吐量和并发服务能力。与此同时，在大模型训练过程中，检查点是保障训练可靠性的关键，但其巨大的文件大小和频繁写入操作给存储系统带来了巨大的 I/O 压力。字节跳动的 ByteCheckpoint [67] 和阿里云的 FlowCheck [68] 等方案，通过增量检查点、异步写入、数据压缩与去重以及分层存储等技术，旨在构建一个能保障大模型训练可靠性和最大限度减少存储开销和性能影响的高效检查点存储平台。

为应对大模型带来的存储挑战，存储平台正积极探索软硬件协同优化路径，通过 GPU 存储直通、新型硬件卸载存储协议栈以及大规模部署高性能 NVMe（Non-Volatile Memory Express）SSD 等技术，显

著提升存储系统的吞吐、延迟和效率，以更好地支撑上层 AI 任务。面对大模型对存储系统提出的极致性能要求，单一的软件优化或硬件升级已难以满足。当前业界普遍共识是通过软硬件协同设计，从数据路径的各个环节进行深度优化，这包括直接利用 GDS（GPU Direct Storage）等 GPU 数据直通技术、引入 DPU 等新型硬件卸载传统 CPU 的存储管理负担，以及大规模部署高性能全闪存储。GPU 存储直通技术旨在消除数据在 CPU 与 GPU 之间传输的瓶颈，使 GPU 能够直接访问存储设备；例如，GoFS [91] 绕过 CPU 构建了一个由 GPU 主导的文件系统，将文件系统的元数据管理和 I/O 操作等关键逻辑都放到 GPU 上运行，支持 GPU 应用以文件接口对 NVMe 存储进行直接的高并发访问，而华为的 GeminiFS [70] 则提供了一种 GPU 原生文件系统，允许智能应用绕过 CPU 以 POSIX 文件语义接口直接访问 SSD，显著提升了 GPU 的存储访问效率和整体计算性能。同时，DPU 等新型计算设备在存储系统中的应用体现了计算卸载的理念，通过将网络、存储和安全等基础设施功能从 CPU 中卸载出来独立处理，从而释放 CPU 资源用于上层应用计算；华为的 HiDPU [71] 便提出了一种面向 DPU 的混合索引方案，用于解耦存储系统，利用 DPU 的并行处理能力加速索引查找和数据管理操作，显著提升了分离式存储系统的性能和效率，降低了端到端的数据访问延迟并提高了存储服务的可扩展性。此外，大规模部署高性能 NVMe SSD 是提升存储介质性能的直接手段，NVMe SSD 凭借其低延迟、高带宽和高 IOPS 的特性已成为高性能存储的首选；Samsung [73] 则探索在新型 CMM-H SSD 上实现目录粒度文件系统日志 [72]，这些都旨在提升文件系统在高性能 SSD 上的元数据操作效率和整体存储性能，为 AI 任务提供了更稳定、更高效的底层存储支持。

面对 AI 场景带来的 EB 级数据规模与极低延迟的双重压力，下一代存储系统正加速架构演进：通过分层元数据管理突破扩展性瓶颈，利用新型编码技术在保障极致性能的同时降低资源消耗，并持续优化存储层面的稳定性。随着云原生与 AI 业务的发展，存储系统的对象数量和元数据规模呈指数级上升，业界正积极探索数据与元数据管理的新范式，以支持千亿级甚至万亿级的文件规模。一方面，为极致优化成本，阿里云弹性块存储深入分析了云上的流量偏斜，揭示了生产环境中的负载不均衡特征 [78]，IBM 则针对跨云和跨地域的数据管理设计了一种智能的数据放置方案 [75]，在满足性能 SLO 的同时最小化存储支出。此外，华为针对内存存储系统提出了无条带数据放置方案 [77]，旨在降低冗余开销的同时实现高吞吐量。另一方面，性能与可扩展性也是下一代存储系统的核心诉求：百度的 Mantle [74] 提出了一种高效的层级元数据管理方案，在目录解析性能与扩展性之间做出了较好的权衡，解决云对象存储服务在大规模场景下的元数据访问延迟与扩展性问题；清华团队则从语义入手，提出了一种将目录语义与元数据索引解耦的创新架构，显著加速了分布式文件系统的元数据服务性能 [92]。为解决分布式存储在动态复杂负载下，扩展性与性能难以保证的严峻挑战，中国电信云计算研究院针对元数据分布式管理的瓶颈，引入智能算法解决元数据子树的负载均衡难题，提出了机器学习驱动的元数据负载均衡框架 Origami [93]。该工作以最小化用户作业完成时间为核心目标，在负载均衡的同时考虑了元数据的局部性特征与层次结构，成功在负载均衡收益与访问开销之间实现较好的权衡。实验结果表明，Origami 框架有效解决了分布式文件系统中因层次化命名空间和动态负载导致的访问热点问题，极大提升了元数据集群的聚合吞吐量，相比传统方案大幅降低了用户端到端操作的完成时延。这些工作共同推动了分布式存储在“高可靠、高性能、低成本”不可能三角上的平衡与突破。

1.4 热点方向三：智能化云运维、可信安全与能效优化

面向下一代云计算形态，智能化云运维、可信安全与能效优化将共同构成云基础设施与云服务体系的核心支撑框架。一方面，智能化云运维依托遥测数据、日志数据、调用链数据等多维数据源的实时采集与融合，通过异常检测、根因分析、容量预测等智能算法，实现对跨层级、跨区域复杂异常的敏锐感知与自适应响应，为自治化故障修复、自动化扩缩容与韧性增强提供可信决策基础。通过将 AIOps、可观测性平台与知识图谱等技术深度结合，云运维体系能够在统一的时空尺度上关联基础设施层、平台层与应用层指标，支持从“事后告警、人工排障”向“事前预警、在线演练、闭环优化”转变，逐步演进为具备持续学习能力的智能运维中枢。另一方面，可信安全通过多层次安全防护体系、异常行为识别、零信任架构

及隐私保护计算等机制，保障在高度动态、多租户、多云/混合云环境下云平台本身及其承载业务的安全性、完整性与隐私性。同时，能效优化依托智能调度、弹性资源管理、异构算力协同与绿色算力策略，在满足服务等级协议（SLA）与安全约束的前提下，实现超大规模集群的高效利用与能源可持续性。

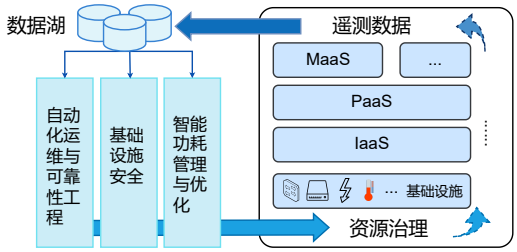


图 1.10: 运维、安全、能效一体化云生态架构

在这一框架中，智能化云运维、可信安全与能效优化并非相互独立的模块，而是通过统一的观测体系、策略引擎与资源治理机制形成闭环协同：运维智能为安全策略与能耗优化提供精细化可观测数据与预测能力，安全机制为运维自动化与资源调度提供可信执行环境和策略约束，能效优化则在性能与安全要求之间进行动态权衡，推动云平台向“高可靠、高安全、高能效”的方向演进。进一步地，这一闭环协同并非停留在策略层面的静态集成，而是通过跨层联动与反馈自适应不断演化：一方面，底层监控与遥测数据经由智能分析与因果推断模型提炼为可操作的运维与安全知识，驱动策略引擎在多租户、多区域和多云环境下进行动态决策；另一方面，策略执行的效果又通过统一观测面被持续评估和量化，为后续的策略迭代与能效调优提供依据，从而实现从“被动响应”向“主动预防”和“持续优化”的演进。图 1.10展示了三者之间的总体关系与协同作用机理。为更好地理解相关技术路径与研究进展，表 1.3重点遴选了部分具有代表性的关键研究成果。

1.4.1 面向大规模集群的自动化运维与可靠性工程

随着云计算基础设施快速扩张和服务规模的持续增长，系统复杂性呈指数级攀升，使得大型云平台正面临前所未有的运维与可靠性挑战。近年来，多起重大云事故（包括 2024 年 10 月 Amazon us-east-1 控制面故障导致美国大半互联网服务中断，2023 年阿里云杭州可用区故障引发全国范围业务停摆，以及滴滴基础设施故障造成数百万司机无法接单）清晰揭示出当前高度集中化云基础架构中的脆弱性：系统耦合度过高、异常传播路径不透明、指标数量庞大且动态变化快、人工与静态规则的运维能力难以应对复杂、跨模态的故障场景。传统自动化运维系统以阈值规则、静态模型和人工巡检为主，其响应速度、鲁棒性与泛化能力已难以满足 99.99% 以上的可用性要求，也难以在分钟级甚至秒级窗口中完成风险识别与故障遏制。更为严峻的是，云平台正从单一数据中心走向多集群、多地域、多云/混合云的异构形态，业务形态也从相对稳定的长周期服务演进为高度动态的弹性计算与事件驱动工作负载。在这一过程中，底层基础设施、服务编排、中间件乃至 SaaS 层服务之间形成了跨层级、跨域的复杂依赖关系，使得故障往往以“级联失效”、“雪崩放大”、“隐性退化”等形式呈现：单点异常极易沿控制面与数据面扩散，局部资源失衡或策略配置偏差可能演化为大规模服务不可用。与此同时，观测数据的维度和噪声水平持续攀升，新指标、新拓扑、新配置版本的迭代速度远超人工团队的认知与建模能力，使得“事后分析式”的被动运维模式日益难以为继。

学术界的相关研究同样指出，大规模云系统的可观测性数据（Metrics、Logs、Traces、Events）呈现高维度、多模态、强关联的特点，且其内部拓扑结构高度动态。已有大量工作试图从不同维度解决这些问题，包括多变量深度异常检测 [113]、跨模态关联分析 [114]、微服务依赖图的根因诊断 [115]、强化学习驱动的自治调度 [116]、以及 LLM 助力的日志结构化与事件语义理解 [117]。这些研究共同形成了智能化运维（Intelligent CloudOps）发展的基础，但在真实云环境中仍面临关键瓶颈：在线性、稀疏标注、高噪声、流式更新、复杂依赖导致的异常放大（Cascade Amplification）仍未被充分解决。

随着云网基础设施规模不断扩大、业务形态持续演进，系统运行状态呈现出高维度、强关联和快速动态变化等特征，传统的告警方式已难以支撑对复杂故障的早期识别。在大规模云环境中，运维面临的核心挑战是如何在几乎没有故障标注的条件下，可靠地识别早期、微小且易被噪声掩盖的异常信号（尤其是 KPI 级别的 Incipient Anomalies）。这些异常通常表现为跨指标、跨服务的微弱偏移，伴随概念漂

表 1.3: 智能化云运维、可信安全与能效优化研究领域热点

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
面向大规模集群的自动化运维与可靠性	在超大规模云基础设施中，实现自动化、智能化的运维决策与故障处置，以提升系统的稳定性、可观测性和整体可靠性，其关键在于构建面向高维指标的智能异常检测、无监督根因分析及自愈闭环机制。	KDD FSE ICSE DAC SRDS	• **智能异常检测：**中国电信云计算研究院团队所提出的 HEIMDALLR 系统 [94]，可动态建模潜空间并识别多指标异常，对复杂业务场景有较强的适应能力。相关研究聚焦于在高维、多关联、非平稳的云环境 KPI 时序中实现早期异常的无监督检测与因果归因。 • **自动化根因分析：**清华大学等团队将服务调用关系建模为大规模有向图或因果图，并结合图表示学习等技术，以支持在缺乏人工规则与标注下进行故障诊断 [95, 96] 。与此类似，相关前沿研究关注基于微服务依赖、调用链和因果图的无监督根因定位，实现对故障源的快速推断与收敛。 • **故障缓解与自愈：** Meta AutoScale、Microsoft Azure 自愈管控系统及阿里云自动化故障演练 [97, 98, 99] 已展示分钟级响应与低人工干预的实践效果。此类现代云平台强调构建检测—分析—自动修复闭环，实现云集群自主高可用性。
云计算环境下的基础设施安全	云计算安全已从传统边界防护发展为多层次、责任共担的综合体系，涵盖供应链安全、容器与 Serverless 等运行时隔离与合规，以及内存安全和零信任架构，以实现复杂共享环境下的全面安全与隐私保障。	OSDI NDSS S&P Security WWW	• **供应链安全：**普渡大学团队通过联合解析工作流配置与源码的静态污点分析技术，自动定位命令注入等关键缺陷，解决云中权限滥用问题 [100]。 • **容器安全：**中科院信工所团队提出的新模型不再默认信任宿主机，而是利用 ARM CCA 硬件构建可信执行环境保障机密性与完整性 [101]。 • **Serverless 计算安全：**北卡罗来纳州立大学团队提出利用图可达性分析技术构建应用权限调用关系，识别潜在的策略配置错误或过度授权 [102]。 • **内存安全:** 佐治亚理工学院团队通过内存分配隔离阻断跨域攻击 [103]。苏黎世联邦理工学院团队发现读干扰阈值会动态波动，这使得静态防御机制难以长期有效 [104]。 • **零信任架构:** Microsoft 团队通过细化服务间认证与最小权限通信，实现流量的零信任访问控制 [105]。阿里巴巴团队通过硬件可信执行环境，确保敏感任务的安全执行 [106]。
云数据中心智能功耗管理与优化	云数据中心智能功耗管理与优化主要集中在提升能源利用效率、优化资源分配与负载调度策略，以及增强系统的绿色可持续性与整体运行可靠性等方面，旨在在既有基础设施条件下最大化算力与能效收益并降低整体运维成本。	ASPLOS ISCA OSDI EuroSys TACO	• **数据中心电力管理:** Microsoft 通过智能分配和管理服务器超频资源，使云数据中心应用成本下降 30%、总能耗减少 10% [107]；Meta 通过智能筛选适合提升频率的服务和硬件、细致管理电力风险，实现了相当于新建半个数据中心额外算力扩容 [108]。 • **数据中心冷却管理与热感知调度:** 清华团队通过贝叶斯优化，实现了数据中心冷却系统在动态负载下的能耗最优与热安全保障，平均节省 10.1% 的冷却能耗 [109]。Microsoft 团队通过热感知调度，实现了 GPU 集群在能效、散热和部署密度上的协同提升 [110]。 • **面向大模型的智能功耗管理与优化：** Microsoft 团队针对云数据中心大语言模型推理的能耗管理挑战，提出了基于功率冗余分析的超分配优化方法 [111]。华为团队在 AI 加速器能效优化领域，基于 Ascend NPU 的 DVFS 机制，提出了算子级的性能与功耗建模及优化策略，实现了高精度、低损耗的能耗管理，推动了 AI 算力的可持续发展 [112]。

移与拓扑演化，要求模型不仅具备对高维、多模态时序关系的精细刻画能力，还要能在线适应、低延迟响应与低误报率。针对这些核心痛点，中国电信云计算研究院的研究团队提出一种新型无监督检测框架 HEIMDALLR [94]，面向云环境中高维、多关联、非平稳 KPI 时序的特点，构建了动态潜空间模型，旨在挖掘隐藏在 KPI 背后的早期微弱异常信号。该方法结合异常归因机制，对潜在因果关系进行刻画与拆解。相比传统方法，HEIMDALL 在动态结构建模、噪声抑制和早期异常识别方面具有更高的灵敏性与鲁棒性。不仅在准确率与误报控制方面表现优越，同时具备低计算开销和高可解释性，更适用于大规模云系统的实时部署需求。

异常检测仅能“发现问题”，然而真正减少故障持续时间 MTTR（Mean Time To Repair）的关键在于快速准确的根因定位。大规模云系统包含复杂的微服务拓扑、依赖链路、网络层与资源层组件，任何一个局部异常都可能沿链路扩散，形成“假异常”和“症状级异常”。因此，自动化根因分析 RCA（Root Cause Analysis）成为可靠性工程的核心能力 [118]。已有研究从多种角度实现 RCA，包括基于服务依赖图的拓扑分析：如 Microscope 通过构建服务依赖图并分析指标协方差或异常传播模式，定位异常组件或服务 [119]。基于调用链和 tracing 数据的路径/分布分析：TraceRCA 通过异常 Trace 检测、可疑微服务挖掘和服务排序，实现高效、准确的无监督根因分析 [95]。基于因果图与图学习的微服务诊断：如 MicroHECL 结合图神经网络和因果建模的微服务诊断方法，通过对依赖图进行嵌入和消息传递，捕获复杂非线性关系进行故障定位 [96]。日志与指标联合分析：通过解析日志事件并与 KPI 异常模式对齐，提高定位精度 [120]。这

些方法的共同目标是从“一个指标异常”推断出“哪个组件、服务、配置或操作导致的根因”。未来，RCA 有望与自动化检测和自愈体系结合，通过异常空间共享、跨模态因果指向和知识图谱融合，形成统一可扩展的诊断框架。

智能化运维的最终目标是将异常检测转化为可自动执行的故障缓解与自愈行为，实现系统自治与高可用性。在大规模云集群中，异常信号若无法快速响应，将可能导致服务级连锁故障，扩大业务影响范围。故障缓解与自愈体系旨在建立从异常检测 → 根因分析 → 自动修复的闭环机制 [121]。自愈策略主要包括自动隔离与流量调度，即当某个节点或服务表现异常时，系统可自动隔离受影响实例，或通过流量切换与负载均衡将请求路由至健康实例，从而避免故障蔓延 [122]。策略驱动的自动修复：根据根因分析结果，系统可执行自动化操作，如回滚异常配置、重启服务、重建容器或虚拟机实例，甚至进行资源弹性伸缩以缓解压力。闭环验证与持续优化：自愈动作完成后，系统持续监控相关 KPI 与依赖链，验证修复效果，并将操作结果反馈到策略引擎，实现自适应优化与经验积累 [97]。现代云平台（如 Meta AutoScale、Microsoft Azure 自愈管控系统、阿里云自动化故障演练平台）均采用类似闭环设计，实现分钟级响应、低人工干预和业务连续性保障 [98, 99, 123]。随着复杂事件处理、因果推断与强化学习等技术的引入，自愈策略将从预定义规则逐步演进为可持续学习和自动演化的策略集合，能够在面对新型故障模式和未知环境扰动时，保持对恢复路径和缓解动作的动态优化。未来，异常检测与自治策略深度结合，实现对复杂异常的主动防护和自愈能力，使云集群在无人值守条件下保持高可靠性和可用性。

1.4.2 云计算环境下的基础设施安全

随着云计算成为数字化转型的基石，其基础设施的安全已成为保障所有上层应用与数据的根本。现代云环境的安全范畴早已超越了传统的边界防护，演变为一个多层次、动态且责任共担的综合性挑战。其研究核心正从宏观的边界防御，深入到工作负载内部与软件生命周期的每一个环节，构建贯穿云原生应用生命周期的纵深防御体系。这要求从供应链的源头保障组件安全，在运行时层面确保容器与 Serverless 等动态工作负载的隔离与配置合规，并依托内存安全与可信执行环境等底层技术为敏感数据和计算提供硬件级强力保护，最终共同实现在复杂共享环境中的全面隐私保护与安全保障。

面向云基础设施的软件供应链安全主要关注源码托管、依赖获取、制品签名与分发等关键组件，以确保从“代码提交”到“云中运行”的整个链路可验证、可审计且不被篡改。其核心机制是在供应链各环节建立最小信任边界，并通过签名、透明日志和策略验证记录构件来源、构建过程与交付状态，为云侧的部署与准入控制提供可信证据。近年的研究聚焦于云场景中两个最关键的节点：云托管链路安全、签名与透明日志作为供应链信任根。在云托管链路安全方面，通过联合解析工作流配置与源码的静态污点分析技术，跟踪不可信输入在跨文件、跨步骤执行路径中的传播关系，从而使命令注入等关键缺陷能够在无需实际运行流水线的前提下被自动定位 [124, 100]。在签名与透明日志构成的供应链信任根方面，为解决传统长期私钥模式无法满足云环境中大规模自动化构建与分发的可信性要求的问题，以 Sigstore 为代表的云原生签名体系通过引入身份绑定、短生命周期证书与公开透明日志，使容器镜像等云基础组件在生产、分发与部署的全流程中具备自动化来源验证与不可抵赖审计能力 [125]。

容器安全研究的核心在于应对其共享内核模型带来的固有风险。容器被广泛部署用于在共享计算基础设施上打包、隔离和复用应用程序，但其安全保障依赖于操作系统。在真实云和边缘计算场景中，容器可能部署在管理员权限不可控的主机上，宿主操作系统一旦被攻破或被恶意操控，传统的容器隔离机制就会失效。为此，有研究提出了一种突破传统假设的容器安全模型：即容器运行时不再默认依赖宿主机和操作系统的可信性，而是将容器自身构建为安全主体，使容器即便运行在不可信系统上，也能维持完整性与机密性。在容器内部运行一个轻量级、高可信的执行组件，负责关键安全决策与监控，对容器中系统状态、执行行为、调用路径等进行持续验证，避免完全依赖宿主操作系统的反馈 [126]。为解决容器与宿主机共享内核带来的根本性风险，有研究提出一种利用 ARM 机密计算架构 CCA（Confidential Compute Architecture）硬件安全原语构建的新型可信容器体系结构。容器在启动阶段通过硬件度量与加密机制验

证自身完整性，确保执行镜像、依赖文件与关键配置未被修改，从而形成从部署、启动、运行的可信链条。同时，轻量级软件框架使容器基本无须大规模重构即可迁移到硬件可信区中。与传统 TEE 迁移方式相比，兼容性更强、开发成本更低 [101]。

Serverless 计算的动态组合、松耦合架构以及高度分布式的函数运行环境在云中引入了新的安全挑战。例如函数间隔离边界模糊、临时容器数据残留、依赖链攻击面扩大，以及跨租户资源调度中的潜在风险，这些都对云服务商与用户共同负责的安全模型提出了更精细的要求。相关研究主要围绕权限滥用、策略配置风险、内存泄漏防护以及函数调用链展开。有学者关注 Serverless 应用运行阶段的实时安全防御，通过引入细粒度、可动态执行的安全策略，在函数执行过程中监控访问行为，实现对异常调用与违规资源访问的阻断 [127]。从策略工程角度出发，研究者将应用的所有函数、资源及其对应的权限策略建模为一个图模型，并利用图可达性分析技术构建 Serverless 应用权限调用关系，识别潜在的策略配置错误或过度授权，从架构层强化函数间的安全边界 [102]。针对 Serverless 平台中普遍存在的内存泄漏攻击，研究者提出一种选择性数据保护机制，能够有效防御内存泄漏攻击，同时保持较低的系统负载，为 Serverless 场景下高效、轻量的内存安全保护提供了新的解决方案 [128]。除此之外，有研究指出基于机密虚拟机 CVM（Confidential Virtual Machines ）的保密容器与 Serverless 的瞬时启动和高并发需求严重不匹配，并导致巨大启动开销、资源低效与过大的可信计算基础 TCB（Trusted Computing Base）。为此作者提出将容器管理与函数执行分离的架构，在保持强机密性的同时显著降低保密 Serverless 的启动与运行开销，实现高效、低 TCB 的函数执行环境 [129]。

在云计算环境中，内存安全的范畴已从单一系统的软件漏洞防护，扩展至对底层共享硬件资源可靠性与隔离性的全局性保障。在这一背景下，作为物理承载的 DRAM 安全变得尤为关键。Rowhammer[130] 和读干扰等电干扰型问题已成为影响内存可靠性和系统安全的重要威胁。这种基于硬件缺陷的攻击，能够绕过传统的软件安全边界，直接威胁到云基础设施的核心信任根基，导致跨租户数据泄露、服务崩溃甚至系统控制权被夺取，从而对云服务所承诺的机密性、完整性与可用性构成了根本性挑战。针对这一类跨单元、跨行的非侵入式攻击或失效现象，已有研究主要从真实硬件测量与物理层实验入手，分析其形成机理、影响因素和可利用性，为后续系统级防护提供基础依据。有研究发现传统 Rowhammer 攻击可以跨越安全域实现内存破坏，因此着眼于“如何从系统层阻断跨域 Rowhammer”的研究问题，通过重新设计内存分配策略，对不同安全域的物理行进行隔离，从根本上破坏攻击所需的受害行与攻击行的空间邻接关系，实现无需硬件修改的系统级缓解 [103]。最新的实验研究揭示了一个关键且棘手的现象：读干扰错误的阈值并非固定不变，而是会随时间发生动态且不可预测的波动。通过对多代真实 DRAM 芯片进行大规模实验，发现读干扰强度会随老化、电压温度条件等在时间维度显著波动，揭示现有静态阈值防御无法长期可靠 [104]。

零信任架构旨在不默认信任任何主体的前提下，通过基于身份、上下文与风险的持续验证机制，实现对访问请求的最小权限控制与动态授权。随着云原生架构日益复杂、服务间交互碎片化、多租户共享与自动化运维普及，云内部的隐式信任链不断被放大，控制平面滥用与横向移动攻击风险显著提升。在这一背景下，零信任架构因能够提供持续验证、细粒度隔离与可信执行等能力，成为提升云基础设施安全性的关键方向。近年的研究主要集中在三个领域：云网络与服务层的零信任隔离、云运维与控制面的最小权限治理，以及云数据与执行环境的零信任保护。在云网络与服务层隔离方面，通过分析海量通信遥测数据自动推断节点角色，在不依赖人工配置的情况下自动导出高覆盖率的微分段策略，以实现大规模云环境中的零信任网络隔离 [105]。同时，通过构建跨多跳服务调用链的策略模型，并使其能够在多种不同数据平面中灵活执行，使得零信任服务间访问控制既具备复杂语义表达能力，又保持较低性能开销 [131]。在云运维与控制面的最小权限治理方面，通常围绕风险建模、操作依赖关系刻画与动态执行约束展开，从根本上减少对管理员或外包运维的信任。通过构建配置变更的依赖图与风险评估模型，在运维执行前对操作进行风险判断，并在执行过程中进行持续监控和范围限制，使运维过程仅限于经过验证的低风险操作 [132]。在云数据与执行环境的零信任保护方面，通过将信任收缩到硬件可信执行环境，避免将云平台或主机管理员纳入信任边界，确保敏感任务在不信任云的前提下仍能被安全调度与执行 [133, 106]。

通过在虚拟机监控层构建可验证的最小可信组件，负责安全执行环节的调度、监控和资源管理，以降低对云端软件栈的假设信任 [134]。

1.4.3 云数据中心智能功耗管理与优化

随着数字经济的高速发展，云数据中心逐渐成为全球信息基础设施的核心支撑。为了减少能源消耗和温室气体排放，各国都制定了碳排放达峰和碳中和的相关政策。数据中心作为互联网的“中枢”，支撑着各种信息服务，但其运行需要大量电力，能源消耗巨大。例如，美国数据中心的电力消耗约占全国总用电量的 1.8% [135]，而在运营成本中，能源支出占比高达 25% - 40% [136, 137]。因此，降低数据中心能耗不仅能节约成本，也是实现全球环保目标的关键。

云数据中心亟需智能化、自动化的能耗优化体系，以兼顾节能减排与业务性能保障，推动绿色可持续发展。当前，云数据中心的业务负载类型极为复杂，既包括延迟敏感型的在线交易、搜索、社交应用，也涵盖了对性能要求相对宽松的数据分析、批量处理等后台服务。多租户环境下，业务动态变化频繁，传统的静态资源分配和人工配置手段已难以满足实际需求。能耗优化与性能保障之间存在天然的矛盾，简单的节能措施往往会导致服务质量下降，影响用户体验和业务稳定性。因此，行业迫切需要一种能够智能感知业务类型、自动进行资源调度、动态优化能耗的新型管理体系。传统的数据中心能耗管理方案主要依赖于静态资源分配、人工标签驱动或基于单一指标的简单调度。这些方法在实际应用中暴露出诸多局限性。首先，静态分配无法应对负载的动态变化，导致大量核心、内存等资源闲置，整体利用率偏低。其次，依赖人工标签或离线分析的负载分类方式，在公有云多租户、黑盒应用场景下难以落地，扩展性和自动化程度不足。再次，单一调度维度（如只关注 CPU 利用率）无法全面反映业务的性能需求，容易造成资源错配和 QoS 违约。更为重要的是，随着业务复杂度提升和用户需求多样化，数据中心亟需实现调度的智能化和自动化，以适应未来的发展趋势。针对数据中心的不同关键环节，研究者们提出了一系列创新方法，包括冷却系统的自适应优化、热感知的任务调度，以及电力管理的智能分配 [138]。这些技术显著提升了数据中心的能效和可靠性。

如何在性能、能效与可靠性之间实现高效权衡，已成为数据中心可持续运行面临的核心挑战。数据中心算力需求的持续攀升和能耗压力的加剧，高效的电力管理成为数据中心可持续运行的核心挑战。传统的电力分配和调度策略往往难以兼顾性能、能效和可靠性。为此，业界提出了多样化的创新方案。Meta 团队通过大规模部署 DVFS Boosting，针对不同服务和硬件类型智能提升 CPU 频率，在保障服务可靠性与电力安全的前提下，实现了数据中心算力的弹性扩容，有效提升了资源利用率和部署密度 [108]。除此以外，通过系统性分析并优化了数据中心的运营碳和设备碳，并结合地理位置、可再生能源、能量存储和负载调度等多维度协同，Meta 近期还探索实现了按小时碳中和和 24/7 绿色运营 [139]。上述方法不仅提升了数据中心的电力利用效率和弹性，还为碳减排和绿色算力基础设施的建设提供了系统性解决方案。不仅如此，Microsoft 团队通过工作负载感知和风险预测，动态分配和管理服务器的功耗预算，开发了 SmartOClock 平台，实现了分布式的过载控制和能耗优化，显著降低了尾延迟、能耗和运维成本。与此同时，碳感知和绿色电力管理也成为数据中心电力调度的重要方向 [107]。

在数据中心热负载持续攀升的趋势下，通过冷却管理与热感知调度实现能耗优化与热安全之间的动态平衡，将成为未来演进的重要方向。随着算力密集型应用和大规模模型在云端的普及，数据中心的热负载持续攀升，对冷却系统的智能管理提出了更高要求。传统冷却策略通常依赖固定设定点，难以应对动态变化的服务器功耗和空间温度分布，导致能耗浪费和热安全隐患。为此，业界逐步采用基于数据驱动的冷却优化方法，通过无线传感器网络实现对数据中心温度的高精度实时监测，并利用深度强化学习 DRL、模型预测控制 MPC 等智能算法，动态调整冷却设定点以匹配实际负载。近期研究表明，模型驱动的 DRL 方法能够显著提升冷却效率，并在保障热安全的前提下实现能耗最优。举例来说，清华团队结合多源温度数据和服务器功耗信息，通过分层控制和可视化决策，实现了冷却资源的动态分配与系统配置的实时重构。该方法不仅提升了冷却系统的能效和可靠性，还优化了资源利用率和运维成本，实现了数

据中心冷却系统在动态负载下的能耗最优与热安全保障，平均节省 10.1% 的冷却能耗 [109]。Microsoft 团队则通过热感知调度，在 GPU 集群的能效、散热和部署密度方面实现了协同提升，并增强了系统应对冷却或电力故障等突发情况的弹性与稳定性 [110]。

智能功耗管理与优化不仅是缓解 LLM 带来能耗激增的关键抓手，也面临在性能、成本与碳排放之间精细权衡的挑战。随着大语言模型在云端的广泛应用，数据中心 GPU 算力需求急剧增长，带来了巨大的能耗压力。由于数据中心通常受到固定电力预算的限制，如何高效管理和优化功耗，成为支撑大规模 LLM 部署的关键挑战。Microsoft 团队系统性分析了云端大语言模型训练与推理的功耗特征，指出推理集群具备显著的功率冗余空间。基于这一结论，他们提出了能够提升服务器部署密度的智能功耗管理方法，通过使用开源模型复制生产中观察到的功耗模式，展示出在现有集群中可以在最小性能损失的情况下增加 30% 的服务器部署密度 [111]。然而，随着推理场景的持续扩展和用户负载的动态变化，单纯依赖静态功耗管理难以充分释放集群的能效潜力。针对推理环境的高度异构性和请求波动性，Microsoft 进一步实现了对 LLM 推理集群的动态能效优化，其能够根据不同请求类型、模型规模和服务 SLO，自动调整实例数量、模型并行度以及 GPU 频率等关键参数，动态划分资源池并实时重构系统配置。通过分层控制与预测调度，不仅显著降低了整体能耗和碳排放，还提升了资源利用率和客户经济效益。实验证明，该方法可在保证服务延迟 SLO 的前提下，可平均节省 52% 的能耗、38% 的碳排放，并将客户成本降低 61%。这一方法为大规模 LLM 云服务的绿色部署和可持续发展提供了系统性解决方案 [140]。

面向未来的大规模云平台，智能功耗管理将进一步与弹性伸缩、负载均衡、任务编排和容错机制深度耦合：通过对应用性能敏感度和延迟容忍度的自动识别，实现跨节点、跨机架乃至跨地域的数据与计算迁移，在保证关键业务性能与可靠性的前提下，主动进行功率封顶、频率调节和资源整形；同时，通过引入强化学习、在线优化与因果推断等智能决策方法，调度系统能够在业务流量波动、硬件老化和能源价格变化等不确定性条件下，实现对能效与性能的持续自适应平衡。由此，数据中心不再仅是“被动用电的算力工厂”，而将演进为具备自主感知、自主优化能力的“能效感知算力基础设施”，为构建绿色、低碳、经济可持续的新型云计算体系提供关键支撑。

1.5 展望与建议

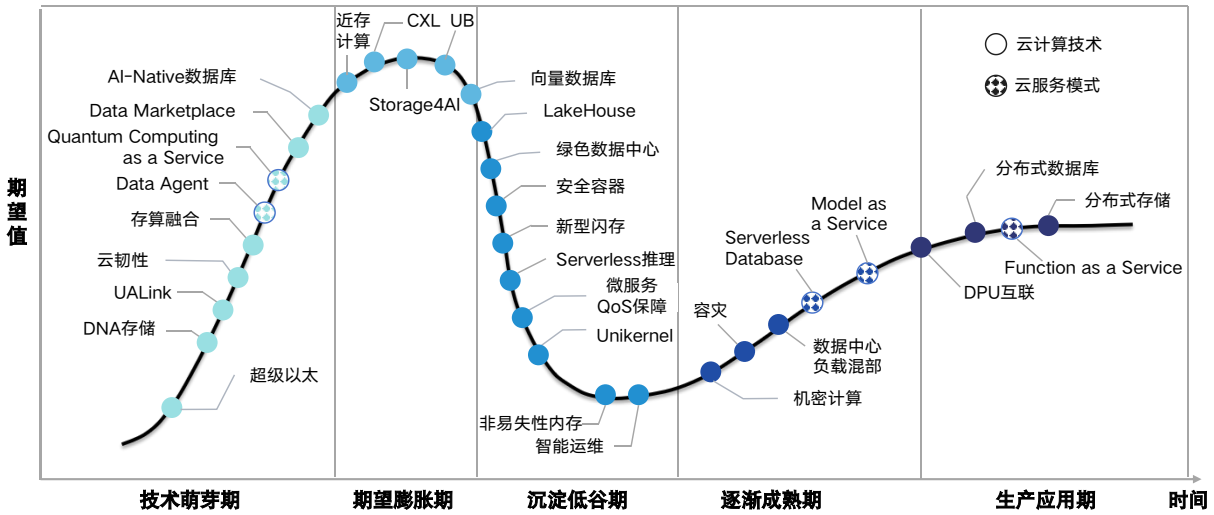


图 1.11: 云计算研究图谱技术成熟度曲线 2025

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，为有关领域的研究与投资提供决策依据。

1.5.1 云计算的未来研究方向和关键技术展望

云计算基础设施正从传统数据中心云向分层多级架构的泛在云架构演进，智能技术的普及将进一步催生云计算行业变革。如今随着 AI 技术的飞速发展，智能化已经与云计算变得密不可分。一方面，泛在云架构、编程模型等系统层的演进将使得其能够有效承载包括大模型训练、自动驾驶等在内的智能化应用和新兴应用的部署需求，充分释放泛在云的计算潜力，并在跨地域、跨异构硬件和多云环境中实现算力的统一编排与弹性供给。另一方面，诸如用户行为预测、智能调优以及智能化编排调度等也能够用于泛在云的业务场景，改善泛在云服务性能，提升资源利用效率与服务稳定性。此外，还可以结合智能运维、智能能效管理等方法优化大规模分布式任务的运行效率与能耗表现，实现性能、成本与绿色低碳目标之间的综合权衡。这些面向人工智能的计算机系统研究和 AI 辅助的系统优化技术共同组成了智能泛在云的内涵，并将推动泛在云架构继续朝向智能泛在云的方向演进，使云平台从被动支撑应用的基础设施逐步演化为可感知、可学习、可自治的智能基础底座。

未来云计算服务模式将以智能化和动态协同为核心，通过平台与工作负载之间的双向实时通信，实现资源管理从“静态分配”向“需求驱动、动态优化”转变。未来云计算服务模式将更加智能化和动态化，平台与工作负载之间的界限逐步模糊，形成高效协同的新型服务生态。工作负载能够主动表达自身特性和需求，包括对延迟、吞吐、可靠性、成本和能耗等多维度指标的偏好，平台则智能感知并自动匹配多种优化措施，如弹性伸缩、算力形态切换、跨层协同调度与差异化保障策略，从而极大提升资源利用效率、成本效益和服务定制能力。在此基础上，平台还可通过在线学习和反馈闭环持续优化决策逻辑，使资源管理策略随业务形态和环境变化自适应演进。这一模式不仅简化了云服务形态，降低了应用开发与运维复杂度，还为新兴应用如大模型训练、实时分析、数字孪生等提供更敏捷、高效且可预测的基础设施支撑，推动云计算向自动化、个性化和可持续发展持续演进。

未来云计算服务模式将不断突破传统资源供给范式，向“行业即服务、智能即服务、场景即服务”等多层次形态演进。依托云网融合和算力网络基础，将分散的算力、算法、数据、应用能力统一封装为标准化服务接口，对外提供覆盖数据、智能、业务与行业场景的一体化“X 即服务”体系，帮助客户以订阅方式快速获取所需能力，实现即开即用、按需伸缩和持续迭代。届时，云平台不再只是输出算力、存储等基础资源，而是面向外部直接提供语义理解、自动化数据分析、实时智能决策、图像识别等各类智能服务，并可根据不同行业如金融、工业互联网、智慧城市、医疗健康等的特定需求，提供深度整合行业知识与场景数据的定制化解决方案。通过在统一技术底座之上沉淀可复用的行业模型、流程模板与安全合规能力，云服务将从“通用能力供给”走向“行业数智操作系统”，极大地提升云服务的灵活性、智能性和创新力，推动云计算成为数字社会的智能基础设施和各行业数字化转型的关键使能者。

1.5.2 云计算的发展建议

加速 AI 原生云平台建设，打造智能化、弹性化的服务能力。随着大模型、智能体等新型业务场景加速普及，云服务商应将 AI 算力、智能调度与自动化运维等能力深度内嵌为平台原生特性。建议重点强化异构算力池化、分离式架构以及智能资源编排与故障自愈机制的研发，推动云平台从传统“资源供给”向“行业即服务、智能即服务、场景即服务”等转型。通过开放 AI 模型服务、行业知识库和自动化工具链等

能力，支持企业和开发者实现业务快速创新与敏捷部署，全面提升服务的智能化水平与定制化能力。同时，应通过构建开放生态和合作伙伴体系，引导更多行业独立软件供应商（ISV）、开发者参与 AI 原生能力的共建共享，形成技术演进与业务创新相互促进的良性循环。在算力计费、服务等级协议（SLA）、数据合规等维度形成适配 AI 原生负载的新型服务与治理框架，降低企业采用门槛与迁移成本。通过在标杆行业率先打造“AI 原生云 + 行业应用”的示范标杆，形成可复制、可推广的解决方案模板，带动上下游生态整体能力跃迁。

完善智能运维与安全防护体系，保障云平台高可用与可信。云服务商需加快智能化运维体系建设，推动多模态异常检测、自动化根因分析、自愈调度等技术在大规模云平台落地。建议在平台层面构建 AI 驱动的安全运营体系，覆盖模型输入、推理过程、资源调用、输出行为等关键环节，实现实时风险发现与响应。同步加强供应链安全、零信任架构、内存安全等多层防护能力，保障云平台和用户业务的可靠性与合规性。建议同步建立统一的运维数据采集与闭环反馈机制，利用运维大数据持续优化算法模型和策略规则，提升平台运维与安全管理的自动化和前瞻性水平。通过发布透明的安全实践报告和合规认证成果，增强用户对云平台的信任度，为关键行业和核心业务上云提供有力支撑。同时，应强化对 AI 本身带来的新型安全风险（如模型窃取、对抗样本、数据投毒等）的研究与防护，将智能安全能力前移到开发、测试与部署全生命周期，实现从“补救式安全”向“内生可信”演进。

推动绿色低碳技术创新，实现云数据中心可持续发展。建议云服务商积极布局智能能效管理、碳感知调度、绿色算力等新技术，优化数据中心能源结构和碳排放。通过产学研协同创新，推动绿色计算标准和最佳实践在行业规模化落地，助力全球数字经济绿色转型。鼓励企业采用可再生能源、智能冷却、能耗优化等措施，提升数据中心运营效率，实现经济效益与环境可持续的双赢。同时，应探索将碳排放、能效指标纳入资源编排与业务调度决策，以算法优化引导业务向低碳资源池和高效机房聚集。通过建立绿色算力评估体系和激励机制，引导用户优先选择低碳云服务产品，推动产业链上下游共同参与绿色转型。进一步地推动绿色算力与金融工具、政策激励相结合，通过绿色认证、差异化定价等方式，增强市场对低碳云服务的内生需求。与此同时，需加强绿色技术指标的监测与透明披露，构建可对比、可验证的能效与碳排放评价体系，为监管部门、行业组织和用户决策提供科学依据。

第 二 章

面向云网融合的研究

云网融合是中国电信自 2016 年提出的数字化转型核心战略，通过整合云计算与通信网络构建智能化数字基础设施。2020 年 11 月发布的《云网融合 2030 技术白皮书》[141] 详细阐述了云网融合的意义和愿景，目标技术架构和发展阶段。2025 年 12 月，配合中国电信“云改数转智惠”的战略升级，《云网融合 2035 技术白皮书》[142] 正式发布。云计算研究院作为中国电信集团承载前沿研究创新使命的专业研究机构，积极承接云网融合的相关技术研究工作，并且深度参与《云网融合 2035 技术白皮书》的撰写，在云网融合科学理论内涵和创新方向上发挥重要作用。

本章将从云计算研究院的视角出发，系统性地探讨面向云网融合的关键研究。首先，2.1 节将以《云网融合 2035 技术白皮书》为基础，提炼和梳理驱动未来发展的层次化、大颗粒度关键趋势，并据此更新面向云网融合的研究图谱。接着，2.2、2.3 和 2.4 节将分别聚焦并详细阐述三项核心热点研究方向：云网一体化调度，面向智算的云网基础设施，以及云边端协同。最后，2.5 节将对本领域未来的发展方向提供前瞻性展望与建议。

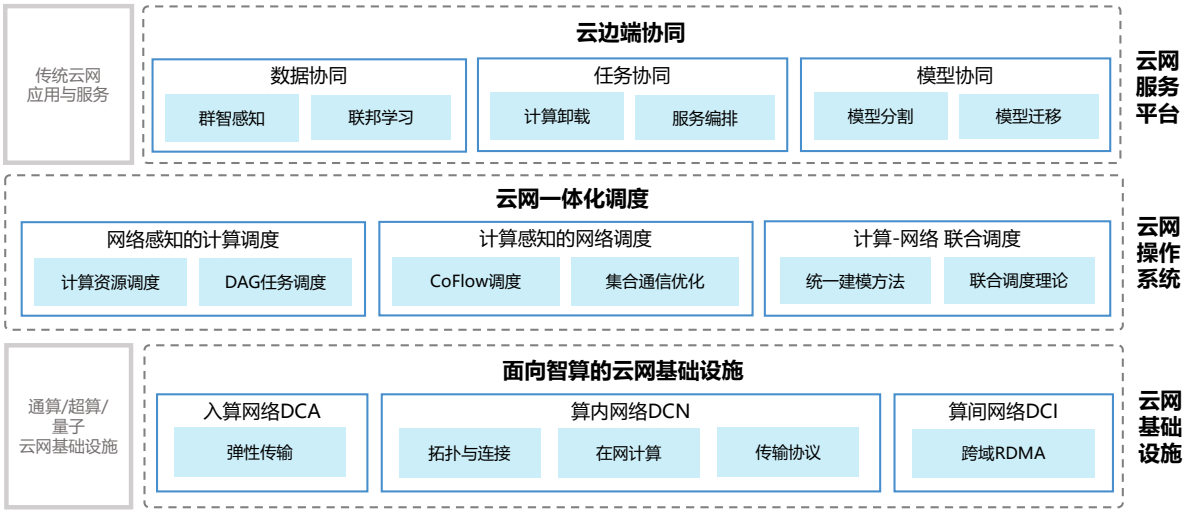


图 2.1: 面向云网融合的研究图谱 (由云计算研究院总结形成)

2.1 研究图谱 2025：战略升级的解读与研究承接

本节简要解读《云网融合 2035 技术白皮书》，围绕“融合”与“融智”两大核心特征，并依托“供给-运营-服务”三层技术架构，分别提炼各层驱动未来发展的关键趋势，并根据趋势总结热点研究方向，更新的研究图谱如图 2.1所示。

2.1.1 趋势分析

本节简要解读《云网融合 2035 技术白皮书》所揭示的战略升级背景，并基于“供给-运营-服务”的三层核心技术体系架构，分析了各层的关键性发展趋势。

2.1.1.1 背景：“云改数转智惠”的战略升级

2025 年，中国电信将战略升级为“云改数转智惠”，构建围绕 AIDC/DC 的高质量网络为基础、云计算为核心、数据与 AI 为引擎，集融合融智、弹性泛在、安全绿色、生态开放等特征于一体的新型基础设施与服务体系。《云网融合 2035 技术白皮书》是基于《云网融合 2030 技术白皮书》的升级，其核心驱动力是智能化时代对数字信息基础设施的迫切需求。AI 大模型技术与应用（如 ChatGPT）的爆发式发展，正在深刻重构计算和网络体系，对算力、带宽和数据安全提出了前所未有的要求。

云网融合在延续“网是基础，云为核心，网随云动，云网一体”的发展原则下，新增“智惠共生”，将高价值的智能云网能力转化为易获取资源与服务，让各类客户平等享受技术红利；以安全内生、绿色低碳为底色，保障生态可持续发展，实现技术价值与社会价值的统一。在此背景下，云网融合的核心特征也随之升级，不再只是资源的简单叠加和统一，而是突出“融合”和“融智”两大特征：融合，体现架构开放融合和云边网业融合，实现架构与资源的深度整合；融智，包括云网内生融智和服务生态融智，推动 AI 要素内生融合到云网体系的各个层面。

云网融合的科学理论内涵是以数据驱动-效能优化-能力进化为核心逻辑链，支撑战略落地。数据驱动为理论基础，通过全域设备采集全量数据，经网络传输至云端存储计算，再依托 AI 挖掘数据价值，形成数据全生命周期闭环，并以此为核心依据，指导基础设施建设；效能优化为理论核心，强调资源最优分配。依托最优化理论体系，基于数据驱动环节形成的全量数据，通过操作系统一体化调度，实现资源供需精准匹配与跨域补位，达成异构资源全局最优；能力进化为理论目标，实现自适应的大规模系统演进。在效能优化形成的高效资源分配基础上，依托复杂系统理论，赋予云网融合 2035 服务体系自感知、自决策、自优化能力，动态适配外部业务需求迭代与内部管理升级。

面向 2035 年，白皮书提出了云网融合愿景架构，形成如图 2.2所示呼应“供给-运营-服务”层次分明的三层体系。其中，智能云网基础设施（供给侧）是整个体系根基，以终端/边缘云结合接入网及边缘云/中心云结合骨干网为核心，构建面向“数据驱动”的算网存一体化资源。智能云网操作系统（运营侧）是核心层，通过对统一调度和编排，实现资源的一体化、智慧化运营，达成“效能优化”。智能云网服务体系（服务侧），面向全球用户提供全栈智能云网服务，实现“能力进化”目标。白皮书中的“供给-运营-服务”三层架构，不仅定义了未来的目标形态，也明确了本章研究热点的三大核心发展趋势。

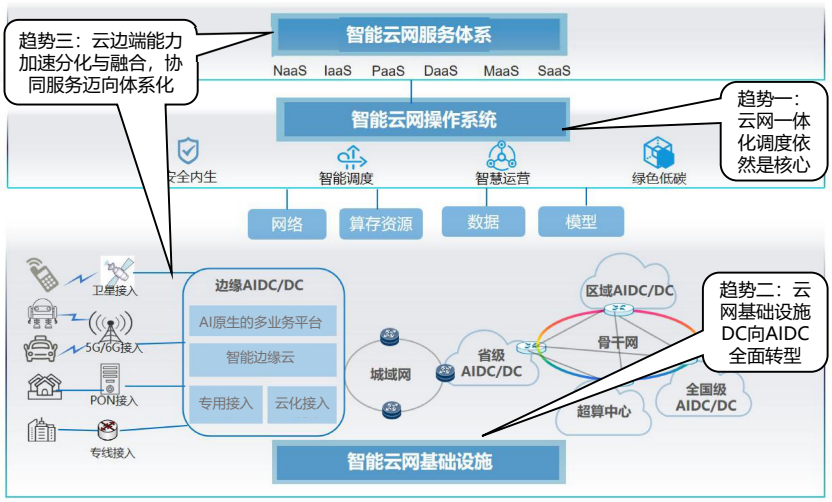


图 2.2: 云网融合三层愿景架构（引用自《云网融合 2035 技术白皮书》，背景风格为适配本文略有调整）

2.1.1.2 云网一体化调度依然是核心

云网一体化调度是《云网融合 2030 技术白皮书》中明确提出的核心理念 [141]，其本质是通过统一的资源视图与调度逻辑，向下实现计算与网络资源的高效整合，向上为多样化服务与应用提供一致的能力支撑。该理念与早期学术界在数据中心研究中提出的任务—流量协同调度 [143, 144]、公有云厂商在数据中心内部探索的算力—带宽联合优化思想一脉相承 [145, 146]，并在此基础上进一步拓展至云、网、边、端的全域资源体系，从而将以往分散的研究与工程实践上升为面向云网基础设施系统化、工程化的方法论。云网一体化调度不仅是中国电信在产业领域的体系化创新，也是学术界相关理念的延展与深化。

随着云网基础设施向智能云网形态演进，计算与网络在资源结构、负载模式与运行特性上都呈现出新的趋势，云网一体化调度依然是支撑未来关键业务的核心能力。一方面，计算侧的任务规模不断扩大，模型训练、推理服务与复杂工作流在执行过程中产生大量细粒度、频繁交互的通信需求 [147, 148]，使得任务性能愈发依赖底层网络状态。另一方面，网络侧在多租户、高并发环境下呈现带宽竞争、突发流量、阶段性峰值等强动态特征 [149, 150, 151]，不同业务的通信行为差异巨大，要求网络调度能够理解并响应计算任务在资源使用上的变化，为系统提供稳定、可预期的传输支持。计算对通信的依赖与网络对计算的敏感共同加深了云网资源之间的耦合 [152]，使单一域内的调度策略难以满足整体性能目标。云网一体化调度将任务需求、流量结构与基础设施状态统一纳入模型，实现面向全域的联合优化，使资源调度从局部最优迈向整体最优。

从行业发展来看，云网一体化调度正从概念验证走向工程化实践。中国电信在云网融合的战略规划下持续加强云网一体化布局，推出了息壤算力服务平台、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 [153, 154]；中国移动提出了一朵云、一张网、一体化服务体系和算网大脑调度体系 [155]；中国联通构建了 Cube Net 和云网一体化资源调度与算网一体化编排调度平台 [156, 157]。三大运营商在云网边端协同、算网大模型、资源图谱等方向形成体系化布局。AWS 在 Lambda、SageMaker 以及 EFA / Nitro 架构上形成了计算–网络闭环优化机制，通过加速网络与任务调度反馈实现可预期性能；Azure 在应用编排、流量工程和虚拟网络中构建了 Host-level 与 Fabric-level 联动的多层调度逻辑；Google 依托 Borg + Jupiter 与 Andromeda 打造跨调度器和网络系统的端到端协同体系；Meta 和 OpenAI 则在超大规模 AI 训练基础设施中持续推进计算资源调度与 RDMA 网络调度的联合优化 [147, 158]。

2.1.1.3 云网络基础设施 DC 向 AIDC 全面转型

随着生成式人工智能的不断发展，大模型在预训练-后训练-推理各个阶段均对网络提出新的需求。在预训练阶段，模型规模遵循 Scaling Law 持续扩大，GPU 数量、互联密度和带宽需求同步增长，使网络成为训练周期的主导瓶颈；在后训练阶段，随着基础模型开源化与大模型行业化落地，企业用户的微调、对齐、蒸馏需求激增，对跨园区传输安全性、数据不落地以及算力就近调度能力提出更高要求；在推理阶段，海量用户与高并发场景推动 PD（Prefill-Decode）分离架构、专家并行和云边协同推理广泛应用，推理链路对低尾时延、高弹性与快速扩缩容的依赖显著增强。在新的需求下，传统云数据中心 IDC（Internet DC）在网络性能上已出现瓶颈，难以高效支撑大规模 AI 网络负载。

AI 算力成为我国算力的主要增长，推动传统云数据中心 IDC 向智算中心 AIDC 加速演进。图 2.3 显示了 2017 到 2023 年我国智能算力规模的快速跃升，其中智能算力占比从不足三成提升至接近三分之二，增长速度远超通用算力。这一趋势表明算力市场的主导需求正在从通用 IT 任务转向 AI 原生任务，进而要求网络具备更高的带宽速度、更强的并发通信能力以及跨园区资源协同能力 [159]。表 2.1 总结了传统 IDC 与 AIDC 的关键差异。传统

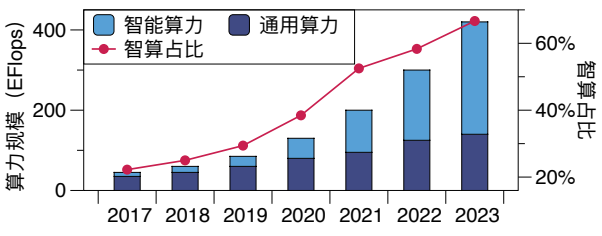


图 2.3: 我国算力规模及增速示意图

数据中心的设计目标以通用业务承载、虚拟化资源管理和成本效率为主，网络流量相对较小、平稳，对性能要求较低。而 AIDC 面向 GPU 主导的并行计算，要求网络能够支撑多维度并行策略（数据并行 [160, 161]、模型并行 [162, 163]、流水并行 [164, 165]、专家并行 [166, 167, 168, 169] 等）下的高频通信，拓扑从树状结构演进为、Torus [170]、Dragonfly [171] 等高带宽、高均衡度的拓扑体系。与此同时，AIDC 对供电密度、散热能力、延迟分布和尾时延控制提出更高要求。随着模型尺寸、训练周期和推理并发数的增长，网络性能正成为影响算力效率和成本结构的核心瓶颈。

表 2.1: 传统云数据中心与智算中心的关键差异

	传统云数据中心 IDC	智算中心 AIDC
理论	虚拟化算法和设计理论	深度学习理论
拓扑	树型拓扑为主	树、Torus、Dragonfly、Slimfly 等
算力	以 CPU 为主	以 GPU 为主，CPU 为辅
架构	冯 · 诺依曼架构；CPU 分配任务给其它部件	全互联对等架构；允许处理器之间直接通信
流量	相对较小、平稳	海量数据、高突发
网络	普通以太网，10G–100G；TCP/IP 协议	无损以太网，200G/400G；RDMA，UEC
机房	分布式、低密度、低功率机柜；风冷为主	集中式、高密度、高功率机柜；液冷为主

2.1.1.4 云边端能力加速分化与融合，协同服务迈向体系化

云边端协同体系如图 2.4所示，是面向未来云边端融合架构提出的关键理念，其本质目标是在不同层级的计算、数据与模型能力之间构建统一的运行抽象与协作逻辑，使系统能够在多层资源环境下实现能力组合、动态调度与连续演化。该理念继承了分布式系统领域关于数据–任务–模型一体化执行优化的研究脉络，也吸收了云计算与边缘计算体系中关于资源协作、近源处理与终端智能的演进趋势，并在此基础上进一步扩展为覆盖云、边缘和终端的全域协作体系。云边端协同体系不仅是与云网融合背景下的体系化升级，也标志着工程实践从分层部署向跨层协同的方法论演进。

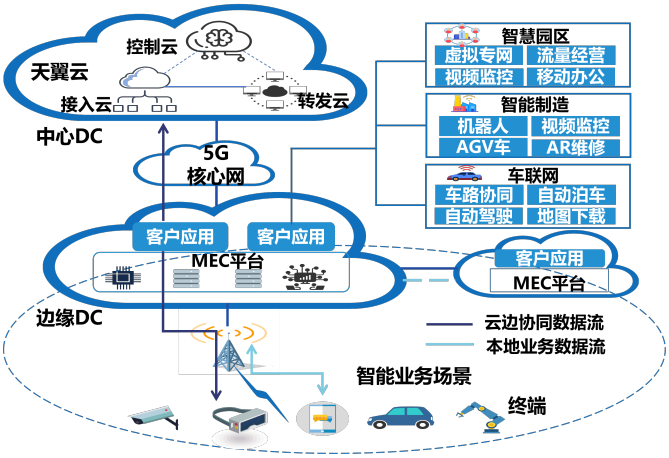


图 2.4: 中国电信云边端协同体系架构

随着终端智能化、边缘节点规模化以及云侧模型能力持续增强，业务在资源形态、执行模式与运行链路上均呈现新的趋势，云边端协同仍将是支撑未来系统能力的核心方向。在数据层面，数据的生产呈现出高频、多模态、多主体并发的特性，大量数据需要在采集端与边缘端进行低延迟处理与智能筛选；任务层面，逐步从云侧集中式执行转向多层分布式执行，许多场景需要在终端与边缘侧完成部分决策、推理或协作处理；模型层面，逐步从静态推送演化为动态更新、持续优化的形态，云侧训练、边侧适配与端侧轻量推理形成全链条联动。这些趋势共同增加了各层之间的耦合复杂性，使得传统以单层为中心的体系难以满足多样业务的性能、效率与智能化需求，亟需统一的跨层协同框架才能实现整体最优。

从行业发展来看，云边端协同正在从探索阶段走向体系化建设，并成为算力基础设施升级的重要方向。国内运营商在算网大模型、边缘智能节点、统一边缘框架等方面持续推进能力建设，云厂商则在多层协同调度、边缘容器运行时、端侧模型推理等领域形成体系化方案。国际云服务提供商在大规模边缘节点调度、终端侧智能增强与跨层模型分发方面积累了丰富实践，终端设备厂商在本地推理、端侧隐私保护与协同学习机制方面持续发展。国内外主要企业边缘节点规模与生态能力对比如表 2.2所示，产业趋势共同推动云边端协同从概念验证走向工程体系化落地。

表 2.2: 国内外主要企业边缘节点规模与生态能力对比

企业	边缘节点规模	覆盖区域	典型产品	生态特点
中国电信	2000+	300+ 城市	天翼边缘云	全国最大边缘部署
中国移动	1800–2000	31 省 / 333 城市	移动边缘云 MEC	站点级覆盖最密
中国联通	900–1200	300+ 城市	联通边缘云	千级规模稳定
阿里云	300+ 城市	全国	ENS / ASK-Lite	云边端协同平台
腾讯云	1500+ 国内	全球 2800+	EdgeOne	全球覆盖强
华为云	千级	300+ 城市	Stack Edge	与运营商深度融合
AWS	600+ POP	全球	Local Zones	云 + 运营商生态
Microsoft Azure	250+	全球	Edge Zones	5G 协同能力强
Google Cloud	180+	全球	GDCE	强 AI + Edge
Akamai	4100+	135 国	EdgeWorkers	全球最大边缘网络

2.1.2 方向聚焦

云网融合的研究重要且宏大，七大战新领域紧密围绕在云网融合核心战略，各自承接相应的研究工作。本白皮书从云计算的视角出发，基于“供给-运营-服务”三层架构中总结的核心发展趋势，聚焦相应的热点方向并进行相应展开。

运营侧：云网一体化调度从传统的计算调度、网络调度相互割裂的领域，收敛为以统一资源视图与联合优化模型为基础的整体性方法论。近几年，研究者围绕有向无环图 DAG（Directed Acyclic Graph）任务调度、集合通信优化、并行任务流水线、算力与带宽的联合分配、跨层调度等方向持续深化 [172, 173, 174, 175, 176]，逐步形成了从单域资源优化到算网联合编排的系统化方法体系。在云边协同、AI 训练/推理、复杂工作流执行、大规模资源管理等真实场景中 [177, 178, 179]，一体化调度相关模型与算法已经不断得到验证并系统化落地。总体来看，相关研究在模型抽象、算法优化和跨域协同机制上不断深化。

供给侧：面向智算的云网基础设施体系正在围绕“入算网络、算内网络、算间网络”形成系统化的三层能力重构。入算网络强调大带宽数据注入、安全可信传输与跨园区数据协同；算内网络面向万卡级训练任务 [180]，以可编程交换机、高维互联拓扑 [170, 171, 181]、智能拥塞控制与光电融合网络 [182, 183] 为核心，实现超低时延、超高带宽的集群内部通信；算间网络构建跨地域算力池化体系，通过长距离 RDMA 技术 [184]、与智能调度实现跨中心算力融合。随着国家“东数西算”布局和中国电信“2+3+7+X”算力体系建设加速推进，这三层网络体系将构成我国 AIDC 的能力底座，为全国范围内的算力互联、智能应用落地与算网一体化发展提供长期支撑。

服务侧：云边端协同正在数据协同、任务协同与模型协同三大方向上不断走向深化与成熟。围绕多模态数据管理、多级任务调度、模型迁移与持续学习等关键问题，研究持续推动智能从单点处理走向跨层协作。在数据层面，分布式采集、语义抽象与隐私前置不断发展，通过轻量化多模态表征 [185, 186]、边缘侧数据压缩与筛选 [187, 188] 以及本地差分隐私与匿名统计机制 [189, 190] 提升数据质量与安全性。在任务层面，调度模式从集中式 DAG 管理扩展到云—边—端协同执行，通过分布式依赖管理与动态调度实现复杂任务的高效运行 [191]。在模型层面，模型切分、端侧适配与跨层迁移成为主要方向，使模型在动态网络环境中保持性能稳定与资源高效 [192, 193, 194]。整体来看，云边端协同研究正由局部优化迈向全链路协同，为构建自适应、可持续的智能基础设施奠定了重要技术支撑。

综上，我们根据智能云网体系的发展趋势在各层进行系统性聚焦。其中，运营侧核心突破计算网络协同的云网一体化调度，供给侧全面构建面向入算、算间、算内的云网基础设施，服务侧持续深化面向数据、任务、模型的云边端协同。下文将围绕以上三项大颗粒热点方向进行系统性展开与深入分析。

2.2 热点方向四: 云网一体化调度

云网一体化调度的核心在于将计算与网络纳入统一模型，通过一致的优化逻辑来实现端到端优化目标。根据调度的侧重点的不同，可以将其划分为三个互补的方向：在网络感知的计算调度中，调度系统以计算任务为中心，通过网络状态指导任务放置、执行和迁移；在计算感知的网络调度中，调度系统以网络资源为中心，通过识别计算阶段和依赖结构来优化带宽分配与路径决策；而在计算–网络联合调度中，同时调度计算资源和网络资源实现全局资源效率的最大化。这三类方法共同构成云网一体化调度体系的核心框架，也形成后续章节展开的主要脉络。

2.2.1 网络感知的计算调度

网络感知计算调度指在制定计算任务放置和资源调度决策时，除了考虑 CPU/GPU/存储等资源信息时，还会明确考虑网络状态和容量等信息。传统的调度器通常优先考虑即时可用的计算资源，例如空闲的 CPU 或 GPU 资源，并将网络视为被动的、稳定的通道。相比之下，网络感知计算调度将网络视为一种动态的、竞争的、异构的资源，可能成为关键的瓶颈，影响调度决策的效果。该技术通过持续监控可用带宽、延迟和拓扑结构等网络指标，智能地部署计算任务和路由数据流，从而最大限度地减少通信开销，并提高应用程序的整体性能和集群效率。具体而言，网络感知计算调度可以分为计算资源调度和基于 DAG 的任务调度这两方面问题，本节将从这两个方面展开，详细介绍目前网络感知计算调度的研究方案和仍存在的问题与挑战。

2.2.1.1 计算资源调度

计算资源调度指在计算集群或超级计算机上将任务分配给适当的计算节点，以充分利用 CPU/GPU 等资源。经典的调度系统包括两类：一类是批处理作业调度（如 HPC 中的 SLURM [232]），另一类是集群调度（如 Kubernetes [233]）。批处理调度器传统采用队列技术，当队首的大型作业等待资源时，小的后续作业可插空先运行，只要不延误大作业的开始时间，这种策略显著提高了资源利用率和集群吞吐量。在线集群调度则需要同时考虑效率和公平。例如，Hadoop YARN 的 DRF (Dominant Resource Fairness) 算法 [234] 确保每个用户在多种资源维度上获得大致公平的份额，同时最大化总体吞吐；Spark 的延迟调度 (Delay Scheduling) 策略 [235] 通过短暂等待来换取任务在数据所在节点上运行的机会，从而兼顾数据本地性和公平调度。总体而言，计算资源调度的核心研究问题在于如何有效地分配有限异构资源来满足多任务需求，同时达到高集群利用率、公平性和服务级别目标 SLO。这一问题属于 NP-hard 的组合优化，涉及多维资源和动态约束，学术界通过启发式算法、优先级规则等手段求解。

在通算领域，网络感知的计算调度可以通过优化服务部署位置来减少延迟和资源浪费，从而满足服务质量要求。在最近 5 年的国际学术会议期刊中，有相当数量的论文关注此方向，旨在减少因网络瓶颈导致的尾延迟和资源浪费。对于通用计算，尤其是在微服务和地理分布式云环境中，基于容器或微服务等实例的服务通常对延迟有严格的服务级别目标 SLO。调度器（例如 Kubernetes）基于 CPU/内存进行装箱调度，通常忽略了将两个通信的微服务放置在不同的机架或区域会引入显著的延迟。网络感知的计算调度从根本上改变了部署策略。调度器不再简单地将微服务打包到任何具有可用 CPU 周期的宿主机上，而是评估服务之间的通信模式。例如，两个频繁进行大容量数据交换的微服务会被部署在同一台物理服务器上或同一机架内，以利用高速本地链路，从而最大限度地减少交换机间的流量并降低延迟。ServiceRouter [196] 阐述了 Meta 的行业解决方案如何演进，以基于位置和网络负载路由请求，从而有效地将计算请求调度到最近的具备网络连接能力的实例。IBM 提出了新型网络感知调度框架 Diktyo [195]，该框架通过确定长期运行应用程序中依赖微服务的位置，从而减少服务的端到端延迟并保证带宽预留。

在智算领域，网络感知的计算调度通过感知通信需求与网络拓扑特征等方式来优化作业放置，减少智算任务中的网络瓶颈，从而降低尾延迟并提升资源利用率。GPU 密集型训练任务可能会被调度到具有

表 2.3: 云网一体化调度研究领域热点

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
网络感知的计算调度——计算资源调度	依据网络状态进行计算资源调度决策，对通信延迟等方面进行显示建模，从而满足计算服务的高吞吐要求。	OSDI NSDI TNSM	• **通算资源调度：** IBM 提出 Diktyo [195] 通过确定微服务的位置，减少服务的端到端延迟并保证带宽预留。Meta 提出 ServiceRouter [196] 基于位置和网络负载将计算请求调度到紧邻实例。 • **智算资源调度：** CMU 团队提出 Pollux [197] 通过显式建模通信延迟实现面向高吞吐量的协同自适应集群调度。
网络感知的计算调度——DAG 任务调度	考虑网络状态和任务间依赖关系，通过将任务关系抽象为 DAG 图，并通过对图的合理划分进行任务放置调度决策。	ASPLOS OSDI TPDS	• **通算任务调度：** Marmara 大学团队提出经典 HEFT [198] 启发式方案，计算任务优先级并依此分配任务。微软联合 UW-Madison 提出 Graphene [199] 通过考虑微秒级网络开销来调度 DAG 任务。 • **智算任务调度：** NVIDIA 联合 CMU 团队提出 GraphPipe [200] 借助 DAG 将模型划分为可并发执行的流水线阶段；MIT 团队提出 Si-PML [201] 实现感知网络拓扑的调度算法，优化训练速度。
计算感知的网络调度——CoFlow 调度	基于流之间的依赖关系，以优化整个传输任务为目标，计算每个流发送速率和流间传输顺序，研究方向聚焦最优算法设计和实际系统实现。	SIGCOMM INFOCOM ICDCS	• **中心式 CoFlow 调度：** Berkeley 大学 Varys [144] 提出最小瓶颈优先算法；香港科技大学 Rapier [202] 联合优化流调度与路由。 • **分布式 CoFlow 调度：** 香港科技大学提出 Optas [203]，通过赋予小任务最高优先级有效降低开销并优化小任务完成时间。D-CAS [204] 提出近似比为 2 的分布式调度算法。 • **信息未知的 CoFlow 调度：** Berkeley 大学提出的 Aalo [205] 基于已发送大小决定流的优先级，近似实现服务最短优先。
计算感知的网络调度——集合通信优化	根据集群网络拓扑、并行策略及通信-计算依赖关系，自动生成最优集合通信算法，并将分解通信操作进行任务内和任务间的精细化调度，以最大化通信-计算的重叠和整体系统吞吐量。	SIGCOMM NSDI SOSP ICNP HotNets	• **集合通信算法：** 百度提出带宽最优的 Ring AllReduce 预定义算法 [206]；微软 SCCL [207] 和 TACCL [173] 求解最优集合通信合成；阿里提出的 SyCCL [208] 利用拓扑对称性进行加速。 • **任务内集合通信调度：** 字节 ByteScheduler [209] 引入统一抽象机制；Meta 公司 SYNDICATE [210] 将通信操作分解为更小的子操作；北大 Centauri [211] 采用三种分割方法精细化弹性调度。 • **任务间集合通信调度：** 北大提出 Muri [212] 和 MIT 提出 Cassini [175] 采用集中调度器（分布式 MLTCP [213]）计算每个训练启动时间偏移；华为和电信云计算研究院提出 Symphony [214] 基于到达时间实现任务交错；阿里团队提出 Crux [151] 引入基于 GPU 强度的优先级调度，最大化 GPU 利用率。
计算-网络联合调度——VNE	通过虚拟节点与虚拟链路的联合映射实现计算–网络资源的统一调度，近年来聚焦于可扩展性、动态在线嵌入与 SLA 保障。	SIGCOMM INFOCOM TNSM	• **可扩展 VNE：** 卢森堡大学团队提出基于并行链路映射的高维嵌入框架，用于提升大规模 VNE 的可扩展性 [215]；中国移动通信联合实验室提出在线 GNN + SLA 感知嵌入机制 [216]。 • **动态与在线 VNE：** 中国科学院团队提出面向 MEC + Optical 的动态 VNE，通过强化学习缓解资源碎片问题 [217]；澳大利亚团队提出分布式 DeViNE 框架，实现自治式虚拟网络嵌入 [218]。 • **跨域 VNE：** 北京邮电大学团队探索能耗优化的跨域 VNE [219]。
计算-网络联合调度——VCE / 软管模型	基于聚合带宽约束的虚拟集群抽象，重点关注可预测网络、骨干鲁棒规划与虚拟机弹性调度。	SIGCOMM INFOCOM ICC	• **可预测网络：** 清华/阿里团队提出 vFabric，在可编程数据面上通过软管抽象实现可预测端到端带宽 [220]。 • **Backbone 软管规划与鲁棒设计：** Meta/Google 团队提出针对不确定流量矩阵的软管规划机制，实现骨干网络容量与调度的鲁棒化优化 [221]；同团队采用 Benders 分解构建跨层软管规划框架，用于算网协同规划 [222]。 • **弹性调度模型：** 中国电信云计算研究院团队提出树/图结构上的最大弹性调度理论，实现虚拟机的可扩展分配 [223, 224]。
计算-网络联合调度——SFC	以 VNF 节点与路径顺序为核心，联合考虑计算负载、带宽/时延与跨节点状态同步，向智能化、动态化方向演进。	INFOCOM TON TNSM	• **联合映射与路由优化：** 纽约州立大团队提出联合资源管理 + 流调度框架，支持混合边缘–云场景 SFC 部署 [225]；同团队提出可证明高效的 Traffic-sensitive 放置与路由算法 [226]。 • **动态与在线 SFC：** 纽约州立大团队基于在线学习的动态部署方法 [227]；IBM 团队提出 Edge-to-cloud 快速近似最优部署机制 [228]。 • **可编程数据面：** 华为团队提出基于可编程交换机的租户级 SFC 加速 [229]；深圳大学等团队提出深度强化学习 SFC 嵌入 [230]；乔治亚理工学院团队提出可证明高效的 SFC 保护 + 嵌入模型 [231]。

空闲 GPU 资源且连接数据源（例如存储服务器）的高带宽、低延迟网络路径的节点上，以确保 GPU 算力能够快速获取数据，避免因为网络拥塞而闲置等待数据传输。尤其是在分布式数据并行 DDP 或模型并行 MP 中，网络通常是瓶颈所在，网络感知调度器在放置和调度作业时会考虑这些因素。例如，Pollux [197]、Centimani [236] 等工作通过显式建模通信延迟、避免网络竞争来选择 GPU 资源，从而将网络通信开销纳入算力资源调度决策。

尽管已有诸多研究进展，在超大规模、异构且动态的计算环境下，计算资源调度在实现兼顾效率、公平性等方面仍面临挑战。计算规模与复杂性不断增长的背景下，在超大规模集群（数万节点）确保调度决策速度的情况下又要追求最优非常困难，次优的决策在大规模上可能累计造成资源浪费。难以预测的工作负载动态性也为资源调度带来挑战，集群混合了短平快的无状态服务、长周期的深度学习训练、突发性的批处理分析等，各具不同瓶颈和优先级，调度器需要在实时变化中做出权衡。多类异构资源耦合问题也是资源调度常面对的难题，任务往往同时消耗 CPU、内存、IO 带宽等资源，如何公平分配并避免资源成为系统瓶颈是持续研究课题。此外，传统计算资源调度忽视网络拓扑与流量带来的影响，这正是网络感知的计算调度兴起的原因。调度器需要避免将大量互通任务分散在网络远端，引发高延迟或拥塞，如何将网络因素融入调度决策而不增加过多复杂度，依然是一项开放挑战。尽管计算资源调度技术已发展出多种经典策略，但在超大规模、异构、动态的环境下，实现快速、公平且全局最优的网络感知的计算调度依然是学术界和工业界共同关注的方向。

2.2.1.2 基于 DAG 的任务调度

基于 DAG 的任务调度聚焦于带有依赖关系的工作流或作业，DAG 既可以表示单个作业（其节点表示任务、边表示数据或控制依赖），也可以表示一个工作流，其中任务之间存在依赖关系。典型场景包括大数据处理中的 MapReduce/Spark 作业、复杂科学计算工作流、以及拆分为子模型阶段的大型 AI 训练任务等。核心研究问题是在满足依赖约束下优化整体完成时间或吞吐，这需要平衡各任务的执行顺序和资源分配。经典方案中，早期系统强调数据本地性，例如 Hadoop 默认在数据块所在节点启动任务，等待具有数据的节点空闲以调度任务，从而减少网络传输。MapReduce 框架按照阶段划分任务组（Map 阶段、Reduce 阶段），Reduce 任务通常会等待大部分 Map 任务完成后再启动，以避免过早产生大量跨节点数据传输。Spark 改进了这一点，通过 DAG 调度器细粒度地管理任务，结合延迟调度和推理执行（对慢任务重启副本）等技术，使典型 Spark 作业比传统 MapReduce 更高效地利用资源，缩短尾延迟。学术界也提出许多调度算法来最小化 DAG 的关键路径。经典的 List Scheduling [237] 及其改进算法（如 HEFT [198]）常被用于 DAG 调度启发式方案，核心思想是计算每个任务的优先级（考虑后续链路的最早结束时间）并依此分配任务至不同处理单元。

针对通算任务中存在的动态性和通信开销等挑战，研究者提出多种方法来优化基于 DAG 的任务调度效率。针对 DAG 任务调度的难题，目前研究者已经探索了多种思路。关键路径优先方案通过找出 DAG 中决定全局执行时间的关键路径任务，优先为其分配最快速的资源或尽量并行执行。例如调度算法会针对关键路径上的任务降低等待时间，非关键路径不会被分配过多资源，从而缩短总长。但是在复杂 DAG 中，关键路径可能随运行进展而动态变化，识别和跟踪仍有挑战。通信感知调度方案不仅考虑任务计算时间，还建模任务间通信成本以优化放置和顺序。离线规划与在线调整相结合的方案将复杂问题分解为离线近似优化及在线执行修正。例如，如果一个 Reduce 任务需要来自集群中多个分散的 Map 任务的数据，则网络感知调度器可能会延迟 Reduce 任务，直到网络拥塞缓解，或者将 Reduce 任务放置在聚合大部分流量的交换机上。可观测性领域的工作 [238] 展示了通过收集观测数据理解 DAG 的关键路径与网络延迟之间的关系。此类工作可与调度器共同部署，为调度器（如 Graphene [199]、CAPSys [239] 等）提供网络状态信息，从而综合考虑任务需求和网络开销通信竞争等方面来调度 DAG 任务。

针对智算任务调度，研究者通过模型划分与流水线并行等策略优化 GPU 资源利用，以减少通信等待时间并提升训练效率。对于机器学习训练/推理任务，模型的规模对于单个 GPU 来说过大，因此它们需要被拆分为子模型的 DAG，调度器需决定如何将模型图划分到各个设备上，以最大限度地减少等待网络数

据的时间（Bubble Time）。SiP-ML [201] 提出感知网络拓扑的并行化算法实现在确保 GPU 的通信度不超过网络限制的情况下最小化训练迭代时间。GraphPipe [200] 借助 DAG 将模型划分为多个可并发执行的流水线阶段，并基于阶段间的依赖关系调度前向与反向计算任务，实现更高效的训练流程。

基于 DAG 的任务调度在大规模复杂场景下面临动态不确定性、全局优化、计算与数据传输权衡及规模复杂性等多重挑战，仍需持续创新以实现鲁棒高效的调度。不确定性和动态性为调度带来挑战，调度器通常假设以任务运行时间和数据大小为基准，但在共享集群中实际可能因资源争用、抖动而环境变化，导致预先规划失效，自适应实时 DAG 执行计划的调整仍属难题。多作业/多用户环境下的全局优化也是亟待解决的难题之一，研究多聚焦于单个 DAG 作业的优化，但在实际集群中往往有多个 DAG 并发。一个作业的理想调度可能损害整体公平性或抢占资源，引发其他作业重大延迟。因此调度需要在作业级优化与集群整体公平之间权衡。计算与数据传输间的效率也需要权衡，经典数据局部性原则强调搬运计算而非数据，但在复杂 DAG 里并非总是最佳，有时复制数据可能比等待计算资源空闲更高效。如何在调度中自动决策输送数据还是迁移计算也是研究方向之一，此外，DAG 规模增加也可能会导致全局优化组合爆炸。总而言之，尽管基于 DAG 的任务调度已有围绕缩短关键路径、提升数据局部性、并行通信优化等一系列经典技术方案，但要在大规模复杂场景下取得鲁棒且高效的表现，仍需持续的创新和实践验证。

2.2.2 计算感知的网络调度

计算感知网络调度是云网一体化调度的另一种表现形式，其本质是将网络资源优化从传统面向流级别（Flow-level）的性能指标（如单个数据流的延迟、吞吐量等）转向面向任务级别（Job-level）的性能指标（如任务完成时间）。在通算场景，如面向 MapReduce/Spark 等分布式计算，网络流量往往表现为大量的中间数据流（例如 Map 阶段输出到 Reduce 阶段输入），这些相关的流集合被称为 CoFlow；网络调度会识别一个 CoFlow 内所有流的依赖关系，进行整体调度，而非将它们视为独立的、不相关的流，从而降低整个作业的完成时间。在智算场景，面向大规模深度学习训练，网络通信通常被抽象为集合通信（Collective Communication）原语，涉及大量并行且多轮迭代的数据交换；网络调度根据 GPU 拓扑和计算进度，实现通信与计算的深度流水线化与重叠，从而降低整体训练时间。

2.2.2.1 CoFlow 调度

在数据中心中，分布式计算任务的执行通常涉及多条具有逻辑依赖关系的数据流并发传输，任务的完成时间取决于最后一条流的接收时间。为优化这类复杂任务的性能，学者提出了聚合流 CoFlow 的概念。CoFlow 将一个任务并发产生的所有相关数据流视为一个整体，其传输时间取决于最后一条流的完成时间。不同于优化单一数据流，CoFlow 调度的目标是减少各个 CoFlow 任务的数据传输时间（如平均 CoFlow 完成时间）。CoFlow 调度为每个流分配适当速率，并决定各流数据传输的顺序，包括单个 CoFlow 内部的调度和多个 CoFlow 间的调度。CoFlow 调度问题是 NP-hard 问题，且在系统实现方面面临诸多挑战。CoFlow 在 2010-2018 年间是数据中心网络研究的重点，其所体现的全局计算感知调度思想，至今也是新兴业务（如智算）调度优化的理论基础。

中心式 CoFlow 调度是最经典的研究范式，通过启发式算法逼近最优解，实现 CoFlow 内和 CoFlow 间调度，同时结合路由进行联合优化。中心式 CoFlow 调度通常从整个集群内收集任务信息，统一计算 CoFlow 的流分配和优先级。早期工作中重点聚焦 CoFlow 内调度，通过加权公平共享算法分配流速率。为进一步提升了效率，此后工作进一步结合 CoFlow 间调度，通过优化瓶颈 flow 优先的方法，最小化 CoFlow 完成时间 [144]。然而，这些调度模型通常将网络抽象为单一的大型交换机（Big-switch），认为网络内部不会出现拥塞，但在真实的大型数据中心网络中，这种抽象是不够的，因此后续的一系列工作均在解决 CoFlow 调度和路由的联合优化和集成 [202]。中心式 CoFlow 调度的优势在于能够基于全局信息进行优化，其缺点在于信息收集、调度策略的计算和实施等环节带来的开销较大。

分布式 CoFlow 调度解决中心式调度中开销问题，对大量的小任务调度更为友好。尽管中心式调度

方案对大型 CoFlow 性能良好，但其高昂的中心化控制开销（如毫秒级的消息批处理间隔）使得无法有效处理小型 CoFlow（如 ≤1 MB 的数据）。为解决这一问题，分布式 CoFlow 调度工作被大量提出 [203, 204]。分布式的 CoFlow 调度机制以每个节点上的局部任务信息为基础，采用小任务优先传输等规则来决定任务的优先级。分布式任务感知调度机制的优势在于开销低、响应速度快。但由于调度策略通常较为简单且仅依赖局部任务信息，这使得它难以实现全局最优的性能。因此，任务信息的准确快速获取是这类分布式流调度机制中的关键挑战。

尽管 CoFlow 调度在学术界已进行了大量深入研究，但在实际大规模商用部署中仍面临着以下诸多挑战。目标单一性：现有机制主要集中于最小化平均 CoFlow 完成时间，却难以优化多样化的目标，如最大化网络利用率或提供差异化服务。可扩展性：中心化方案因开销过大无法处理小型 CoFlow，而分布式方案虽然可处理小型流，但由于缺乏及时全局信息而导致次优性能，在保证可扩展性的同时提高最优性能是未来的系统部署的优化目标。部署难度：将 CoFlow 机制部署到云环境面临挑战，包括如何在不修改应用程序的情况下获取 CoFlow 信息，以及如何在不修改应用和硬件的前提下实施调度等 [205]。

2.2.2.2 集合通信优化

集合通信是分布式并行计算领域中对数据交换模式的一种高级抽象，指的是在一组计算节点（通常是 GPU 或 CPU）之间，为了实现同步或数据聚合目标而进行的一系列结构化、协调一致的数据交换操作，如 AllReduce（所有节点贡献数据并接收聚合结果）、AllGather（所有节点收集所有其他节点的数据）和 Broadcast（主节点将数据分发给所有从节点）等通信原语。随着模型规模（如万亿级参数）和训练集群规模（如万卡 GPU）的爆炸式增长，网络通信已成为限制整个分布式训练速度和能耗的主要瓶颈。集合通信的优化对大规模分布式训练性能十分重要，已成为近年网络研究领域的焦点之一，在近年 SIGCOMM、NSDI 等顶级网络会议中占据相当大的比例。关于集合通信优化的研究集中在以下几个方面：

集合通信算法关注如何在给定网络拓扑和硬件约束下，设计最小流量或最短延迟的通信算法（如 Ring、Tree），是集合通信库优化的核心。集合通信算法包括预定义算法，如 Ring AllReduce 算法、Double Binary Trees 算法等，以及适配网络和硬件的动态合成算法。合成算法的搜索空间巨大，如何快速找到近优解，是当前研究的重点，也是各大厂商私有集合通信库的主要竞争力之一。例如，SCCL [207] 将合成问题编码为 SMT (Satisfiability Modulo Theories) 进行求解，TACCL [173] 将寻找最优通信算法的问题建模为混合整数线性规划问题进行求解，SyCCL [208] 利用拓扑的对称性进行搜索空间压缩。

任务内集合通信调度旨在优化单个大规模训练任务内部的通信效率，通过重组调度通信操作，使通信和计算最大化重叠，以降低迭代时间。通信调度的核心概念在于根据并行训练的数据依赖关系重新排序通信操作，并动态调整集合通信的路由、数据切分大小，实现计算与通信的深度并行。传统的通信调度基于 FIFO，但 FIFO 调度方案往往与最优差距较大，因为后向传播阶段的通信与前向传播阶段的计算存在差异，导致通信阻塞计算。因此，业界提出采用基于优先级的方法来调度通信操作，以消除计算间的 bubble [209]。除了优先级调度，基于分解的调度是进来热门的解决方案，通过将通信操作分解为更小的子操作（如 SYNDICATE [210] 中的 Motif，Centauri [211] 中的 Partition），从而进行精细化弹性调度。

任务间集合通信调度是解决智算中心多租户资源共享的核心，也是未来实现 ML as a Service 的重要组件。随着智算中心多租户和资源共享成为常态，如何在多个并发的训练任务之间高效、公平地分配网络带宽资源尤为重要。传统的通信调度器（如 Colow）未考虑分布式学习流量特有的特征（如多轮重复迭代、通信与计算的重叠特性），因此在智算多租户场景难以得到较好的表现 [213]。近年来阿里巴巴、华为、北大、MIT 等团队提出了多租户的通信调度系统 [212, 175, 213]，均聚焦在面向平台的优化目标（如最大化 GPU 利用率 [151]、最小化平均完成时间 [214] 等），而忽略了面向用户的优化目标（如公平性）。

中国电信云计算研究院自研提出 Dike 调度系统，**业界首个关注最大最小公平性的多租户通信调度器**。该系统针对不同训练任务间的公平性，创新性提出“最大化最小训练任务进展率”的优化目标，并提出一个轻量的 lazy 贪婪调度算法，只需要周期性对流进行排序，从实现最大化最小训练任务进展率。该

算法可兼容主流任何支持严格优先级调度的传输协议，理论上可以保证 1/2 的最优近似比，实际性能接近于最优，相比业界主流的调度系统，可以大大提升最小任务进展率和任务间公平性。

2.2.3 计算-网络联合调度

计算–网络联合调度是在统一视图下同时规划计算资源、通信资源与映射关系，使系统能够在全局范围内实现最优的资源配置与性能表现。它不仅为任务放置、带宽分配、路径选择等提供一致的优化逻辑，也为可预测性、弹性、跨域协同等能力奠定理论基础。因此，本节将从“统一建模方法”与“联合调度理论分析”两个层面展开，系统阐述计算–网络联合调度的基础框架、核心思想与理论价值。

2.2.3.1 统一建模方法

计算–网络联合调度建模方法是对计算资源与网络资源，并刻画两者映射关系的统一抽象。现有研究普遍采用以图为核心的建模方式，将计算节点、通信需求、网络拓扑和资源约束统一纳入一个结构体系中，从而支持将虚拟资源映射到物理基础设施的联合优化。在这一通用框架下，不同模型通过对节点属性、边属性和约束表达方式的差异化定义来适配不同应用场景。当前主流包括虚拟网络嵌入 VNE（Virtual Network Embedding）、虚拟集群嵌入 VCE（Virtual Cluster Embedding）、软管模型（Hose Model）、服务功能链 SFC（Service Function Chain）等。

VNE 以“虚拟节点 + 虚拟链路”到“物理节点 + 物理链路”的映射实现算网联合抽象。 VNE 是最早将计算资源与网络资源纳入统一模型的研究框架，其核心思想是通过对虚拟节点与虚拟链路的联合映射，在满足节点计算能力、链路带宽与路径连通性等多重约束的前提下，将上层业务需求嵌入到底层物理基础设施，从而保证计算可用性、链路可达性以及端到端性能。VNE 最早由 Chowdhury 等人系统化提出，定义了节点映射与链路映射的双层资源约束结构，奠定了后续算网联合调度研究的基础 [143]。近年来，随着数据中心规模扩大和业务负载多样化，VNE 研究逐步从静态、小规模嵌入演进至更复杂的在线与大规模场景，研究重点围绕可扩展性（如图神经网络驱动的高维映射）、动态性（面向在线请求的增量嵌入）与性能保障（SLA / 时延敏感嵌入）持续推进，包括利用 GNN 进行并行特征学习的嵌入方法 [215]、基于强化学习的动态 VNE 决策框架 [217]、面向时延和路径约束的链路映射优化模型 [216]，以及多域算网协同 VNE [219] 与面向 AI 负载的自适应嵌入方法 [218] 等方向。

VCE/软管模型是基于聚合带宽约束描述计算–通信关系的经典建模框架，使系统能够在动态通信矩阵下仍然保证可预测的网络性能。其思想最早由 Duffield 等人在 SIGCOMM’99 中提出，通过软管模型接口统一描述端点向全体节点的聚合带宽需求，为虚拟专网与多租户网络隔离奠定了理论基础 [240]。在数据中心场景中，Ballani 在 SIGCOMM’11 的 Oktopus 中首次基于软管模型提出了虚拟集群抽象，以 (N, B) 描述 VM 数量与聚合带宽需求，形成可直接用于任务通信与租户隔离的建模体系 [146]。随后 [241] 提出了更通用的扩展模型，使 embedding 更高效、接受率更高。近年研究主要集中在面向数据中心骨干的软管约束规划与鲁棒优化 [221, 222]，面向可预测网络服务的虚拟网络抽象 [220]，以及围绕在树型拓扑和图拓扑下的弹性调度理论 [223, 224]。这些研究共同推动了软管模型在可预测网络、任务编排和算网联合调度中的体系化发展。

SFC 将计算型网络功能与传输路径作为联合决策变量，是运营商场景中应用最广的算网联合模型。SFC 的本质是将一系列虚拟网络功能按顺序组织并部署在网络中，并确保流量能够依序通过这些功能节点。在统一建模框架下，SFC 不仅需要决定各网络功能实例的部署位置，还需联合考虑功能间的转发路径、链路带宽及时延约束，以及跨节点状态维护与同步等系统性因素，使其成为一个典型的算网联合优化问题。近年来，SFC 调度逐步从离线、静态配置演进为面向复杂业务和动态流量的在线优化过程，研究主要集中在面向任务与流量结构的联合映射与路由算法 [225, 226]，面向在线与动态需求的自适应优化方法 [227, 228]，以及依托可编程交换机等新型平台提升系统可扩展性的探索 [229, 230, 231, 242]。这些研究推动 SFC 从静态映射走向动态、智能、跨域协同，为计算–网络联合调度提供了重要模型支撑。

2.2.3.2 联合调度理论分析

联合调度的理论基础是在于通过对计算与网络资源进行联合决策，实现全局最优的资源分配，从而获得比传统模式更高的效能。使用函数 U（公式 (2.1)）来描述总体效能，其值由计算能力 x 与网络能力 y 带来的综合收益 $R(x,y)$（如用户体验、业务转化、收入等）与成本 $C(x,y)$（如建设成本、运维成本、融合带来的额外开销等）共同决定：

$$U = \max_{x,y\in\Phi} \left[R(x,y) - C(x,y)\right], \quad \Phi \subseteq \Omega. \tag{2.1}$$

$$y' = \max_{x\in\Omega} \left[R(x,y) - C(x,y)\right], \tag{2.2}$$

$$x' = \max_{y\in\Omega} \left[R(x,y) - C(x,y)\right]. \tag{2.3}$$

在传统的单独调度模式下，计算与网络能力各自独立优化，往往采取“先算后网”或“先网后算”的策略，如下式 (2.2) 和 (2.3) 所示：首先固定网络能力 y，对计算能力 x 进行最优求解得到 x'；随后固定计算能力 x'，再从网络维度对 y 进行优化得到 y'。如图2.5所示，这种“逐维优化”的方式本质上属于局部最优，无法捕获算网之间的协同效应，也无法保证最终解 (x', y') 为全局最优解。

一体化调度的效能可以通过提升整体收益、降低系统成本来实现。在收益层面，一体化调度打破了云、网、智资源之间的壁垒，可通过联合调度显著提升服务质量 QoS 与用户体验 QoE，例如高清视频、云游戏中减少卡顿与画质波动；同时扩展业务创新空间，如在工业互联网中利用跨域算网协同支撑复杂实时仿真，在大模型训练和推理中实现算力及带宽的协同分配以提升整体处理效率。在成本层面，一体化调度能够避免云、网、智算资源的闲置或重复建设；减少跨平台、跨域的运维管理开销；通过能力共享提升资源复用率；但也可能引入额外的系统开销，例如跨域协同带来的管理复杂性。

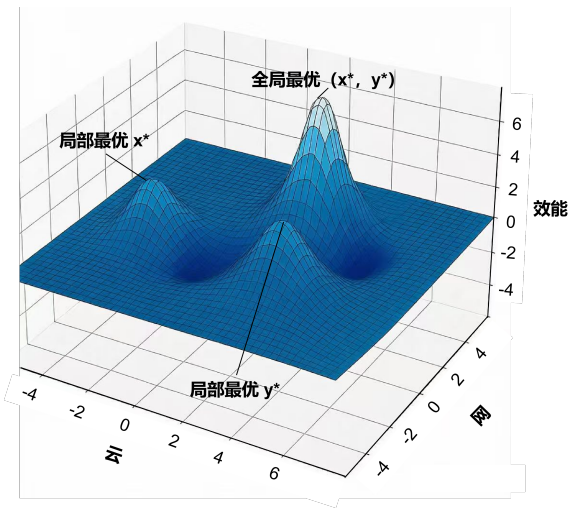


图 2.5: 联合优化与单独优化的最优解

中国电信云计算研究院在理论与算法两方面开展了深入研究，围绕数据中心典型的树型拓扑与一般图拓扑下基于软管模型的 VM 放置与带宽调度问题，系统分析了其最优性、求解复杂度与调度机制。针对树型拓扑，提出了可在线性时间内求得最大可接受负载的严格最优解。对于一般图拓扑，最大弹性调度被视为 NP-hard 问题，研究院据此提出基于最大流构造切片的 D-Slice/Elastic 框架，并通过 short-cut 路由降低路径跳数与链路占用，从而提升整体带宽利用率与通信效率。在动态批量请求场景中，提出 Tailor 裁缝式调度，通过利用最大切片进行按需裁剪与碎片整理，使整体调度效率相较传统方法显著提升。

2.3 热点方向五: 面向智算的云网基础设施

随着智算体系向更大规模、更高并发与更强协同方向演进，网络基础设施的角色正在从传统的数据传输通道转变为决定算力效率的核心要素，网络能力需要在拓扑结构、传输方式等多个维度系统适配智算任务的同步与通信需求。图 2.6 所示的入算-算间-算内三层协同网络体系为这一演进提供了总体架构：入算网络负责 TB 级数据的实时导入与安全入算，算间网络支撑跨智算中心算力资源的高效调度与传输，算内网络则通过专用拓扑、光电融合与在网计算等关键技术保障万卡级集群的训练通信效能。本章将围

表 2.4: 面向智算的云网基础设施研究领域热点

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
算内网络 DCN —— 拓扑与连接	智算任务带来的高强度通信需求，要求网络拓扑从通用结构转向为智算定制的专用设计，同时引入光电融合拓扑架构，以提供高带宽、低时延和可重构能力。	SIGCOMM NSDI TON INFOCOM ICDCS	• **面向智算的专用拓扑设计：**阿里巴巴 HPN [150]，通过双 ToR 接入和分层互联减少通信抖动；以及华为 UB-Mesh [243, 244]，利用高维互联结构提升带宽利用率和时延可预测性。以上代表性工作通过结构化互联、路径确定性和高带宽密度，适配大模型训练中的高通信负载。 • **光电融合网络与可重构互联：**Google Lightwave Fabrics [245]、阶跃星辰/北京大学 InfiniteHBD [246]、香港科技大学 MixNet [247]、复旦大学 MUSE [248]、以及上海交通大学 TROD [249]。以上是几个代表性系统，主要关注利用光交换降低功耗、增加带宽并支持拓扑动态重构。
算内网络 DCN —— 在网计算	通过把聚合、规约等核心算子下沉到交换机数据平面，让网络从单纯传输通道变成参与计算的关键组件，以减少通信成本，提升训练效率，并为大规模集群提供更高的资源利用率与可扩展性。	NSDI ISCA SC ICNP INFOCOM	• **功能实现落地：**Illinois 提出的 FPGA 原型 iSwitch [250]，KAUST 提出的 SwitchML [251]，以及清华大学提出的 ATP [252]。在工业界，NVIDIA 推出了 SHARP 在网聚合协议，ETH Zurich Flare [253] 提供专用硬件，共同证明在网计算在大规模集群中的性能优势。以上代表性工作验证了在交换芯片中执行聚合类操作的可行性。 • **资源管理优化：**华为提出了 NetReduce [254]，复用 RoCE 控制面以降低交换机协同开销；中科大提出 GOAT[255] 通过跨交换机的梯度分区提升内存与带宽利用率；清华大学的研究 INAlloc [256] 和香港科技大学的研究 DSA [257] 通过动态和抢占式内存管理提升资源效率。以上代表性工作缓解交换芯片内存受限、跨设备协同困难、任务动态性强等问题。
算内网络 DCN —— 传输控制协议	聚焦万卡规模训练中的高并发、低熵、强同步通信模式，探索新型拥塞控制、乱序容忍与协议内建可靠性机制，以支撑高带宽、低尾延的智算集群内部通信。	SIGCOMM NSDI ATC	• **商用部署模式：** Google Falcon [258, 259] 通过硬件部署拥塞控制协议、负载均衡协议以及丢包恢复协议构建高效数据中心网络；华为 DCP [260] 通过有损数据面 + 无损控制面的方式摆脱对 PFC 的依赖；Meta 采用框架驱动的接收端准入控制保障集合通信的效率 [261]；阿里巴巴 HPN [262] 通过双平面和路径空间优化增强拓扑带宽的一致性。 • **学术创新研究：**斯坦福大学 Bolt（亚 RTT 控制）、BFC（逐跳反压）[263, 264] 结合可编程交换机与端侧控制构建亚 RTT 级调节。
算间网络 DCI	面向跨域算力互联，研究广域 RTT、链路抖动与丢包条件下的 RDMA 扩展机制。	SIGCOMM NSDI ATC	• **跨域 RDMA 部署实践：** Microsoft Azure 在区域级数据中心上采用 DCQCN 参数调优与缓冲优化，使 RDMA 在数百微秒 RTT 下保持可控性能 [265]；Swing [266] 通过在边界引入 PFC-Relay，提前转发 PFC 暂停报文；ATC [267] 通过 DCI-switch 聚合 ECN/INT 实现跨域速率整形，提升跨域稳定性。
入算网络 DCA	面向训推任务中样本规模大、传输频繁和实时性要求高的问题，提供高弹性带宽、稳定传输。	SIGCOMM NSDI	• **弹性传输：**阿里云 Solar-RDMA 协议栈与 vFabric 选路技术 [268, 269]，中国电信“息壤”跨地域算力调度平台 [270]，以及“智云上海”城市级智算网络 [271]。以上代表性实践通过部署高速链路、实现无损 RDMA 传输、端网协同调度和数据直达算力机制，缩短 TB 级数据的入云周期。

绕该三层架构展开，系统阐述智算网络在能力构建与技术演进上的核心方向。三层体系共同构筑了智算时代网络系统的能力主轴，并构成本章后续讨论的逻辑脉络。

2.3.1 算内网络构建 AI 数据中心 DCN

算内网络指的是集群内网络，包括单节点内 Scale-Up 网络和节点间 Scale-Out 网络，实现超大规模算力互联。算内网络是整个智算云网基础设施的核心，涵盖了多项关键技术研究点。本章将重点围绕算内网络中的拓扑与连接、在网计算以及传输控制协议这三个主要研究点展开深入介绍。

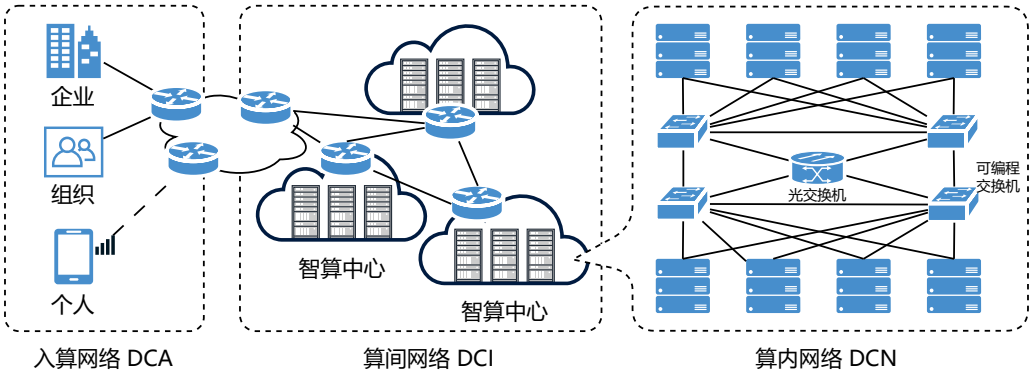


图 2.6: 入算、算间、算内网络示意图

2.3.1.1 拓扑与连接：从通用结构走向面向智算的专用设计

随着智算任务规模的持续扩张，算力系统的集群规模与节点间通信强度均呈现快速增长态势。大模型训练与推理、向量数据库检索、多模态任务等智算业务逐渐占据主导地位，网络流量特征发生显著变化，呈现出集中式带宽需求突出、通信行为呈周期性突发等新特征。在此背景下，传统通用型拓扑结构已难以适应智算任务所特有的多并发、高突发与强耦合通信模式。当前算力集群普遍具备节点密度高、训练任务集中、多维流量交织复杂等特点，网络不再仅是数据通道，而是与计算过程深度耦合，成为影响训练效率的关键要素。随着算力规模每提升一个量级，网络瓶颈所导致的训练吞吐下降、收敛延迟增加、带宽利用率不均等问题愈发凸显，传统规则化、通用化的拓扑设计在路径多样性、端到端带宽保障与通信可预测性之间难以兼顾。

数据中心网络从传统通用拓扑向智算专用拓扑转变。面对大模型训练对高并发、强同步通信的严苛要求，传统 Clos/Fat-Tree 等通用拓扑在带宽保障、路径确定性与延迟方面已显现结构性不足，难以支撑万亿参数规模的模型并行。网络架构的设计重点因此从“通用连接”转向“通信感知”，通过对拓扑的深度优化，实现对智算流量特征的精准适配。业界实践已印证这一趋势：阿里巴巴提出的 HPN 架构 [150]，通过双 ToR 接入、Rail-Optimized 结构及分层互联，将拓扑与 GPU 间同步通信深度绑定，显著降低了 AllReduce 等集合通信中的路径不确定性与带宽波动，为万卡级集群提供了关键网络支撑。华为的 UB-Mesh 等新型拓扑 [243, 244]，借助高维度全互联与分区化路由相结合的设计，在维持高带宽密度的同时有效控制复杂度，更契合智算任务对低时延、高可预测性的本质需求。这些演进共同标志着数据中心网络正超越传统 Clos 的通用范式，走向以“结构化互联、通信感知、局部高带宽、全局可管控”为核心的新型设计范式。

光电融合网络正在成为下一代智算集群的核心架构。智算任务规模扩大令传统电交换网络在带宽、功耗与扩展性上遭遇瓶颈，亟需引入新互联范式。光电融合凭借高带宽、低时延、低功耗及动态可重构优势，正在演进为支撑万卡级及以上算力系统的核心架构。其光交换能力与智算任务中强模式化、高同步性通信特征高度契合，能优化集合通信关键操作，提升训练吞吐与能效比 [272]。以光为底座、光电协同的新型网络体系正在成为全球领先企业的战略共识，光电融合已在智算场景取得多项关键突破：Google 的 Lightwave Fabrics 系统 [245] 通过自研 OCS 与波分复用实现数千节点间动态光路调度；北京大学与 Lightelligence 的 InfiniteHBD 方案 [246] 将 OCS 集成至收发器内部，实现节点级拓扑重构与点到多点通信，可动态构建 K-Hop 环形网络以降低组网复杂性与故障影响；面向 MoE 模型的 MixNet 系统 [247] 在区域级部署可重构光交换设备，实现专家流量按需调度，在测试环境中接近非阻塞 Fat-Tree 性能；MUSE [248] 通过动态增量重构算法缓解非均匀流量下拥塞，大幅加速应用完成时间；TROD 架构 [249] 以光交换替代传统 Clos 核心层，在过载条件下优于传统 Clos 拓扑。这些实践共同表明，光电融合不仅带来带宽数量级提升，更通过拓扑可重构性实现网络与任务高度耦合，为智算集群的下一代演进奠定坚实技术基础。

2.3.1.2 在网计算：从技术验证到系统化演进

随着智算集群规模持续扩大，传统依赖拓扑与带宽优化的方式已难以系统性化解通信瓶颈。网络架构的演进方向正从“传输加速”迈向“参与计算”，推动在网计算发展为新一代智算网络的核心技术。其本质是通过在网络设备中部署专用硬件或可编程逻辑，将训练中的聚合、规约等关键算子下沉至数据转发平面，实现通信与计算在拓扑与时序层面的深度融合，从而根本性提升系统协同效率与资源利用率。

在网计算技术已经从原型验证迈向商业化大规模部署。早期围绕在网计算原型系统的研究聚焦于技术可行性验证：iSwitch [250] 采用 FPGA 构建可编程数据平面，将聚合能力拓展至强化学习等细粒度场景，展现了在网计算的潜力；SwitchML [251] 在可编程交换机流水线中实现梯度累加，证明了无需修改主机协议栈即可显著降低网络负载与延迟；ATP [252] 面向多租户环境，通过片段化内存共享与基于 ACK 的反馈机制，首次系统化验证了复杂集群环境下在网聚合的可行性与效能。产业界也正在推进规模化落地，例如 NVIDIA 的 SHARP 协议通过 InfiniBand 交换芯片实现硬件级分层聚合，优化了大规模 AllReduce 操作的性能与带宽利用率；新一代定制化芯片方案 Flare [253] 进一步拓展能力边界，支持更复杂的模型并行与动态调度，标志着在网计算从学术探索走向工程实践。

优化资源消耗与提升系统鲁棒性是在网计算下一步的研究热点。面对模型规模扩大导致的交换芯片内存瓶颈，研究重点转向资源精细化管理与系统鲁棒性提升。NetReduce [254] 通过将聚合操作与 RoCE 协议深度集成，复用现有网络的可靠传输机制，降低了对交换芯片协议处理能力的要求。GOAT [255] 设计了多交换机间的梯度分区与调度策略，在异步到达场景下实现负载均衡与内存协同，提升了跨设备聚合效率。针对内存资源利用，INAlloc [256] 将交换机内存抽象为可动态分配的资源池，并引入一致性更新协议以支持任务运行中的平滑迁移；DSA [257] 则提出抢占式内存调度机制，通过优先级区分实现更细粒度的资源分配。此外，Rina [273] 创新地将聚合能力引入 RingAllReduce 同步架构，构建起“在网计算+RingAllReduce”的系统化融合路径。这些工作共同推动了在网计算从单一的聚合加速向高资源效率、强鲁棒性的全网计算-通信协同平台演进。

2.3.1.3 传输控制协议：从无损网络向面向智算的协议体系演进

随着智算任务规模进入万卡乃至十万卡级别，底层传输控制协议逐渐成为制约智算系统整体效率的核心因素之一。在万卡级甚至十万卡级训练场景中，网络负载呈现高度同步、周期性突发、低熵分布与极端尾延敏感等特征，导致传统面向通用云业务的 RDMA + PFC 体系逐渐暴露出结构性瓶颈。AllReduce、AlltoAll 等集合通信会放大任意节点的微小抖动；低熵通信流量难以通过 ECMP 分散路径，从而产生交换机长排队与突发拥塞；而 PFC 暂停将进一步加剧队头阻塞，使局部瞬时拥塞演化为全局暂停传输的问题。随着链路速率从 100 Gbps 迈向 400/800 Gbps，调优与运维成本成倍增加，传输协议正从网络调优问题上升为智算集群保障训推系统稳定性的关键因素。

在工业界，传输协议正从依赖无损底层网络转向协议内实现可靠性，形成可丢包、快恢复的核心设计原则。PFC 的传统避免丢包思路难以适应多跳、爆发性强的智算通信场景，企业开始采用主动容错机制重构协议体系。Google Falcon [258] 融合多种传输协议，通过传输延迟感知来调整速率控制、硬件级 pacing 与多路径传输机制在有乱序、网络抖动与拥塞的以太网上保持稳定且高效的吞吐；华为 DCP [260] 构建有损数据面 + 无损控制面的架构，通过包头报文驱动 RDMA 网卡实时重传并实现接收端的乱序写入操作，有效改善 RDMA 流在大规模并发下的性能损失。Meta 在大规模训练中采用由 NCCL 直接驱动的接收端准入控制 [261]，避免 DCQCN 算法的被动响应机制带来的速率振荡；阿里巴巴 HPN [262] 则通过双平面网络、多路径传输机制，使训练流量能够充分使用网络带宽。这些实践共同表明：工业界已经从底层网络保证无损转向协议承担高效可靠性。

在学术界，传输控制研究正突破 RTT 级反馈机制的限制，迈向端–网协同与亚 RTT 级实时调节。新一代研究不再依赖端到端 ECN 或 RTT 的长反馈周期，而是通过交换机级轻量信号与亚 RTT 决策实现高精度控制。BFC [264] 利用逐跳反压机制降低交换机队列长度；Bolt [263] 通过亚 RTT 级拥塞信号反馈实

现微秒级速率收敛；Poseidon [274] 借助轻量带内遥测信号获取瓶颈链路准即时状态，避免多跳拓扑下的误判；ACC [275] 则通过 ACK 时序构建高效网络传输模型，使协议能够在一个 RTT 内完成排队清空与快速恢复。这些研究展示了一个明确趋势：协议控制回路正在从端到端向端–网协同演进，并通过更丰富的反馈信号（INT 等）实现对突发流量与尾部延迟的高效控制。

总体来看，面向智算的传输协议体系正在从无损网络迈向协议保障可靠的弹性体系。未来传输协议设计将以容忍丢包、快速恢复、拓扑协同为核心特征，通过智能拥塞控制、多路径友好策略、亚 RTT 信号以及与训练框架的深度耦合，实现对智算流量更高效的性能保障。随着智算中心规模持续扩大、网络速率持续提升以及训练任务更趋复杂，传输协议的能力将直接决定模型训练周期、资源利用效率与系统稳定性，并成为未来智算基础设施建设与算网融合中的关键战略方向。

2.3.2 算间网络实现跨数据中心互联 DCI

随着智算中心从单集群形态走向多集群、多园区、跨地域的体系化布局，模型训练流、参数同步流和跨中心调度流逐渐突破单一数据中心边界，延伸至区域级乃至省际级网络。跨域链路的 RTT 从微秒级上升至数百微秒甚至毫秒级，链路带宽呈现明显的时变特征，路径抖动与随机丢包更加普遍，传统依赖低时延和域内无损的通信模式在新环境下面临失效风险。同时，PFC 在长链路中的传播距离大幅增加，带来缓冲占用急剧上升和链路阻塞放大等问题，使端到端无损难以在跨中心环境中维持。算间网络因此成为新型算力基础设施中的关键承载点，其网络设计需要从物理链路、传输协议到调度策略全方位适配新型跨域环境特征。

在跨数据中心场景中部署 RDMA，对传输层提出了与域内完全不同的技术需求。大的带宽时延积链路要求发送端维持更大的拥塞窗口以充分利用带宽，而长 RTT 又显著拉长反馈回路，使传统基于快速反馈的拥塞控制难以稳定。同时，链路级丢包不可避免，且恢复周期被大幅延长，传统依赖 PFC 机制抑制丢包的方式难以持续。因此，跨域 RDMA 需要具备应对大 RTT 的稳态控制能力、对丢包可快速恢复的传输机制、对路径变化更强的适应能力，并在架构层面构建域内与域间差异化的拥塞控制机制。同时，需要在跨区域的边界形成可观测、可编排的调度平面，以缓解跨域流量对域内网络的影响。这些需求推动 RDMA 从低 RTT、无损依赖的设计迈向高 RTT、可控优先的体系重塑。

产业界正在通过增强控制能力与构建分层治理机制推动 RDMA 的跨域化落地实践。Microsoft Azure [265] 在区域级部署中，通过数十公里光纤互联多个数据中心，使 RDMA 提升至更大地理范围。工程实践表明，长 RTT 使 DCQCN 的控制周期和收敛过程显著变形，而 PFC 范围扩大带来缓冲区成本上升与链路暂停的扩散风险。为此，Microsoft 通过调优 DCQCN 参数、优化交换机缓冲策略、提升链路稳定性等多维措施，使 RDMA 在跨机房链路中保持可控性能。经验说明，RDMA 并非无法跨域，但跨域部署必须依赖更强的跨域边界控制能力、更细粒度的链路调优与跨层次治理机制。

学术界围绕跨域 RDMA 面临的结构性挑战提出了多项体系化创新。SWING [266] 在数据中心边界构建轻量级 PFC 传播技术，使 PFC 信号提前传输至远端节点并提前暂停处理，以缓解跨域环境中的缓冲放大和排队扩散。ATC [267] 则从系统架构层面提出将控制环路从端侧迁移至跨域边界，由边界交换机统一汇聚 ECN、RTT、INT 等多源拥塞信息并进行快速速率调整，将端到端的长回路拆解为多个低延迟的局部回路，从而在跨域场景中实现吞吐稳定、队列长度可控，并减少跨域流量对域内业务的干扰。这些研究逐步形成共识：跨域 RDMA 的关键在于构建可控的跨域边界，以替代端到端的长回路控制。

总体来看，跨数据中心 RDMA 网络正从无损延伸迈向分层可控的体系化架构，是未来算力网络协同的关键方向。未来跨域 RDMA 将以域内保持无损语义，域间采用可控流控，边界承担增强调度为核心框架，同时结合多种拥塞信号和跨域路径选择机制，实现大 RTT 环境下的稳定性能与高的带宽利用率。随着跨园区训练、跨地域推理和全国算力资源池调度成为常态，跨数据中心 RDMA 的稳定性、可控性将直接决定国家算力网络的协同效率、训练规模上限与算力成本结构，其重要性将持续提升并成为智算基础设施竞争力的核心组成部分。

2.3.3 入算网络支撑用户算力接入 DCA

随着云网基础设施从传统 IDC 演进至面向大模型训练、推理与持续数据供给的 AIDC，数据中心接入网络不再仅仅承担南北向访问入口，而是成为数据能否快速进入算力体系的决定性环节。从企业到边缘产生的 TB 级样本、长周期迭代训练所需的持续数据流、推理服务中跨区域的 KV 缓存同步，都使数据入算路径呈现出高并发、高通量、高时效的新特征。如果入算链路的吞吐、时延抖动与路径调度能力不足，会直接拖慢训练迭代速度、降低算力利用率，甚至使大规模推理服务无法稳定运行。因此，新一代 DCA 必须从传统接入通道升级为具备高带宽供给、可预测性能与业务感知能力的战略入口，以支撑数据从入云走向真正的入算。在这一演进趋势中，弹性传输正成为入算网络技术体系的核心特征。

产业界已围绕弹性传输展开体系化实践。头部云厂商与运营商均在探索面向入算场景的高弹性、高可保障的传输能力。在协议侧，阿里云 Solar-RDMA [268] 等高性能协议栈实现微秒级链路利用调节，使数据注入在负载突增情况下依然保持稳定吞吐；其 vFabric [269] 技术进一步实现可预测选路与可保障带宽，为企业数据入算提供确定性通道。在算网融合方向，中国电信“息壤”算网平台构建跨地域 20 ms 级的算力调度网络，使新疆哈密绿色算力可与京沪万卡算力池协同，体现了跨域弹性传输的能力 [270]；“智云上海”项目则通过全互联架构将近千节点边缘集群与临港智算中心打通，实现 TB 级数据分钟级入算 [271]。腾讯云在架构层面基于专线 + 云联网 + 私网构建可伸缩、可自控的企业入算路径，并在网卡层引入智能卸载与虚拟化能力，以应对流量瞬时波峰。这些实践共同展示了产业界正从提升链路带宽转向构建可扩展、可调度、可保障的弹性传输体系，其目标是让数据注入不再成为训练与推理任务规模化的瓶颈。

总体来看，入算网络正从连接型基础设施走向具备弹性调度能力的算力入口。弹性传输能力使 DCA 能够在多业务、多来源、高通量的数据注入需求下保持可控性能，为大模型训练提供持续稳定的数据供给，并在推理服务中保障跨区域状态同步与流量分层调度。它不仅关乎带宽扩容，更关乎算力体系能否高效运转。未来 DCA 将成为 AIDC 的前置调度层，其功能将从带宽承载进一步扩展至任务感知、跨域编排与算网联动，成为智算中心中优先投入、必须重构的关键基础设施。

2.4 热点方向六: 云边端协同

云边端协同旨在构建覆盖数据、任务与模型全链路的跨层级协同框架，实现云–边缘–终端资源与能力的结构化组合与动态协作。通过在体系内形成多层资源的可编排组织方式，支撑数据的分级处理与共享、任务的多点协同执行以及模型的持续更新与分发，从而使不同层级的能力能够以一致的方式被调用、组合与演进。围绕这一体系，下文将从数据协同、任务协同和模型协同三个方面展开，分别讨论数据流动的组织治理机制、任务执行的跨层调度机制以及模型生命周期管理与演进机制。

2.4.1 数据协同构建跨层级数据流通体系

在云边端一体化计算体系中，数据协同作为承载智能服务的基础能力，其核心使命在于实现数据的高效获取、分布式处理、安全共享与跨域协同训练。伴随物联网终端数量的急剧增长、多模态数据的普遍生成、隐私保护法规的日趋严格、以及大模型驱动下对海量高质量数据的需求不断增强，数据协同的研究开始从传统的数据汇聚与存储问题，逐步转向分布式智能、实时处理、多源协作以及可持续学习等更具前瞻性的方向。在这一背景下，群智感知与联邦学习构成了当前数据协同研究中最具代表性的两大技术体系。

2.4.1.1 群智感知：多模态感知与边缘智能的融合演进

群智感知已从早期以用户终端参与为主的移动众包模式，扩展为面向人、物、车、环境等多主体协同的广域数据获取体系。随着智能手机、摄像头、多模态传感器、可穿戴设备、工业终端和车联网设备的普及，数据在边缘侧呈现出规模大、模态多、频率高和隐私敏感性强等特点，使得传统中心化数据采集

表 2.5: 云边端协同研究领域热点

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
数据协同	面向云边端多源数据的分布式感知与协同学习需求，通过多模态对齐、数据治理与隐私前置机制提升感知效率与数据质量，并在联邦学习框架下结合个性化优化、通信压缩与安全聚合，实现高效、可信的跨端协同。	CVPR NeurIPS ICLR SIGCOMM NSDI S&P CCS	• **群智感知：** Meta、Google、字节跳动等在多模态融合方向提出对齐与统一表征方法 [276, 277, 186, 185]；MIT、UIUC、Stanford 等在数据治理方向通过语义压缩等提升协同效率 [187, 188, 191, 278]；Cornell、Stanford 等在隐私前置方向提出可控扰动、本地差分隐私等机制 [279, 189, 190]；中国电信云研院与吉大提出多尺度补全模型，Google 在安卓体系中部署端侧数据分析体系 [280, 281]。 • **联邦学习：** NUS、UIUC、CMU 等研究在非独立同分布下通过表示约束、个性化优化等提升多源数据收敛稳定性 [192, 193, 194]；MIT、Google 等研究通过梯度压缩、稀疏编码与分裂优化减少通信规模 [282, 283, 284, 285]；在隐私与安全方面，Stanford、Harvard 等提出差分隐私、鲁棒聚合机制等 [286, 287, 288, 289]；Apple 在 iOS 生态部署结合差分隐私等端侧协作学习体系 [290]。
任务协同	在云、边缘和终端各层之间，根据任务特性、资源状态和服务需求，动态决定任务的分发、卸载、迁移与联合执行方式，实现整体系统的性能最优。	SIGCOMM INFOCOM ICDCS IPDPS ISIT	• **计算卸载：** 香港科技大学、哥廷根大学提出任务分层与动态卸载策略，提升边缘与云端资源利用率及服务效率 [291, 292]；上海交通大学、香港科技大学通过智能决策与多主体协同算法，实现自适应卸载优化 [293, 294]；大连理工大学、东南大学在多维安全机制和迁移连续性保障方面推动云-边-端大规模落地 [295, 296]。 • **服务编排：** 比萨大学、吉林大学通过多层架构下的服务分解与自动化调度，推动云-边-端高效资源协同 [297, 298]；清华大学、香港城市大学结合智能优化算法，提升系统弹性和任务流转效率 [299, 300]；清华大学、东北大学提出多域安全与流程治理方案，保障复杂业务场景下的可信服务交付 [301, 302]。
模型协同	通过对深度学习模型的合理分割与迁移适配，充分利用云边端的异构资源，协同实现模型推理过程的优化与智能服务能力的跨域迁移。	INFOCOM ICDCS MobiCom SenSys KDD TMC	• **模型分割：** 中国电信云研院团队针对网络波动对特征传输的影响，提出了基于冗余编码和特征恢复算法的高可靠特征传输算法 [303]。中科大团队针对端侧有限的算力资源，结合量化与早退机制设计了自适应的模型压缩算法 [304]；天普大学团队针对多样化的异构端侧资源，设计了自适应的模型分割方法 [305]。 • **模型迁移：** 北理工团队面向端云资源差异，设计了模型分解与集成算法实现高效推理 [306]；北邮团队面向边缘任务种类的复杂多变，设计了端云协同的模型微调方法快速适配新兴应用 [307]；上交团队面向推理任务差异化的能力需求，设计了模型路由算法，灵活地选择任务的执行位置 [308]。

模式无法满足实时性与安全性要求。因此，近三年的研究重点逐渐从广泛采集转向智能、高效、可信的分布式感知。

多模态融合已成为近年来群智感知系统演进的核心驱动力，其关键目标是在端–边分布式环境中实现对视频、语音、雷达、序列信号等多源数据的统一表达。随着传感器类型的丰富与规模的增长，系统需要在设备本地进行多模态间的轻量融合与语义对齐，以降低带宽压力并提升实时响应能力。相关研究通过结构轻量化的跨模态表征方法增强分布式场景下的感知一致性，并探索统一表示空间以支持多源异构模态在边缘环境的协同表达 [276, 277, 186]。与此同时，自监督跨模态重建与模态补全机制被用于应对模态缺失、噪声干扰等复杂条件，使系统在非完备数据输入下仍具备稳健的语义理解能力 [185]。这些进展推动群智感知从多源数据汇聚迈向语义一致的多模态智能表达，显著增强了数据协同的上层基础。

数据质量控制与结构化前处理正在成为群智感知支撑数据协同的关键环节。该技术核心在于将数据治理能力从云端前移到端–边侧，使进入协同链路的数据更加紧凑、准确与可信。在实际群智感知系统中，数据质量参差不齐、缺失严重和冗余度高等问题普遍存在，因此提升数据在源头侧的有效性成为必要前提。近期研究表明，通过对数据可信度、噪声水平、空间覆盖和任务相关性进行联合建模，可有效提升数

据可用性并减少协同过程中的不一致性 [278]。与此同时，边缘节点上的关键帧提取、语义压缩与结构化表示生成等轻量机制能够减少上行数据量，并在多节点之间形成一致的数据表达，为后续协同分析提供更高质量的输入 [187, 188, 191]。

中国电信云计算研究院与吉林大学联合团队在移动群智感知数据稀疏问题上提出了基于时空金字塔结构的多尺度数据补全框架，该工作针对移动群智感知数据稀疏性、异构性和多尺度特征等问题，通过构建多尺度嵌入、时空金字塔结构以及跨尺度注意力机制，有效捕捉群智感知中不同空间尺度与采集密度之间的复杂关联关系，为交通管理、环境监测和灾害响应等场景提供了更高质量的数据输入。这类研究进一步验证了数据治理能力前移的必要性，也是边缘侧数据协同从被动采集走向智能补全的产业实践案例 [280]。国际产业实践中，Android 系统中部署的端侧数据筛选机制通过本地完成数据清洗、质量评估与结构化处理，使统计特征在无需上传原始数据的前提下即可被利用，从而有效支撑了大规模移动设备的安全数据协同 [281]。这些案例表明，数据治理能力前移正在成为学术界与产业界的共同趋势。

隐私前置与可信数据贡献机制正在成为群智感知应对数据安全风险与合规需求的关键路径。随着视觉、语音和位置类数据的敏感性不断提升，系统需要在采集端完成隐私保护，使数据在进入协同链路前即具备安全属性。相关研究在端侧隐私扰动、可逆模糊化与动态屏蔽等方向上取得进展，通过在本地对敏感区域进行结构化模糊与扰动，在不显著牺牲任务性能的情况下降低隐私泄露风险 [279]。同时，可验证的匿名贡献机制通过安全求和协议支持跨机构协同分析，确保在无需暴露个体信息的前提下实现数据贡献的真实性与可审计性 [189]。进一步的设备侧差分隐私机制表明，在源头进行随机化处理可有效抵御推断攻击，提升协作场景下的隐私保障能力 [190]。这些隐私原生设计正推动群智感知从传统的上传后保护转向采集端保护，为数据协同的安全性和可持续运行奠定基础。

2.4.1.2 联邦学习：数据不出域条件下的协同智能建模

联邦学习作为一种在数据不出域条件下实现跨节点协同建模的机制，正在成为云边端体系中构建智能应用的重要基础能力。随着数据隐私监管强化、终端数量增长以及跨域协作需求提升，联邦学习从传统集中式学习的替代方案逐步演进为具备多源协作、隐私保护与持续优化能力的智能系统。当前研究主要聚焦于非独立同分布数据适配、通信效率提升与可信安全机制三个方向。

非独立同分布适配与个性化建模能力已成为联邦学习鲁棒性提升的核心基础。在云边端体系中，各终端侧数据常呈现标签偏移、特征差异与采样不均等问题，导致本地模型难以直接对齐全局模型，从而影响整体收敛性能。为此，研究提出了多种提升分布异构条件下训练稳定性的机制，包括在本地训练中引入全局表示约束以缓解特征偏移，通过个性化优化方式同时兼顾全局共享能力与本地适配能力，以及通过在优化目标中加入近端正则项提升模型在强异构场景下的稳定性 [192, 193, 194]。这些方法从表示、结构与优化层面增强了模型对终端侧差异化数据的适应性。产业实践中，移动端协作学习框架通过在本地构建个性化模型并在服务器侧进行安全聚合，实现了在大规模异构设备间的稳定部署 [309]。

通信效率优化与协同调度技术正在成为联邦学习规模化部署的关键保障。模型更新通常具有高维特征，其在大规模系统中的频繁传输会带来带宽压力与较高的通信成本。针对这一问题，相关研究提出了多种更新压缩机制，通过梯度稀疏化、量化与结构化编码显著减少单轮通信量，并在弱网条件下提升模型更新的可靠性 [282, 283]。与此同时，通过将训练过程拆分为局部子问题求解与周期性交互的方式，可有效降低通信频次并改善训练效率 [284]。面对终端掉线、网络波动等通用场景，异步协同机制允许各节点按照自身节奏上传更新，从而提升系统对动态网络条件的适应性与整体可用性 [285]。通信侧的一系列优化推动联邦学习从高频同步通信逐步演化为更加弹性与自适应的协作模式。

隐私保护与可信协作机制已成为联邦学习满足安全合规要求的关键前提。尽管联邦学习避免了原始数据的直接共享，但模型更新仍包含潜在敏感信息，在缺乏保护的情况下可能被用于重建输入数据或推断用户属性。为应对这些风险，研究引入了多种保护机制，例如通过在，模型更新中添加噪声抑制敏感信息泄露，通过鲁棒聚合过滤恶意或异常更新，并通过可验证的安全聚合协议确保不同参与方的更新在

跨组织协作中的真实性与可信性 [286, 287, 288, 289]。在实际应用中，设备端联邦学习框架通常结合安全硬件、差分隐私与安全聚合机制，为隐私敏感型任务提供可信的模型训练环境 [290]。

总体来看，联邦学习正从统一模型的集中式优化范式向个性化建模、通信高效、可信协作的综合体系演进。非独立同分布适配提升了模型在真实环境下的普适性，通信优化增强了系统的可扩展性，隐私与安全机制确保了跨主体协作的可信运行。随着行业对跨域数据协作与本地化智能的需求持续增长，联邦学习将在未来云边端智能体系中发挥愈加关键的基础性作用。

2.4.2 任务协同实现多点协作与动态调度

在云边端一体化计算体系中，任务协同是指在云、边缘和终端各层之间，根据任务特性、资源状态、网络状况和服务需求，动态决定任务的分发、卸载、迁移与联合执行方式，实现整体系统的性能最优。任务协同作为实现智能服务和高效资源利用的核心机制，其使命在于实现计算任务在不同层级间的合理分配、动态迁移与高效协作。随着终端智能化水平的提升、边缘计算基础设施的完善，以及 AI 驱动的多样化应用需求不断增长，任务协同的研究范式正向云边端分层自治、协同优化的方向演进。

2.4.2.1 计算卸载：弹性调度任务以提升系统性能

计算卸载作为云边端协同计算体系的核心环节，通过将计算密集型与延迟敏感型任务从终端动态分发到边缘节点或云端，有效提升了整体资源利用率与用户体验。近年来，随着网络基础设施升级与智能算法的广泛应用，计算卸载正从传统的静态分发向自适应、智能协同和多目标优化加速演进，成为移动应用、大规模物联网、工业控制等场景的重要支撑 [291, 292, 310]。

任务分层与动态卸载策略是提升卸载效率的核心基础。在实际部署中，任务分层与动态卸载策略通过对任务类型、数据依赖关系和终端资源状态的细致分析，实现任务在本地、边缘、云端三层间的灵活迁移。以移动设备为例，简单的数据采集与控制逻辑可由本地处理，资源消耗较大的视觉推理、语音识别等任务则按需卸载到邻近的边缘节点 [311, 312]。动态卸载策略要求系统实时感知终端负载、网络带宽、延迟波动等多维状态，并结合服务质量目标和能耗约束进行任务调度。在多无线接入、节点异构的复杂环境下，任务 DAG 分解、资源动态映射等方法已被广泛应用，极大提升了多任务、多用户场景下的系统吞吐和公平性 [313, 314, 315]。此外，智能终端与边缘节点之间的协同调度，以及对历史运行数据的智能分析，进一步推动卸载策略向自适应与智能化方向演进。

智能决策与多主体协同算法推动卸载机制自适应和高效化。面对复杂动态的网络环境和多样化的应用需求，传统基于规则的静态卸载策略难以满足实际需求。近年来，基于马尔可夫决策过程（MDP）[293]、深度强化学习（DRL）[294] 等先进智能算法被广泛应用于卸载决策优化，能够根据系统当前状态和历史经验进行自适应学习与调整。多主体协同算法为大规模系统带来全局最优的卸载决策，支持终端与边缘、边缘与云之间的实时负载均衡和资源共享 [316, 317]。这些算法不仅提升了系统的吞吐率和资源利用率，还增强了应对突发流量和网络波动的鲁棒性。结合用户行为预测、数据分析与进化优化等手段，卸载决策的实时性、精确性和能耗控制能力不断增强，有效支撑了智能制造、智慧医疗和车联网等高要求场景的落地应用 [318]。

2.4.2.2 服务编排：多层协同提升任务流转效率

服务编排是实现云边端多层异构计算资源统一调度和敏捷服务交付的关键技术。通过将复杂业务流程拆解为可组合、可迁移的微服务，服务编排能够动态地将任务分配到最合适的云、边、端节点，极大提升了系统的灵活性和资源利用率。随着物联网、智慧城市、智能制造等应用对实时性和弹性的要求日益提高，服务编排体系持续拓展，涵盖了自动化管理、智能感知和多域安全等多个前沿方向 [297, 319]。

多层架构的服务分解与端到端编排机制。在多层次云边端混合环境下，服务编排的首要挑战是如何合理分解复杂应用为服务功能链，并通过高效描述与自动化调度机制实现端到端流转 [298]。YAML、TOSCA

等标准化工作流描述语言，以及 Kubernetes、OpenFaaS 等开源平台，为多层服务分解、部署与弹性伸缩提供了基础支撑。QoS 感知的自适应编排框架，能够实时感知网络延迟、节点算力和服务依赖，根据实际运行状况动态调整服务链部署位置，实现负载均衡和高可用性。此外，函数即服务 FaaS、微服务和服务功能链 SFC 等理念的引入，为大规模分布式系统的敏捷编排和自动化运维提供了坚实理论与实践基础[320]。通过这些技术的协同，服务编排能够支持跨云、边缘、终端的业务连续流转和弹性扩展，满足不同场景下的多样化需求。

智能化编排与自适应调度优化。面对动态变化的业务负载、资源约束和网络状态，传统静态编排方法已难以兼顾系统性能和资源利用的最优平衡。智能化编排技术通过引入图神经网络、深度强化学习、多目标进化优化等先进算法，能够结合网络状态、节点能耗、服务优先级等多维特征，实现任务的自适应映射与弹性伸缩 [299]。如基于强化学习的云边端编排系统，能动态感知环境变化，实时优化服务链布局，在保障延迟和资源约束的前提下，提升全局系统效能 [300]。此外，结合流量预测、异常检测与自愈机制，智能编排系统能主动调整任务流转路径和资源分配，提升系统鲁棒性和服务连续性 [321]。这些技术的融合有效支撑了智慧交通、应急调度、工业互联网等对高可靠性和强弹性的场景需求。

2.4.3 模型协同支撑智能能力演进

在云边端一体化智能体系中，模型协同作为打通云端算力优势与端边部署需求的核心纽带，其核心使命是通过深度学习模型的分割、迁移、适配等方法，在资源异构、网络动态、任务多变的复杂环境下，实现性能无损、时延可控、资源适配的智能服务交付。随着端边设备算力的持续提升、大型深度模型的快速发展以及实时智能需求的日益迫切，模型协同已从传统的云端下发、端边执行单向模式，演进为端云协同推理、边缘按需迁移、模型动态适配的双向互动体系。当前，模型分割与模型迁移构成了支撑这一体系的两大核心技术路径，分别聚焦推理过程的协同优化与模型能力的跨域迁移，共同推动云边端智能的高效落地。

2.4.3.1 模型分割：端云协同推理的过程优化与资源适配

端侧算力有限，难以承载参数规模庞大的大型模型，但作为离用户最近的计算节点，其本身具备一定本地计算能力，能够快速响应简单推理需求；而云端虽拥有海量算力与存储资源，可高效运行大型模型以保障推理准确率，却受限于数据传输的带宽约束与时延开销，无法满足实时性服务需求。为兼顾智能服务的高准确率与低时延响应，模型分割成为充分整合端、边、云资源优势的核心方案——即将深度学习模型按照层结构合理切分为多个部分，分别部署在端侧设备、边缘服务器和云服务器上。通过这种分层部署与协同推理模式，模型分割既发挥了端侧“就近响应”的低时延优势，又借助云边的算力支撑实现了高准确率，有效平衡了资源约束、传输开销与服务质量三者之间的关系。该领域当前的主要研究主要围绕下列关键点展开。

针对端云协同推理过程中存在的网络波动问题，优化特征传输方法，提供高可靠、低时延的服务能力。针对网络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 [322]。针对无线网络传输中随机丢包造成的干扰，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 [323]。

近期，中国电信云计算研究院团队提出了高可靠的端云协同推理方法 [303]。如图 2.7所示，在端侧通过不均衡冗余编码机制，有效保护移动设备提取出的重要特征的传输，在云端设计额外的特征重构模块，恢复传输中可能丢失的特征。在真实测试平台上的实验结果表明，该方法能够在高丢包率的无线环境下提供高可靠的协同推理服务。针对有限的网络带宽资源，中国电信云计算研究院团队进一步提出了基于扩散模型先验的生成式编码方法 [324]。该方法在编码端提取轻量化特征，在云端利用扩散先验完成高保真重建，在极低码率下有效提升重建质量。作为端云协同的一种新范式，该框架突破了传统编码的效率–

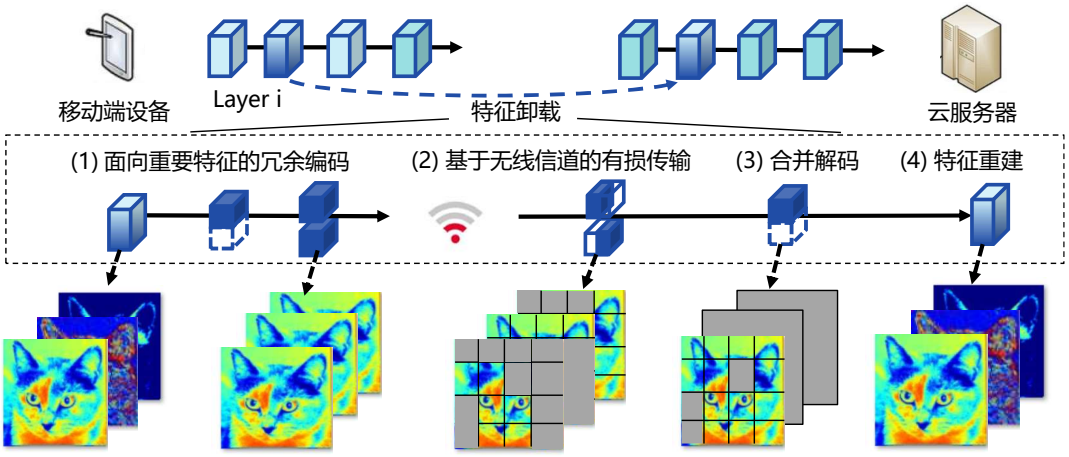


图 2.7: 面向无线传输环境下的高可靠端云协同推理方法

质量权衡，为资源受限场景下的视觉传输与智能处理提供了极具潜力的技术路径。

面向端侧设备有限的算力资源，通过自适应模型压缩实现端侧推理加速。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 [325]。研究工作 [304] 在端云协同的推理任务中，依据时间相近任务的相似性，为不同任务设置不同比特的定制化量化方法，并且结合早退机制在任务调度中减少资源空置时间，提高多任务处理的并行度，从而在保障推理任务整体性能的同时，提高了推理任务的整体吞吐量，降低了推理时延。

针对端侧异构计算资源和网络带宽差异，设计自适应的模型分割方法减小整体推理时延。在实际端云协同场景下，多个异构的端侧设备将同时运行不同的深度学习模型，其差异化的计算资源和网络资源将导致不同的时间延迟。对此，研究人员 [305] 联合考虑多个推理任务的联合调度与模型分割位置，设计了最优化算法，有效降低了并发推理任务的整体时延。与此同时，针对端云协同的 Transformer 推理任务，研究工作 [326] 通过自适应的令牌合并技术使得模型的计算量和特征通信量逐层降低，并设计由网络带宽决定的动态模型分割方法，灵活变更模型的分割位置，在动态网络环境下提升吞吐量，降低延迟。

2.4.3.2 模型迁移：云端能力向边缘的跨域赋能与按需适配

云端大模型经海量数据训练，具备优异的推理性能与广泛的泛化能力，但参数规模庞大、计算开销高昂，难以直接部署在资源受限的边缘 / 端侧设备；而边缘 / 端侧设备作为服务落地的关键载体，虽能就近响应需求、降低传输时延，却受限于算力、存储资源，其原生小模型往往存在性能弱、适配复杂任务能力不足的问题。为克服这一矛盾，模型迁移成为打通云端能力与边缘需求的关键方案——即通过知识迁移、轻量化适配将云端大模型的核心能力迁移至边缘 / 端侧小模型，或根据任务需求灵活选择模型的执行节点。这种跨域赋能模式既复用了云端大模型的训练成果，又适配了边缘 / 端侧的资源约束，实现了智能服务性能达标、时延可控、部署可行的多重目标。该领域当前的研究主要围绕下列关键点。

将云端的大型深度学习模型通过知识蒸馏等方式将能力迁移至边缘设备上的小型模型，提供就近的低时延服务。为了在资源受限的边缘设备上部署深度学习算法，云端服务器模型能够作为教师模型，监督在边缘设备上轻量级的学生模型的训练，在加速边缘模型训练过程的同时，降低边缘模型由于模型压缩带来的性能损失 [327, 328]。最近，研究人员进一步设计了一种基于知识蒸馏的模型分解-集成算法 [306]，将云端模型分解为多个子任务相关的小型模型，分布式部署在多个边缘设备上高效地执行并行推理，随后通过特征聚合实现接近云端模型的良好性能。

面向灵活多变的边缘推理任务，利用云端的基础进行高效的模型微调，快速发布边缘模型适配新兴

业务。云端服务器上部署的深度学习模型通常在固定的预训练数据集上具有良好的性能，然而直接应用于边缘设备上复杂多变的下游新兴任务时，即面临严重的性能下降问题 [329]。研究人员提出通过低秩适配的模型微调方法 [330]，不直接更新原始预训练模型的高维权重矩阵，仅添加少量低秩分解的可训练的参数数量实现易于部署的快速模型适配。近期的最新研究提出了边缘环境可感知的端云协同模型定制方法 [307]，通过获取边缘设备的上下文信息和无标签数据，从云端获取知识指导边缘小模型的定制化微调，提升对本地数据的适配性，同时减少性能损失。

面向边缘场景中难易程度不同的推理任务需求，灵活地选择云端大模型或者边缘小模型，实现资源与性能的最优平衡。当边缘设备直接承接全部推理任务时，受到有限计算资源的约束，难以保持复杂任务的推理精度; 而将任务全量上传至云端使用大型模型推理，会造成资源浪费和由传输时延导致的额外开销。模型路由的方法能够根据任务的难易程度灵活地将用户请求分配至云端或边缘模型执行。例如，研究工作 [331] 针对目标检测任务中微小目标更难以被准确检测的情况，设计了依据微小目标的数量和尺寸的轻量化传输决策器，仅传输少量困难样本至云端，在保持整体精度的同时，避免了大量传输造成的时间延迟。研究工作 [308] 针对文生图任务的差异化需求，通过多项图像生成指标决策哪些任务需要上传至云端执行，在资源预算成本的约束下，最大化了图像的生成质量。

2.5 展望与建议

借鉴 Gartner 成熟度曲线，本节构建了包含云网融合领域近 30 项关键技术的成熟度曲线（如图2.8所示），并给出本领域的研究展望和发展建议。

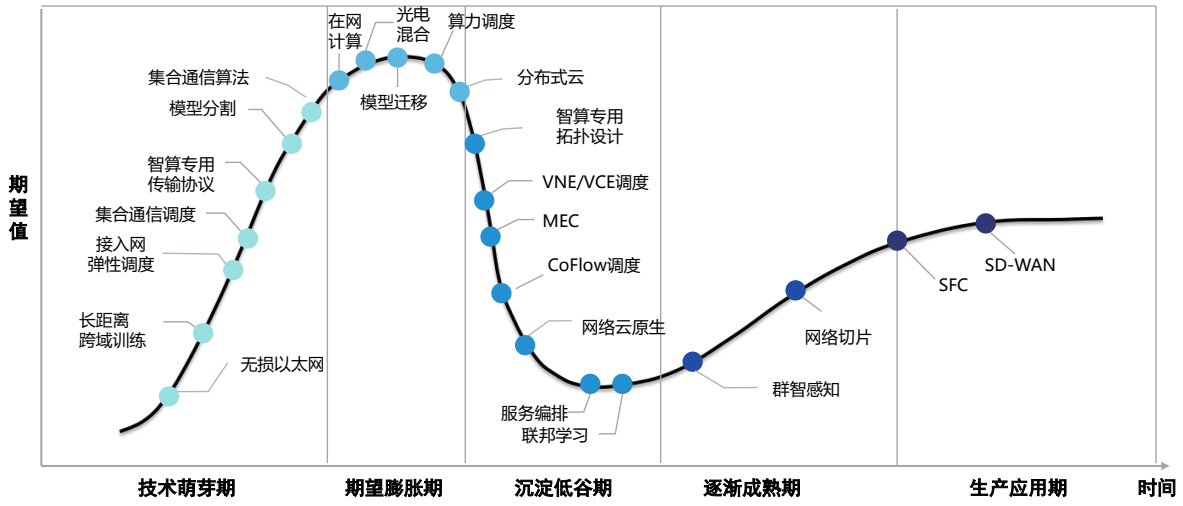


图 2.8: 云网融合研究图谱技术成熟度曲线 2025

2.5.1 云网融合的未来研究方向和关键技术展望

面向智能云网体系，云网一体化调度将进一步从“资源整合”走向“资源融合”，在更大空间、更高维度实现计算、网络、存储等资源的统一感知、统一决策与统一优化。其核心趋势包括：调度模型将朝着统一描述任务结构、通信关系与资源状态的多模态方向发展，实现对算网全域行为的精准建模；调度方法将基于最优化、组合优化等理论，引入在线学习、近似最优等策略，进一步降低大规模联合优化的计算复杂度；调度系统将基于跨地域、多集群与多运营域的协同需求，以及面向大模型训推、智算服务等场景低时延、高可靠的性能要求，进一步扩展到跨域资源协调与弹性流量治理，实现端到端性能的全局优化。在这些趋势下，云网一体化调度将升级为面向全域的智能协同引擎，逐步具备在全域范围内提高资源利用效率、提升系统性能并支撑多业务自适应优化能力，成为智能云网持续演进的核心驱动力。

未来智算云网基础设施将演化为面向“AI 超级计算平台”的超大规模、异构、统一互联的云网基础

设施，并从单域优化走向全域互联的系统化演进“AI 超级计算平台”作为 Gartner 发布的 2026 年十大战略技术趋势之一，是集成了高性能计算、专用处理器等的超大规模算力系统 [332]。根据预测，到 2028 年将有 40% 的企业工作负载采用该混合计算架构。未来，智算云网基础设施将演化为面向“AI 超级计算平台”的云网基础设施，以实现支持数万乃至数十万计算单元的超大规模扩展性，高效整合 CPU/GPU/AI 专用加速器（AI ASIC）等多种异构计算资源，支持计算/存储间、节点内外统一互联的低延迟网络。同时，网络传输将从单域优化走向全域互联的系统化演进。其中，智算中心网络内部研究将持续火热，包括 Scale-Up/Out 互联协议、异构通信库互通、智能流量调度等；另一方面，跨智算中心、跨地域算力调度研究或将成为新的热点，以实现多智算中心算力资源的高效互联、协同调度与全域优化。

云边端协同体系正从静态分层的资源组织转向面向任务与语义的智能协同架构，其目标是实现跨层级能力的统一抽象与高效编排，使数据、任务与模型在不同层之间实现更可控的流动与优化。随着大模型与在线学习技术的发展，协同体系将进一步强化基于任务意图的自动化决策，通过对任务特征、数据分布、模型复杂度和网络状态的综合感知，选择跨云、边、端的最优执行路径。同时，隐私计算与可信执行环境将作为底层基础能力深度融合，为跨域数据使用与协作调度提供安全保障。围绕这些能力，协同体系将更加注重可解释的策略生成、可验证的全链路协作安全以及可动态演化的模型部署方式，从而在复杂场景中稳定支撑实时任务执行、高质量数据流动与模型持续更新，构建具有韧性与可演进性的协同智能底座。

2.5.2 云网融合的发展建议

面向智能云网未来形态，需围绕联合建模、高效求解与跨域协同三个方向深化基础研究，形成可解释、可扩展、可验证的调度理论体系。首先，在建模层面，应持续推进任务结构、通信关系与资源状态的统一抽象研究，进一步加强对异构计算、网络与存储的刻画能力，为后续调度奠定更强表达力与可验证性的模型基础。其次，在求解层面，应面向大规模联合优化中的复杂度瓶颈，开展分布式求解、图分解、剪枝搜索以及在线学习驱动的快速近似策略等关键技术攻关，使调度在多约束场景下具备实时性与可扩展性。第三，在跨域协同层面，重点研究跨域资源映射一致性、分布式调度稳定性，同时增强算网观测体系、可编程网络能力与云网统一调度控制能力，为跨域资源编排与弹性治理提供可执行的底层支撑。通过在这些方向上的持续推进，可进一步提升云网一体化调度的理论深度与系统刻画能力，为未来智能云网的复杂业务场景提供坚实的算网联合优化基础。

面向智算云网基础设施，需加大推进产业标准化与开放性，并尽早关注跨界融合和交叉领域的研究。标准化和开放性是智算云网基础设施发展的重要推动力。应进一步推动企业、科研机构共同参与制定技术标准和规范，如统一集合通信库、统一通信协议、接口标准等，以确保不同设备和系统之间的兼容性和互操作性；倡导开放硬件与软件平台，建立智算云网基础设施开源社区，允许企业与高校共享集群资源，加速技术创新。同时，必须关注跨界融合和交叉领域的研究。未来新型网络架构、量子通信、边缘计算等前沿技术有望进一步提升智算云网基础设施的综合性能；通过与人工智能、大数据、物联网等跨领域融合，将形成更加丰富的智算服务生态系统。

推动智能模型在云边端的协同部署框架，并促进其规模化的应用落地。首先，开发轻量化跨平台智算协同框架，兼容云边端异构算力资源的接入。依托中国电信“息壤”等智算服务调度平台的技术底座与生态优势，聚焦技术适配与标准化攻坚，破解异构环境集成难题，实现模型在云边端间的无缝迁移与动态适配。其次，完善前沿技术栈，提供一站式算力接入与模型部署方案。通过提供模型压缩、边缘适配、性能调优等轻量化开发组件，降低中小企业的技术使用门槛，缩短从技术开发到实际应用落地的周期。最后，构建协同框架开放共享的生态体系，激活全产业链参与活力。面向生态伙伴开放标准化接口与适配工具链，支持第三方开发者、算力服务商便捷接入智算协同框架，扩大其影响力和用户规模。

第三章

围绕智能算法的研究

云计算深刻改变了信息服务的开发、部署、运维和计费方式，依托互联网构建起按需供给、弹性伸缩和统一运维的云环境，使用户能够随时随地获取和管理算力、存储与网络等关键资源。在这一模式下，企业通过虚拟化与按需付费机制显著降低了基础设施投入与运维成本，云服务也从单一行业扩展到医疗、金融、制造、社交网络等广泛领域。行业分析机构普遍预计，在未来数年内，云计算将持续成为企业保持竞争力和实现数字化转型的核心基础。与此同时，《新一代人工智能发展规划》[333] 等国家战略文件，**将人工智能明确为带动经济社会转型升级的重要方向，并提出到 2030 年建成世界主要人工智能创新中心的目标**，反映出智能算法正从通用技术工具逐步演化为推动产业升级和公共服务创新的关键支撑能力，实质性地影响着经济结构和日常生活。

在此背景下，云计算与智能算法的深度融合正在重塑云—网一体化基础设施的技术格局。一方面，云平台通过大规模算力集群和统一资源池，为大模型训练、复杂强化学习和多模态推理提供了可扩展的运行环境，使得智能算法的算力需求在不依赖单一专用硬件投资的前提下得以释放；产业界也普遍认为，智能算法的算力需求呈现出远超传统摩尔定律的增长趋势，进一步强化了云基础设施在智能时代的基础性地位。另一方面，智能算法正嵌入资源编排、任务调度、流量工程、故障诊断和运维治理等关键环节，将原本高度依赖人工经验的管理过程，逐步转化为数据驱动、自适应和可自动优化的决策过程。围绕这一趋势，本章聚焦云计算和云网融合场景中的关键智能算法，从**运筹优化**、**深度学习**、**强化学习**、**大模型**、**智能体**几个维度，系统梳理其问题建模思路、方法体系及在资源管理、性能保障和新型业务支撑中的作用，为后续章节展开面向具体场景的技术路径与应用模式分析奠定基础。

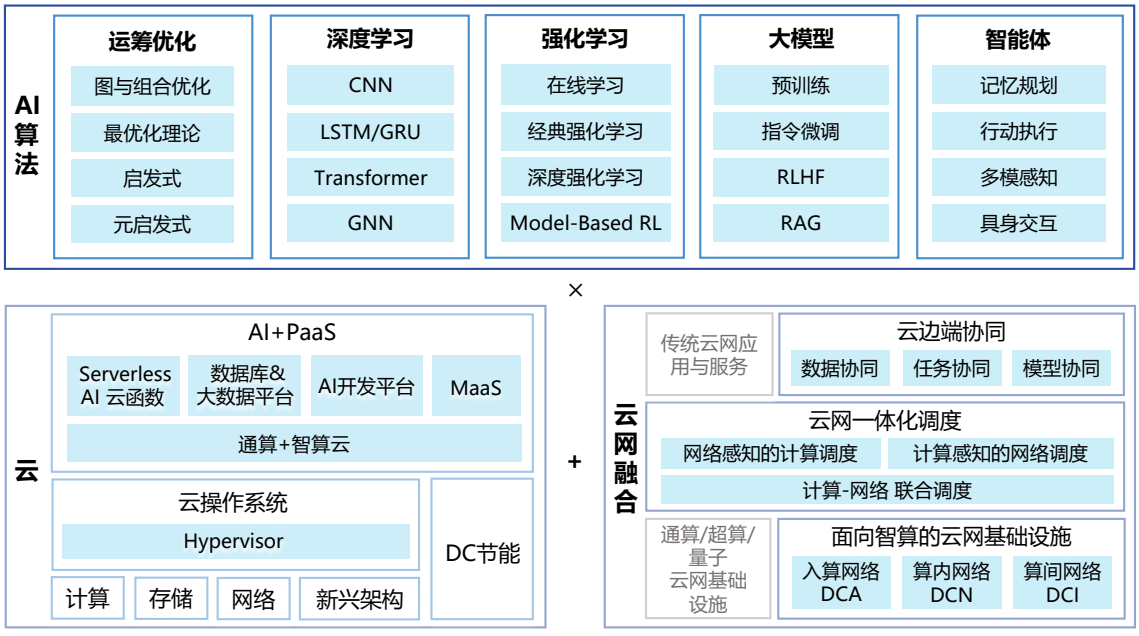


图 3.1: 围绕智能算法的研究图谱 (由云计算研究院总结形成)

3.1 研究图谱 2025：云计算与云网融合中的智能算法

在云计算和云网融合的背景下，随着规模和复杂度的不断增长，传统依赖人工经验和静态规则的运维方式在多租户、大规模和高动态环境下逐渐难以满足对性能、可靠性与运维效率的要求，亟需采用运筹优化、深度学习和强化学习等方法构建可建模、可学习、可自适应的智能控制机制。同时，基于大语言模型的新一代智能算法正在迅速发展，多模态与具身智能、LLM Agent 与 Agentic AI 等新兴领域的研究正逐步渗透到实际应用中，推动应用开发范式的转型和人机协同模式的演进。本章旨在探讨智能算法在资源调度、网络控制、自治运维和智能体构建等关键领域的作用，为后续关于安全、治理与产业实践的分析提供技术基础。

在这一背景下，建立系统化的“研究图谱”以对云计算与云网融合相关的智能算法进行整理和动态更新显得尤为重要。一方面，运筹优化、深度学习、强化学习以及 LLM Agent 和多模态与具身智能等技术方法，共同构成了支持智能基础设施与智能体应用的算法体系：前者主要应用于资源编排、流量工程和能效优化，构成“算法赋能云计算 (AI for Cloud)”的基础能力；后者则面向任务规划、工具调用与环境交互，代表了“Agentic AI”的发展方向。另一方面，随着算力需求、网络架构和模型范式的不断发展，智能算法的研究重点也在不断演化。因此，需要通过图谱化的方式，持续跟踪和更新关键技术方向、方法演化和应用趋势。图 3.1 展示了这一研究图谱，帮助我们更清晰地了解各类技术的关系和发展脉络。本节在此基础上，介绍了面向智能算法的云计算与云网融合研究图谱，并为后续的运筹优化、深度学习、强化学习及 Agent 相关内容的展开提供了框架。

3.1.1 趋势分析

在云计算和云网融合的演进过程中，智能算法正在从辅助工具转变为支撑资源编排、业务调度、网络控制与运维治理的关键机制。围绕这一转变，可以从若干互相关联的技术谱系来梳理其研究图谱：运筹优化为云—网系统提供可建模、可解释的资源配置与流量工程框架，是刻画约束、成本与服务等级目标的数学基础；深度学习面向监控指标、日志序列与拓扑结构等多源数据，构建从时序建模到结构表征的特征学习体系，为异常检测、性能预测与智能调优提供支撑；强化学习将云—网视作动态环境，在交互过程中学习调度、路由与弹性伸缩策略，推动资源管理走向闭环自适应优化；大模型与智能体则把大规模预训练模型与工具调用、任务规划相结合，增强云平台的跨任务、跨系统协同决策与自治运维能力；面向空天地一体与低空场景，多模态与具身智能通过融合多源感知并形成感知—行动闭环，为自动驾驶、无人机协同与虚拟实体控制等应用提供统一的云端算法基础。上述方法在资源优化、数据表征、在线决策与具身交互等层面形成互补，构成后文讨论的云计算和云网融合场景下智能算法的主要研究方向。

运筹优化作为支撑云—网系统资源配置、任务调度与网络流量工程的核心数学工具体系，是推动大规模云基础设施从经验驱动走向结构化、可建模、可解释决策范式的关键引擎。与依赖统计关联或深度表征的学习式方法不同，运筹优化以目标函数与约束条件为中心，通过显式建模系统结构、资源容量、任务依赖与成本收益，形成了一套可精确求解或近似求解的数学优化框架。在这一体系中，围绕云计算与网络运营的离散化、连续化与多目标的问题形态，逐渐形成了三大技术谱系：一类以图优化与组合优化为核心 [334, 335]，聚焦路径选择、拓扑规划、任务调度、资源匹配等离散决策问题，依托最短路径、最大流、多商品流、整数规划等方法，为超大规模分布式训练拓扑设计、路由问题、任务分配与资源调度和流量工程等场景提供理论可证的优化结果；第二类以线性规划、非线性优化及其扩展为基础，通过连续变量刻画带宽分配、能耗优化、跨地理分布式调度等问题，在可行域具有凸性或可松弛时获得全局最优解，为绿色数据中心、路由优化与分布式通信加速提供稳定、高效的求解机制 [336]；第三类则由启发式与元启发式算法构成 [337]，通过遗传算法、模拟退火、粒子群与贪心策略等智能搜索手段应对高维、非凸、动态环境下的复杂优化任务，在分布式推理调度、地理分布式数据中心成本优化与 GPU 资源编排等场景中展现了卓越的灵活性与实时性。三类方法分别在“结构精确建模—离散问题优化-连续可解优化—智能近似搜索”技术路径上形成互补，为云网系统在性能、成本、能效和可靠性之间的多目标权衡提供

了强有力的决策支撑。随着云网规模爆炸式增长、AI 训练需求激增以及系统动态性显著增强，单一的优化范式难以应对高度非平稳与高维耦合的复杂环境，推动运筹优化逐步迈向“优化模型 + 学习机制”的混合求解时代。通过引入模型松弛、问题分解、分布式求解以及在线自适应学习策略，运筹优化正成为支撑未来智能云网基础设施的关键数学基础。

深度学习作为驱动新一代人工智能与培育新质生产力的关键引擎，正引领云计算体系从传统的资源供给型基础设施，向融合智能算力、数据与算法的国家新型基础设施体系加速演进。在这一进程中，围绕云场景中监控指标、日志序列、业务流量与基础设施拓扑等不同数据形态，逐渐形成了彼此衔接、相互支撑的两大技术体系：一方面聚焦于欧式空间数据，技术脉络从早期的卷积神经网络 CNN（Convolutional Neural Network）与循环神经网络 RNN（Recurrent Neural Network），逐步演进至以 Transformer 与大语言模型为代表的统一时序建模与语义理解框架；另一方面则面向非欧式图结构数据，形成了以图神经网络 GNNs（Graph Neural Networks）及 GTs（Graph Transformers）为核心的技术路线，实现对系统拓扑结构与数据流动关系的统一表征。这两大技术谱系分别在序列建模、语义理解与结构认知层面形成能力互补，共同构建起支撑 AI for Cloud 的完整深度学习方法体系，为推进云计算从“可观测”走向“可理解、可协同、可自治”的智能新阶段奠定坚实的算法基础。

强化学习是一类以“试错互动—奖励反馈”为核心机制、通过环境在线反馈自主学习最优策略的序列决策技术体系。相较于依赖离线样本和静态模型的传统优化方法，强化学习将云平台与网络系统视作动态环境，在“感知状态—选择动作—获得回报—更新策略”的闭环中持续迭代，将调度、路由、资源控制等问题统一到长期收益最大化的框架下。从方法演进看，在线学习与多臂老虎机为轻量级的探索—利用平衡提供理论基石，价值函数与策略梯度方法支撑了在复杂马尔可夫决策过程中的稳定优化，深度强化学习借助神经网络突破了高维状态与复杂场景的建模瓶颈，而基于模型的强化学习则通过构建“数字孪生”与系统模型在样本效率与安全性上取得优势。在云计算与网络基础设施中，强化学习正从作业调度、弹性伸缩、流量工程、能耗优化等关键环节入手，推动资源管理从规则驱动走向数据驱动、自主演化，使大规模云网系统具备更强的自适应性、鲁棒性与智能运维能力，成为建设新一代智能云网基础设施的重要技术路径之一。

LLM Agent 是以大语言模型为核心、能够通过自然语言自主理解目标、规划任务并调用工具执行行动的自治智能体。相比传统 Agent，它在知识获取、泛化能力与交互模式上实现代际跃迁，通过模型推理、工具使用和记忆机制的结合构建起“感知—推理—决策—行动”的语义闭环。围绕这一能力基础，LLM Agent 形成了由构建、协作与演化组成的方法论框架：构建层面强调角色定义、记忆管理、任务规划与行动执行；协作层面涵盖集中式、去中心化与混合式多智能体组织；演化层面通过自反学习、群体共演化和外部知识反馈不断提升策略稳定性与适应性。同时，工具生态为智能体提供知识检索、程序执行与应用操作等能力扩展，评测体系则从通用任务、行业场景到多 Agent 协作全面衡量系统的可靠性与智能水平。借由这些能力的收敛，LLM Agent 正成为构建可扩展、可协作、可演化智能系统的关键技术路径。

多模态与具身智能主要围绕如何融合多源感知与行动能力，构建能够在复杂环境中感知、理解并执行任务的智能系统。多模态智能系统通过融合视觉、听觉、触觉等不同类型的感知数据，相较于单一模态系统更有条件获得更丰富和精细的环境表征与信息处理能力 [338]。例如，在机器人领域，通过结合视觉、语音、力觉等多种感知输入，机器人可以在复杂环境中完成目标识别、状态估计与交互决策等任务 [339]。具身智能强调通过实体存在与环境交互来获得和更新知识，凸显感知与行动的耦合关系，使系统在物理世界中具备自主感知与行为生成能力，从而执行结构较为复杂的操作任务 [340]。这一研究路径的核心在于将感知与行动统一到闭环中，使智能体不仅能够对多源数据进行理解，还能够通过自身动作和反馈不断调整对环境的表征与策略 [341]。云计算在多模态与具身智能系统中提供算力与数据管理方面的重要支撑，尤其是在处理来自多个传感器的大规模、近实时数据时，云平台可为感知数据融合、模型训练与推理部署提供弹性资源 [342]。借助云端的集中计算与协同处理，多模态与具身智能体能够在动态环境中高频更新决策与策略，为自动驾驶、智能机器人及虚拟/混合现实等应用提供关键技术支撑。

3.1.2 方向聚焦

在上述智能算法谱系的基础上，本章后续将沿两条主线展开讨论：一条是基础设施智能化，3.2 节以 AI for Cloud 为核心，聚焦云计算与云网融合基础设施，将运筹优化、深度学习和强化学习嵌入资源编排、流量工程、能效管理与自治运维等关键环节，构建覆盖算力、网络与数据中心的一体化智能底座；另一条是智能体形态演进，3.3 节以 AI Agent 与 Agentic AI 为主线，重点分析以大语言模型为核心的 LLM Agent 及其与多模态和具身智能的结合，如何在云端算力与工具生态的支撑下，演化为面向具体业务与场景的自治主体与群体智能。前文归纳的五类算法中，运筹优化、深度学习和强化学习在 3.2 节中构成 AI for Cloud 的三大方法论支柱，主要向下依托云基础设施实现可扩展的训练与部署；而 3.3 节中讨论的 LLM Agent 以及多模态与具身智能，则作为 Agentic AI 的主要能力载体与交互形态，向上支撑多智能体决策与具身交互需求。下文将围绕这两条主线开展系统梳理与深入分析。

3.2 热点方向七：算法赋能云计算

自 20 世纪中叶人工智能概念 [343] 正式提出以来，其发展脉络历经了从符号推理与专家系统为代表的“第一代智能” [344, 345]，到以统计学习与大规模特征工程为核心的“第二代智能” [346, 347]，再到以深度学习与大模型为标志的“数据驱动智能”阶段的演进 [348, 349]。与此同时，云计算于 2006 年前后以 Amazon S3/EC2 等服务为标志实现商业化落地 [350]，推动计算与存储完成从本地部署向按需供给、弹性扩展的范式转型，逐步构建起支撑全球数字经济发展的关键基础设施。

这一技术演进路径在我国获得系统的政策响应与战略牵引。自“互联网 +”行动实施及《新一代人工智能发展规划》 [351] 发布以来，国家层面持续推动人工智能与云计算、网络基础设施的统筹布局，明确将 AI 定位为驱动经济结构转型与培育新质生产力的战略引擎，并设定 2030 年建成全球主要人工智能创新中心的目标。随后，《算力基础设施高质量发展行动计划》 [352] 等政策进一步将“智能算力”“算力网络”纳入新型基础设施体系，设定总算力规模与智能算力占比等具体目标，系统构建面向大模型的智算中心与全国一体化算力调度体系。在国际层面，我国积极提出人工智能能力建设与治理合作倡议 [353]，倡导以智能基础设施互联互通与“AI+”融合应用推动实体经济发展，参与构建开放、安全、可信的全球智能治理生态。这一完整的政策体系为“Cloud for AI”范式奠定了制度基础，使云平台成为支撑 AI 规模化训练与推理的可靠算力底座 [354]。

随着云原生与多云架构的广泛普及，资源形态日趋异构、服务依赖关系愈加复杂，系统治理难度显著提升 [355, 356]。在此背景下，AI 与云的关系逐步从“Cloud for AI”所体现的单向算力支撑，演进为双向赋能、深度融合的共生体系：云作为 AI 训练与推理的算力基座持续发挥作用，而 AI 技术则反向驱动云系统在运维、调度与治理层面的全面智能化，催生出以 AIOps、智能调度与自治云平台为代表的“AI for Cloud”新范式 [357, 358]。该范式基于对监控、日志、拓扑等多维运行数据的统一建模，实现从资源管理、故障定位到能效优化的系统级智能 [359, 360]，推动云基础设施从传统资源供给模式演进为具备自感知、自优化与自适应能力的智能实体。“Cloud for AI”与“AI for Cloud”由此共同构成智能基础设施发展的双轮驱动，分别回应“AI 如何在云上运行”与“云如何运行更优”的核心命题。

在此基础上，**本节从算法视角将 AI for Cloud 划分为运筹优化、深度学习与强化学习三大方法论体系，系统阐释“算法驱动—数据驱动—决策驱动”在云环境中的协同演进逻辑**。运筹优化 [361, 362] 基于明确目标函数与约束条件，为资源调度与网络规划提供可解释的理论基准；深度学习 [363, 364, 365] 依托云环境中高维、多源、大规模数据，构建系统感知与预测能力，是实现环境可观测与模式识别的核心支撑；强化学习 [366, 367] 则面向动态环境中的连续决策问题，通过交互学习实现自适应控制与策略执行，形成从感知到决策的闭环。三者层层递进、互为补充：运筹优化奠定结构化建模基础，深度学习提供表征与推断能力，强化学习完成在线优化与自主控制。基于这一框架，后续将依次阐述三类方法的基本原理、关键技术及其在云与网络系统中的典型应用，并展望其在云–网–算一体化设施中的融合趋势。

表 3.1: AI for Cloud 的主要研究领域

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
运筹优化	运筹优化算法是一套基于约束条件、目标函数极值求解的系统化数学方法论，在云网系统关键任务中具有核心作用。通过将多维资源约束、通信拓扑与任务依赖建模为可计算模型，并借助松弛等策略，将高维离散与连续变量映射至可优化空间，实现高效求解。依托其可验证性、可解释性与理论最优性，该方法体系能够为资源管理、任务调度、网络拓扑设计、流量工程和数据库等复杂问题提供最优或近似最优解。	SIGCOMM NSDI INFOCOM ICDCS IEEE JSAC TC NeurIPS	• **图优化与组合优化：** 清华大学团队 [368] 基于图模型构建高带宽低时延的高效互联结构，提出 ZCube 拓扑结构，旨在提升高性能计算环境中集合通信的性能与成本效益。华盛顿大学团队 [369] 探讨了直连拓扑结构，并使用多商品流算法对直连网络的带宽利用、流量分布和拥塞管理进行优化。剑桥大学团队 [370] 研究了直连拓扑结构上的 AlltoAll 集合通信调度，并提供接近最优吞吐量的拓扑结构。Google [371] 的 B4 系统将流量工程问题抽象多商品流优化问题，展示了在大规模 WAN 资源调度的工程可行性。微软 [372] 通过构建图结构索引以实现向量数据库近似查询问题。香港科技大学团队 [373] 采用混合整数线性规划优化分布式作业调度和资源分配，以满足包括截止时间和延迟的调度目标。 • **最优化方法：** 北京航空航天大学团队 [374] 利用内点法将利润最大化问题建模为凸优化问题，提出一种地理感知任务调度方法。北京邮电大学团队 [375] 提出一个带有延迟约束的优化模型并通过 Gibbs 采样方法平衡任务卸载和资源分配，提高任务处理能力和资源利用率。 • **启发式与元启发式算法：** 新泽西理工学院团队 [376] 结合遗传算法、模拟退火和粒子群优化将能源成本最小化建模为非线性约束优化问题，提出利用地理分布数据中心的时空任务调度算法。上海交通大学团队 [377] 采用遗传算法优化 6G 移动通信系统中的任务安排，以最小化推理延迟，同时保持系统可靠性。
深度学习	以多层非线性表征与端到端特征学习为核心，通过从大规模多源异构数据中自动提取多尺度表示，将监控指标、日志序列、业务流量与拓扑结构等映射到低维稠密的可计算特征空间，为负载预测、异常检测、根因定位与能效建模等任务提供可复用的建模范式，支撑复杂云环境下的精细感知与全栈智能分析。	NeurIPS EMNLP AAAI ICML SIGCOMM TPDS	• **传统深度学习方法：** 多伦多和蒙特利尔大学等团队 [363, 378, 379, 380] 确立了 CNN、RNN 及其变体在非线性特征提取中的核心地位；墨尔本大学等团队 [381, 382] 利用序列建模挖掘日志规律以预测工作负载；清华和南开大学等团队 [383] 在智能运维方向提出 KPI 联合预测与异常检测框架，构筑可观测云的感知基石； • **Transformer 与大模型：** Google 和 OpenAI 等团队 [365, 384, 385] 基于自注意力机制奠定了大规模预训练范式；北航和清华大学等团队 [386, 387] 利用该架构统一长跨度时序建模，实现流量与能效的高精度预测；Microsoft 和阿里等团队 [388, 389] 将其拓展至语义运维，部署 Cloud/Ops Copilot 辅助自动化决策； • **图神经网络与 Graph Transformers：** DeepMind 等团队 [390, 391, 392] 提出消息传递机制以显式刻画非欧式空间结构；卡尔顿大学等团队 [393, 394, 395] 将其引入虚拟网络嵌入与资源动态分配任务，解决复杂映射难题；南开大学等团队 [396, 397] 针对微服务调用链构建依赖图谱，实现复杂故障的根因定位，确立了云基础设施的结构化建模范式。
强化学习	通过智能体与环境的交互进行学习，以试错互动和奖励反馈为核心机制，不依赖标注数据，通过在线反馈来学习最优决策序列，在“感知状态—选择动作—获得回报—更新策略”的闭环中持续迭代，将调度、路由、资源控制等问题统一在长期收益最大化的框架下。	ICML NeurIPS ICLR CoLT AAMAS RLC	• **在线决策与老虎机算法：** 塞吉巴黎大学在 K8S 资源调度、边缘计算资源分配等实时决策场景中应用老虎机算法 [398, 399]，将决策过程视作一轮轮的重复博弈，进行探索—利用的权衡 [400, 401]； • **经典强化学习算法:** NVIDIA 团队在能耗管理、负载均衡等状态空间规模有限的场景中基于马尔可夫决策过程描述环境变化 [366, 402]，对系统负载、资源情况等状态进行刻画 [403, 404]； • **深度强化学习算法:** 华为、Google 等团队在多维资源打包与作业调度中使用强化学习算法 [360, 405, 406, 407]，利用深度网络处理图像、时序信号等系统状态； • **基于模型的强化学习:** 谢里夫理工大学团队为了降低环境交互成本，避免系统风险，在资源分配问题中学习环境模型，并利用所学模型模拟决策结果、规划决策序列，提高了策略安全性和样本效率 [408, 409, 410, 411]。

3.2.1 运筹优化算法及其应用

运筹优化算法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，广泛应用于云网系统中的资源管理、任务调度、网络设计、流量工程、数据库等问题。随着计算规模和网络环境的复杂性增加，传统的基于规则或静态优化的方法在处理动态和高维问题时往往力不从心。相比之下，运筹优化算法通过构建精确或近似精确的数学模型，能够提供优化理论解决方案。这些算法主要可分为三类：图优化与组合优化、最优化（线性规划、非线性优化等）、以及启发式与元启发式算法（如贪心算法、遗传算法、粒子群优化等）。在图优化与组合优化问题中，利用图论和排列组合等方法解决复杂的离散优化问题 [412]。

一些任务可通过经典图算法（如最短路径、最大流、最小割等）在多项式时间内高效求解，这类问题通常具有明确的图结构与离散对应形式 [413, 414]。然而，更多实际场景下的组合优化问题（如任务调度、节点选择、资源匹配、虚拟网络嵌入等）属于 NP-hard，其复杂性来自离散决策空间的指数级增长 [415]。此类问题常用方法包括整数线性规划、混合整数规划等，通过分支定界、割平面等方式求解，但在最坏情况下，求解时间随问题规模呈指数级增长 [416]。在大规模云网系统中，传统精确算法往往难以满足实时性要求，因此通常采用松弛、分解、图划分、匹配近似等方法获得可行的近似最优解 [417]。在非线性优化问题中，若目标函数和约束条件具有凸性，则能确保全局最优解的唯一性，通常可以通过梯度下降、牛顿法等经典算法高效求解 [361]。而非线性非凸优化问题则可能存在局部最优解，可通过松弛、近似等方法转化为凸优化或易于处理的形式 [418]。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 [419, 420]。这些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，已广泛应用于多种问题。每类算法适用于不同类型的优化问题，具有各自的优势与挑战。

在云计算与网络系统中，运筹优化算法与系统的结构化特征高度契合，是支撑资源管理、任务调度与网络流量工程等核心机制的重要数学基础。云数据中心及其网络本质上由多维资源（计算、存储、带宽、内存、能耗）、复杂拓扑结构（多层 Clos/Fat-Tree/BCube/光交换架构）以及具有依赖关系的任务图（DAG 工作流、微服务调用链）构成 [421]，其运行目标通常涉及吞吐量最大化、时延最小化、能效优化、可靠性保证等多目标权衡。运筹优化算法能够在明确的目标函数与约束条件下给出可解释、可验证且具有理论保证的近似最优或最优解，特别在系统结构清晰、资源可量化的云网场景中具有独特优势，并且天然适配云网系统中多维资源、复杂拓扑与任务图结构。

3.2.1.1 图优化与组合优化算法

图优化与组合优化算法广泛应用于云计算和网络系统中的任务调度、资源分配、路由选择、数据库查询等问题。这类问题通常涉及大规模的离散决策和复杂的约束条件。图算法通常应用于网络拓扑设计、路径选择、流量工程等方面，在优化网络性能、提高资源利用率及实现负载均衡等方面发挥着重要作用。图优化与组合优化算法在云计算与网络系统中具有广泛应用，涵盖任务调度、资源分配、路由选择及数据库查询等场景。它们能够处理大规模离散决策与复杂约束，在网络拓扑设计、路径选择和流量工程中有效提升网络性能、资源利用率与负载均衡能力。随着大模型进入万亿参数时代，AI 训练和推理逐渐演进为跨数据中心的大规模分布式协同计算。模型和数据规模的持续增长使得节点间梯度交换与模型同步更加频繁，对网络带宽、端到端时延、拓扑结构与通信模式优化提出更高要求。在此背景下，图优化方法与智算网络的结构性需求高度适配，成为构建高效 AI 网络基础设施与资源调度体系的核心数学工具。在数据中心网络设计中，拓扑结构直接影响通信效率、带宽利用率和可扩展性。图优化算法常用于评估不同通信模式下的链路负载与流量分布，为拓扑选择与链路容量规划提供依据，从而实现低拥塞与高带宽利用。这类方法适用于 Fat-Tree、Dragonfly 等传统拓扑及可重构数据中心网络，为网络–计算协同奠定基础 [422, 423]。集合通信是分布式训练中的关键操作，可通过图优化建模以设计低延迟、高吞吐、可扩展的通信拓扑 [370, 368]。进一步地，结合优化调度方法，可显著提升大规模分布式训练的通信效率 [369]。在广域网场景中，Google B4 系统将跨数据中心流量工程抽象为带约束的多商品流优化问题，通过线性规划与迭代重优化实现带宽按优先级分配，并达到接近 100% 的链路利用率，体现了优化方法在大规模 WAN 调度中的工程可行性 [371]。经典图算法如最短路径、图匹配、网络流与图割算法等，可用于求解最短路由和最大流，从而提升网络资源效率并降低延迟 [424]。在向量数据库中，为快速执行 Top-K 相似度搜索，通常构建具备“小世界”结构的图索引，并通过贪心搜索在有限跳数内逼近目标邻域 [372, 425]。在图数据库分析中，实时查询主要依赖图遍历与路径发现算法 [426]；离线分析则包括路径查找 [427]、链路预测 [428] 和社区发现等算法 [429]，可从复杂网络中提取结构信息，为数据理解和应用提供支持。

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 [430] 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。例如，国防科技大学研究团队 [431] 通过基于超图划分的调度策略提升了云工作流的效率，结合云状态并使用带斐波那契堆的 Dijkstra 算法，显著降低了任务执行时间和能耗，推动了云计算资源的更加平衡和可持续的运作。最近，中国电信云计算研究院对超图在云计算场景中的理论与应用进行了系统梳理。研究首先阐明超图相较传统图在表达高阶依赖关系上的优势，随后概要回顾了超图划分、着色、同构等核心理论，以及谱方法与超图神经网络的发展方向，并分析其在数据中心网络拓扑、任务调度等问题中的适用性。然后，进一步总结了超图在网络拓扑优化、流量预测、资源调度、数据管理、异常检测、云安全等典型云计算场景的最新实践，展示了超图在模型表达能力和预测精度上的优势。同时指出当前在算法复杂度、可扩展性与跨场景迁移方面的瓶颈，并提出未来改进方向，以推动超图算法在云计算系统和行业应用中的更广泛落地。

组合优化算法可在资源容量、延迟、成本与功耗等约束下求解任务分配、资源配置与负载均衡问题。典型问题包括（混合）整数规划、背包问题与排队模型。云计算与网络任务调度通常以最优资源分配以降低延迟、提升吞吐量为目标，并常将调度建模为 0–1 背包或整数规划问题求解 [432]。其中，混合整数规划通过线性方程与离散变量刻画组合结构，可保证可行性并处理调度复杂性。在虚拟机/容器放置、任务—节点映射与工作流调度等场景中，问题常被构建为整数规划或多目标组合优化，通过联合求解离散资源选择与连续资源分配，以满足 QoS/SLA 要求。边缘云与地理分布式计算中，任务拆分、数据迁移与节点选择则常需结合图算法、混合整数规划与松弛–舍入方法，以平衡延迟、带宽与跨区域成本。例如，研究 [433] 提出混合整数规划方案，联合优化计算负载调度、碳排放控制、微电网运行特性，以降低地理分布式数据中心的电力成本与碳排放。分布式训练调度关注资源利用率、能源消耗与成本的综合权衡，并确保训练性能。例如，有研究将具有不同 GPU 数量的节点视为独立虚拟机，将调度建模为混合整数规划以降低租赁成本并控制作业延迟 [434]。在具备截止期约束的分布式训练中，调度需确保作业在截止时间前完成：典型做法是在混合整数规划框架下联合优化作业分析、调度与资源分配，以同时满足截止时间 SLO 与延迟等目标 [373]。当调度涉及连续量资源的决策时，凸优化等连续最优化方法尤为关键，它们能对资源精细调控，并有效平衡网络资源、优化性能指标，确保云与分布式系统中的资源利用最优。

3.2.1.2 最优化方法

最优化方法主要适用于带宽分配、工作负载调度、任务卸载等连续决策问题。在云计算与网络系统中，许多调度与资源分配的目标与约束本质上是连续变量，使得最优化方法能够高效求解并提升资源利用率。线性规划因模型结构简单、约束与目标均为线性，可借助单纯形法与内点法快速求得全局最优，适用于大规模资源分配与网络流量工程。凸优化则因其全局最优可保证性，在多维资源调度与复杂约束场景尤为重要，常用于云环境中高维资源分配问题。在工作负载调度方面，北京航空航天大学团队将跨地理绿色数据中心的利润最大化建模为凸优化问题，并基于内点法提出地理感知调度 [374]。新泽西理工学院团队则利用 G/D/1 排队模型刻画工作负载分布，将跨区域 SLA、电价与负载分配建模为凸优化问题，从而获得可验证的最优解 [435]。在移动边缘计算中，任务卸载常采用凸优化与混合非线性整数规划协同求解。例如，北京邮电大学团队构建带延迟约束的优化模型，并通过 Gibbs 采样协调任务卸载与资源分配，以提升系统处理能力与资源利用效率 [375]。在网络资源分配方面，分布式深度学习中的带宽调度可通过线性规划建模通信依赖关系并优化带宽缩放，以最小化通信时间 [436]。在 WAN 与数据中心的流量工程中，线性规划与凸优化广泛用于带宽分配、负载均衡与路径优化，有效缓解高动态流量下的拥塞与尾时延。随着云系统复杂性提升，高维、动态、非线性优化问题难以由传统方法在有限时间内求解，启发式与元启发式算法（如遗传算法、模拟退火、粒子群优化）因其随机化搜索能力，可为实时调度、负载均衡与多目标资源分配提供高质量近似解。

3.2.1.3 启发式与元启发式算法

启发式与元启发式算法适用于大规模、复杂且具有动态性、非线性与多目标特征的优化问题，常用于工作负载预测、任务调度与资源分配场景。这些方法通过近似搜索、随机扰动与群体智能等机制快速获得可接受解，特别适合传统精确优化难以高效处理的场景。在工作负载预测方面，研究提出结合在线与离线策略的启发式方法，通过任务依赖图实现虚拟机放置，以提升负载均衡与降低成本 [437]。为解决任务顺序难以保证、截止日期不明确等问题，另一类研究结合最早完成时间预测与蚁群优化，将任务按最小成本和截止期排序并分配至最优虚拟机，同时利用人工蚂蚁在虚拟机间迁移负载以实现平衡 [438]。在任务调度方面，新泽西理工学院团队构建能源成本最小化的非线性优化模型，并基于地理分布数据中心时空差异提出 STTS 算法，将遗传算法、模拟退火与粒子群优化结合，实现跨区域能源价格感知的任务调度最优 [376]。在分布式深度学习推理调度中，系统通常基于请求密度主动扩展资源，并通过任务完成时间预测与遗传算法优化任务排序，以最小化推理延迟并保证系统可靠性 [377]。为提升推理吞吐量，调度器还通过启发式方法动态调整批次大小，以提高资源利用率并降低调度开销 [439]。在资源分配方面，贪心算法常被用于根据作业优先级对网络流进行排序，并将速率分配交由底层网络执行 [440]。此外，部分研究会先确定合适的批处理规模与 GPU 资源下界，再通过贪心策略为推理任务分配 GPU，以获得最小性能干扰的设备配置 [441]。

总体来看，运筹优化理论与方法的优势在于其建模严谨性、可解释性、可验证性，当目标函数与约束条件清晰定义时，它可以输出全局或近全局最优解，并且天然适配云网系统中多维资源、复杂拓扑与任务图结构，为系统资源部署、拓扑规划与任务调度提供数学保证。然而，随着 AI 集群规模增大、云网系统高度动态、多租户数量与业务形态持续复杂膨胀，单一的运筹优化算法也面临可扩展性、实时性和多目标耦合带来的建模与求解挑战。因此，未来趋势将更多聚焦于“最优化理论与方法 + 学习方法”驱动的混合求解框架，通过模型松弛、问题分解、分布式求解、在线智能决策机制，实现新一代云–网系统在大规模、高负载与非平稳环境下的高效、可靠一体化调度和管理。

3.2.2 深度学习及其应用

深度学习作为驱动新一代人工智能与培育新质生产力的关键引擎，正引领云计算体系从传统的资源供给型基础设施，向融合智能算力、数据与算法的国家新型基础设施体系加速演进 [442]。在这一进程中，围绕云场景中监控指标、日志序列、业务流量与基础设施拓扑等不同数据形态，逐渐形成了彼此衔接、相互支撑的两大技术体系：一方面聚焦于欧式空间数据，技术脉络从早期的卷积神经网络 [363] 与循环神经网络 [378]，逐步演进至以 Transformer 与大语言模型 [365, 384, 385] 为代表的统一时序建模与语义理解框架；另一方面则面向非欧式图结构数据，形成了以图神经网络及 Graph Transformers [443, 444] 为核心的技术路线，实现对系统拓扑结构与数据流动关系的统一表征。这两大技术谱系分别在序列建模、语义理解与结构认知层面形成能力互补，共同构建起支撑 AI for Cloud 的完整深度学习方法体系，为推进云计算从“可观测”走向“可理解、可协同、可自治”的智能新阶段奠定坚实的算法基础。

3.2.2.1 面向欧式数据的序列与语义建模方法

基于欧式数据的传统深度学习方法，为 AI for Cloud 构建了初代的自动化感知与预测能力，是实现“可观测云”的重要技术基石 [445, 446]。（1）局部模式建模：CNN 适用于具有局部相关性与平移不变性的监控数据，通过将多维指标与时序信息映射为“资源热力图”或“时序图像”，在数据中心温度场建模、硬件故障图像识别及多指标联合异常检测等任务中实现自动特征提取 [447, 448, 449]，为容量规划与风险预警提供高层抽象表征。（2）长序列预测：RNN 及其变体长短期记忆网络（Long Short-Term Memory，LSTM）与门控循环单元（Gated Recurrent Unit，GRU） [379, 380] 擅长刻画云环境中的长程时间依赖，被广泛用于工作负载预测、多变量 KPI 联合预测与日志序列异常检测等场景 [381, 382, 383]，支撑弹性扩缩容策略与资源预留决策，从而在分钟至小时尺度上提升云系统的预测性管理能力。（3）能力边界与演进：

随着云系统规模与复杂度的增长，传统 CNN 和 RNN 在应对跨组件、跨业务的全局依赖与多源异构数据融合时，逐渐显现表达能力有限 [450]、对新业务模式泛化不足的局限，特别是在需要同时处理长时间跨度 [451]、多尺度波动 [386] 与多指标耦合关系 [452] 的任务中，模型往往依赖大量人工特征与规则补充，这为后续基于注意力机制的 Transformer 架构及其在云场景中的结构化扩展提供了演进动力。

以 Transformer 为核心的序列建模方法及其在大语言模型中的应用，推动 AI for Cloud 在长序列建模与语义层决策两个维度上实现了从“可观测”到“可理解、可协同、可执行”的能力跨越 [365, 453]。(1) 长序列建模：在数值时序建模层面，基于自注意力机制的时序 Transformer 能够在统一框架内处理长时间跨度、多变量的云监控数据 [386]，显式捕捉不同时间片与指标间的远程依赖，已应用于跨数据中心流量预测、全局流量调度及能效曲线建模等任务 [387, 454]，为跨集群资源编排提供更精细的统计先验。(2) 语义决策协同：在语义与决策层面，以 Transformer 为骨架的大语言模型将运维工单、手册、告警与脚本等非结构化文本嵌入统一语义空间，并借助检索增强、工具调用与多智能体协同，形成 Cloud Copilot 和 Ops Copilot 类运维助手 [388, 389]。运维人员可通过自然语言描述故障或治理意图，由大语言模型自动解析、构造查询、调用 API 或基础设施即代码工具，生成具备可解释性的诊断与执行方案 [455, 456]，从而在“监控—分析—处置”之间构建闭环。(3) 结构增强探索：针对云系统中普遍存在的拓扑依赖，一系列研究开始在 Transformer 框架中显式引入结构信息，例如通过邻接矩阵或路由路径构造注意力偏置 [457]，对服务或节点引入结构化位置编码，或采用“时序序列 + 拓扑视图”的多视角联合建模 [458]，从而在保持长序列建模能力的同时提升对网络结构与依赖关系的刻画精度。这类结构增强 Transformer 为后续 GNNs 和 GTs 在云场景中的系统性应用奠定了模型与特征基础，并在复杂资源编排、跨域调度与多目标优化任务中展现出优于传统序列模型的性能潜力。

3.2.2.2 面向非欧式数据的图建模方法

面向云系统中普遍存在的非欧式图结构数据，GNNs 及 GTs 通过显式建模拓扑关系与数据流动，为 AI for Cloud 提供了从“局部异常检测”迈向“系统级认知与全局优化”的关键支撑 [443, 457]。(1) 拓扑一致建模：在统一建模云基础设施与业务依赖方面，数据中心网络、虚拟网络拓扑、任务有向无环图、微服务调用链、虚拟网络功能组件关系、工作负载相似性图及攻击路径等，均可自然表示为图结构 [391]。GNNs 通过消息传递机制迭代聚合邻域信息，其计算过程与“云中请求跨多层资源与服务流动”的物理过程高度同构 [390]。在 IaaS 层，GNNs 被用于虚拟网络嵌入 [393]、网络配置综合 [459] 与资源分配 [394]，通过结合强化学习或组合优化方法，求解满足约束的近似最优部署策略；在 PaaS 层，面向微服务与无服务器架构的图模型统一处理服务依赖图与链路追踪因果图 [397]，应用于自动扩缩容 [395]、服务拆分重构与链路级根因定位 [396]，显著增强对复杂调用拓扑的治理能力。(2) 时空依赖刻画：为刻画大规模复杂拓扑中的长程依赖与动态演化，GTs [392] 将多头自注意力引入图结构，在保留局部聚合能力的同时强化对远距离节点与多尺度模式的捕捉。在跨数据中心路由、跨域流量工程及算力—能耗—制冷耦合系统等场景中，结合时间编码的时空 GTs 能够同步表征拓扑演化与负载波动 [460]，为全局路由、节能控制 [387] 与多目标调度提供统一决策模型。进一步引入结构学习 [461] 与对比学习 [462] 后，此类模型可在观测数据不完整或噪声环境中推断隐含依赖，提升运维与安全分析中的鲁棒性。

在此基础上，图–序列融合范式进一步打通结构表示与语义建模的边界，为构建具备迁移能力与通用推理能力的云系统智能体提供了新的发展路径 [463, 464]。近期图基础模型 GFM（Graph Foundation Models）[463, 464] 与图增强大模型 [465] 的研究表明，通过在大规模图数据上进行自监督预训练，并与文本、大语言模型进行联合对齐 [466]，可以在统一框架中同时编码结构信息与语义知识，为跨场景迁移与零样本推理提供基础 [467]。在 AI for Cloud 场景中，这一方向使得“拓扑—指标—日志—文本知识”可以映射到统一的表示空间：一方面，图编码器或 GTs 为大语言模型提供结构感知的上下文，使运维助手在生成诊断与变更方案时能够显式考虑服务依赖与网络约束；另一方面，大语言模型通过指令微调与工具调用能力，驱动图模型完成拓扑级推理、关键节点识别与策略搜索。二者结合形成面向云系统的图–序列联合智能体框架，为在复杂云环境中实现可泛化、可解释、可迁移的系统级优化提供了新的技术路径。

3.2.3 强化学习及其应用

强化学习是通过智能体与环境交互，在试错过程中利用奖励信号学习最优决策策略的机器学习方法。主要关注在未知、动态的环境中选择行动，使得长期收益最大化的问题，适合于对连续决策与控制问题进行建模。强化学习可以涵盖从简单到复杂的一系列决策建模，主要方向包括在线学习与老虎机问题、基于马尔可夫决策过程 MDPs（Markov Decision Processes）的强化学习、深度强化学习以及基于模型的强化学习 MBRL（Model-Based Reinforcement Learning）。除了这些常见的研究领域，多智能体强化学习、离线强化学习、安全强化学习等方向的研究也在快速发展并逐步进入工程应用。在云计算与网络系统中，强化学习天然契合动态、复杂、难以精确建模的运维与调度场景。相比于基于规则或静态优化的方法，强化学习在云网场景中的主要优势在于能够根据运行数据中不断自我优化，可以在不完全了解系统机制的前提下给出接近最优的决策，在非平稳工作负载与复杂约束下也能保持较好的适应性。在工程落地中，强化学习技术也面临着样本效率、安全性、可解释性以及与现有运维体系的兼容性等挑战。由于模型复杂度及落地效率的不同，各类强化学习技术有着不同的适用场景。

3.2.3.1 在线决策与老虎机算法

在线决策及多臂老虎机问题是强化学习中的一类基础而轻量级的问题形式，适用于云网平台中需要实时高效决策的应用场景。其将决策过程视作一轮轮的重复博弈，算法在每一轮根据历史信息选择一个决策，观察到即时收益或损失，目标是在长期内最大化累计收益 [400, 401]。多臂老虎机问题通常不涉及复杂的状态转移，更强调探索—利用权衡以及计算与实现的高效性，这与云网平台的实时性约束和高并发决策需求高度吻合。在云计算与边缘计算中，多臂老虎机算法已经被用于 Kubernetes 资源调度、边缘计算资源分配、网络防御资源分配、5G 网络频谱分配等多种资源管理任务中 [398, 399, 468, 469]。从研究趋势上看，多臂老虎机与在线学习在云网环境中的进一步发展，主要集中在以下几个方向。（1）结构化与约束化的老虎机模型：如带容量/能耗约束的老虎机、能并行拉取多臂的资源分配老虎机等，贴近多资源共享的实际云平台 [470, 471, 472]。（2）考虑互相干扰与系统结构的老虎机：例如多个服务器在同一网络或存储子系统上产生的互扰效应，需要在探索臂期望收益的同时估计干扰结构 [473, 474]。（3）面向大规模在线服务的分布式与并行老虎机算法：在多数据中心、多边缘节点的部署下，通过分布式反馈与局部决策实现快速收敛与鲁棒性能 [475, 476]。

3.2.3.2 经典强化学习算法

经典强化学习更多地基于马尔可夫决策过程进行建模，适用于云网平台中存在状态变化和复杂系统演变的决策场景。MDPs 显式的对状态空间、动作空间、状态转移概率及奖励函数这四个要素进行建模，智能体的目标是在给定折扣因子下，最大化长期回报 [366, 402]。在云网系统中，状态可以刻画系统负载、资源使用率、队列长度、网络拓扑与链路利用率等，动作则对应调度、路由或配置决策。经典的强化学习算法可以大致分为值函数方法（Value-based Methods）和策略梯度方法（Policy-based Methods）两类。值函数方法估计状态值函数或动作值函数，并利用值函数导出策略。在模型未知但可交互的场景，常用的有蒙特卡洛方法、时序差分学习以及 Q-learning、SARSA 等。这些方法在状态空间较小或可以进行线性函数逼近的情况下具有较好的理论收敛性和可解释性，是很多深度强化学习算法的理论基础。策略梯度方法直接对策略函数的参数进行优化，通过估计梯度来更新策略，典型算法包括 REINFORCE 等。Actor-Critic 方法将值函数估计器（Critic）与策略更新器（Actor）结合，利用 Critic 提供的优势估计减少方差，提高样本效率，是众多现代深度强化学习算法的原型。上述经典强化学习算法的理论基础比较成熟、实现复杂度相对较低。在云网场景中应用这些算法对算力和存储的要求相对可控，适合在状态空间规模有限、可观测变量较少的子系统中部署。在数据中心能耗管理、云边协同与负载均衡、数据中心资源扩所容等场景中，这些经典的强化学习方法得到了广泛的应用 [403, 477, 478, 404, 479, 480]。但随着云网系统状态维度、业务类型和耦合关系的快速膨胀，单纯依赖表格型或线性近似的经典 RL 方法难以全面刻画复杂环境，这也是深度强化学习在云网领域快速兴起的重要动因。

3.2.3.3 深度强化学习

深度强化学习将深度神经网络引入强化学习框架中，用于近似高维状态空间下的值函数、策略函数或环境模型，能够处理图像、时序信号、高维特征等系统状态，适用于复杂决策场景。深度强化学习的代表性算法包括 DQN、DDPG、A3C/A2C、PPO 等 [360, 405, 406, 407]。在云计算资源管理方面，可以将多维资源打包与作业调度问题建模为强化学习问题，利用深度网络学习高维资源利用率和队列状态下的调度策略，相关工作包括任务调度/工作流调度及资源扩缩容与配置优化 [481, 482]。在网络与通信系统方面，可以利用深度神经网络感知全局网络状态，将流量工程、网络拥塞控制、频谱资源分配等场景建模为强化学习问题，学习得到优于传统规则及启发式协议的决策策略 [483, 484, 485, 486]。深度强化学习在云网系统中的研究热点主要包括：(1) 多智能体强化学习与分布式决策：利用多智能体对多个服务器、交换机或边缘节点进行协同控制，在保证局部自治的同时实现全局优化 [487, 488]。(2) 样本效率与安全性：在真实云网环境中直接进行大量试错成本高且存在安全风险，因此利用仿真环境、离线数据与安全约束设计深度强化学习算法，是工程实践中的关键问题。(3) 可解释性与可运维性：将深度 RL 策略与现有的运维规则、监控告警体系整合，提供可解释的决策依据和回滚机制，方便运维人员理解与调试 [489]。

3.2.3.4 基于模型的强化学习

基于模型的强化学习会显式或隐式地学习环境模型，估计状态转移和奖励函数。在进行决策之前，智能体会利用学习到的状态转移以及奖励模型进行模拟和规划，从而在有限真实交互样本下实现更高的样本效率和更安全的策略改进。在云计算和网络系统中，真实环境交互往往昂贵或风险较高。例如，在真实数据中心中频繁尝试新的调度和流量策略，可能导致 SLA 违约或大规模性能波动；在运营商网络中测试新路由策略，可能带来大范围的用户体验下降。在这种场景下，MBRL 可以利用历史日志或仿真器构建系统模型，对潜在危险策略进行仿真环境评估，也更容易与编排系统和规则库结合，通过“模型 + 规则 + RL”的方式构造层次化控制架构，具有天然的优势。近年来，已有研究尝试将自动规划技术与 MBRL 结合，用于网络管理规划与决策、联邦学习资源分配、网络性能优化等领域 [408, 409, 410, 411]。相较于无模型强化学习，MBRL 在云网平台中的应用仍处于较早探索阶段，但由于其在样本效率、可解释性和安全性上的潜在优势，预计会成为未来云网智能调度与运维中的一个重要研究方向。

3.3 热点方向八：AI Agent 与 Agentic AI

AI Agent 概念的形成贯穿了人工智能的发展主线，其思想源头可追溯至 1956 年达特茅斯会议对人工系统能力的设想，即机器应能够从环境获取信息并采取有用行动 [491]。随后，经典人工智能研究通过符号推理、规划与知识表示对智能体进行了形式化建模，形成以感知—决策—执行为核心的结构框架 [344]。在这一阶段，AI Agent 多依赖明确的任务建模与静态世界假设，主要应用于受控场景中的规划求解、自动控制与博弈决策。然而，传统 AI Agent 的能力受到三项瓶颈制约：其一，知识获取方式依赖人工构建与领域工程；其二，泛化能力受限于特定环境与任务设定；其三，缺乏跨任务连续性与自主适应能力，难以在开放世界中保持稳定行为。随着数据驱动方法兴起，强化学习与深度学习推动了决策自动化的发展，但仍未突破知识迁移与长期自主性的关键障碍。

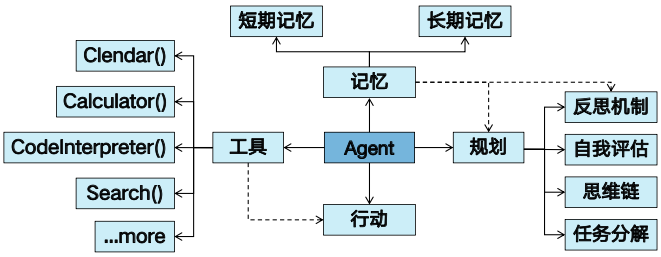


图 3.2: 基于大语言模型的自主智能体系统 [490]

2023 年后，“Agentic AI”概念开始在学术与产业层面被系统化讨论。OpenAI 发布的治理白皮书将 Agenticness 定义为“系统在有限监督下于复杂环境中适应性地实现复杂目标的能力”，并指出其核心特征包括目标复杂度、环境不确定性、适应性与独立执行，而并非对系统进行拟人化理解 [492]。与此同时，吴

恩达在 2024 年提出 “Agentic workflow” 框架，强调通过规划、执行、反思与迭代改进，使人工智能从被动预测式转向具备持续自主性的主动执行式系统，并将该过程概括为可复用的设计范式 [493]。两项工作共同推动了 Agentic AI 从概念认识走向方法论收敛，为后续研究提供了统一的能力语言与分析视角。

在此背景下，大语言模型的出现提供了实现 Agentic AI 的关键能力支撑，OpenAI 系统结构总结了基于大语言模型的自主智能体框架 [490]（见图 3.2），将智能体能力抽象为记忆、规划、行动与反思四个核心模块，并提出以语言作为统一决策接口的范式。该框架的关键在于，使智能体能够在无需重新训练的条件下完成任务分解、工具调用与自我改进，为后续方法（如 ReAct[494] 与 Generative Agents[495]）提供了统一的概念基底，并成为当前 LLM Agent 研究的共识性结构。

然而，能力扩展也伴随相应的治理挑战。有研究指出 [492]，随着系统的 Agenticness 增强，风险重心正由传统的模型输出错误转向持续行动可能引发的系统性危害，其中包括可预测性失效、用途滥用、责任归属不明确以及 “过度委托” 所带来的运行偏移等问题 [496]。针对这一趋势，Shavit 等人提出了面向 Agentic AI 的治理实践框架，强调在系统全生命周期内引入行动空间约束、可审计性机制、可中断能力与责任追踪等要求，以确保此类系统在具备更高自主性的同时仍保持可控、安全与可问责的运行特性 [492]。

总体而言，AI Agent 正从封闭任务求解迈向持续自主与环境适应，在此背景下，本节将依照能力形成与应用外延展开说明：第一部分聚焦 LLM 与 Agent 方法体系，解析角色定义、记忆机制、规划与行动执行等核心能力；第二部分讨论多模态与具身智能体应用，覆盖机器人、交互式 Agent 与跨媒体任务；第三部分面向未来展望，包括智能体互联网与超级人工智能等方向，回应产业体系从单体智能迈向群体涌现的趋势。通过这一结构，本节旨在构建从理论基础、技术能力到未来生态的完整认知框架。

3.3.1 LLM 与 Agent

LLM Agent 是以大语言模型为核心、能够通过自然语言自主理解目标、规划任务并调用工具执行行动的自治智能体。相比传统 Agent，它在知识获取、泛化能力与交互模式上实现代际跃迁，通过模型推理、工具使用和记忆机制的结合构建起 “感知—推理—决策—行动” 的语义闭环。围绕这一能力基础，LLM Agent 形成了由构建、协作与演化组成的方法论框架：构建层面强调角色定义、记忆管理、任务规划与行动执行；协作层面涵盖集中式、去中心化与混合式多智能体组织；演化层面通过自反学习、群体共演化和外部知识反馈不断提升策略稳定性与适应性。同时，工具生态为智能体提供知识检索、程序执行与应用操作等能力扩展，评测体系则从通用任务、行业场景到多 Agent 协作全面衡量系统的可靠性与智能水平。借由这些能力的收敛，LLM Agent 正成为构建可扩展、可协作、可演化智能系统的关键技术路径。

3.3.1.1 Agent 构建方法

LLM Agent 的构建由角色定义、记忆机制、规划能力与行动执行四个核心模块组成，这一体系决定了智能体的基础行为模式与任务执行能力。（1）角色定义：确立智能体的身份与行为框架。基于不同场景，角色可以以两种方式构建：静态角色 [497, 506, 518] 通过人工预设固定身份与规则，使 Agent 在专业任务中保持一致、可控的行为表现，适用于对稳定性与专业性要求高的研发流程和协作系统；动态生成角色 [495, 498] 利用模型生成细粒度的人格与背景，从而支持社会行为建模、用户模拟与开放式场景中的自适应行为。（2）记忆机制：支撑 Agent 长期任务执行与知识持久化。在角色确定之后，记忆机制让 Agent 具备跨轮次任务、状态追踪与知识积累的能力，是 “从一次性对话” 走向 “持续行动” 的关键。记忆机制一般包括三类，短期记忆 [494, 518] 用于维持对最近上下文的理解，是即时推理的基础；长期记忆 [519, 503] 将技能与经验结构化存储，使 Agent 能跨任务复用知识；检索增强记忆 [520, 521] 则为智能体提供可动态扩展的外部知识源，弥补模型时效性与容量的限制。（3）规划能力：驱动智能体 “理解任务、分解目标、监控进度”。在角色与记忆的基础上，规划能力将 “静态知识” 转化为 “可执行步骤”，为 Agent 的行动执行建立清晰路径。链式推理 [522] 和结构化规划 [523] 适用于线性任务分解；树式推理 [524, 502] 适合复杂决策；而基于环境、人类或多 Agent 反馈的迭代规划 [494, 503] 则进一步提升了策略优化能力。（4）行动

表 3.2: Agent 方法体系主要类别与代表性工作

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
Agent 构建	LLM Agent 的构建围绕角色定义、记忆机制、规划能力与行动执行四个核心模块展开，决定智能体从“一次性对话”走向“持续行动”的基础行为模式与任务执行能力。通过语义闭环的“感知–推理–决策–行动”设计，使 Agent 能在复杂环境中稳定完成任务。	NeurIPS ACL EMNLP AAAI ICLR NAACL ICML TACL	• **角色定义：** KAUST 团队 [497] 提出 CAMEL，以人工预设身份与规则构建静态角色；腾讯团队 [498] 提出 HPD 模型生成细粒度人格与背景实现动态角色设定。 • **记忆机制：** ETH 团队 [499] 提出 Graph of Thoughts，用于维护近期上下文的短期记忆；港科团队 [500] 提出 COMEDY，以结构化形式存储技能与经验，实现长期记忆；中科院团队 [501] 提出 DeepRAG，通过外部知识库检索增强上下文，实现检索增强记忆。 • **规划能力：** KAIST 团队 [502] 提出 ReAcTree，以任务分解将复杂目标拆解为可执行步骤；Princeton 大学团队 [503] 提出 Reflexion，基于环境/人类/多 Agent 反馈进行迭代规划与自我改进。 • **行动执行：** 港科团队 [504] 提出 TIP，面向程序执行、搜索与 API 操作等工具调用；清华大学团队 [505] 提出 Toolink，支持自动生成脚本/函数的工具创造。
Agent 协作	Agent 协作机制决定智能体在群体任务与复杂系统中的整体表现，通过集中式、去中心化与混合式三类组织形态，将单体能力扩展为系统级智能与群体智慧。	NeurIPS ACL EMNLP AAAI ICLR AAMAS	• **集中式协作：** DeepWisdom 团队 [506] 提出 MetaGPT，以中央控制器进行任务拆解与角色分配；Rochester 大学团队 [507] 提出 Coscientist，体现由中心化调度组织专家能力的协作范式。 • **去中心化协作：** Yale 大学团队 [508] 提出 MedAgents，以多 Agent 平等主体的讨论与辩论形成分布式群体智慧。 • **混合式协作：** 浙大团队 [509] 提出 KnowAgent，结合集中式可控性与去中心化灵活性，并支持按任务动态调整协作拓扑。
Agent 演化	演化机制使 LLM Agent 能在长期交互中自我优化与适应环境变化，通过自主演化、多智能体共演化与外部资源驱动等，从“静态性能”迈向“持续成长”的自治智能体系。	NeurIPS ACL EMNLP AAAI ICLR TMLR	• **自主演化：** CMU 团队 [510] 提出 Self-Refine，以反思/验证等机制实现自我改进；港科团队 [511] 提出 ControlMath，以自反馈与控制策略提升长期推理稳定性。 • **多智能体共演化：** 港中文团队 [512] 提出 ProAgent，通过协作或对抗交互提升系统鲁棒性与整体能力。 • **外部资源驱动演化：** 微软团队 [513] 提出 CRITIC，利用外部工具与反馈信号持续纠错与能力扩展；上交团队 [514] 提出 SelfEvolve，借助环境交互与工具执行实现可迁移的长期积累。
Agent 评测	评测体系从通用任务、领域场景到多 Agent 协作，系统性衡量智能体在推理、规划、工具使用、环境操作与长期自治等维度的综合表现，是检验 Agent 是否“可用、可信、可落地”的关键标准。	NeurIPS ACL EMNLP ICLR ICML TPAMI ICRA	• **通用评测：** 清华大学团队 [515] 提出 AgentBench，关注推理、规划、工具使用、网页/环境操作与长期任务执行等基础能力；并提供标准化协议与指标以衡量成功率、步骤正确性与鲁棒性。 • **领域评测：** 斯坦福团队 [516] 提出 MedAgentBench，面向垂直行业检验专业知识、决策质量与风险控制能力；强调安全合规与证据支撑）。 • **多智能体评测：** 浙大团队 [517] 提出 CrewAI 评测基准，衡量群体智慧、协作稳定性与策略一致性；同时刻画沟通效率、共识收敛与典型失效模式（角色漂移/循环对话）。

执行：决定智能体是否真正能完成任务。行动执行是规划能力的落点，使智能体从“会说”走向“能做”，覆盖从工具调用到多模态交互的全链条能力。工具调用 [525]（Python、API、搜索、计算工具）使 Agent 能处理模型无法直接完成的任务；Web 操作 [526] 使 Agent 能在真实环境执行点击、操作、填表等具体动作；多模态与具身行动 [527] 进一步扩展到物理世界。

3.3.1.2 Agent 协作机制

LLM Agent 的协作机制决定智能体在群体任务、复杂环境与多角色系统中的整体表现，是从单体能力向系统能力扩展的关键。协作通常呈现集中式、去中心化与混合式三类模式，分别适应不同的任务结构与组织需求。(1) 集中式协作：由控制器统一规划与调度。集中式模式以单一控制器为核心，负责任务拆解、角色分配与流程管理，确保系统整体的一致性与可控性。MetaGPT [506]、Coscientist [507] 等系统通过中心 Agent 统筹全流程，使多 Agent 协作更接近工业流水线与研发过程。这一模式适合高结构化任务，如软件开发、科研项目管理与分阶段生产流程。(2) 去中心化协作：依赖群体讨论与分布式协同。在去中心化模式下，Agent 之间平等交流，通过讨论、辩论或角色自治形成群体智慧。AutoGen [528]、Multi-Agent

Debate [529] 等框架通过多 Agent 的平行对话、观点碰撞与相互验证，在开放式问题求解中表现突出。该模式特别适合复杂推理、多解任务、社会行为模拟与需要多样观点融合的场景。(3) 混合式协作：兼具集中式的可控性与去中心化的灵活性。混合模式将集中式组织架构与分布式自治机制结合，能够在不同任务阶段动态调整协作结构。CAMEL [497]、AFlow [530] 采用固定拓扑结构实现稳定协作；DyLAN [531] 等系统则允许智能体根据任务状态动态重组互联结构。混合式能够平衡“可控性”“灵活性”与“扩展性”，适合多阶段任务、动态环境与复杂策略空间。

3.3.1.3 Agent 演化机制

演化机制让 LLM Agent 能在长期任务中自我优化、积累经验并适应环境变化，是迈向自治智能的重要能力层。演化通常由自主演化、多智能体共演化与外部资源驱动三部分构成。(1) 自主演化：通过反思、纠错与奖励自我提升。自主演化让 Agent 不依赖人工干预即可持续改进。Self-Refine [510]、Self-Verification [532] 与 Self-Rewarding [533] 等机制使智能体能够自检输出、纠正错误并基于模型评估优化策略。这一机制强化了 Agent 的稳定性与长期表现，是持续自治的基础。(2) 多智能体共演化：通过协作或对抗提高系统能力。共演化强调多个智能体之间的交互促进整体能力提升。协作式共演化 [512] 通过信息共享与策略互补提高整体效率；对抗式共演化 [534, 529] 通过“攻防训练”增强系统鲁棒性。这种机制接近生态系统中的“互惠—对抗—竞争”结构，使 Agent 在复杂环境中更稳健。(3) 外部资源驱动演化：借助知识与工具反哺能力增长。外部资源为 Agent 提供持续扩展的能力边界。KnowAgent [509]、CRITIC [513] 等方法通过外部知识库、执行反馈与环境模拟不断增强智能体能力，使其具备可迁移性与长期积累能力。演化机制最终让 LLM Agent 从“静态性能”迈向“长期成长”，实现可持续、自适应的智能体系。

3.3.1.4 Agent 工具体系

工具体系是扩展 LLM Agent 能力边界的关键基础设施，使智能体能够超越语言生成，完成真实环境中的操作、计算与信息获取。工具生态一般由“Agent 使用工具”“Agent 创造工具”与“部署框架”三部分组成。(1) Agent 使用的工具：为智能体提供外部能力接口。这一类工具直接提升 Agent 的行动能力、信息获取能力与精准计算能力，是最基础的能力扩展方式。检索工具（WebGPT [535]、GraphRAG [521]）为智能体提供外部知识补全；执行工具（Python Executor、Toolformer [536]）让 Agent 能进行正式计算与程序执行；API 工具（RestGPT [537]）使其能操作互联网服务与应用系统。这些工具共同构建 Agent 与外部世界的交互界面，使其“能行动、能查询、能执行”。(2) Agent 自主创造工具：增强智能体的自适应与可扩展性。当任务超出现有工具能力，Agent 可以生成新的工具来补足系统能力缺口。CRAFRT [538]、Toolink [505] 等系统让 Agent 拥有“工具制造能力”，能够自动构建函数、脚本或任务专用工作流，从而动态扩展系统能力边界。这类能力让 Agent 具备“进化式扩展”，对应演化机制中的工具强化路径。(3) Agent 部署与工作流工具：构建可落地、可管理的 Agent 系统。工具生态的第三层是构建完整 Agent 产品所必需的运行时环境。LangChain [539]、AutoGen [528]、LlamaIndex [540]、Dify [541] 与 MCP [542] 提供任务编排、上下文管理、工具协议、任务路由与多 Agent 协作框架，是构建产业级 Agent 的基础设施。这些工具共同构成智能体的“应用层骨架”，支持从模型推理到任务执行的全链路运行。

3.3.1.5 Agent 评测体系

评测体系用于衡量 LLM Agent 在推理、规划、协作、环境操作与长期自治能力等维度的综合表现，是检验智能体是否“可用、可信、可落地”的核心标准。现有评测框架通常分为通用评测、领域评测与多智能体评测三类，它们共同构成对 Agent 全链路能力的系统性测量。(1) 通用评测 [515, 543, 544, 545]：衡量 Agent 在真实任务与环境中的基础能力。通用评测关注智能体的推理质量、任务规划、工具使用、网页操作、环境交互与长期任务执行能力，是判断 Agent 是否具备基础“可执行智能”的标准。(2) 领域评测：检验 Agent 在专业场景中的可靠性与安全性。领域评测关注智能体在垂直行业场景中的专业知识、决策质量与风险控制能力，是评估 Agent 是否能够在高要求领域落地的必要条件。医疗 [516]、自动驾驶

[546]、数据科学与工程 [547, 547] 等评测框架，覆盖诊疗咨询、自动驾驶策略、数据分析流程与模型构建等专业场景。这些评测强调准确性、安全性、可解释性与任务合规性。(3) 多智能体评测：衡量群体智慧、策略一致性与协同稳定性。多智能体评测用于检验 Agent 在协作、辩论、任务分配与资源协调中的表现，是构建大规模智能体组织的核心标准。TheAgentCompany[1] [548]、MLRB [549] 等评测体系关注多 Agent 协作效果、冲突处理、角色履行与系统稳定性。这些评测反映群体智能的涌现能力，带来比单体测试更复杂的动态性。三类评测共同构成 LLM Agent 从底层能力、专业能力到系统性协作能力的立体测量框架，不仅是检测智能体性能的工具，更是推动 Agent 向可信、可控、可部署方向发展的核心驱动力。

3.3.2 多模态与具身 Agent

多模态与具身 Agent 在近年来的人工智能研究中愈发受到关注，被视为推动具身智能与通用智能体系发展的重要研究方向之一 [551]，具有广泛的应用场景（图 3.3）。此类智能体通常以“感知–认知–行动”的交互循环为基础，将语言模型、视觉语言模型、强化学习与模仿学习等方法置于统一的决策框架中进行建模 [552, 553]。在多模态学习方面，Agent 能够整合视觉、语言、语音等多源信号，形成语义化环境表征，并依托 LLMs 与视觉语言模型在视觉内容与指令表达之间建立关联 [554, 555]。在具身智能方面，Agent 通过持续与环境交互获得状态、动作与反馈信息，从而在物理或虚拟场景中更新策略 [494]。此外，多模态信号（如手势、姿态与语言输入）的联合使用，使得智能体能够在交互过程中处理更丰富的上下文信息 [556]。基于以上背景，本节从多模态感知与行动的建模机制、具身智能中的交互与策略学习过程，以及面向不同应用场景的泛化与迁移能力三个方面，对多模态与具身 Agent 的研究内容进行介绍。

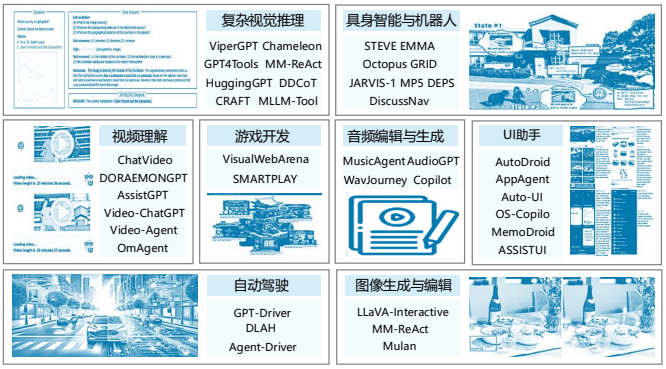


图 3.3: 多模态与具身智能体的广泛应用 [550]

3.3.2.1 多模态感知建模

多模态感知与行动的建模机制侧重于通过跨模态表征与对齐，将视觉、语言等感知信息转化为可驱动 Agent 行为的语义表示。多模态 Agent 通常依赖视觉语言模型与大规模语言模型实现跨模态信息表示与对齐，其目标在于从图像、文本、语音等多源输入中提取相互关联的语义线索 [557]。相关研究通过图像–文本对齐、跨模态注意力建模以及多模态预训练，使得 Agent 能够在接收到视觉数据时结合语言描述进行语义理解，从而支持图像描述、视觉问答、视频分析与指令跟随等任务 [558, 559]。随着跨模态预训练数据和模型规模的扩大，Agent 在目标识别、场景理解与时序推理等任务中能够构建更丰富的环境表征 [560]。此外，多模态建模常与决策模块结合，例如在导航或人机交互场景中，通过视觉特征提取、文本指令解析与语义映射，将多模态信息转换为可用于动作生成的结构化表示 [561, 562]。该类方法在自动驾驶、机器人感知与交互式系统中得到广泛应用 [563]。

3.3.2.2 具身交互学习

具身智能中的交互与策略学习主要围绕利用传感器输入和环境反馈构建感知–动作–学习的闭环展开。具身 Agent 侧重于利用传感器输入与环境反馈完成感知、动作与学习的闭环过程 [564]。强化学习与模仿学习是具身智能体中常见的训练范式：前者基于试错机制学习动作策略，后者通过专家示范构建初始策略或约束策略空间 [367]。在具身交互中，Agent 通常利用视觉、触觉、运动反馈等信息估计环境状

[1] https://the-agent-company.com/

态；同时，电机控制、路径规划与时序动作生成等模块负责执行决策输出 [565]。研究中还结合模型预测控制、状态估计与策略优化方法，以在动态、不确定或部分可观环境中提高学习效率 [566]。围绕复杂操作任务（如抓取、装配、导航等），具身智能研究形成了包括操作技能分解、反馈调节机制、模拟到现实（Sim-to-Real）迁移等在内的技术体系，用于支持智能体在不同交互条件下进行策略学习 [567, 568]。

3.3.2.3 泛化与迁移能力

多模态与具身 Agent 的泛化与迁移能力研究关注智能体在跨任务、跨环境与跨模态条件下保持性能稳定与适应性的机制。多模态与具身 Agent 的研究涉及到跨任务、跨环境与跨模态的泛化能力 [569]。相关研究探索在不同数据分布、感知条件和任务要求下的策略适应方法，包括跨域学习、知识迁移、策略微调与分层任务规划等 [570]。视觉–语言–行动任务中，Agent 可利用预训练模型获得跨领域的表征能力，再通过适量任务特定数据进行调优，实现从一个场景向另一个场景的迁移 [554]。对于具身学习，任务与动作规划框架通过将复杂任务结构化为子任务，使得智能体能够在长时间序列任务中保持可解释的执行路径 [571]。多模态输入（例如视觉、语言、雷达或激光雷达数据）的联合使用，也使得 Agent 能够在自动驾驶、智能家居与医疗场景中适应不同感知条件与环境变化 [572, 573, 574, 575]。上述方法共同构成了多模态与具身 Agent 在实际应用中实现泛化能力的主要技术途径。

3.3.3 2025 年的 AI 发展

虽然麦肯锡等咨询机构的调研显示 AI Agent 的应用尚处在早期 [576]，2025 年围绕 AI Agent 的空前讨论热度则是肉眼可见。

2025 年 AI 成为最广泛的共识，模型能力继续发展，应用开始大范围落地，研究、产业、资本等各方面都聚焦 AI。从年初 Deepseek-V3/R1 的横空出世，到年底 Google 带着 Gemini-3 王者归来，2025 年 AI 大模型在开源和多模态方面继续带来新的惊喜。2025 年最大的特点也许是 AI 达成了前所未有的广泛共识，几乎所有人都坚信 AI 是确定性的未来，NVIDIA 在 2025 年成为人类历史上首个市值超过 4 万亿美元的公司。人们不再满足于 AI 技术的点滴进步，而是畅想 AI 将会带来怎样的未来。

“从通用人工智能（AGI）走向超级人工智能（ASI）”开始被提出，标志着以大语言模型为代表的 AI 能力开始接近甚至某些方面已经达到人类水平。2025 年，超级人工智能（Artificial Super Intelligence，or ASI）开始被越来越多提及并且严肃讨论。年初，OpenAI 的 Sam Altman 在个人博客中提到 AI 的关注点开始从 AGI 向 ASI 升级 [577]。6 月，Meta 公司成立超级智能实验室，并且逐渐成为 Meta 的 AI 研发主线。9 月，阿里巴巴的云栖大会以“云智一体. 碳硅共生”为主题，主旨演讲中抛出了观点“通用人工智能已成为确定性事件，但这只是 AI 发展的起点，行业终极目标是实现能自我迭代、全面超越人类的超级人工智能”。11 月，微软 AI 团队提出 Humanist Super Intelligence 的目标 [578]。

“碳硅共生”开始被讨论，当 AI 的能力与人类相当，基础设施将面临新的“硅基”主体。2025 年底，AI 已经有能力生成让人类肉眼无法分辨真假的图片 [579]。不管是通用人工智能还是超级人工智能，当 AI 不再只是工具，而是具备和人类类似，甚至超越人类的能力，哪怕只是在某些方面，如何与未来的 AI 共生成为无法回避的问题。7 月，中国移动发布《2025 智能体互联网络白皮书》，讨论 Agent 成为人类之外的互联网主体之后，互联网基础设施该如何演进。

AI 安全与治理仍然是广泛关注的问题，AI 对社会伦理和人类生存安全的冲击需要有合理应对。于此同时，AI 带来的风险依然受到广泛的关注 [580]。当 AI 的定位不再只是工具，那么该怎么保障 AI 不会破坏人类社会的安全和伦理？“先发展，还是先保障安全”的问题也仍然存在着争议。本文第 4.3章节将展开讨论数据和 AI 的安全和隐私保护问题。

AI 接下来的发展会更像互联网还是元宇宙？ 2025 年，生成式 AI 的应用高速发展，chatgpt 等工具已经成为日常工作和生活的一部分。Agentic AI 概念开始被广泛接纳，为智能体的长足发展奠定了认知层面

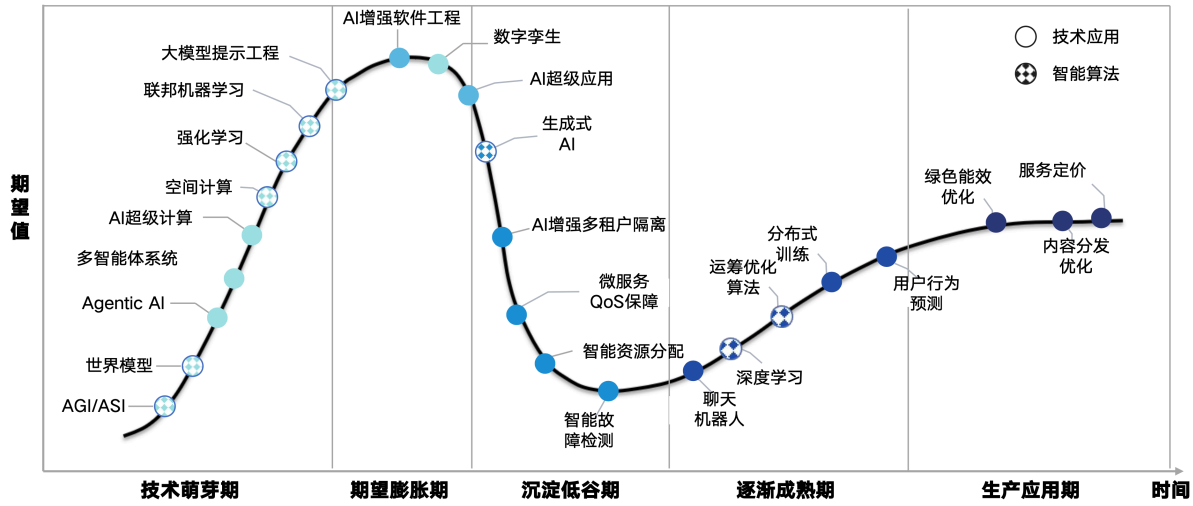


图 3.4: 智能算法研究图谱技术成熟度曲线 2025

的共识基础。AGI 和 ASI 则为 AI 发展指明了方向。那么，接下来 AI 的发展是不是一片坦途？参考 Gartner Hper Cycle 为新技术提供的发展规律曲线，现在的 AI 是在预期多于应用价值的发展早期（Peak of Inflated Expectations），还是已经处在稳步落地的应用推广期（Plateau of Productivity）？同样是普惠型基础创新，互联网的发展也许是最接近的参考。以史为鉴，2000 年的“互联网泡沫”前后，是互联网基础设施的如火如荼建设和互联网应用的长足稳步发展。也许不算一帆风顺，但是前景广阔、影响深远。当然，科技发展过程中，我们也在见证不一样的案例，例如元宇宙（Metaverse），今天的 AI 已经有些相似，两者同样有着清晰的目标和可以用科幻电影具象展示的未来，例如《头号玩家》之于元宇宙，《钢铁侠》的贾维斯和《终结者》中的天网之于 ASI。元宇宙离大范围应用还有差距，ASI 是不是也需要更长的时间来兑现？

3.4 展望与建议

基于技术成熟度曲线的分析方法（如图 3.4 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。

3.4.1 智能算法的未来研究方向和关键技术展望

智能算法在云-网-智算一体化系统的研究将沿着“运筹优化的结构化数学能力”与“深度学习模型的数据驱动优势”双线融合演进。随着算力规模爆炸式增长、业务负载呈现强动态性与不确定性，单一的传统优化方法难以应对高维资源、多层拓扑与复杂工作流之间的快速演化关系。未来研究将重点聚焦三个方向：发展面向超大规模离散结构的高效近似组合优化技术，为复杂任务图、任务分配、资源调度提供高效可扩展的优化决策基础理论；构建能够覆盖跨资源维度、异构拓扑与复杂约束条件的可扩展凸/非凸优化算法，以适配云网系统持续增长的数据规模与决策复杂性；发展具备快速搜索、动态调整与强鲁棒性的启发式与元启发式智能优化方法，使系统能够在非平稳、高动态环境下保持高效运营。整体而言，“优化 + 学习”的深度融合将成为未来智能云管理的核心路径，使云网系统具备端到端的预测能力、自主优化能力与跨场景迁移能力，为构建下一代自治化、智能化云基础设施提供关键技术支撑。

以图算法、图神经网络和深度神经网络为代表的结构化建模与表征学习技术，将持续支撑云-网系统在复杂拓扑、多维依赖和异构数据上的智能演进。传统图算法通过图划分、路径规划与流量分配等机制，

在 AI 分布式计算、负载均衡和网络性能优化中提供了高效可靠的解决方案；超图与动态图算法则面向高阶关系和时变拓扑，为数据中心多维资源管理、复杂任务分解和高频变化场景下的实时调度提供新的建模能力。在此基础上，将 GNNs 与传统图算法深度融合，可通过端到端学习节点、边和子图表征，实现对流量模式、异常行为、任务依赖的精准建模，支撑智能流量预测、异常检测、任务调度与故障定位等关键能力；而更通用的深度神经网络则在多模态指标建模、性能预测与策略参数化方面提供了强大的函数逼近能力。未来，图算法 + GNNs + 深度网络的协同，将推动云–网系统从结构可解析走向结构可学习，在保持工程可控性的同时，释放大规模数据驱动优化与智能运维的潜力。

未来 AI Agent 的关键突破或聚焦于自主智能的体系化演进、新型智能基础设施、以及面向复杂社会环境的安全可信治理体系三条主线。AI Agent 的发展将从模型能力竞争转向系统能力建设，重点形成可感知、可推理、可行动、可进化的自主智能体系。随着大模型逼近 AGI、迈向 ASI，Agent 将在多模态理解、具身交互、长期规划与自我改进方面持续增强，并从“智能工具”转变为“智能主体”。这一转变要求计算系统实现从云到边到端的全栈重构，使 AI 能在大规模、低时延、高安全场景下协同运行，真正迈向“碳硅共生”的技术范式。同时，智能体数量的指数级增长也将把安全治理推向核心议程，在身份可信、行为可控、价值对齐、隐私保护与风险隔离等方面提出系统级要求。总体来看，未来 AI Agent 的研究重点将围绕自主智能、智能基础设施与安全治理三大方向展开，共同支撑可信、可控、可持续的智能社会形态。

3.4.2 智能算法的发展建议

重视算法理论的基础研究，以形式化方法、复杂性分析、最优化理论等为核心，构建面向大规模云网系统的科学理论框架，为资源调度、负载均衡、容量规划等关键机制提供可靠的理论支撑。扎实的算法基础不仅能够提升调度策略的精准性、可解释性与可验证性，还可显著增强云平台在异构算力池、多租户环境以及跨区域集群中的决策一致性与调度稳定性。此外，在业务需求高度动态、多模态并发持续增长的趋势下，算法理论与 AI 模型的深度融合将成为构建下一代智能决策体系的核心驱动力。依托算法理论与 AI 模型的支撑，可以对超大规模、异构化资源（CPU/GPU/FPGA、存储介质、网络带宽）进行更为系统的分析与预测，包括利用时序建模、流量预测、拓扑感知优化等技术提前识别业务负载趋势，为资源池预配置、多集群联邦调度和边缘-云协同布局提供决策依据，从而实现资源的智能编排与最优配置。这不仅能降低整体运营成本（TCO），也能强化服务质量（QoS）、响应延迟与系统可用性的保障能力。

构建可控可信的自治智能与治理体系。未来云网中的智能算法应从单点模型能力扩展为“系统中心”的自主智能框架，将可验证性、可解释性、行为预测与安全沙箱等机制前置设计，把工具调用边界、跨租户隔离、行为可追溯等要求固化为平台能力。需主动对接国家与国际组织在人工智能伦理与监管上的共识，围绕益处最大化、伤害最小化、公平性与责任可追溯，形成覆盖模型训练、部署到运行全生命周期的评估与审计体系。尤其是在大规模 Agentic AI 和多智能体协同场景中，要从单体可控扩展到群体可预测与可干预，通过数据安全、隐私保护、内容可信与责任边界等制度化安排，在技术加速与社会稳态之间建立长期平衡，使云平台真正成为大规模自治智能系统的运行与治理基础设施。

推动绿色高效的云—边—端协同与具身智能基础设施建设。面对大模型、Agent 与具身智能带来的算力与能耗压力，云网中的智能算法设计应将能效视为基本约束，把绿色人工智能与国家“双碳”战略深度融合：在算法层面通过剪枝、蒸馏、参数共享等技术降低训练与推理开销，在系统层面依托云—边—端一体化调度、冷热分级存储与异构算力编排实现全栈节能。云和边缘基础设施将为多模态、具身 Agent 的训练与在线决策提供弹性算力支撑，通过将长期规划与记忆托管于云侧、把实时感知与控制下沉到边端，实现性能、成本、能耗与体验的协同优化。通过绿色、高效、协同的基础设施演进，使智能算法既成为云网基础设施智能化升级的核心引擎，也真正成为传统高碳行业节能减排和可持续转型的重要工具。

第 四 章

面向新兴技术的研究

在强大的云计算和算力网络支撑下，新兴技术产业呈现出融合创新、多元发展的态势。以低空智能、6G、AI 与量子技术为代表的新兴技术，为工业互联网、智慧金融、视联网及各行业数字化创新应用提供坚实基础，催生出一系列具有战略价值的新兴业态。本章将深入分析上述重点技术领域的发展现状、挑战、关键技术与研究热点，并提出未来发展建议。

4.1 研究图谱 2025：新兴产业布局中的技术生态与发展脉络

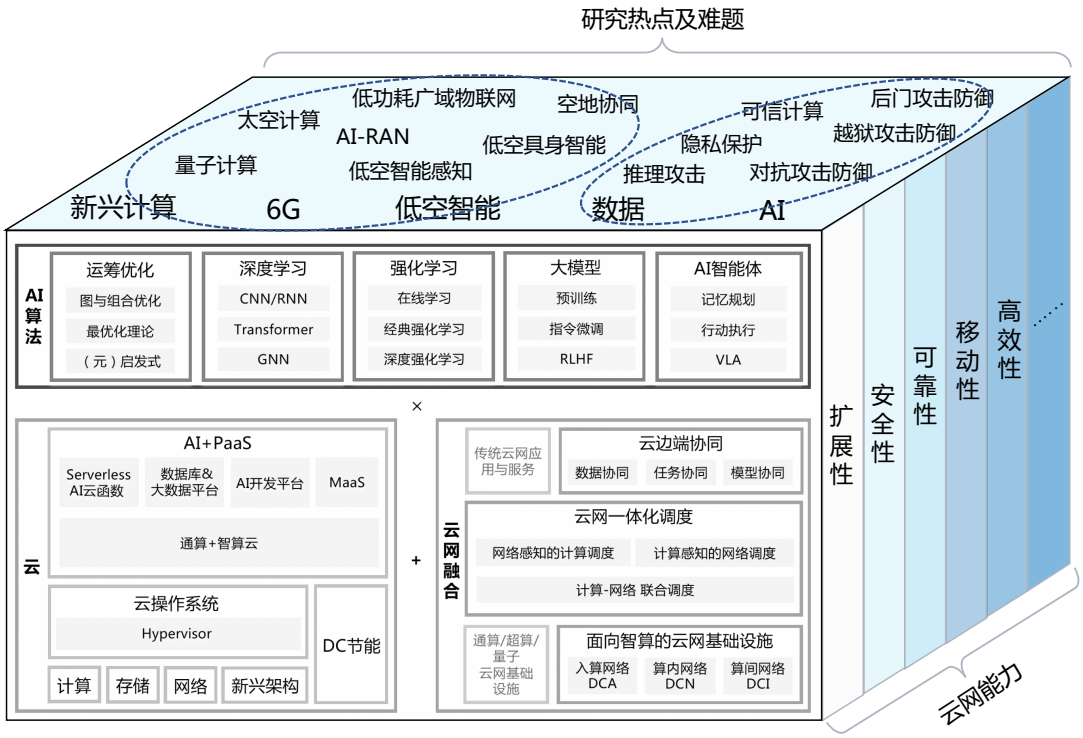


图 4.1: 面向新兴技术的研究图谱 (由云计算研究院总结形成)

当前，云计算与云网融合技术已在一系列新兴产业中得到广泛应用，包括工业互联网、智慧金融、量子计算、视联网等。这些产业的发展不仅依赖于前沿技术，如 6G 通信、低空智能计算、数据隐私保护、AI 系统防护，同时也在实践中推动了云网系统能力的提升，使其在安全性、可靠性、移动性、高效性和扩展性等方面得到进一步保障。本节在前三章研究架构之上，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。为系统刻画新兴技术背景下云计算与云网融合的发展脉络，有必要在应用场景与技术能力的双重驱动基础上，对相关研究内容进行体系化梳理。一方面，新兴产业的快速演进持续提出更高的算力调度、网络协同、安全可信与智能化管理需求；另一方面，前沿技术的涌现也不断重塑云网一体化的体系结构与演进方向。因此，我们需要从更全局的视角重新审视云网融合的内在逻辑、关键挑战及未来趋势。

4.1.1 趋势分析

当前，新一轮科技革命与产业变革正处于交汇叠加阶段，国家层面围绕数字经济、先进制造、未来产业等领域密集出台战略部署，为新兴技术的发展提供了明确方向。**与此同时，云网基础设施持续演进，算力体系、网络架构、调度机制及安全框架正面向智能化、融合化方向加速升级**。在此背景下，系统研判新兴产业发展态势与关键技术突破路径，对于构建高质量数字底座、支撑产业数字化与数字产业化协同发展具有重要意义。从云网基础设施视角观察，新兴技术的迭代正在深刻影响算力供给方式、网络承载能力与智能服务体系。例如，工业互联网在政策推动下加速向高可靠、可解释、场景化的智能系统演进；面向 6G 的技术体系与 AI-RAN 架构推动云网深度协同成为趋势；智慧金融对安全可信计算、模型风险管理及高稳定性基础设施提出新要求；量子计算对现有密码体系和高性能计算模式形成潜在影响；低空经济的兴起带动时空信息网络、边缘节点与通信保障体系的重构；视联网的发展则进一步推动泛在感知、视频处理与实时交互能力的普及。

基于此，本小节将围绕工业互联网、视联网、智慧金融、低空经济、6G 及量子计算等重点方向，从政策动向、技术趋势、产业需求与基础设施适配性等维度展开前瞻性分析，为后续云网能力规划与产业生态布局提供参考。

工业互联网正通过政策引领与厂商实践双向驱动，重构云基础设施的技术内核与产业生态。从“连接设备与数据采集”的早期阶段，演进为推动制造业实现智能化、柔性化与高质量发展的关键基础设施。国家”十四五”规划及《制定国民经济和社会发展第十五个五年规划的建议》明确将工业互联网定位为制造业数字化关键基础设施，要求促进实体经济与数字经济深度融合。这一战略跃迁对云设施提出更高要求：不仅需要大规模算力、存储与实时处理能力，还须支持边缘智能、数据安全隔离及标准化接口，以满足多租户、多场景的复杂需求。当前产业呈现三大趋势：云厂商与设备制造商深度合作、AI 与数字孪生技术规模化应用、中小企业加速上云与边缘部署。2025 年，国内外云厂商在工业互联网领域展现出深度垂直整合与平台化服务能力提升的显著趋势：中国电信依托天翼云在柳钢集团热轧厂部署 5G 融合架构，实现 AI 质检与产线无人化作业 [581]；Microsoft Azure 获评 Gartner 全球工业物联网平台领导者，通过 Azure Arc 与 Copilot AI 推动工业智能规模化生产 [582]；Amazon 云科技展示 AWS IoT Greengrass 预测性维护方案，以边缘计算实现故障提前预警，标志着从”卖产品”向”卖结果”的服务转型深化 [583]。这些实践印证，工业互联网已成为驱动云基础设施创新的核心引擎与明确方向。

作为支撑数字政府、智慧城市与产业智能化转型的新型视频基础设施，视联网正从传统安防网络加速迈向“云网融合、视频融云、云智一体、安全可信”的系统化升级阶段。《视联网云化技术白皮书（2024）》指出，视联网已从模拟、数字化、网络化演进至智能化阶段，并进入以云化为核心的关键转折期，产业链正在形成“标准体系建设—云化技术突破—区域规模部署—行业生态培育”的整体路径。当前，云网融合基础设施加速构建，分布式云网、SDN/NFV、SRv6 等技术实现资源统一编排和低时延接入；视频融云推动智能编码 [584]、超低时延传输、直存技术落地，支撑视频能力从监控走向实时业务；云智一体带动视觉大模型与视频理解在政务、交通、园区等场景规模化应用，形成从数据采集到智能决策的闭环能力；同时，量子加密、隐私计算、区块链等安全可信技术体系逐步完善，为视联网跨域共享和合规流通提供基础保障。在这一演进过程中，一批代表性应用场景已经展现出产业加速成型的趋势：如天翼视联网依托“一个平台、一朵云、一张网”打造全国最大的视频监控数字化平台，实现 EB 级视频资源统一纳管与跨域调度；基于 WebAssembly[585] 的无插件视频通信内核突破浏览器限制，为政务指挥、应急调度提供低门槛的实时视频通信能力；端云协同的“数字视网膜”架构在交通与城市治理中实现实时识别、结构化分析和态势研判；量子加密链路与区块链存证应用已在重点区域落地，构建面向公共安全与关键行业的可信视频流通体系。总体来看，视联网产业呈现云化加速、标准体系完善、智能能力融入、应用场景扩展的态势，预计将在 2025–2027 年进入从区域试点向全国性规模化部署的关键窗口期，并成为国家级数据底座与新型基础设施的重要组成部分。

在数字经济全面跃迁与金融业务加速数字化的背景下，智慧金融体系正向智能化、可信化快速迈

进。《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》明确指出："加快建设金融强国"，强调发展金融科技、数字金融，并建设安全高效的金融基础设施，加强构建风险防范化解体系。近年来，中国电信聚焦金融科技核心能力，在数据治理、智能风控等方向取得了系列突破 [586]。在数据安全治理方面，中国电信翼支付构建企业级可信数据空间，以区块链、隐私计算、数据沙箱及后量子密码技术实现数据"可用不可见"，覆盖存储、处理与跨域流通过程，为跨平台数据交互提供高强度加密与严格访问认证，形成面向未来算力威胁的隐私防护体系。依托自研"密流安全计算平台 PrivTorrent"，中国电信进一步实现跨机构场景下的数据可控融合计算，在监管、风控、反欺诈等任务中实现高性能、可计量的可信数据流通。在智能风控体系建设上，翼支付推出自研 RiskX 2.0，通过"智能体为核心、大小模型协同"架构将图智能、风险语义理解与多模态识别深度融合。国内外云厂商同样关注智慧金融领域，并持续加强面向高敏感金融场景的技术基础能力建设。在国内，阿里云以机密计算支撑核心账务、交易清算与风控系统上云 [587]。在国外，Microsoft Azure 通过机密计算提供高隔离金融计算环境，支持金融机构开展联合风控、风险分析；Google Cloud 通过机密计算空间支持跨机构数据协同与生成式 AI 风险管控 [588, 589]。伴随业务向云化和智能化深化，数据安全与生成式 AI 安全的重要性进一步凸显，可信的数据处理方式与可控的模型能力成为保障金融应用可靠落地的关键。

近年来，低空经济作为培育新质生产力的重要战略性新兴产业，正逐步上升为国家重点布局的发展方向 [590]。随着无人机、空中出租车、无人货运等技术的突破，低空经济正迅速改变传统产业格局，并推动着新一轮产业变革。《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》[591] 明确指出："加快新能源、新材料、航空航天、低空经济等战略性新兴产业集群的发展，推动量子科技、生物制造、氢能和核聚变能、脑机接口、具身智能、第六代移动通信等成为新的经济增长点。"低空经济的发展不仅为交通运输、物流配送、农业植保等多个领域带来了创新机遇，同时也为云计算和智能计算技术的应用提供了广阔的舞台。以无人机为代表的低空飞行器需要强大的实时数据处理能力和高效的智能计算支持，这促使低空智能计算成为技术研究的焦点。无人机在飞行过程中产生的大量数据，需要通过云计算和边缘计算的协同处理来实现实时决策和精准控制。尤其是在低空经济的核心领域——智慧物流与智能城市建设中，低空智能计算技术的应用将极大提升运营效率、降低成本，并推动智能交通系统的进一步发展 [592]。随着国家政策的支持和技术的持续进步，低空经济的产业链将更加完善，数字化转型的步伐也将加速，云计算将在其中扮演至关重要的角色。

作为支撑未来数字文明和智能社会的关键基础设施，6G 成为我国重点关注和发展的重点方向之一。《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》将 6G 作为新的经济增长点 [591] 。目前，6G 的产业发展现状正在从概念研发布局进入关键技术攻关与预商用验证的过渡期，全球产业链已逐步形成"标准预研—关键技术突破—试验网络验证—生态体系培育"的整体路径。各大头部企业积极推动产业化进程，爱立信、诺基亚、三星、华为、中兴等厂商已展示 140GHz 以上太赫兹原型机、AI-RAN 架构、星地融合网络样机和通感一体化设备 [593]。同时，运营商与产业链企业形成了较为完整的研发体系，中国电信将 6G 视为未来网络发展的核心战略方向，并率先提出"全域智惠网络"技术体系，围绕天地一体化、通感融合、智能超表面（RIS）、星地融合通信等关键技术方向展开系统布局。中国电信牵头的"6G 系统计费研究"项目获批通过，实现中国电信在 3GPP 国际标准组织的 6G 牵头立项突破，能够为 AI、通感、天地一体等 6G 关键服务提供可行的计费方案。此外，中国电信在 6G 网络架构、星地融合、近域网络、无线智能化等方向开展攻关，牵头"6G 网络架构及关键技术"国家项目，提出"三层四面"网络服务框架和数据驱动分布自治的新型网络架构，发布了《6G 网络架构展望白皮书》、《6G 分布式组网技术白皮书》、《网络节能技术白皮书》等多部白皮书，共享自身最新科研成果，推进 6G 创新与产业进程。整体来看，6G 产业正呈现出技术快速收敛、产业链加速联动、星地融合场景拓展、AI 原生化趋势强烈的格局，预计 6G 将在 2029–2030 年进入首批商用，产业生态现已从基础理论研究迈向系统验证与预商用阶段的关键窗口期。

量子计算硬件与软件取得快速发展，正逐步融入云计算和云网融合产业生态。在过去十年中，量子计算硬件和软件取得了迅速发展。中国电信积极布局量子计算领域，推动量子技术与云服务的深度融合。

近期，中国电信在量子通信和量子计算领域取得重大突破，成功完成超百公里的空芯光纤量子与经典信号共纤传输实验，创下新纪录 [594]。该创新通过频谱协同分配机制，有效解决了传统实芯光纤中量子与经典信号相互干扰的难题，为量子通信与现有网络的融合部署提供了低成本、高效率的解决方案。此外，中国电信自主研发的超导量子计算机“天衍-287”已正式运行，具备“量子优越性”，在特定问题上的处理速度远超传统超级计算机。这些突破不仅推动了量子通信和量子计算的产业化进程，也为云计算和云网融合相关新兴技术产业注入了强劲创新动能，加速了数字经济和未来信息网络的升级发展。目前，量子计算已成为全球主要国家之间开展综合国力竞争，维护国家技术主权的关注焦点之一。近年来，我国政府相关部门出台了《关于推动未来产业创新发展的实施意见》《元宇宙产业创新发展三年行动计划（2023-2025年）》《新产业标准化领航工程实施方案（2023 - 2035 年）》等多项政策，支持量子计算、量子通信等量子技术的发展 [595, 596, 597]。

4.1.2 方向聚焦

在全球科技革命与产业深度变革加速推进的背景下，新兴技术产业正成为推动经济高质量发展和数字化转型的核心驱动力。我国高度重视新兴技术产业发展，2025 年发布的《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》明确提出，**要加快推进 6G、人工智能、量子计算、低空智能网、算力网络等关键新兴技术在各领域的应用，构建安全、可靠、高效的数字经济基础设施**。政策的持续引导与投入，为新兴技术产业的创新发展营造了良好的生态环境。中国电信作为通信领域的领军企业，积极响应国家战略，加快在新兴技术产业的布局与落地。2025 年，中国电信提出以“**云改数转智惠**”为核心的战略升级方案，重点推进低空智能感知、边缘算力、量子通信与 AI 赋能的数字化解决方案，加快推动工业互联网、智慧交通、智慧医疗、智慧金融等多行业数字化转型。在政策与技术双轮驱动下，中国电信持续扩大算力网络与云服务能力，构建覆盖 **“中心-区域-属地-边缘”** 的全场景云网基础设施体系，同时积极探索国际合作与海外业务拓展，提升全球竞争力。

2025 年新兴技术领域在新型计算架构、空天地一体化体系与智能网络等方向持续突破，技术演进与产业应用的深度耦合正加速未来基础设施体系的重构。新兴技术正围绕“下一代智能计算范式的演进”、“空天低空等新场景的广域协同”以及“智能系统的安全可信与风险防控”等主题形成发展共识。在此背景下，新一代计算体系与空天地低空的泛在智能应用逐步形成技术演进的主线，同时，伴随数据规模的激增与模型复杂度的提升，数据与 AI 的安全可信体系也成为支撑产业可持续发展的关键基础。基于此，下文将围绕新兴技术应用的发展脉络，以及数据与智能系统的安全与治理需求展开系统性梳理与深入分析。

4.2 热点方向九：新兴技术及应用

在数字经济全面跃升和新质生产力加速形成的关键阶段，全球信息基础设施正从“地面—空天—量子—低空”多域协同的体系化演进进入纵深。为支撑未来社会的泛在连接、实时智能和极限可靠，创新范式亟需突破传统计算与通信体系的边界：太空计算通过在轨分布式算力与智能协同，扩展了地外信息能力空间；量子计算以颠覆式算力优势，为复杂优化、材料设计与安全体系重构提供新路径；6G 作为下一代国家战略型信息底座，构建天地空海一体化的超宽带、超低时延和原生智能通信网络；而低空智能计算则在低空空域中实现感知、通信与边缘智能的深度融合，支撑无人机集群、低空交通与城市安全等新场景的规模化落地。四类技术相互促进、协同演进，共同构成未来信息基础设施的新核心能力，是推动产业变革和社会智能化升级的关键引擎。

4.2.1 智能时代下的新兴计算范式

随着数字基础设施不断向更高维度延伸，传统计算范式已难以满足未来信息系统在规模、异构性和实时性上的极限需求。面向深空探测、全球互联和极端环境智能化任务，太空计算正在成为构建新型算力体系的重要方向；同时，为解决传统计算在复杂优化、材料模拟与安全算法上的瓶颈，量子计算正以

表 4.1: 新兴技术应用领域热点

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
新兴计算	与地面计算设施相比，太空算力平台受到有限计算资源、未知太空环境、动态无线链路等因素的影响。现有研究重点关注星载算力平台中的边缘计算与 AI 推理两个方向，旨在提升太空计算效率。	ATC MobiCom INFOCOM RTSS TSC	• **星载边缘计算：** 北京邮电大学的相关团队探索 RUST 语言与 linux 内核的融合，提升星载操作系统的响应速度和实时性 [598]，并对计算设备在太空环境下的性能进行测量，分析辐射、温度等关键因素对卫星计算系统的性能影响 [599]。 • **太空 AI 推理：** 北京邮电大学的相关团队在卫星能量采集系统动态性和无线环境不确定性的影响下设计系统能耗优化算法，提升卫星运行寿命 [600, 601]。同时，在算力和存储受限的情况下，该团队设计星地协同的图像处理算法 [602]，用于提升系统整体运行效率。
6G	随着通信网络的进一步发展，智能化成为 6G 技术发展的新趋势。现有研究重点聚焦智能化网络管理和网络架构演两大方向，旨在全面提升网络服务能力。	SIGCOMM MobiCom NSDI INFOCOM	• **智能化网络管理：** 欧洲电信学院的相关团队利用大模型提升 6G 网络管理自动化 [603]，构建智能化运维体系。北京大学相关团队设计面向 6G 边缘智能的高效视频分析系统 [604]。 • **网络架构演进：** 谷歌和格拉纳达大学的相关团队针对 6G 驱动的网络架构演进展开研究，针对 O-RAN [605] 的架构设计和多路径广域传输 [606] 展开分析。来自米兰理工大学、埃因霍芬理工大学的相关团队则关注 6G 开放研究基础设施，包括 6G 测试平台、可复现实验系统、AI-native 网络架构 [607, 608, 609]。
低空智能计算	低空智能计算是面向无人机等低空载体，融合感知、通信与云边协同算力，以支持低时延、高可靠的实时环境理解与安全智能决策的综合计算体系，支撑多模态感知与复杂任务规划与执行，提升全面自主性。	CVPR/ICCV NeurIPS/ICML AAAI/IJCAI ACM MM TPAMI/TNNLS IJCV/JMLR	• **低空智能感知：** 云研究院联合中电信无人科技公司提出一种面向低空场景的多模态感知增强算法 [610]，提升了无人机的感知鲁棒性；山东大学研究团队 [611] 引入跨尺度特征融合和区域注意力，缓解了目标与背景遮挡带来的检测偏差。 • **低空大模型：** 香港科技大学研究团队提出了低空感知多模态基座大模型 RemoteCLIP[612]，利用对比学习，将视觉特征和与文本嵌入进行特征对齐。 • **无人机具身智能：** Skoltech 研究团队提出了 UAV-VLA [613]，结合视觉语言模型和 GPT，能够根据卫星图像和语言描述生成 UAV 飞行路径和动作计划。 • **空-地协同：** Western 大学团队 [614] 提出目标驱动的信任感知机制，在动态 UAV-UGV 系统中，将卸载建模为图匹配与任务聚类，实现高效可信协作分配。

突破性的并行性和指数级性能潜力推动计算模式革新。二者共同构成未来智能基础设施演进的关键驱动力，并将在算力组织方式、任务卸载模式及系统架构设计上带来深远影响。

4.2.1.1 太空计算

太空计算是指在卫星、探测器、空间站等航天平台上部署具备抗辐射、高可靠和低功耗特性的计算与智能处理能力，使其能够在太空环境中自主完成数据处理、任务规划、智能识别与协同决策。随着地轨星座规模化部署、星间链路的普及以及航天级 AI 芯片的发展，太空计算正从单星的独立计算演进为星座级的分布式智能体系，实现星上 AI 推理、在轨自治运行、跨星协同计算以及深空探测的高自主化，最终形成“天基边缘计算 + 太空云 + 深空智能”融合的发展方向 [615, 616]。目前，卫星计算领域重点关注星载边缘计算和太空 AI 推理与自主决策两个方向 [617, 618, 619]。星载边缘计算是指在卫星、探测器等航天器上部署具备抗辐射、高可靠、低功耗特性的计算硬件，并结合航天级操作系统、容器化与虚拟化等软件技术，使其能够在轨执行高性能的数据处理、AI 推理与自主决策，从而减少对地面站和地面算力网络的依赖、提升任务实时性并降低星地链路下行传输压力。其体系结构通常由三部分构成：一是硬件层，包括航天级处理器（如 RAD750、RAD5545、LEON3/4）、散射和辐射防护升级的商业 AI 芯片（如 NVIDIA Jetson AGX）、低功耗存储器等；二是系统层，涵盖 VxWorks、RTEMS 等航天级操作系统，星上容器化、虚拟化、OTA 在轨升级机制以及任务调度；三是应用层，例如深空通信、地面目标识别、地面目标追踪、灾害预警等典型应用。太空 AI 推理与自主决策，是指在空间环境中利用先进的人工智能技术，使航天器、探测器、卫星等能够在复杂的环境下实现自主感知、理解、推理、规划和决策。与地面系统相比，太空环境内存在通信延迟大、计算资源受限、任务环境极端等挑战，这对 AI 推理和自主决策提出了更高的要求 [599]。AI 推理主要涵盖知识表示、逻辑推理、概率推理、因果关系建模以及复杂事件推断等技术，能够让航天器从感知数据中提取环境状态信息、预测未来可能事件并生成合理的推论。目前，太空 AI 推理

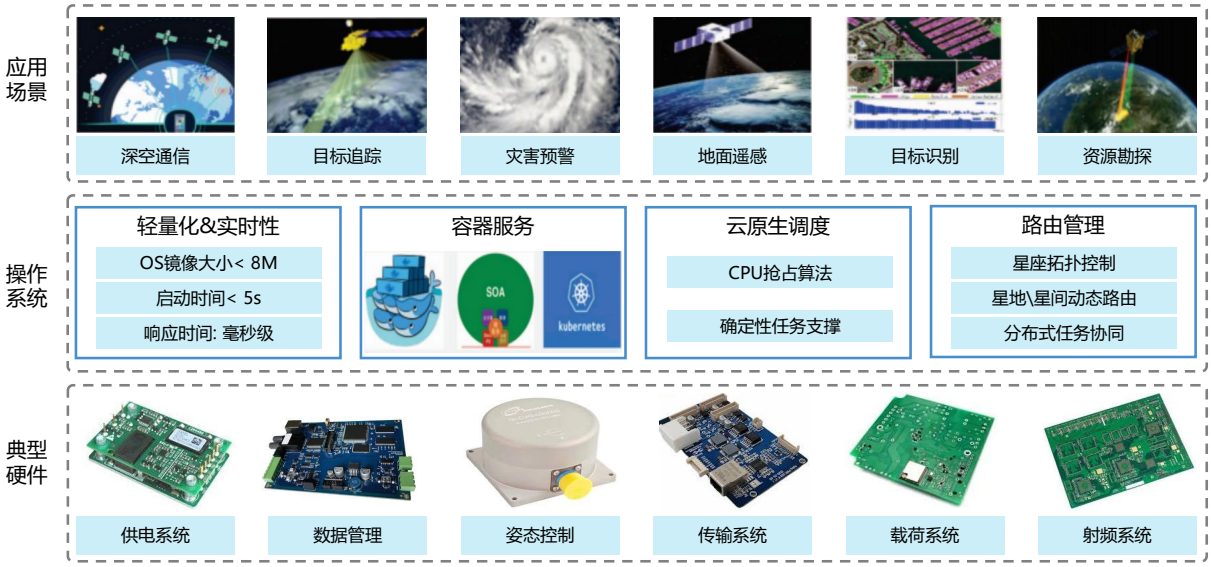


图 4.2: 太空计算系统组成结构 [616]

与自主决策正处于从基础理论研究向工程应用快速发展的阶段，通过智能化手段实现航天器在复杂太空环境中高效、安全、可靠的自主运行，为未来人类太空探索与长期自主运营奠定智能化基础。

4.2.1.2 量子计算

量子计算是指在量子芯片、量子处理器等新型计算平台上，利用量子力学中的叠加、纠缠和干涉等基本原理，赋予系统远超经典计算机的信息处理与智能推理能力。随着超导量子比特、离子阱、光子等多种硬件技术的突破，以及量子算法、量子编程语言和量子误差校正机制的不断发展，量子计算正从实验室原型逐步迈向实用化和产业化。当前，量子计算领域重点关注高保真度量子比特的制备与操控、量子纠错与容错机制、量子操作系统与编程工具的完善，以及面向化学模拟、优化、人工智能等方向的专用量子算法创新。其体系结构通常由三部分构成：一是硬件层，包括超导量子芯片、离子阱处理器、光子量子器件等，重点攻关量子比特的相干时间提升、门操作精度和大规模集成能力；二是软件层，涵盖量子操作系统、量子编程框架（如 Qiskit、Cirq、Paddle Quantum 等）、量子云平台及算法库，支持量子程序的设计、调试与优化；三是模型与应用层，涉及量子模拟、量子机器学习、量子优化等前沿领域，推动量子计算在化学材料、人工智能、金融分析等行业的实际落地。在实际应用方面，量子计算已经在多个行业展现显著价值。在实际应用方面，量子计算已在金融、医药、制造、能源等行业展现出显著价值。例如，汇丰银行与 IBM 合作验证了量子计算在企业债券算法交易中的优化能力 [620]，富士通则通过量子模拟器推动化学、工业优化和图像处理等落地 [621]。量子技术不断突破药物发现、材料设计、供应链管理、密码学与网络安全，成为数字化转型的重要驱动力。与此同时，量子计算的发展也对数据安全和加密体系提出了挑战。由于 Shor 算法等量子算法可能破解现有加密技术，NIST 等机构正加快抗量子加密标准制定，企业也在积极提升加密敏捷性。未来，量子计算与人工智能、云计算等技术的融合，将持续推动产业创新，但也要求企业提前布局安全策略，以应对量子时代的新机遇与挑战。

4.2.2 面向泛在互联的第六代移动通信系统

6G（第六代移动通信系统）通常被认为是继 5G 之后的下一代信息基础设施，将在智能化、泛在化、实时化和空间化通信方面实现根本性跃升。它不仅关注更高的峰值速率（Tbps 级）、更低的时延（亚毫秒级）和更大的设备连接规模，更重要的是提出了“空天地海一体化网络”“感知–通信–计算深度融合”“原生智能通信”“可信与安全全域覆盖”等新型能力，使通信从单纯的“数据传输”演进为“万物智能互联”的基础平台 [622, 623]。当前国际对 6G 的定义尚未完全统一，但普遍共识是：6G 将成为一个融合通信、感

知、计算、智能和安全的综合系统，使设备不仅能高速通信，还能实时感知环境、协同推理决策，并在跨场景、跨介质、跨维度的网络中自由移动。目前，各国已经开始面向 2030 年左右的 6G 商用展开系统性布局。美国发布“Next G Alliance”路线图，强调以人工智能原生网络、极高频段（如太赫兹）通信、智能表面等为重点方向；欧洲通过 Hexa-X、Hexa-X II 项目构建 6G 愿景框架，强调网络可持续性、可信性和泛在连接；中国则提出天地融合、智能原生、安全可信的整体方向，积极开展太赫兹器件、RIS 可重构智能表面、空天地海一体化网络、AI 原生无线系统等研发，并已完成多项重大验证实验 [624, 625, 626]。6G 网络将从以连接为中心向以智能与服务为中心转变，从地面通信系统扩展为空天地海全域覆盖系统，从人和物的连接演进为万事万物的实时协同。

4.2.2.1 AI-RAN

AI-RAN（AI-Native Radio Access Network）是面向 6G 的新一代无线网络体系结构，其核心思想是让 AI 从外部附加功能演进为网络的内生能力。通过在协议栈全层深度嵌入学习、推理、预测和自主优化能力，使无线网络具备持续学习、自我进化、自主决策与实时优化的能力 [627]。AI-RAN 强调以数据驱动、端–网协同、在线学习和智能调控为核心，使基站、边缘节点、终端能够共享状态信息，基于环境变化动态调整频谱、功率、波束、调度、切片和缓存策略，从而显著提升网络服务质量。AI-RAN 的发展不仅旨在解决当前蜂窝网络在超密集连接、极端动态性和高维复杂性下难以通过传统方法优化的问题，更希望构建一个“可感知、可学习、可预测、可闭环”的智能无线网，为 6G 的超低时延海量业务、空天地海融合通信、智能终端群协作提供基础支撑 [628, 629, 630]。从发展现状来看，AI-RAN 已成为 6G 国际研究的核心方向之一。学术界围绕大模型驱动的物理层智能、基于强化学习的无线资源调度、基于图神经网络的拓扑优化、基于联邦学习的隐私友好型网络智能等展开深入研究，出现了 AI-native PHY/MAC、智能波束管理、数据驱动链路自适应、智能覆盖预测等一系列关键成果。产业界方面，3GPP、O-RAN Alliance、ETSI ZSM、Next G Alliance、Hexa-X II 等国际标准组织陆续启动 AI-RAN 相关机制的预研 [631]，多家设备与芯片厂商（如华为、爱立信、诺基亚、高通、英伟达等）推出了 RAN 专用 AI 加速引擎、边缘推理芯片和针对物理层的大模型，推动 AI-RAN 从研究走向可部署验证阶段。

4.2.2.2 低功耗广域物联网

低功耗广域物联网是指面向大规模连接、远距离覆盖与超低功耗需求设计的新型物联网通信技术体系。能够以毫瓦级功耗、公里级覆盖、十年级电池寿命和低成本模组，实现数以亿计的传感器、计量设备、智慧城市终端等设备的长期稳定运行 [632, 633]。与传统蜂窝物联网或短距离无线技术相比，低功耗广域物联网的核心特征包括：低速率但覆盖范围极大、节点能耗极低、终端成本极低、网络部署灵活、能在地面、地下、室内、偏远地区等多类型场景下实现稳定连接。其技术路线主要分为两大类：一类是基于运营商网络的方案，如 NB-IoT 与 LTE-M，通过许可频段提供高可靠连接与全网覆盖；另一类是基于自组织的方案，如 LoRa 与 Sigfox 等，通过灵活和轻量级广域网络方式实现低成本部署 [634, 635, 636]。为了解决 LoRa 网络在高密度部署中出现的数据包冲突问题，云计算研究院联合清华大学相关团队，提出了基于无线信道特征的冲突解调方案，相关成果发表于计算机网络领域顶级会议 IEEE ICNP。论文提出的 CD-LoRa 框架包含三项关键创新：第一，通过构建精细的硬件与负载耦合相位模型来分离硬件和载荷调制导致的相位畸变，恢复真实信道特征；第二，设计线性相位拟合策略，在极低信噪比下稳定提取信道特征；第三，结合信道时变特性设计基于轨迹模型的动态聚类算法，使系统能够在动态场景下实现可靠解调。

4.2.3 面向低空经济的智能计算

随着无人机技术的迅速发展，低空智能计算已经成为推动低空经济和智能化应用的重要支柱。低空智能计算不仅仅指飞行器的感知、决策与执行能力，还涉及如何通过高效的计算和智能算法，提升飞行器在复杂低空环境中的自主性和任务执行效率。低空空域通常指海拔 1000 米以下的区域，这一高度范围内的飞行器面临着多样的挑战，例如气象变化、动态障碍物、复杂的地形地貌以及对实时性和精度的严

格要求 [637]。低空智能计算的核心目标是利用人工智能、大数据和高性能计算能力，推动低空飞行器在巡检、物流、农业、应急救援等多个领域的应用创新。为了实现这一目标，低空智能计算依赖于多项关键技术的协同作用，包括低空感知、低空大模型、无人机具身智能和空地协同等。

4.2.3.1 低空智能感知

低空智能感知技术通过集成多种传感器（如视觉、激光雷达、红外热成像等）帮助无人机在复杂环境中进行目标识别、跟踪和环境感知，确保飞行器能够在各种动态和复杂的低空场景下做出实时响应 [638]。在这些感知任务中，底层视觉技术作为基础层，发挥着至关重要的作用，主要负责从图像数据中恢复和增强环境信息，提升飞行器在低空环境中的感知能力。底层视觉涵盖了图像去雾、超分辨率重建、去模糊、低光增强等任务，通过多源数据融合技术，如将视觉和红外图像进行融合，有效解决低光照、雨雾、沙尘等环境中的图像质量问题，为目标检测和路径规划等应用提供可靠的数据支持 [639]。底层视觉与目标检测、跟踪等任务紧密结合，典型的深度学习算法如 YOLO [640] 和 Faster R-CNN [641] 被广泛应用于目标检测，能够高效实时地识别多类目标。随着低空经济加速发展，无人机在复杂环境中持续运行的能力面临更高要求，尤其在雾霾等恶劣天气条件下，视觉退化问题成为制约其部署与应用的关键技术瓶颈。对此，云计算研究院联合中电信无人科技公司，提出了一种面向无人机平台的雾浓度感知跨模态数据融合方法 HDCFN [610]，成功发表在多媒体领域顶级国际会议 ACM MM 2025。该方法利用红外模态在结构感知方面的优势，通过自适应融合机制动态增强可见光模态的视觉特征，有效提升了边缘设备在复杂气象环境中的场景感知能力，性能达到了国际领先水平。该方法不仅增强了端侧设备在实际场景中的感知鲁棒性，也为实现端云协同的智能感知奠定了技术基础。

4.2.3.2 低空大模型

低空大模型是用于处理低空环境数据的大规模预训练模型，能够跨模态理解低空环境中的复杂数据，从而提供更强的感知与决策能力。这些模型通常包含视觉大模型、语言大模型、遥感大模型以及多模态大模型，旨在提升无人机的跨域任务执行能力。低空大模型的核心任务包括视觉处理、语言理解、遥感分析和多模态推理等。这些任务通过深度学习框架优化，使得无人机能够在复杂环境下做出更精准的决策。针对低空环境，视觉大模型如 ViT [642] 和 MAE [643] 等被广泛应用，用于改进无人机的目标检测、视觉定位和图像分割等任务。在语言理解方面，BERT 系列和 GPT 系列模型可以优化无人机对环境描述和命令的理解，增强自主飞行能力。此外，遥感领域的模型如 SatMAE [644] 和 RemoteCLIP [645]，通过处理遥感数据帮助无人机实现更精准的环境感知和动态监测。多模态大模型，如 CLIP [646] 和 GeoCLIP [647]，则结合视觉与文本信息，通过对比学习和跨模态表示学习，提高了无人机对低空复杂环境的理解能力，特别是在低空巡检和监控等场景中的应用。

4.2.3.3 无人机具身智能

无人机具身智能是指无人机通过与物理环境的交互，具备感知、决策和执行的闭环能力，实现自主操作。其系统包括感知、决策和执行模块：感知模块集成视觉、激光雷达和惯性导航单元等传感器，实时获取环境信息；决策模块基于感知数据进行路径规划、任务分配和避障；执行模块根据决策指令控制飞行动作。具身智能面临的主要挑战是应对动态的六自由度环境、飞行姿态控制和气动效应，尤其在低空环境中，还需考虑气象变化和突发障碍。动态路径规划与避障技术，如 A* [648] 和 EGO-Planner [649]，通过算法优化保证飞行安全，深度强化学习算法如 FASTER [650] 和 TOOR-MPCC [651, 652] 能够应对动态障碍、优化飞行路径。多机协同任务分配通过 DRL 算法（如 MADER [653] 和 DREAM [654]）协调任务执行，提高效率。视觉-语言-行动 VLA（Vision-Language-Action）模型结合视觉、语言和行动生成，使无人机在复杂任务中做出高效决策。强化学习同样用于飞行路径和任务策略的动态调整，特别在避障中帮助无人机应对复杂环境，减少碰撞风险。具身智能技术通过感知、决策和执行模块的协同推动无人机在低空环境中的自主飞行，广泛应用于智慧城市、灾害救援和环境监控等领域。

4.2.3.4 空-地协同

空-地协同技术是在任务执行过程中，由空中平台（如无人机集群）与地面平台（如地面机器人、无人车、固定监测站等）形成异构多智能体系统，通过信息共享和协同控制，以显著提升任务执行的效率、鲁棒性和空间覆盖能力 [655]。系统架构通常包括空中平台、地面平台和通信与计算支撑层：空中平台负责大范围、全局性的感知与监控，可快速获取宏观态势信息；地面平台则承担高精度操作、近距离交互与环境干预等精细任务；通信与计算层依托车联网/专网、边缘计算与云端服务，实现多源异构数据的实时融合与下行控制指令的可靠分发 [656]。空地协同的典型任务可分为协同感知、协同态势理解与决策、协同导航与路径规划以及协同执行与反馈闭环等。多智能体强化学习、博弈论驱动的任务分配、分布式规划与一致性控制算法是常见的空地协同方法体系 [657]，可在存在通信时延、局部观测与资源受限的条件下优化任务分解、角色分工与执行调度，并支持在灾害救援、低空物流、巡检安防等复杂场景中的在线重规划和自适应协作。

4.3 热点方向十：数据与 AI 的安全

在新兴技术快速落地与跨场景融合的背景下，数据与 AI 的安全已经从单纯的风险防范转向支撑业务可信、生态协同与社会可持续的重要基础能力。其中，数据安全成为智能系统可信运行的前提，在跨云、跨域、跨设备场景中，数据采集、存储、交换与使用过程面临隐私泄露、数据滥用与越权共享等风险。而 AI 安全则成为智能决策可靠性的核心保障，模型受到对抗样本、越狱攻击、后门攻击等威胁，可能导致误判、错误执行甚至引发现实风险。**对数据与 AI 的安全体系化研究，是未来智能应用真正可信落地的根本支撑**。具体而言，后续两个小节将分别探讨面向数据隐私、面向 AI 系统的攻击方法与防御策略等关键问题。为更好地理解相关技术路径与研究进展，表4.2重点遴选了部分具有代表性的关键研究成果。

4.3.1 面向数据隐私的安全威胁与保护机制

随着云计算的普及，数据与模型的部署范式正迅速由本地计算转向云端托管。在典型的云 AI 服务框架中，用户数据被上传至云侧，由集中托管的模型完成推理处理，并通过 API 或其他形式对外提供能力。在这一架构下，业务日志、用户交互记录、领域知识库等高敏感数据长期驻留并流转于云端，使隐私保护成为支撑云服务可信性的关键前提。云环境的资源集中性与跨租户共享特征，使攻击者一旦突破边界便可能访问大规模敏感数据；同时，大模型的推理行为具有强依赖性和可诱导性，若缺乏有效隔离与约束，则可能泄漏训练样本、交互上下文乃至系统提示信息，显著放大隐私风险。当前围绕云端智能服务的数据安全，攻击手段已覆盖数据从存储、传输到使用的完整生命周期。围绕这些风险面，目前在研究与工业实践中形成了三类主要的技术防护路线，如图4.3所示。下面将讨论各类攻击与防护方案的适用性与技术特点。

4.3.1.1 数据生命周期内的安全威胁

数据存储中的安全威胁主要源于静态数据长期存留与权限配置复杂化带来的攻击面。其核心风险在于：一旦存储资源的控制权限被攻破，攻击者即可直接读取或篡改其中的敏感内容。主要可分为未授权访问 [677]、身份滥用 [678] 和权限配置错误 [679] 三类。未授权访问源于数据库、对象存储或镜像仓库的认证缺失或接口暴露，使攻击者无需凭证即可直接读取敏感数据。身份滥用指攻击者通过已泄露的密钥或未轮转的长期令牌获取合法身份，在无需攻击存储系统的情况下完成数据窃取或篡改。权限配置错误则来自云环境的自动化与多级权限继承，导致新建资源意外获得访问快照、备份或日志等敏感静态数据的权限，从而形成隐蔽且难以察觉的暴露面。

数据传输中的安全威胁主要源于跨服务通信链路复杂化与网络路径可被中断与劫持带来的攻击面。其核心风险在于：一旦传输路径被监听、篡改或重定向，攻击者即可获取明文数据，或注入恶意载荷操

表 4.2: 数据与 AI 的安全研究领域热点

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
面向数据隐私的安全威胁与防护机制	随着云计算与大模型服务的高速演进，数据在跨场景中的高频流动显著增加了隐私泄漏风险。当前研究正以多角度攻击建模为驱动，以隐私增强计算为核心，构建面向智能服务的数据安全体系。	S&P Security NDSS CCS NeurIPS ICML TIFS TDSC	• **面向数据隐私的推理攻击：**字节跳动的团队提出基于生成式模型的上下文依赖特性，利用多轮诱导触发模型隐式记忆，使其复述其他用户历史对话的上下文泄漏攻击 [658]。洛桑联邦理工的团队提出通过参考样本校准决策边界，实现对成员样本记忆的成员推理攻击 [659] 。 • **具有理论保证的防御：**阿里巴巴的团队基于安全多方计算支持多机构在不共享原始数据的前提下联合完成分析任务 [660]。Google 的团队提出向参数梯度注入噪声，使结果不依赖个体的差分隐私机制 [661]。多伦多大学的团队提出移除特定样本影响，使模型在统计上恢复至未见该数据的遗忘学习方法 [662]。 • **基于数据最小化暴露的防御：**北京邮电大学的团队提出基于拆分学习，仅上传中间激活以弱化输入可逆性 [663]。Meta 的团队提出使用生成数据替代真实敏感样本用于训练或分析，以减少原始隐私数据的直接暴露 [664, 665]。
面向 AI 系统的攻击方法与防御策略	随着 AI 系统深入关键领域，其攻击面逐步扩大，对抗样本攻击、越狱攻击和后门攻击成主要威胁，相应的防御机制以及安全治理方法也在不断推陈出新。	NeurIPS ICML ICLR S&P CVPR ACL MM WWW AAAI	• **对抗样本攻击与防御：**哥伦比亚大学的团队通过词嵌入空间生成人类难以辨别的对抗文本，使 LLM 生成文本检测器的性能暴跌 [666]。NVIDIA 的团队提出对抗攻击先发制人防御机制，确保在给定强度范围内任何针对当前输入的攻击都会失败 [667]。 • **越狱攻击与防御:** 清华大学的团队通过排版和扩散模型将违禁内容转换为图像，从而绕过安全对齐机制 [668]。香港中文大学联合阿里巴巴的团队发现视觉-语言大模型在处理不安全提示时会表现出独特的激活模式，这些模式可用于检测和缓解对抗性输入，而无需进行大量的微调 [669]。 • **后门攻击与防御：**清华大学联合腾讯的团队提出的攻击方法针对 token 化层，植入一个切换 token，使得模型可以在良性和恶意行为之间动态切换 [670]。香港科技大学的团队通过模拟触发并定位后门行为，再结合多轮消除与校准步骤有针对性地撤销大语言模型中的后门触发能力 [671]。 • **AI 安全治理:**UCSB 等的工作聚焦人工智能生成内容的检测问题，即区分人类和大模型生成的内容（文本和图像），以推动 AI 安全治理 [672, 673, 674, 675, 676]。

纵业务逻辑。主要可分为未加密或弱加密通信、传输路径劫持和会话凭证截获三类。未加密或弱加密通信指服务之间、用户与云平台之间或微服务内部存在明文传输，使攻击者可在链路层直接窃听、重放或篡改数据 [680]。传输路径劫持来自代理链路被控制或接口调用链遭篡改，使数据被重定向至攻击者控制的节点 [681]。攻击者不仅可获得传输内容，还可注入恶意响应，进一步移动至控制平面或其他服务。会话凭证截获是指攻击者在传输路径中窃取 Token、短期会话密钥，从而冒充合法用户或服务执行操作。由于云环境依赖大量接口调用与自动化认证流程，这类威胁可迅速放大，导致数据大规模泄露。

数据使用中的安全威胁主要源于模型行为对其训练数据或上下文数据具有依赖性，使攻击者能够通过查询模型的输出来推断敏感信息。其核心风险在于：攻击者无需直接访问数据，只需观察模型输出来推断训练样本、敏感属性或上下文内容。主要可分为训练数据推理攻击与上下文数据窃取攻击两类。训练数据推理攻击旨在仅观察模型外部行为的情况下推断训练数据的成员关系、敏感信息，或从整体上推测训练数据的统计特征。其中，成员推理攻击 MIA（Membership Inference Attack）通过分析输入对模型响应的敏感性差异判断某样本是否属于训练集。早期方法依赖小规模模型的对训练数据的过拟合特性 [682]，但随着预训练大模型的泛化性加强，近期研究开始关注通过生成参考样本校正边界 [659]、在模型训练环节注入轻量级扰动以放大成员记忆 [683]，或从用户群体的生成风格推理 [684]。属性推断攻击 AIA（Attribute Inference Attack）通过分析模型行为与训练特征之间的相关性，推断本不公开的个体属性或群体统计特征 [685, 686]。最新研究从放大属性与输出之间的微弱关联实现更高重建精度 [687]，以及在联邦学习等分布式训练场景中从局部梯度、残差或参数更新中反推出个体的敏感属性 [688]。

上下文数据窃取攻击利用大模型推理阶段对历史对话、系统提示、知识库内容的依赖，通过诱导或操控生成式模型的响应来获取其他用户的对话内容或知识库片段。其中，上下文泄漏攻击通过多轮次构造诱导性问题、反事实提示或伪装任务，使其复述其他用户的历史对话 [689]，或通过设计示例检测模型

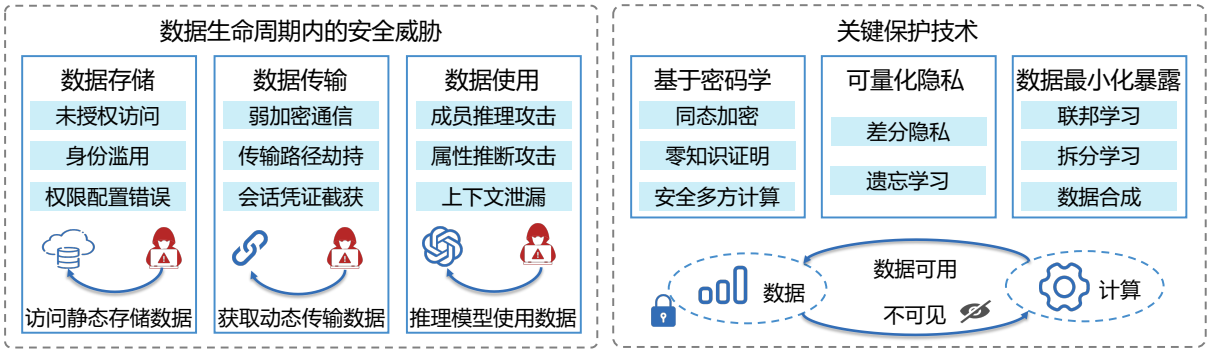


图 4.3: 面向数据隐私的安全威胁与保护技术

是否保留提示中的敏感片段 [690]。研究还表明上下文规模越大，泄漏概率越高 [658]。面向 RAG 的知识库泄露攻击则通过操控检索路径，使模型在回答中直接复述召回文档中的敏感内容 [691]，或通过遮蔽并测试补全行为判断目标文档是否存在于知识库中，从而暴露文档级数据隐私 [692]。

4.3.1.2 数据隐私保护技术

基于密码学的防护方法在协议层切断了原始数据与云端计算环境的直接接触，使任何计算模型都只能在密文或隐藏表示上运行，从而从根本上避免训练与推理阶段的数据泄漏。这一类方法的核心优势在于与模型结构、推理框架、任务类型无关，提供可证明的隐私保护。当前研究与实践主要包含两条技术路线：全同态加密和多方安全计算。全同态加密 FHE（Fully Homomorphic Encryption）提供密文域可计算能力，使云端系统无需访问明文，即可执行线性模型、部分复杂模型的推理。代表性进展有：HECO 展示了首个可将高级程序自动转换为高效 FHE 运行形式的编译器，使传统机器学习算法、神经网络算子都能在密文上执行 [693]；AThFHE 将任意门限 FHE 构造为一种近似秘密共享任务，使多参与方在密文域协作推理或训练结构化模型时的通信和计算大幅降低 [694]。安全多方计算 MPC（Secure Multi-Party Computation）适用于多个组织需要共同使用模型，但彼此不暴露数据的场景，是跨机构合作服务的核心技术路线。代表性进展有：Squirrel 优化梯度聚合协议，适用于金融等传统服务场景 [660]。TVA 支持滑动窗口、会话窗口等复杂时序分析，使云侧服务可在不暴露时间戳、行为序列的前提下执行分析任务 [695]。

基于可量化隐私损失上界的防护方法在算法层面约束数据对模型的影响，使训练与推理结果在统计层面不强依赖于某一个体，从而在不改变模型结构的前提下提供可量化的隐私保证。当前研究与实践主要包含两种机制：差分隐私与遗忘学习。差分隐私通过在训练梯度或模型输出中加入精心控制的噪声，使单个样本的加入或删除不会显著改变模型行为，从而提供可量化的隐私保证 [696]。在训练阶段，通过在梯度上的噪声注入抑制模型对个体样本的记忆 [661]，是当前深度学习中最常用的差分隐私机制，可用于保护语言模型 [697]、视觉模型 [698] 等。在数据分析与统计阶段，差分隐私也常用于云端的统计分析与数据服务接口的计数保护，在无需共享原始数据的情况下生成均值 [699]、三角计数 [700] 等统计结果。对于图等结构化数据，可通过节点级 [701, 702, 703, 704]、边级 [705, 706] 和图级 [707] 等不同隐私定义约束在模型发布或图统计中的泄露风险，分别对应保护节点、边关系或整体图信息。云计算研究院联合上海交通大学团队面向在统一隐私预算下的过度保护与精度损失问题，提出了节点重要性分级自适应图神经网络 NAP-GNN。该方法通过拓扑感知的节点重要性估计与分级拉普拉斯扰动，实现隐私预算的自适应分配，并结合自适应残差聚合机制削弱噪声在多跳消息传递中的积累。实现结果表明，在满足节点级差分隐私约束与抵抗成员推断攻击的前提下，NAP-GNN 在给定隐私预算下显著提升了模型精度与鲁棒性 [704]。

遗忘学习通过在不重新训练整个模型的前提下，从已训练模型中高效移除特定样本、特征或标签的影响，使模型在统计意义上与未使用该数据训练的状态保持一致，是满足被遗忘权与云端模型合规性的重要技术路径。现有方法分为精确遗忘和近似遗忘两类。精确遗忘通过训练流程设计或结构化重训练机

制，完全消除目标数据点对模型的影响，使遗忘后的模型在理论上与未使用该数据训练时等价 [708]。代表性方法通过对训练集进行分片与切片，使每个样本只影响部分子模型 [662]。当需要遗忘某个数据点时，只需重训练其所在的少量分片并重新聚合，而不需对整个模型进行重训练，大幅降低了计算成本。近似遗忘在可接受误差范围内削弱目标数据对模型的影响，以在效率与性能之间取得平衡。代表性方法包括基于影响函数的调整 [709]、重优化 [710] 与梯度更新 [711]。

基于数据最小化暴露的防护方法改变数据在云端训练与推理流程中的组织方式，减少原始数据的集中暴露或以替代性数据结构替换真实敏感样本，从结构层面降低攻击面。当前研究与实践主要包含三条路线：（1）联邦学习通过将训练数据分散到不同数据持有方，使原始数据在本地完成前向计算和梯度更新，云端仅负责聚合模型参数或梯度，从而避免敏感数据在云侧直接暴露 [712]。为防止遭到梯度反演攻击等隐私威胁，实际部署中常结合安全聚合、差分隐私噪声注入和局部剪裁等机制增强鲁棒性 [713]。（2）拆分学习将模型按层拆分为本地段与云端段，使前向传播的激活由本地计算后再传输到云端，敏感原始输入无需上传 [714]。在部署中，拆分学习常结合加密、扰动或降维操作，进一步降低激活泄露原始输入的风险 [663]。（3）隐私合成数据通过生成模型学习数据的总体分布，而不保留可识别的个体样本，用以替代真实数据参与模型训练或分析任务 [664]。对于无法将原始数据直接暴露给云服务的场景，合成数据能在保持整体统计结构的同时隐藏敏感个体。典型应用包括生成匿名图数据用于网络分析 [665]、合成日志用于安全监测、生成仿真用户行为用于推荐系统建模等。

4.3.2 面向 AI 系统的攻击方法与防御策略

在云计算成为 AI 系统核心部署环境的背景下，AI 安全性正加速演进为云原生架构中不可或缺的生产级基础能力。随着大模型及智能系统在政府、金融、交通等关键和新兴领域依托云平台快速落地与规模化服务，其安全性直接关系到云上业务系统的整体可信与持续稳定。近年来，大量研究表明，对抗样本攻击、越狱攻击、后门攻击等典型威胁，已构成 AI 系统可控性、可靠性与可信决策的现实挑战。因此，如图4.4所示，本节围绕上述三类主流攻击及对应防御机制，结合 AI 安全治理的最新研究，系统梳理相关进展，为构建面向未来的云上 AI 安全治理体系提供可落地的技术参考。

4.3.2.1 对抗样本攻击与防御

对抗样本攻击是指在输入上施加微小但特定的扰动，导致人工智能模型产生错误输出的攻击 [715]。自发现以来，这类攻击已覆盖视觉、语音、文本与多模态系统，能在不改变人类感知的前提下，严重削弱模型的可靠性与安全性。在视觉领域，对抗样本攻击研究焦点从早期的一次性扰动 [716] 转向基于优化的方法 [717, 718] 以提升对抗样本的可迁移性，以及使用生成模型的方法，例如通过 GAN、扩散模型等模型生成对抗样本 [719, 720]。在文本领域，早期研究针对文本数据的特性直接通过插入、删除等文本编辑方式生成对抗样本 [721]，这类方法简单直接，但容易通过拼写检查等方式识别或纠正。近年来针对文本的对抗攻击研究呈现出攻击场景更现实、目标更多元、策略更精巧的趋势。通过操作词嵌入空间生成人类难以辨别的对抗文本，可以成功使 LLM 生成文本检测器的性能暴跌 [666]。随着多模态数据的爆发式增长，当前的前沿已进入多模态领域，攻击不再局限于单一模态，而是通过跨模态协同的方式展开，例如开发能够同时扰动图像与文本 [722, 723]、或生成通用的与输入无关的通用对抗扰动或补丁，可以叠加在任意图像上对多模态模型产生误导，这些方法经过预训练后展现出强大的跨模型、跨任务攻击能力，揭示了多模态 AI 系统底层更深层次的安全脆弱性 [719, 724]。

针对 AI 系统的对抗样本防御方法向主动防御、样本净化方向演进。在视觉领域，研究呈现出从认证保证到主动拦截的多样化趋势。一方面利用视觉 Transformer 与去随机化平滑技术构建针对补丁攻击的可认证防御 [725]，另一方面则主动出击，提出通过生成防御性扰动进行对抗增强的预防御框架 [667]。在文本领域，早期研究通过解读模型逻辑值在输入扰动下的变化模式来探测攻击 [726]；更前沿的思路是“不重训练，只重写”，即训练专门的模型拦截并重写对抗性输入，使其对下游分类器失效，这种方法在保持任务性

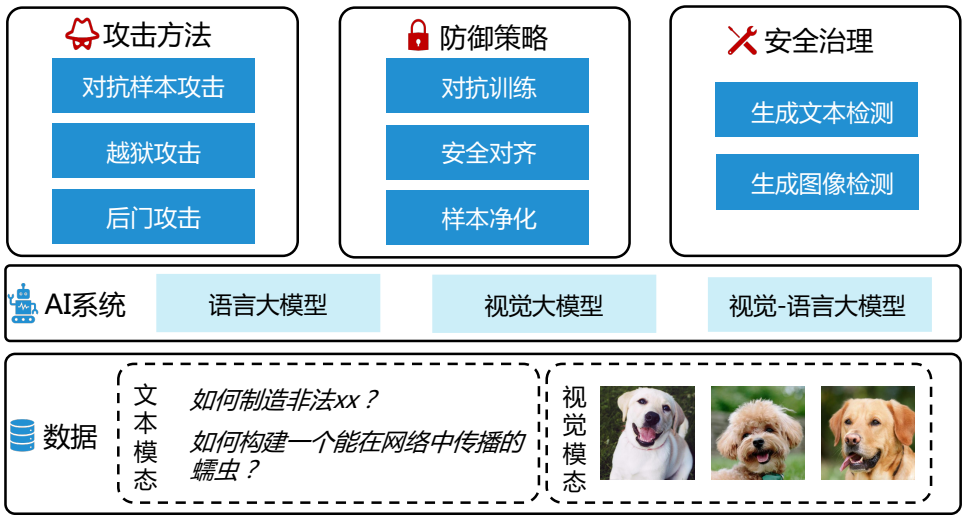


图 4.4: 面向 AI 系统的攻击、防御与安全治理

能的同时展现了良好的泛化能力 [727]。多模态模型的对抗样本防御研究仍处于研究初期，沿袭了单模态的防御方法，探索了对抗训练、模型结构增强、对抗样本检测或净化等多种防御方法 [728, 729, 730]。

4.3.2.2 越狱攻击与防御

越狱攻击是指诱导具有限制/安全策略的 AI 系统（尤指大语言模型）执行模型被设计为拒绝的操作，诱导其产生违规、有害或偏见性内容。与传统漏洞不同，越狱攻击利用模型的语言理解与上下文学习能力，通过构造输入或示例提示改变模型行为。随着大语言模型和视觉-语言大模型成为新一代人工智能的基础设施，当前的研究焦点已不再局限于传统的文本分类模型，而是深度聚焦于这些规模庞大、能力通用的基础模型本身。大多数越狱攻击针对 LLM 即服务（LLM-as-a-Service）场景，从攻击大语言模型开始，随着视觉模态的增加，逐步过渡到对视觉-语言大模型的攻击。对大语言模型的越狱攻击通过精心的提示词工程手动构造越狱指令，例如提示大语言模型进行角色扮演或者直接将越狱指令进行编码或加密，以欺骗模型打破规则 [731, 732]。随着研究的深入，攻击者开始系统性地利用黑盒查询或模型白盒知识，自动生成难以被预先防御的对抗性提示。一类研究是基于贪心搜索的梯度引导攻击，通过不断替换输入 tokens 并评估其诱导模型输出有害内容的效果，以优化出高成功率的对抗后缀 [733, 734]。另一类研究尝试使用一个参数化的生成模型迭代地对每个提示进行定制化修改 [735, 736]。视觉-语言大模型通过预训练的图像编码器和对齐模块扩展了大语言模型，也为越狱攻击提供了额外的途径。其中，白盒越狱攻击利用梯度信息扰动输入图像或文本，以诱导模型产生特定类型的有害输出 [737]。黑盒越狱攻击不需要直接访问目标模型的内部参数，而是利用外部漏洞进行越狱攻击。例如通过开源图像编码器制作对抗图像，再加上干净的文本提示来破坏模型的安全对齐 [738]；手动设计将恶意文本排版转为图像，通过视觉通道使得模型理解恶意指令 [668]；结合扩散模型自主生成恶意的图像-文本对 [739]。

对越狱攻击的检测和防御研究分为两条防线，第一道防线是识别有害输入或输出以进行拒绝或净化，第二道防线是通过安全对齐提升模型内在鲁棒性。在第一道防线中，对大语言模型的输入防御可以通过输入重述实现，例如使用语义相似的描述等方式来过滤掉提示的恶意意图 [740]。输出防御则通过大模型内部的拒绝损失函数来识别不安全输出 [741]。视觉-语言大模型的输入或输出防御需要增加对视觉模态的越狱攻击检测，有研究将检测过程分成了两个阶段，先验证输入文本的不安全性，再防范基于图像的攻击 [742]。在第二道防线中，安全对齐防御是通过微调预训练模型，增强其内部安全能力，经典的对齐算法包括监督微调 SFT（Supervised Fine-Tuning）[743]、人类反馈强化学习 RLHF（Reinforcement Learning from Human Feedback）[744]、直接偏好优化 DPO（Direct Preference Optimization）[745]。针对视觉-文本大模型，研究者将不安全图像转换为文本描述，以激活这类大模型内在的语言模型的安全对齐 [746]。或

者利用模型在处理不安全提示时会表现出独特的激活模式来检测和缓解对抗性输入 [669]。

4.3.2.3 后门攻击与防御

后门攻击是指在训练数据中植入少量带触发器的样本，使得模型在遇到触发器时按攻击者意图输出错误，而在常规输入上保持良好性能。后门攻击的恶意行为难以通过常规评估被发现，且可能通过开源模型发布在供应链中广泛传播。后门攻击的研究从视觉、文本等单模态模型，显著扩散到多模态模型，展现出更为复杂的威胁维度。在视觉领域，现有的后门攻击大多数是基于数据投毒的。图像块级攻击主要通过在图像块级别植入触发器，仅需要少量数据即可将模型的焦点从分类相关的图像块重定向到对抗触发器 [747]。Token 级攻击针对模型的 Token 层，攻击者在模型的学习过程中植入一个开关 Token，通过控制这个开关 Token，可以使模型在良性和恶意行为之间动态切换 [670]。在文本领域，后门攻击的关键步骤是触发器注入，通常通过数据投毒、训练操纵或者参数修改将后门触发器注入目标模型。数据投毒使用预先设计的后门触发器毒化一小部分训练数据，然后在污染数据集上训练一个后门模型 [748]。训练操纵将后门注入视为多任务学习，通过控制梯度大小和方向，巧妙地改变模型优化过程来注入后门 [749]。参数修改通过修改模型一小部分模型参数嵌入后门 [750]。在多模态领域，后门攻击是通过嵌入视觉或文本输入中的触发器完成攻击的。有研究为自动驾驶引入一种物理后门，通过生成带有恶意行为的后门训练样本（例如红气球）触发不安全行动 [751]。

针对后门攻击的防御旨在识别并打破触发器模式与目标类别之间的关联，同时保持模型准确性。在视觉领域，代表性的防御策略是图像块处理和图像阻断。图像块处理通过破坏图像块的完整性（例如随机丢弃或乱序）来对抗基于图像块的后门攻击 [752]。图像阻断是利用可解释性机制（例如注意力图）来定位和后门触发器 [753]。在文本领域，防御手段包含后门检测和后门移除。后门检测侧重于检测可能触发后门行为的输入，例如利用梯度和自注意力分数识别导致异常预测的关键 token [754]。后门移除通常通过修改模型的参数来覆盖或者抑制后门映射，例如在干净数据集上微调模型 [671]。视觉和文本领域的后门防御策略同样适用于多模态领域，但其复杂性和挑战性显著增加。多模态防御不仅要分别检测各模态中的异常触发器，更需要深入理解模态间的协同攻击机制。

4.3.2.4 AI 安全治理

国内外产业界、学术界已经涌现许多尝试，聚焦于人工智能生成内容的溯源检测问题，即区分人类和大模型生成的内容（文本和图像），以推动 AI 安全治理 [672, 674, 675]。AI 生成文本的溯源方法可以分成三类。第一类是基于微调范式的方法，通过微调 RoBERTa [755] 等模型来学习不同模型生成文本的语义特征分布差异。中国电信云计算研究院针对大模型生成文本检测的研究被自然语言处理顶级国际会议 EMNLP 2025 主会接收并发表 [676]。该研究发现，尽管大模型在模仿人类写作风格方面表现突出，但仍存在事实幻觉问题，表现为文本中的实体关系与现实世界不一致。因此，研究提出基于微调范式的一个事实感知模型，通过比较文本抽取到实体-关系图与事实知识库中的实体-关系图差异来识别机器生成文本。为了全面分析上下文信息，研究采用带门控单元的分层特征提取方式，自适应融合实体、句子、以及文档级别的多粒度特征，显著提升 AI 生成文本的识别准确率。第二类是基于风格特征的方法，需要提取文本的词法、句法和结构特征训练分类器来实现生成文本检测 [?]。第三类是基于概率特征的方法，利用语言模型倾向于生成高概率的单词或字符这一现象进行概率特征的计算 [673]。

生成图像的溯源方法大体可分为三类。第一类是基于模型水印的方法，该方法需要在深度伪造模型训练或部署阶段预先嵌入水印，使水印信息能够随模型生成过程传递到最终图像中，从而在溯源时通过检测水印实现来源判定 [674]。该方法的适用范围仅限于已加水印的模型，且水印的引入在一定程度上可能影响生成图像的质量。第二类是基于模型反演的方法，其利用模型末层参数信息对生成过程进行反演，将溯源问题转化为一个凸优化问题 [?]。这种方法充分利用了模型参数本身的信息，因此在溯源精度上具有优势，但通常依赖于模型参数已知的白盒假设，因而在实际应用中面临较大的计算开销和效率限制。第三类是基于模型指纹的方法，该方法通过挖掘生成内容中存在的细微统计特征或隐含痕迹来识别其来

源模型。这些特征能够反映模型的结构设计和参数分布等内在属性，为不同生成模型提供类似“身份标识”的区分依据 [756]。

4.4 展望与建议

本节构建了新兴领域关键技术的 Gartner 成熟度曲线，如图 4.5 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，新兴技术的演进与发展建议主要聚焦四个方面：加快构建可信数据体系、推动 AI 安全体系化升级、推进云网算一体化融合发展、布局低空智能计算基础设施。

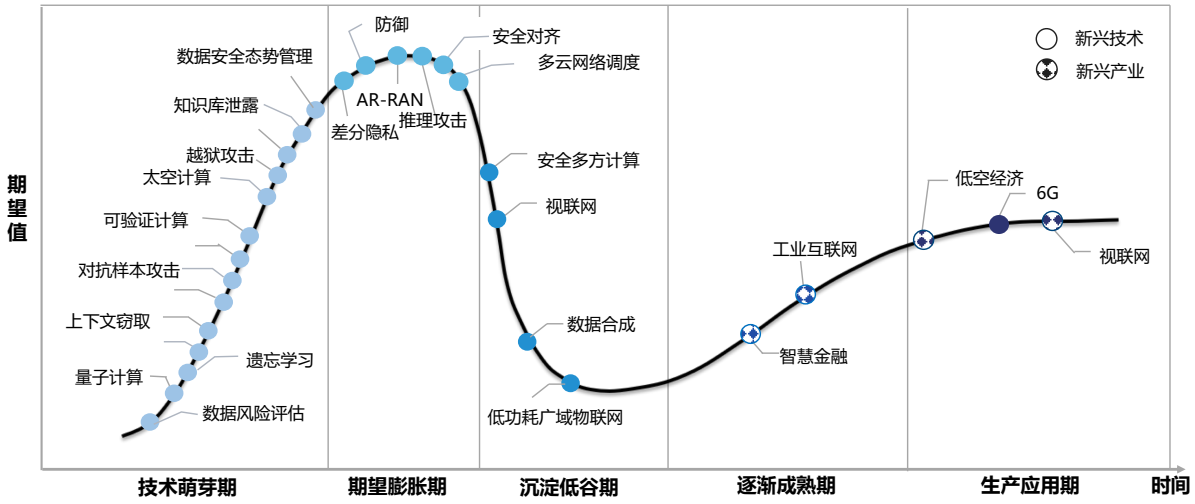


图 4.5: 新兴技术研究图谱技术成熟度曲线 2025

4.4.1 新兴技术的未来研究方向和关键技术展望

随着数据跨域流动加速、隐私保护需求提升，数据安全治理正向以隐私保护与可验证操作为核心的可信数据体系演进。数据在跨区域、跨环境的流动规模持续扩大，访问主体、访问链路与治理边界愈加复杂，传统以封闭边界和静态策略为核心的数据安全治理模式已难以满足智能化业务对可信数据环境的要求。未来，隐私增强的跨域访问控制技术将成为基础能力，从传统规则式授权升级为基于语义理解、行为建模与情境感知的动态零信任决策，实现对多主体、多链路访问行为的实时验证与细粒度管控。其次，可验证的数据要素流通技术将成为核心支柱，通过隐私计算、可验证计算与版权保护等机制，确保数据在跨域共享、模型训练与模型推理环节实现可用不可见、可控可审计。最后，智能化的数据治理技术体系将成为能力上限，由大模型驱动的风险推理、策略生成与自动化处置将使数据安全从静态规则走向自适应演进，支撑新兴技术环境下的可信数据空间构建。

AI 安全将从单一防护向体系化、全生命周期和智能化演进。随着 AI 与云计算、6G、低空经济、机器人、AI Agent、量子计算等融合发展，威胁模型从数据与模型攻防扩展到跨环境、跨设备、跨智能体的复合式安全挑战。未来 AI 安全研究的重点，将围绕可信计算环境下的模型隐私保护、硬件级隔离与加密推理、跨层多模态威胁检测、AI 自动攻防博弈与免疫自适应防御、可追溯与可问责的模型决策解释、以及 AI 原生的安全治理体系展开。同时，状态感知与风险量化将逐渐成为关键技术，使 AI 系统能够实时评估自身安全态势并动态调整能力边界；多智能体协同将推动从“安全的 AI”走向“用 AI 构建安全”。最终，AI 安全将形成集可信、可审计、可防御、可自治、可治理于一体的新型全栈安全体系，支撑未来智能社会的安全与可持续发展。

建设超高速率、超低时延、超大连接的 6G 通信网，并与云计算和算力网络深度融合。为了提供更强的接入能力，6G 将聚焦于太赫兹通信与可见光通信技术的高频谱利用、超大规模 MIMO 与分布式天线系统的空间复用能力提升、网络切片与端到端服务质量保障机制的智能优化，以及空天地海一体化网络的协同调度与自组织管理。此外，6G 还将重点探索人工智能驱动的网络自适应与自优化、安全增强技术，以及能效优化和绿色通信技术，以应对海量物联网设备、智能终端以及自动化工业、虚拟现实/增强现实等新兴应用对网络的高带宽、低延迟和高可靠性需求，实现通信网络从信息传输向智能感知、计算与决策的全面升级。

随着低空空域逐步开放和无人系统规模化部署，低空智能计算将成为支撑低空经济安全运行与业务创新的关键底座。未来研究将围绕“云–边–端–空”一体化算力体系展开，重点攻关在受限功耗与不稳定链路条件下的算力编排、任务卸载与协同容错机制，实现感知、通信与计算的深度融合。其次，需要面向复杂低空场景构建多模态环境理解与预测模型，结合具身智能和多智能体强化学习，支撑无人机集群、空–地协同平台在动态空域中的自主规划、避障与协同决策。此外，低空智能计算还需引入可信计算、数字孪生与形式化验证等技术，构建从仿真测试、在线监测到事后追溯的全流程安全与合规框架，推动低空交通管理、物流配送、巡检安防等典型场景形成可复制、可推广的技术标准和产业体系。

4.4.2 新兴技术的发展建议

建设可信数据基础设施，强化云服务商的数据安全底座与治理能力。随着智慧金融、政务协同等场景对跨域数据流通的依赖增强，云服务商正成为可信数据体系的核心建设方。一方面，通过语义感知、行为基线建模和零信任动态授权，实现跨主体、多链路访问的实时验证与细粒度控制，提升跨域数据流通的可控性；另一方面，依托隐私计算、可验证计算、机密计算、版权保护等技术，构建云端数据可用不可见、可控可审计的可信数据环境。同时，云服务商应将策略管理、数据分类分级、多域治理协同等能力作为平台原生能力嵌入，以数据安全治理为牵引，在智慧金融的数据要素流通、联合建模、跨机构监管等场景中提供系统化的安全底座。

云服务商需要从传统基础设施防护转向构建可信、可监测、可治理、可自适应演进的 AI 原生安全体系。首先，在云平台层面推进可信执行与密态推理能力建设，通过硬件隔离、模型加密、多方安全计算等技术，为企业与开发者提供可证明可信的 AI 运行环境。其次，云服务商应将 AI 风险检测从简单日志分析扩展到跨层多模态监控，覆盖模型输入、推理过程、资源调用、输出行为等关键环节，实现模型越狱、数据隐泄、异常调用、自动化滥用等威胁的实时发现与响应。同时，应构建 AI 自适应防御与云原生安全运营体系，使 AI 模型具备在线自诊断、自修复和动态能力收缩能力，真正实现“安全随模型演化”。

运营商需对现有云网基础设施进行体系化升级，以支撑 6G 所要求的泛在连接、空天地一体化和智能化原生网络能力。在网络层面，运营商需推进全光承载和智能可编排的 IP+ 光网络融合，实现毫秒级调度、端到端确定性传输与按需切片的资源保障，同时通过网络数字孪生和自治驱动控制平面，实现跨域、跨层、跨地域的全局资源动态优化。在云基础设施方面，需构建“中心云 + 区域云 + 边缘云 + 行业专属云”的多层云化体系，并通过云原生技术（容器、无服务器计算、微服务、安全沙箱）实现 6G-AI 原生业务的弹性部署。为满足 6G 的可信传输需求，云网需全面升级零信任体系、可信执行环境与跨域安全认证机制，从而使云、网、边、端在安全框架下实现可管控的协同演进。

低空经济的发展以低空智能计算为牵引，运营商在其中发挥基础设施与平台建设的主导作用。围绕“云–边–端–空”一体化能力，在重点城市和示范区域布局融合通信、导航、感知与算力调度的低空智能底座，为无人机集群调度、空管协同和行业应用提供统一的平台服务。同步完善低空数据采集、传输、存储与处理的安全合规规范，构建覆盖飞行审批、态势监测和事后追溯的治理闭环。通过开放 API、算法市场与仿真测试环境，联动设备厂商和行业客户打造巡检安防、应急救援、低空物流等标杆应用，将安全与可信机制前置集成到低空智能计算架构中，加强对飞行行为、模型决策和数据流动的持续监测与风险评估，支撑低空业务在可观测、可管控、可持续的轨道上发展。

第 五 章

智能泛在云和白皮书总结

本章作为本年度白皮书的收尾，首先介绍云计算研究院在 2025 年提出的研究愿景 —— 智能泛在云，然后对 2025 年度云计算研究白皮书作整体总结。智能泛在云立足于泛在融合的云网基础设施，依托于云计算系统与 AI 算法的深度融合，体现了云计算研究院对于云计算技术演进方向的研判。智能泛在云的研究深度结合前四章讨论的十个热点子方向，具体研究课题通过识别各个子方向中的关键挑战和创新机会来产生。图 5.1展示了智能泛在云的整体架构，最下方是泛在融合的云网基础设施，最上方以 AI 赋能的各类代表性应用作示例，中间的核心部分是云计算系统、AI 算法以及两者的双向融合。

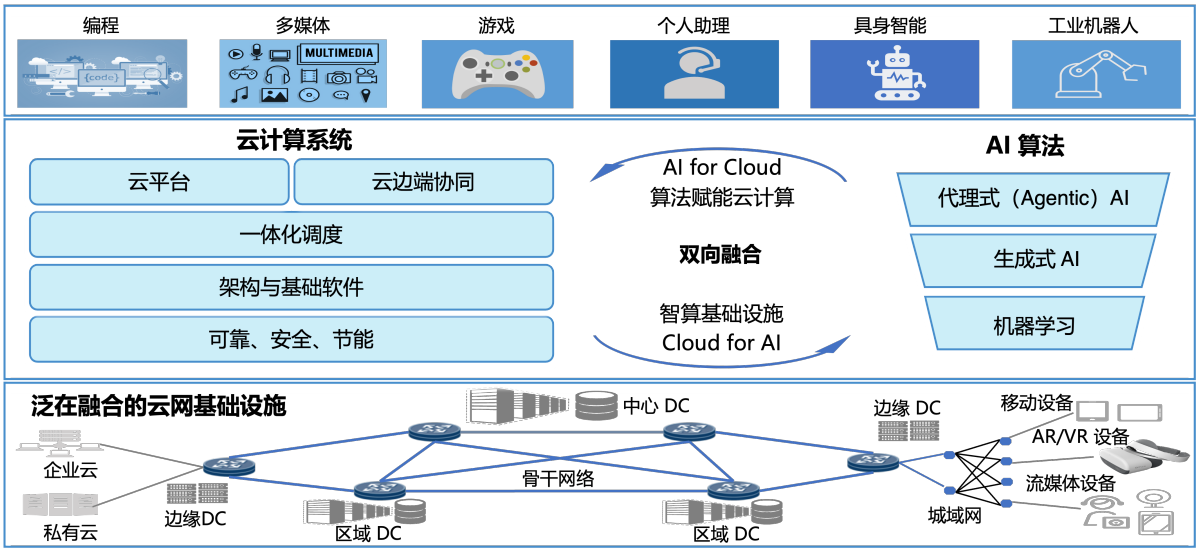


图 5.1: 智能泛在云架构立足于泛在融合的云网基础设施依托于云计算和 AI 的深度融合

5.1 智能泛在云

智能泛在云提出的基础包括两方面，一方面是云网融合的泛在基础设施，另一方面是云计算系统与 AI 算法的双向融合。两方面对于云计算技术栈的各层都带来了新的挑战。本节介绍智能泛在云的背景与特征、技术挑战与创新机会以及定位与展望。

5.1.1 智能泛在云的背景与特征

泛在融合的云网基础设施包含了层次化广覆盖的数据中心布局和新型城域网等创新的网络架构设计。中国电信近年来持续推动“2+4+31+X”以及“一城一池”等数据中心布局，各级数据中心协同配合不仅广泛覆盖全国大部分地理区域，同时发挥东西部区域各自的差异化优势，满足不同区域的差异化需求。层次化广覆盖的数据中心布局也发挥了中国电信网络能力完善的优势。骨干网方面，CN2 为数据中心间互联提供了高带宽和虚拟化支撑。城域网方面，叶-脊（Leaf-Spine）架构的新型城域网增强了城域内的东西向带宽，提升了云接入和边缘云互联的能力。

参考第 3.3章节对 AI 算法发展的详细分析，随着 AI 算法在近年来的突飞猛进，特别是大语言模型和 AI 智能体的高速发展，云计算系统与 AI 算法的双向融合成为重要而紧迫的命题。一方面，云网基础设施需要为即将到来的 AI 应用高速增长提供随处可用的高性能算力和软件工具。另一方面，AI 算法的发展也为云计算的系统效率优化，运维功能增强和服务能力提升提供新的强大工具。

基于泛在融合的云网基础设施和云计算系统/AI 算法的双向融合，智能泛在云将展现如下几个未来云计算所需要的特征：

- 无处不在的低延迟高性能算力。得益于层次化广覆盖的数据中心布局以及智算所需的高性能算力的广泛部署，智能泛在云上的 AI 应用在任何地理区域都可以就近找到可用的低延迟高性能算力。
- 可靠的弹性高性能算力。得益于边缘、区域、中央数据中心之间的网络能力，当高性能算力的负载需求短时间超出就近的数据中心能力，可以在应用性能影响不大的前提下弹性扩展到其他数据中心。
- 可演进的原生 AI 能力。一方面，智能泛在云为将来无处不在的 AI 应用提供持续优化迭代的软件栈和各类工具。另一方面，随着 AI 算法的不断进步，AI 算法对云计算的赋能也持续加强。
- 融合的云网系统技术栈。智能泛在云支持网络虚拟化与云化。虚拟网元功能与普通云计算应用部署在同样的数据中心，应用同样的 AI 算法赋能，同时共享同样的 AI 算力资源与软件栈。

5.1.2 智能泛在云的技术挑战与创新机会

智能泛在云愿景的实现需要云计算技术栈各层的全面创新，本小节将基于前四章四大研究方向中聚焦的十个热点子方向，讨论智能泛在云在各个子方向上的技术挑战与创新机会。

计算任务和网络流量的调度直接影响云网系统的整体效能，云网一体化调度可以从全局视角寻找最优方案。泛在融合云网基础设施的复杂度巨大，求解一体化调度问题需要从理论建模、算法设计、系统实现的各个环节入手。第 2.2章节从网络感知的计算调度、计算感知的网络调度、计算–网络联合调度等多个角度开展了分析讨论。未来研究中将围绕联合建模、高效求解与跨域协同三个方向深化基础研究，形成可解释、可扩展、可验证的调度理论体系。

数据中心架构的最新演进有机会为数据中心效能优化带来大幅提升，分离式数据中心将 CPU、GPU、内存、硬盘等功能单元从服务器中分离，形成各自的资源池，有利于面向不同应用的资源灵活分配，避免由应用的多元化需求导致的服务器资源碎片问题。第 1.2章节从弹性可扩展的云数据中心资源优化、面向资源池化的分离式数据中心架构、支持分离式数据中心架构的软件栈等多个角度开展了分析讨论。未来研究中将与业界学术界同步推动创新，将最新的架构和基础软硬件技术应用到智能泛在云。

云边端协同是发挥泛在融合基础设施能力的关键手段，也是为 AI 应用提供无处不在、及时高效、弹性稳定的高性能算力和软硬件基础设施的必要方式。第 2.4章节从数据协同构建跨层级数据流通体系、任务协同实现多点协同与动态调度、模型协同支撑智能能力演进等多个角度开展了分析讨论。未来研究中在优化传统应用云边端协同的基础上，重点聚焦 AI 应用的云边端协同方案，推动 AI 算法云边端协同部署的框架设计，为 AI 应用的规模化落地提供高效能的基础设施保障。

AI 应用对云计算提出了新的要求，在软硬件各个层面都带来了新的挑战。智算基础设施是云计算支撑 AI 应用的关键，是智能泛在云研究的重要部分。第 2.3章节聚焦在基础设施中的网络部分，从算内网络构建 AI 数据中心 DCN、算间网络实现跨数据中心互联 DCI、入算网络支撑用户算力接入 DCA 等多个方面开展了分析讨论。未来研究将在数据中心智算网络组网设计、集合通信优化、智算软件栈中的弹性训练、PD 分离/AF 分离推理优化等方面开展。

算法是计算机系统的核心能力之一，云计算系统的效能优化同样依赖于高效的算法。近年来 AI 算法的突飞猛进在传统数学优化算法基础上增加了数据驱动的新手段，而 AI 智能体则有望带来自适应能力，在系统运维和用户服务中发挥重要作用。第 3.2章节完整讨论了运筹优化、深度学习、强化学习等各类传

统数学优化和 AI 算法，全面分析了各类算法在云计算中的应用场景。未来研究在探索各类算法应用的同时，将突出基础研究，以形式化方法、复杂性分析、最优化理论等为核心，构建面向大规模云网系统的科学理论框架，为资源调度、负载均衡、容量规划等关键机制提供可靠的理论支撑。

云计算 PaaS 层为应用提供平台抽象，包括数据库、大数据计算、Serverless、AI 训练推理等。PaaS 层也为数据在云上的流转、存储、分析处理提供了工具。第 1.3章节从面向智能应用的 Serverless 计算平台技术、面向大模型的智能数据平台技术、支撑智能任务的高性能存储平台技术等多个方面开展了分析讨论。未来研究中，PaaS 层一方面将从传统的数据中心内部向跨地域的泛在融合基础设施扩展，另一方面将聚焦 AI 应用的支撑，让大模型、智能体的开发与部署成为云平台的原生能力。

可靠性、安全性、绿色节能是云计算的基础能力和长期命题，泛在融合基础设施和 AI 应用都将引入新的挑战。第 1.4章节从面向大规模集群的自动化运维与可靠性、云计算环境下的基础设施安全、云数据中心智能功耗管理与优化等多个方面开展了分析讨论。未来研究将围绕两方面开展：完善智能运维与安全防护体系，保障云平台高可用与可信；推动绿色低碳技术创新，实现云数据中心可持续发展。

科技创新大潮下新兴技术层出不穷，AI 智能体、低空智能、6G、量子计算等的发展日新月异。云网基础设施需要为新兴技术的应用提前布局，起到推动新兴技术快速推广的作用。第 4.2章节从智能时代下的新兴计算范式、面向泛在互联的第六代移动通信系统、面向低空经济的智能计算等多个方面开展了分析讨论。未来研究将持续聚焦新兴技术对云网基础设施提出的新需求与新挑战。

5.1.3 智能泛在云的定位与展望

智能泛在云体现了中国电信云计算研究院对于云计算发展方向的研判，也代表了云计算研究中各个子课题之间的组织逻辑。智能泛在云与中国电信息壤和云网操作系统的关系可以用泛在操作系统与鸿蒙操作系统的关系 [757] 类比，前者聚焦在基础理论应用，核心算法设计，系统原则凝炼，创新技术突破等研究环节，而后者需要为技术应用、产品设计、工程实现与业务落地等产品环节负责。研究与产品的深度协同将有益于双方各自的长足发展和目标实现，因此也是开展智能泛在云研究过程中的重要方式。

云计算研究院深信智能泛在云的研究不仅有助于持续提升中国电信在云计算领域的技术能力，也将起到推动整个云计算行业发展的作用。云计算研究院将围绕智能泛在云开展各个子方向上的课题研究，研究过程一方面与产品团队紧密配合，确保研究与产品的目标一致、过程协同，另一方面也充分挖掘产学研结合的作用，与国内外高校开展广泛交流和深度合作，共同攻关智能泛在云中的关键技术难题。

5.2 云计算研究白皮书 2025 的总结

表 5.1: 白皮书聚焦的十个热点方向

1. 分离式数据中心架构及关键技术	2. 面向 AI 场景的 PaaS 数据平台层技术
3. 智能化云运维、可信安全与能效优化	4. 云网一体化调度
5. 面向智算的云网基础设施	6. 云边端协同
7. 算法赋能云计算	8. AI Agent 与 Agentic AI
9. 新兴技术及应用	10. 数据与 AI 的安全

作为云计算研究院研究洞察和研究工作的年度总结汇报，本年度云计算研究白皮书延续一贯的内容风格，以国际国内的最新行业趋势为导向，以详尽的产业数据分析和全面的学术界进展梳理为主要论述依据，共引用行业白皮书等文献 60 余篇，高水平论文近 700 篇。内容结构方面，继续基于“三个面向一个围绕”的四大研究方向，包括面向下一代云计算的研究、面向云网融合的研究、围绕智能算法的研究和面向新兴技术的研究。新内容方面，今年首次提炼了十个热点子方向（表 5.1），也首次阐述了云计算研究院 2025 年提出的研究愿景–智能泛在云。研究成果总结方面，十个热点方向的详细论述中也介绍了

云计算研究院 2025 年度研究成果中的 16 项，其中各项成果均已在高水平会议/期刊发表，或者已被接收录用。

最后，本年度白皮书的趋势展望和发展建议总结如表：

表 5.2: 白皮书提出的趋势展望与发展建议

	趋势展望	发展建议
下一代云	• 云计算基础设施正从传统数据中心云向分层多级架构的泛在云架构演进，智能技术普及将进一步催生云计算行业变革。 • 未来云计算服务模式将以智能化和动态协同为核心，通过平台与工作负载之间的双向实时通信，实现资源管理从“静态分配”向“需求驱动、动态优化”转变。 • 未来云计算服务模式将不断突破传统资源供给范式，向“行业即服务、智能即服务、场景即服务”等多层次形态演进。	• 加速 AI 原生云平台建设，打造智能化、弹性化的服务能力。 • 完善智能运维与安全防护体系，保障云平台高可用与可信。 • 推动绿色低碳技术创新，实现云数据中心可持续发展。
云网融合	• 面向智能云网体系，云网一体化调度将进一步从“资源整合”走向“资源融合”，在更大空间、更高维度实现计算、网络、存储等资源的统一感知、统一决策与统一优化。 • 未来智算云网基础设施将演化为面向“AI 超级计算平台”的超大规模、异构、统一互联的云网基础设施，并从单域优化走向全域互联的系统化演进。 • 云边端协同体系正从静态分层的资源组织转向面向任务与语义的智能协同架构，其目标是实现跨层级能力的统一抽象与高效编排，使数据、任务与模型在不同层之间实现更可控的流动与优化。	• 面向智能云网未来形态，需围绕联合建模、高效求解与跨域协同三个方向深化基础研究，形成可解释、可扩展、可验证的调度理论体系。 • 面向智算云网基础设施，需加大推进产业标准化与开放性，并尽早关注跨界融合和交叉领域的研究。 • 推动智能模型在云边端的协同部署框架，并促进其规模化的应用落地。
智能算法	• 智能算法在云–网-智算一体化系统的研究将沿着“运筹优化的结构化数学能力”与“深度学习模型的数据驱动优势”双线融合演进。 • 以图算法、图神经网络和深度神经网络为代表的结构化建模与表征学习技术，将持续支撑云–网系统在复杂拓扑、多维依赖和异构数据上的智能演进。 • 未来 AI Agent 的关键突破或聚焦于自主智能的体系化演进、新型智能基础设施以及面向复杂社会环境的安全可信治理体系三条主线。	• 重视算法理论的基础研究，以形式化方法、复杂性分析、最优化理论等为核心，构建面向大规模云网系统的科学理论框架，为资源调度、负载均衡、容量规划等关键机制提供可靠理论支撑。 • 构建可控可信的自治智能与治理体系。 • 推动绿色高效的云—边—端协同与具身智能基础设施建设。
新兴技术	• 随着数据跨域流动加速、隐私保护需求提升，数据安全治理正向以隐私保护与可验证操作为核心的可信数据体系演进。 • AI 安全将从单一防护向体系化、全生命周期和智能化演进。 • 建设超高速率、超低时延、超大连接的 6G 通信网，并与云计算和算力网络深度融合。 • 随着低空空域逐步开放和无人系统规模化部署，低空智能计算将成为支撑低空经济安全运行与业务创新的关键底座。	• 建设可信数据基础设施，强化云服务商的数据安全底座与治理能力。 • 构建可信、可监测、可治理、可自适应演进的 AI 原生安全体系。 • 云网基础设施体系化升级以支撑 6G 的泛在连接、空天地一体化和智能化。 • 以低空智能计算为牵引，发挥基础设施与平台建设的主导作用。

参考文献

[1] Yuhong Zhong, Daniel S. Berger, Carl A. Waldspurger, et al. Managing memory tiers with CXL in virtualized environments. In 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.

[2] Mingxing Zhang, Teng Ma, Jinqi Hua, et al. Partial failure resilient memory management system for (cxl-based) distributed shared memory. In Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023, pages 658–674. ACM, 2023.

[3] Guowei Liu, Laiping Zhao, Yiming Li, et al. 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.

[4] Wei Liu, Kun Qian, Zhenhua Li, et al. Mitigating scalability walls of rdma-based container networks. In Theophilus A. Benson and Radhika Niranjan Mysore, editors, 22nd USENIX Symposium on Networked Systems Design and Implementation, NSDI 2025, Philadelphia, PA, USA, April 28-30, 2025, pages 1049–1065. USENIX Association, 2025.

[5] Ao Xiao, Bangzheng He, Baoquan Zhang, et al. xdeepserve: Model-as-a-service on huawei cloudmatrix384. CoRR, abs/2508.02520, 2025.

[6] Inc. Gartner. Forecast: Public cloud services, worldwide, 2023-2029, 3q25 update, 2025.

[7] 中国电信云计算研究院. 《云计算研究白皮书（2024）》, December 2024.

[8] 国际数据公司 (IDC). 2023 年全球公共云服务收入, 2023.

[9] Borui Wan, Gaohong Liu, Zuquan Song, et al. Robust llm training infrastructure at bytedance. In Proceedings of the ACM SIGOPS 31st Symposium on Operating Systems Principles, pages 186–203, 2025.

[10] Wei An, Xiao Bi, Guanting Chen, et al. Fire-flyer ai-hpc: A cost-effective software-hardware co-design for deep learning. In SC24: International Conference for High Performance Computing, Networking, Storage and Analysis, pages 1–23. IEEE, 2024.

[11] 信通院. 2025 信通院云计算蓝皮书, 2025.

[12] Krzysztof Rzadca, Pawel Findeisen, Jacek Swiderski, et al. Autopilot: workload autoscaling at google. In Angelos Bilas, Kostas Magoutis, Evangelos P. Markatos, et al., editors, EuroSys '20: Fifteenth EuroSys Conference 2020, Heraklion, Greece, April 27-30, 2020, pages 16:1–16:16. ACM, 2020.

[13] China Electronics Standardization Institute (CESI) and China Electronic Information Industry Development Academy (CEIDA). 中国电子技术标准化研究院发布《异构计算资源池化技术白皮书》（征求意见稿）. 电子信息产业网, October 2023.

[14] Johannes Weiner, Niket Agarwal, Dan Schatzberg, et al. TMO: transparent memory offloading in datacenters. Commun. ACM, 68(9):102–110, 2025.

[15] Padmapriya Duraisamy, Wei Xu, Scott Hare, et al. Towards an adaptable systems architecture for memory tiering at warehouse-scale. In Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3, ASPLOS 2023, page 727–741, New York, NY, USA, 2023. Association for Computing Machinery.

[16] Yuhong Zhong, Daniel S. Berger, Carl A. Waldspurger, et al. Managing memory tiers with CXL in virtualized environments. In 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.

[17] Wenda Tang, Yiduo Wang, Yanwen Wang, and Jie Wu. Leave no one behind: Towards fair and efficient tiered memory management for multi-applications. In 54th International Conference on Parallel Processing, ICPP 2025, San Diego, CA, USA, September 8-11, 2025, New York, NY, USA, 2025. Association for Computing Machinery.

[18] Yawen Wang, Kapil Arya, Marios Kogias, et al. Smartharvest: harvesting idle cpus safely and efficiently in the cloud. In EuroSys '21: Sixteenth European Conference on Computer Systems, Online Event, United Kingdom, April 26-28, 2021, pages 1–16. ACM, 2021.

[19] Benjamin Reidys, Pantea Zardoshti, Íñigo Goiri, et al. Coach: Exploiting temporal patterns for all-resource oversubscription in cloud platforms. In Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1, ASPLOS 2025, Rotterdam, The Netherlands, 30 March 2025 - 3 April 2025, pages 164–181. ACM, 2025.

[20] Alexander Fuerst, Stanko Novakovic, Iñigo Goiri, et al. Memory-harvesting vms in cloud platforms. In ASPLOS '22: 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Lausanne, Switzerland, 28 February 2022 - 4 March 2022, pages 583–594. ACM, 2022.

[21] Ruoyu Qin, Zheming Li, Weiran He, et al. Mooncake: Trading more storage for less computation—a kvcache-centric architecture for serving llm chatbot. In 23rd USENIX Conference on File and Storage Technologies (FAST 25), pages 155–170, 2025.

[22] Ao Xiao, Bangzheng He, Baoquan Zhang, et al. xdeepserve: Model-as-a-service on huawei cloudmatrix384. arXiv preprint arXiv:2508.02520, 2025.

[23] Gabriele Oliaro, Xupeng Miao, Xinhao Cheng, et al. Flexllm: A system for co-serving large language model inference and parameter-efficient finetuning. arXiv preprint arXiv:2402.18789, 2024.

[24] Xiaoyang Wang, Yongkun Li, Kan Wu, et al. Finemem: Breaking the allocation overhead vs. memory waste dilemma in fine-grained disaggregated memory management. In 19th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2025, Boston, MA, USA, July 7-9, 2025, pages 57–74. USENIX Association, 2025.

[25] Yang Zhou, Hassan M. G. Wassel, Sihang Liu, et al. Carbink: Fault-tolerant far memory. In 16th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2022, Carlsbad, CA, USA, July 11-13, 2022, pages 55–71. USENIX Association, 2022.

[26] Cong Guo, Rui Zhang, Jiale Xu, et al. Gmlake: Efficient and transparent gpu memory defragmentation for large-scale dnn training with virtual memory stitching. In Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2, pages 450–466, 2024.

[27] Huaicheng Li, Daniel S Berger, Lisa Hsu, 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.

[28] Xinjun Yang, Yingqiang Zhang, Hao Chen, et al. Unlocking the potential of cxl for disaggregated memory in cloud-native databases. In Companion of the 2025 International Conference on Management of Data, SIGMOD/PODS '25, page 689–702, New York, NY, USA, 2025. Association for Computing Machinery.

[29] Yuxin Ren, Mingrui Liu, Hongbo Li, et al. Towards rack-as-a-computer in memory interconnect era with coordinated operating system sharing. In Proceedings of the 17th ACM Workshop on Hot Topics in Storage and File Systems, HotStorage 2025, Boston, MA, USA, July 10-11, 2025, pages 77–85. ACM, 2025.

[30] 天翼云. "分离""聚合"两手抓, 天翼云聚合计算赋能多元化应用场景!, 4 2024. 访问日期: 2025-11-22.

[31] Cunchen Hu, Chenxi Wang, Sa Wang, et al. 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 2023, Providence, RI, USA, June 22-24, 2023, pages 94–102. ACM, 2023.

[32] Quanxi Li, Hong Huang, Ying Liu, et al. Beehive: A scalable disaggregated memory runtime exploiting asynchrony of multithreaded programs. In 22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25), pages 167–187, 2025.

[33] Paul Barham, Aakanksha Chowdhery, Jeff Dean, et al. Pathways: Asynchronous distributed dataflow for ml. Proceedings of Machine Learning and Systems, 4:430–449, 2022.

[34] Patrick Damme, Marius Birkenbach, Constantinos Bitsakos, et al. Daphne: An open and extensible system infrastructure for integrated data analysis pipelines. In Conference on Innovative Data Systems Research, 2022.

[35] Lexiang Huang, Anjaly Parayil, Jue Zhang, et al. Workload intelligence: Workload-aware iaas abstraction for cloud efficiency. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC '25, page 2203–2215, New York, NY, USA, 2025. Association for Computing Machinery.

[36] Xupeng Miao, Chunan Shi, Jiangfei Duan, et al. Spotserve: Serving generative large language models on preemptible instances. In Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2, pages 1112–1127, 2024.

[37] Jiangfei Duan, Runyu Lu, Haojie Duanmu, et al. Muxserve: Flexible spatial-temporal multiplexing for multiple llm serving. arXiv preprint arXiv:2404.02015, 2024.

[38] Emmanuel Amaro, Christopher Branner-Augmon, Zhihong Luo, et al. Can far memory improve job throughput? In EuroSys '20: Fifteenth EuroSys Conference 2020, Heraklion, Greece, April 27-30, 2020, pages 14:1–14:16. ACM, 2020.

[39] Chenxi Wang, Yifan Qiao, Haoran Ma, et al. Canvas: Isolated and adaptive swapping for multi-applications on remote memory. In 20th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2023, Boston, MA, April 17-19, 2023, pages 161–179. USENIX Association, 2023.

[40] Zhenyuan Ruan, Malte Schwarzkopf, Marcos K. Aguilera, and Adam Belay. AIFM: high-performance, application-integrated far memory. In 14th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2020, Virtual Event, November 4-6, 2020, pages 315–332. USENIX Association, 2020.

[41] Yuxing Xiang, Xue Li, Kun Qian, et al. Aegaeon: Effective gpu pooling for concurrent llm serving on the market. In Proceedings of the ACM SIGOPS 31st Symposium on Operating Systems Principles, pages 1030–1045, 2025.

[42] Jiale Xu, Rui Zhang, Yi Xiong, et al. ellm: Elastic memory management framework for efficient llm serving. arXiv preprint arXiv:2506.15155, 2025.

[43] Bin Lin, Chen Zhang, Tao Peng, et al. Infinite-llm: Efficient llm service for long context with distattention and distributed kvcache. arXiv preprint arXiv:2401.02669, 2024.

[44] Yibo Huang, Haowei Chen, Newton Ni, et al. Tigon: A distributed database for a CXL pod. In 19th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2025, Boston, MA, USA, July 7-9, 2025, pages 109–128. USENIX Association, 2025.

[45] Mingxing Zhang, Teng Ma, Jinqi Hua, et al. Partial failure resilient memory management system for (cxl-based) distributed shared memory. In Proceedings of the 29th Symposium on Operating Systems Principles, SOSP '23, page 658–674, New York, NY, USA, 2023. Association for Computing Machinery.

[46] Yizhou Shan, Yutong Huang, Yilun Chen, and Yiying Zhang. Legoos: A disseminated, distributed OS for hardware resource disaggregation. In 13th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2018, Carlsbad, CA, USA, October 8-10, 2018, pages 69–87. USENIX Association, 2018.

[47] Ankit Bhardwaj, Chinmay Kulkarni, Reto Achermann, et al. Nros: Effective replication and sharing in an operating system. In 15th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2021, July 14-16, 2021, pages 295–312. USENIX Association, 2021.

[48] Ho-Ren Chuang, Karim Manaouil, Tong Xing, et al. Aggregate VM: why reduce or evict vm's resources when you can borrow them from other nodes? In Proceedings of the Eighteenth European Conference on Computer Systems, EuroSys 2023, Rome, Italy, May 8-12, 2023, pages 469–487. ACM, 2023.

[49] Xingguo Jia, Jin Zhang, Boshi Yu, et al. Giantvm: A novel distributed hypervisor for resource aggregation with dsm-aware optimizations. ACM Trans. Archit. Code Optim., 19(2):20:1–20:27, 2022.

[50] Chris Lattner, Mehdi Amini, Uday Bondhugula, et al. Mlir: Scaling compiler infrastructure for domain specific computation. In 2021 IEEE/ACM International Symposium on Code Generation and Optimization (CGO), pages 2–14. IEEE, 2021.

[51] Jeffrey Pound, Floris Chabert, Arjun Bhushan, et al. Micronn: An on-device disk-resident updatable vector database. In Companion of the 2025 International Conference on Management of Data, pages 608–621, 2025.

[52] Tian Jin, Gheorghe-Teodor Bercea, Tung D Le, et al. Compiling onnx neural network models using mlir. arXiv preprint arXiv:2008.08272, 2020.

[53] Microsoft Azure. xla. https://github.com/openxla/xla, 2024.

[54] Intel. mlir-extensions. https://learn.microsoft.com/en-us/azure/container-apps/gpu-serverless-overview, 2023.

[55] Minchen Yu, Ao Wang, Dong Chen, et al. Torpor: Gpu-enabled serverless computing for low-latency, resource-efficient inference. In Proceedings of the 2025 USENIX Annual Technical Conference, USENIX ATC 2025, Boston, MA, USA, July 7-9, 2025, pages 597–612. USENIX Association, 2025.

[56] Microsoft Azure. Using serverless gpus in azure container apps. https://learn.microsoft.com/en-us/azure/container-apps/gpu-serverless-overview, 2025.

[57] Artjom Joosen, Ahmed Hassan, Martin Asenov, et al. Serverless cold starts and where to find them. In Proceedings of the Twentieth European Conference on Computer Systems, EuroSys 2025, Rotterdam, The Netherlands, 30 March 2025 - 3 April 2025, pages 938–953. ACM, 2025.

[58] Xiaohu Chai, Tianyu Zhou, Keyang Hu, et al. Fork in the road: Reflections and optimizations for cold start latency in production serverless systems. In Lidong Zhou and Yuanyuan Zhou, editors, 19th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2025, Boston, MA, USA, July 7-9, 2025, pages 199–218. USENIX Association, 2025.

[59] Rong Kang, Yanbin Chen, Ye Liu, et al. Abase: the multi-tenant nosql serverless database for diverse and dynamic workloads in large-scale cloud environments. In Volker Markl, Joseph M. Hellerstein, and Azza Abouzied, editors, Companion of the 2025 International Conference on Management of Data, SIGMOD/PODS 2025, Berlin, Germany, June 22-27, 2025, pages 471–484. ACM, 2025.

[60] Jeff Swenson, Andy Kimball, Raphael 'kena' Poss, et al. Cockroachdb serverless: Sub-second scaling from zero with multi-region cluster virtualization. In Volker Markl, Joseph M. Hellerstein, and Azza Abouzied, editors, Companion of the 2025 International Conference on Management of Data, SIGMOD/PODS 2025, Berlin, Germany, June 22-27, 2025, pages 648–661. ACM, 2025.

[61] Junhao Hu, Jiang Xu, Zhixia Liu, et al. DEEPSERVE: serverless large language model serving at scale. In Deniz Altinbüken and Ryan Stutsman, editors, Proceedings of the 2025 USENIX Annual Technical Conference, USENIX ATC 2025, Boston, MA, USA, July 7-9, 2025, pages 57–72. USENIX Association, 2025.

[62] Yangshen Deng, Zhengxin You, Long Xiang, et al. Alayadb: The data foundation for efficient and effective long-context llm inference. In Companion of the 2025 International Conference on Management of Data, pages 364–377, 2025.

[63] Alibaba Cloud Native Community. Alibaba cloud bailian open source nl2sql intelligent framework for java developers. https://www.alibabacloud.com/blog/alibaba-cloud-bailian-open-source-nl2sql-intelligent-framework-for-java-developers$_{6}$02307, 6 2025. Accessed: [2025-12-15]; Open-source NL2SQL framework based on Spring AI, integrated with Alibaba Cloud XiYan GBI and Tongyi large model.

[64] Sunil Chakkappen, Shreya Kunjibettu, Daniel McGreer, et al. Automatic indexing in oracle. Proceedings of the VLDB Endowment, 18(12):4924–4937, 2025.

[65] Xinjun Yang, Yingqiang Zhang, Hao Chen, et al. Unlocking the potential of cxl for disaggregated memory in cloud-native databases. In Companion of the 2025 International Conference on Management of Data, pages 689–702, 2025.

[66] Panagiotis Antonopoulos, Mansi Chauhan, Shailender Dabas, et al. Md-mvcc: Multi-version concurrency control for schema changes in azure sql database. Proceedings of the VLDB Endowment, 18(12):4791–4803, 2025.

[67] Borui Wan, Mingji Han, Yiyao Sheng, et al. Bytecheckpoint : A unified checkpointing system for large foundation model development. In 22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25), pages 559–578, 2025.

[68] Zimeng Huang, Hao Nie, Haonan Jia, et al. Flowcheck: Decoupling checkpointing and training of large-scale models. In Proceedings of the Twentieth European Conference on Computer Systems, pages 1334–1349, 2025.

[69] Jiayi Yao, Hanchen Li, Yuhan Liu, et al. Cacheblend: Fast large language model serving for rag with cached knowledge fusion. In Proceedings of the Twentieth European Conference on Computer Systems, pages 94–109, 2025.

[70] Shi Qiu, Weinan Liu, Yifan Hu, et al. Geminifs : A companion file system for gpus. In 23rd USENIX Conference on File and Storage Technologies (FAST 25), pages 221–236, 2025.

[71] Wenbin Zhu, Zhaoyan Shen, Qian Wei, et al. Hidpu : A dpu-oriented hybrid indexing scheme for disaggregated storage systems. In 23rd USENIX Conference on File and Storage Technologies (FAST 25), pages 271–285, 2025.

[72] Seung Won Yoo, Joontaek Oh, Myeongin Cheon, et al. Djfs : Directory-granularity filesystem journaling for cmm-h ssds. In 23rd USENIX Conference on File and Storage Technologies (FAST 25), pages 35–51, 2025.

[73] Michael Allison, Arun George, Javier Gonzalez, et al. Towards efficient flash caches with emerging nvme flexible data placement ssds. In Proceedings of the Twentieth European Conference on Computer Systems, pages 1142–1160, 2025.

[74] Jiahao Li, Biao Cao, Jielong Jian, et al. Mantle: Efficient hierarchical metadata management for cloud object storage services. In Proceedings of the ACM SIGOPS 31st Symposium on Operating Systems Principles, pages 928–943, 2025.

[75] Shu Liu, Xiangxi Mo, Moshik Hershcovitch, et al. Skystore: Cost-optimized object storage across regions and clouds. arXiv preprint arXiv:2502.20818, 2025.

[76] Ziheng Wang, Junyu Wei, Alex Aiken, et al. Logcioud: Fast search of compressed logs on object storage. Proceedings of the VLDB Endowment, 18(8):2362–2370, 2025.

[77] Jian Gao, Jiwu Shu, Bin Yan, et al. Stripeless data placement for Erasure-Coded In-Memory storage. In 19th USENIX Symposium on Operating Systems Design and Implementation (OSDI 25), pages 821–838, 2025.

[78] Haonan Wu, Erci Xu, Ligang Wang, et al. Hey hey, my my, skewness is here to stay: Challenges and opportunities in cloud block store traffic. In Proceedings of the Twentieth European Conference on Computer Systems, pages 736–752, 2025.

[79] NVIDIA. Multi-process service (mps). https://docs.nvidia.com/deploy/mps/index.html, 2025.

[80] Keeping Functions Warm. https://docs.aws.amazon.com/lambda/latest/dg/lambda-concurrency.html. Referenced 2022.

[81] Alexander Fuerst and Prateek Sharma. Faascache: keeping serverless computing alive with greedy-dual caching. In ASPLOS '21: 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Virtual Event, USA, April 19-23, 2021, pages 386–400. ACM, 2021.

[82] Mengzhao Wang, Xiaoliang Xu, Qiang Yue, and Yuxiang Wang. A comprehensive survey and experimental comparison of graph-based approximate nearest neighbor search. Proc. VLDB Endow., 14:1964–1978, 2021.

[83] Ilias Azizi, Karima Echihabi, and Themis Palpanas. Graph-based vector search: An experimental evaluation of the state-of-the-art. Proceedings of the ACM on Management of Data, 3:1 – 31, 2025.

[84] Suhas Jayaram Subramanya, Devvrit, Rohan Kadekodi, et al. Diskann : Fast accurate billion-point nearest neighbor search on a single node. 2019.

[85] Jinhao Zhu, Liana Patel, Matei Zaharia, and Raluca Ada Popa. Compass: encrypted semantic search with high accuracy. In 19th USENIX Symposium on Operating Systems Design and Implementation (OSDI 25), pages 915–938, 2025.

[86] Jiayi Wang and Guoliang Li. Aop: Automated and interactive llm pipeline orchestration for answering complex queries. CIDR, 2025.

[87] Liang Shi, Zhengju Tang, Nan Zhang, et al. A survey on employing large language models for text-to-sql tasks. ACM Comput. Surv., 58(2), September 2025.

[88] Biao Ouyang, Yingying Zhang, Hanyin Cheng, et al. Rcrank: Multimodal ranking of root causes of slow queries in cloud database systems. arXiv preprint arXiv:2503.04252, 2025.

[89] Yibo Huang, Haowei Chen, Newton Ni, et al. Tigon: A distributed database for a cxl pod. In 19th USENIX Symposium on Operating Systems Design and Implementation (OSDI 25), Boston, MA, 2025.

[90] Hubert Mohr-Daurat, Xuan Sun, and Holger Pirk. Boss-an architecture for database kernel composition. ACM SIGMOD Record, 54(1):37–46, 2025.

[91] Shaobo Li, Yirui Eric Zhou, Yuqi Xue, et al. Managing scalable direct storage accesses for gpus with gofs. In Proceedings of the ACM SIGOPS 31st Symposium on Operating Systems Principles, pages 979–995, 2025.

[92] Wenhao Lv, Hao Guo, Qing Wang, et al. Accelerating distributed filesystem metadata service via decoupling directory semantics from metadata indexing. In ACM Symposium on Cloud Computing (SoCC '25), 2025.

[93] Yiduo Wang, Wenda Tang, Meng Linghang, et al. Origami: Efficient ml-driven metadata load balancing for distributed file systems. In 54th International Conference on Parallel Processing, ICPP 2025, San Diego, CA, USA, September 8-11, 2025, New York, NY, USA, 2025. Association for Computing Machinery.

[94] Yanwen Wang, Wenda Tang, and Jie Wu. Nip it in the bud: Unsupervised kpi incipient fault detection via dynamic latent feature ensembling. In 44rd International Symposium on Reliable Distributed Systems, SRDS 2025, Porto, Portugal, September 29 - Oct. 2, 2025, pages 1–11. IEEE, 2025.

[95] Zeyan Li, Junjie Chen, Rui Jiao, et al. Practical root cause localization for microservice systems via trace analysis. In 29th IEEE/ACM International Symposium on Quality of Service, IWQOS 2021, Tokyo, Japan, June 25-28, 2021, pages 1–10. IEEE, 2021.

[96] Dewei Liu, Chuan He, Xin Peng, et al. Microhecl: High-efficient root cause localization in large-scale microservice systems. In 43rd IEEE/ACM International Conference on Software Engineering: Software Engineering in Practice, ICSE (SEIP) 2021, Madrid, Spain, May 25-28, 2021, pages 338–347. IEEE, 2021.

[97] Facebook Production Engineering Team. Auto-remediation at facebook. SRECon 技术演讲, 2019. 可访问: https://engineering.fb.com/.

[98] Microsoft Azure. Azure automanage：自动化与自愈能力. 微软官方文档, 2020. 可访问: https://learn.microsoft.com/.

[99] 阿里巴巴工程实践. 阿里云混沌工程平台 chaosblade. 阿里巴巴开发者社区, 2021. 可访问: https://developer.aliyun.com/.

[100] Siddharth Muralee, Igibek Koishybayev, Aleksandr Nahapetyan, et al. Argus : A framework for staged static taint analysis of github workflows and actions. In 32nd USENIX Security Symposium (USENIX Security 23), pages 6983–7000, 2023.

[101] Qihang Zhou, Wenzhuo Cao, Xiaoqi Jia, et al. Rcontainer: A secure container architecture through extending arm cca hardware primitives. NDSS, 2025.

[102] Isaac Polinsky, Pubali Datta, Adam Bates, and William Enck. Grasp: Hardening serverless applications through graph reachability analysis of security policies. In Proceedings of the ACM Web Conference 2024, pages 1644–1655, 2024.

[103] Anish Saxena, Walter Wang, and Alexandros Daglis. Citadel: Rethinking memory allocation to safeguard against inter-domain rowhammer exploits. In Proceedings of the 58th IEEE/ACM International Symposium on Microarchitecture®, pages 1117–1131, 2025.

[104] Ataberk Olgun, F Nisa Bostancı, İsmail Emir Yüksel, et al. Variable read disturbance: An experimental analysis of temporal variation in dram read disturbance. In 2025 IEEE International Symposium on High Performance Computer Architecture (HPCA), pages 849–866. IEEE, 2025.

[105] Sathiya Kumaran Mani, Kevin Hsieh, Santiago Segarra, et al. Securing public cloud networks with efficient role-based micro-segmentation. In 22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25), pages 1033–1048, 2025.

[106] Yu Zou, Yiran Li, Sheng Wang, et al. Salus: A practical trusted execution environment for cpu-fpga heterogeneous cloud platforms. In Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 4, pages 252–266, 2024.

[107] Jovan Stojkovic, Pulkit A. Misra, Íñigo Goiri, et al. Smartoclock: Workload- and risk-aware overclocking in the cloud. In 51st ACM/IEEE Annual International Symposium on Computer Architecture, ISCA 2024, Buenos Aires, Argentina, June 29 - July 3, 2024, pages 437–451. IEEE, 2024.

[108] Leonardo Piga, Iyswarya Narayanan, Aditya Sundarrajan, et al. Expanding datacenter capacity with DVFS boosting: A safe and scalable deployment experience. In Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1, ASPLOS 2024, La Jolla, CA, USA, 27 April 2024- 1 May 2024, pages 150–165. ACM, 2024.

[109] Hanfei Geng, Yi Sun, Yuanzhe Li, et al. TESLA: thermally safe, load-aware, and energy-efficient cooling control system for data centers. In Proceedings of the 53rd International Conference on Parallel Processing, ICPP 2024, Gotland, Sweden, August 12-15, 2024, pages 939–949. ACM, 2024.

[110] Jovan Stojkovic, Chaojie Zhang, Íñigo Goiri, et al. TAPAS: thermal- and power-aware scheduling for LLM inference in cloud platforms. In Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2, ASPLOS 2025, Rotterdam, Netherlands, 30 March 2025 - 3 April 2025, pages 1266–1281. ACM, 2025.

[111] Pratyush Patel, Esha Choukse, Chaojie Zhang, et al. Characterizing power management opportunities for llms in the cloud. In Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3, ASPLOS '24, page 207–222, New York, NY, USA, 2024. Association for Computing Machinery.

[112] Zibo Wang, Yijia Zhang, Fuchun Wei, et al. Using analytical performance/power model and fine-grained dvfs to enhance ai accelerator energy efficiency. In Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1, ASPLOS '25, page 1118–1132, New York, NY, USA, 2025. Association for Computing Machinery.

[113] Vincent Jacob and Yanlei Diao. Unsupervised anomaly detection in multivariate time series across heterogeneous domains. Proc. VLDB Endow., 18(6):1691–1704, 2025.

[114] Kaiqi Ding, Yuanmu Ma, Zijian Song, and Kaigui Bian. Enhancing microservices anomaly detection via multimodal data fusion in the wavelet domain and spatiotemporal graph-based diffusion probabilistic model. In Luiza Antonie, Jian Pei, Xiaohui Yu, et al., editors, Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025, pages 510–520. ACM, 2025.

[115] Yidan Wang, Zhouruixing Zhu, Qiuai Fu, et al. MRCA: metric-level root cause analysis for microservices via multi-modal data. In Vladimir Filkov, Baishakhi Ray, and Minghui Zhou, editors, Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering, ASE 2024, Sacramento, CA, USA, October 27 - November 1, 2024, pages 1057–1068. ACM, 2024.

[116] Wen Gao, Zhiwen Yu, Tian Wang, et al. Gnn-based deep reinforcement learning for computation task scheduling in autonomous multi-robot systems. J. Syst. Archit., 168:103534, 2025.

[117] Fei Teng, Haoyang Li, and Lei Chen. Llmlog: Advanced log template generation via llm-driven multi-round annotation. Proc. VLDB Endow., 18(9):3134–3148, 2025.

[118] Yichen Li, Yulun Wu, Jinyang Liu, et al. COCA: generative root cause analysis for distributed systems with code knowledge. In 47th IEEE/ACM International Conference on Software Engineering, ICSE 2025, Ottawa, ON, Canada, April 26 - May 6, 2025, pages 1346–1358. IEEE, 2025.

[119] JinJin Lin, Pengfei Chen, and Zibin Zheng. Microscope: Pinpoint performance issues with causal graphs in micro-service environments. In Claus Pahl, Maja Vukovic, Jianwei Yin, and Qi Yu, editors, Service-Oriented Computing - 16th International Conference, ICSOC 2018, Hangzhou, China, November 12-15, 2018, Proceedings, volume 11236 of Lecture Notes in Computer Science, pages 3–20. Springer, 2018.

[120] Jiahao Shi, Sihang Jiang, Bo Xu, and Yanghua Xiao. Serverrca: Root cause analysis for server failure using operating system logs. In 34th IEEE International Symposium on Software Reliability Engineering, ISSRE 2023, Florence, Italy, October 9-12, 2023, pages 486–496. IEEE, 2023.

[121] Victor Dumitriu, Lev Kirischian, and Valeri Kirischian. Run-time recovery mechanism for transient and permanent hardware faults based on distributed, self-organized dynamic partially reconfigurable systems. IEEE Trans. Computers, 65(9):2835–2847, 2016.

[122] Wenbin William Dai, Laurynas Riliskis, Peng Wang, et al. A cloud-based decision support system for self-healing in distributed automation systems using fault tree analysis. IEEE Trans. Ind. Informatics, 14(3):989–1000, 2018.

[123] Microsoft Azure. Azure 虚拟机自动修复（virtual machine automatic repairs）. Azure 官方文档, 2020. 可访问: https://learn.microsoft.com/.

[124] Igibek Koishybayev, Aleksandr Nahapetyan, Raima Zachariah, et al. Characterizing the security of github ci workflows. In 31st USENIX Security Symposium (USENIX Security 22), pages 2747–2763, 2022.

[125] Zachary Newman, John Speed Meyers, and Santiago Torres-Arias. Sigstore: Software signing for everybody. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, pages 2353–2367, 2022.

[126] Alexander Van't Hof and Jason Nieh. Blackbox : a container security monitor for protecting containers on untrusted operating systems. In 16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22), pages 683–700, 2022.

[127] Deepak Sirone Jegan, Liang Wang, Siddhant Bhagat, and Michael Swift. Guarding serverless applications with kalium. In 32nd USENIX Security Symposium (USENIX Security 23), pages 4087–4104, 2023.

[128] Maryam Rostamipoor, Seyedhamed Ghavamnia, and Michalis Polychronakis. Leakless: Selective data protection against memory leakage attacks for serverless platforms. In Proceedings of the Network and Distributed System Security Symposium (NDSS), San Diego, CA, 2025.

[129] Jiacheng Shi, Jinyu Gu, Yubin Xia, and Haibo Chen. Serverless functions made confidential and efficient with split containers. In 34th USENIX Security Symposium (USENIX Security 25), pages 1091–1110, 2025.

[130] Jeremie S. Kim, Minesh Patel, A. Giray Yağlıkçı, et al. Revisiting rowhammer: An experimental analysis of modern dram devices and mitigation techniques. In 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA), pages 638–651, 2020.

[131] Divyanshu Saxena, William Zhang, Shankara Pailoor, et al. Copper and wire: Bridging expressiveness and performance for service mesh policies. In Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1, pages 233–248, 2025.

[132] Yuejie Wang, Qiutong Men, Yongting Chen, et al. Heimdall: Towards risk-aware network management outsourcing. In NDSS, 2025.

[133] Mark Zhao, Mingyu Gao, and Christos Kozyrakis. Shef: Shielded enclaves for cloud fpgas. In Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, pages 1070–1085, 2022.

[134] Adil Ahmad, Alex Schultz, Byoungyoung Lee, and Pedro Fonseca. An extensible orchestration and protection framework for confidential cloud computing. In 17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23), pages 173–191, 2023.

[135] Md Abu Bakar Siddik, Arman Shehabi, and Landon Marston. The environmental footprint of data centers in the united states. Environmental Research Letters, 16(6):064017, 2021.

[136] Kostis Kaffes, Dragos Sbirlea, Yiyan Lin, et al. Leveraging application classes to save power in highly-utilized data centers. In SoCC '20: ACM Symposium on Cloud Computing, Virtual Event, USA, October 19-21, 2020, pages 134–149. ACM, 2020.

[137] Wenda Tang, Yutao Ke, Senbo Fu, et al. Demeter: Qos-aware cpu scheduling to reduce power consumption of multiple black-box workloads. In Proceedings of the 13th Symposium on Cloud Computing, pages 31–46, 2022.

[138] Shaohong Li, Xi Wang, Xiao Zhang, et al. Thunderbolt: Throughput-Optimized, Quality-of-Service-Aware power capping at scale. In 14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20), pages 1241–1255. USENIX Association, November 2020.

[139] Bilge Acun, Benjamin Lee, Fiodar Kazhamiaka, et al. Carbon explorer: A holistic framework for designing carbon aware datacenters. In Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2, ASPLOS 2023, Vancouver, BC, Canada, March 25-29, 2023, pages 118–132. ACM, 2023.

[140] Jovan Stojkovic, Chaojie Zhang, Íñigo Goiri, et al. Dynamollm: Designing LLM inference clusters for performance and energy efficiency. In IEEE International Symposium on High Performance Computer Architecture, HPCA 2025, Las Vegas, NV, USA, March 1-5, 2025, pages 1348–1362. IEEE, 2025.

[141] 中国电信集团公司. 云网融合 2030 技术白皮书. 2020.

[142] 中国电信集团公司. 云网融合 2035 技术白皮书. 2025.

[143] N. M. Mosharaf Kabir Chowdhury and Raouf Boutaba. A survey of virtual network embedding. Computer Networks, 54(5):862–876, 2010.

[144] Mosharaf Chowdhury, Yuan Zhong, and Ion Stoica. Efficient coflow scheduling with varys. In Proceedings of the 2014 ACM conference on SIGCOMM, pages 443–454, 2014.

[145] Albert Greenberg, James R Hamilton, Navendu Jain, et al. Vl2: A scalable and flexible data center network. In Proceedings of the ACM SIGCOMM 2009 conference on Data communication, pages 51–62, 2009.

[146] Hitesh Ballani, Paolo Costa, Thomas Karagiannis, and Antony Rowstron. Towards predictable datacenter networks. In ACM SIGCOMM, pages 242–253, 2011.

[147] Abhijit Gangidi et al. RDMA over Ethernet for Distributed AI Training at Meta Scale. In ACM SIGCOMM, 2024.

[148] Hao Wang, Han Tian, Jingrong Chen, et al. Towards domain-specific network transport for distributed dnn training. In USENIX Symposium on Networked Systems Design and Implementation (NSDI 2024), 2024.

[149] Mohammad Javed et al. Understanding and Optimizing Communication in Distributed Machine Learning Training. In USENIX ATC, 2022.

[150] Kun Qian, Yongqing Xi, Jiamin Cao, et al. Alibaba hpn: A data center network for large language model training. In Proceedings of the ACM SIGCOMM 2024 Conference, pages 691–706, 2024.

[151] Jiamin Cao, Yu Guan, Kun Qian, et al. Crux: Gpu-efficient communication scheduling for deep learning training. In Proceedings of the ACM SIGCOMM 2024 Conference, pages 1–15, 2024.

[152] Hong Liu et al. An integrated optimization method to task scheduling and vm placement in energy-efficient datacenters. Applied Soft Computing, 2024.

[153] 中国电信集团公司. 中国电信云网自智白皮书 2.0, 2022.

[154] 中国电信集团公司. 云算网一体化调度产品, 2024.

[155] 中国移动通信集团有限公司. 算网大脑白皮书. Technical report, 中国移动, 2022. “算网融合技术与产业白皮书 (2022)”发布.

[156] 中国联合网络通信有限公司研究院, 中国联合网络通信有限公司广东省分公司, and 华为技术有限公司. 云网融合向算网一体技术演进白皮书. Technical report, 中国联通, March 2021. 基于 CUBE-Net 3.0，总结云网融合实践并面向算网一体演进，强调云网/算网一体化资源与业务调度.

[157] 中国联合网络通信有限公司. 中国联通自智网络白皮书（2025 年）. Technical report, 中国联通, 2025. 提出“云网一体化资源调度”和“算网一体化编排调度-智枢平台”等理念，是联通面向自智网络阶段的云网/算网调度架构.

[158] Samyam Rajbhandari et al. ZeRO-Infinity: Breaking the GPU Memory Wall for Extreme Scale Deep Learning. In ACM SC, 2021.

[159] 中国信息通信研究院. 中国信息通信研究院：算力发展白皮书（2023）. https://www.caict.ac.cn/english/research/whitepapers/202311/P020231103309012315580.pdf, 2023.

[160] Baidu Research. Baidu allreduce: A high-performance ring-allreduce implementation. https://github.com/baidu-research/baidu-allreduce, 2017. Software repository, Baidu Research.

[161] Yimin Jiang, Yibo Zhu, Chang Lan, et al. A unified architecture for accelerating distributed {DNN} training in heterogeneous {GPU/CPU} clusters. In 14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20), pages 463–479, 2020.

[162] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, et al. Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053, 2019.

[163] Lianmin Zheng, Zhuohan Li, Hao Zhang, et al. Alpa: Automating inter-and {Intra-Operator} parallelism for distributed deep learning. In 16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22), pages 559–578, 2022.

[164] Yanping Huang, Youlong Cheng, Ankur Bapna, et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. Advances in neural information processing systems, 32, 2019.

[165] Deepak Narayanan, Aaron Harlap, Amar Phanishayee, et al. Pipedream: Generalized pipeline parallelism for dnn training. In Proceedings of the 27th ACM symposium on operating systems principles, pages 1–15, 2019.

[166] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, et al. Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668, 2020.

[167] 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.

[168] Samyam Rajbhandari, Conglong Li, Zhewei Yao, et al. Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale. In International conference on machine learning, pages 18332–18346. PMLR, 2022.

[169] Chenggang Zhao, Shangyan Zhou, Liyue Zhang, et al. Deepep: an efficient expert-parallel communication library. https://github.com/deepseek-ai/DeepEP, 2025.

[170] Narasimha R Adiga, Matthias A Blumrich, Dong Chen, et al. Blue gene/l torus interconnection network. IBM Journal of Research and Development, 49(2.3):265–276, 2005.

[171] John Kim, Wiliam J Dally, Steve Scott, and Dennis Abts. Technology-driven, highly-scalable dragonfly topology. ACM SIGARCH Computer Architecture News, 36(3):77–88, 2008.

[172] Yifan Zeng, Ruiting Zhou, Lei Jiao, and Renli Zhang. Online scheduling of edge multiple-model inference with dag structure and retraining. In IEEE INFOCOM 2025-IEEE Conference on Computer Communications, pages 1–10. IEEE, 2025.

[173] Aashaka Shah, Vijay Chidambaram, Meghan Cowan, et al. {TACCL}: Guiding collective algorithm synthesis using communication sketches. In 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23), pages 593–612, 2023.

[174] Biao Sun, Ziming Huang, Hanyu Zhao, et al. Llumnix: Dynamic scheduling for large language model serving. In 18th USENIX symposium on operating systems design and implementation (OSDI 24), pages 173–191, 2024.

[175] Sudarsanan Rajasekaran, Manya Ghobadi, and Aditya Akella. {CASSINI}:{Network-Aware} job scheduling in machine learning clusters. In 21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24), pages 1403–1420, 2024.

[176] Bichen Wang, Jingzhou Wang, Yu-E Sun, and He Huang. Copo: Joint cost and performance optimization for task placement in geo-distributed clouds. In 2025 IEEE 33rd International Conference on Network Protocols (ICNP), pages 1–12. IEEE, 2025.

[177] Yunfeng Zhao, Chao Qiu, Xiaoyun Shi, et al. Cur-coedge: Curiosity-driven collaborative request scheduling in edge-cloud systems. In IEEE INFOCOM 2024-IEEE Conference on Computer Communications, pages 411–420. IEEE, 2024.

[178] Rui Han, Shilin Wen, Chi Harold Liu, et al. Edgetuner: Fast scheduling algorithm tuning for dynamic edge-cloud workloads and resources. In IEEE INFOCOM 2022-IEEE Conference on Computer Communications, pages 880–889. IEEE, 2022.

[179] Marius Eriksen, Kaushik Veeraraghavan, Yusuf Abdulghani, et al. Global capacity management with flux. In 17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23), pages 589–606, 2023.

[180] Ziheng Jiang, Haibin Lin, Yinmin Zhong, et al. {MegaScale}: Scaling large language model training to more than 10,000 {GPUs}. In 21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24), pages 745–760, 2024.

[181] Maciej Besta and Torsten Hoefler. Slim fly: A cost effective low-diameter network topology. In SC'14: proceedings of the international conference for high performance computing, networking, storage and analysis, pages 348–359. IEEE, 2014.

[182] Nathan Farrington, George Porter, Sivasankar Radhakrishnan, et al. Helios: a hybrid electrical/optical switch architecture for modular data centers. In Proceedings of the ACM SIGCOMM 2010 Conference, pages 339–350, 2010.

[183] Kai Chen, Ankit Singla, Atul Singh, et al. Osa: An optical switching architecture for data center networks with unprecedented flexibility. IEEE/ACM Transactions on networking, 22(2):498–511, 2013.

[184] Radhika Mittal, Alexander Shpiner, Aurojit Panda, et al. Revisiting network support for rdma. In Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication, pages 313–326, 2018.

[185] Reto Bachmann et al. Multimae: Multi-modal masked autoencoders. In Advances in Neural Information Processing Systems (NeurIPS), 2022.

[186] Rohit Girdhar et al. Imagebind: One embedding space to bind them all. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[187] Lin Jiang et al. Reducto: Edge-based semantic compression for efficient video analytics. In USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2022.

[188] Tian Li et al. Filterforward: Semantic keyframe filtering for edge video analytics. In Proceedings of the Annual International Conference on Mobile Computing and Networking (MobiCom), 2022.

[189] Henry Corrigan-Gibbs and Dan Boneh. Prio: Private, robust, and verifiable aggregation for federated systems. In USENIX Security Symposium, 2017.

[190] Andre Kasper et al. Local differential privacy for mobile sensing: Theory and practice. In IEEE Symposium on Security and Privacy (S&P), 2023.

[191] Daniel Kang et al. Cova: Collaborative video analytics at the edge. In USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2022.

[192] Qinbin Li, Bingsheng He, et al. Model-contrastive federated learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[193] Canh T. Dinh, Nguyen H. Tran, and Wei Hong. Personalized federated learning with moreau envelopes. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

[194] Tian Li, Anit Sahu, Ameet Talwalkar, and Virginia Smith. Fedprox: Federated optimization in heterogeneous networks. In Proceedings of Machine Learning and Systems (MLSys), 2020.

[195] José Santos, Chen Wang, Tim Wauters, and Filip De Turck. Diktyo: Network-aware scheduling in container-based clouds. IEEE Trans. on Netw. and Serv. Manag., 20(4):4461–4477, December 2023.

[196] Harshit Saokar, Soteris Demetriou, Nick Magerko, et al. ServiceRouter: Hyperscale and Minimal Cost Service Mesh at Meta. 2023.

[197] Aurick Qiao, Sang Keun Choe, Suhas Jayaram Subramanya, et al. Pollux: Co-adaptive cluster scheduling for goodput-optimized deep learning. In 15th USENIX Symposium on Operating Systems Design and Implementation (OSDI 21), pages 1–18. USENIX Association, July 2021.

[198] H. Topcuoglu, S. Hariri, and Min-You Wu. Performance-effective and low-complexity task scheduling for heterogeneous computing. IEEE Transactions on Parallel and Distributed Systems, 13(3):260–274, 2002.

[199] Robert Grandl, Srikanth Kandula, Sriram Rao, et al. GRAPHENE: Packing and Dependency-Aware scheduling for Data-Parallel clusters. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), pages 81–97, Savannah, GA, November 2016. USENIX Association.

[200] Byungsoo Jeon, Mengdi Wu, Shiyi Cao, et al. Graphpipe: Improving performance and scalability of dnn training with graph pipeline parallelism. In Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1, ASPLOS '25, page 557–571, New York, NY, USA, 2025. Association for Computing Machinery.

[201] Mehrdad Khani, Manya Ghobadi, Mohammad Alizadeh, et al. Sip-ml: high-bandwidth optical network interconnects for machine learning training. In Proceedings of the 2021 ACM SIGCOMM 2021 Conference, SIGCOMM '21, page 657–675, New York, NY, USA, 2021. Association for Computing Machinery.

[202] Yangming Zhao, Kai Chen, Wei Bai, et al. Rapier: Integrating routing and scheduling for coflow-aware data center networks. In 2015 IEEE Conference on Computer Communications (INFOCOM), pages 424–432. IEEE, 2015.

[203] Ziyang Li, Yiming Zhang, Dongsheng Li, et al. Optas: Decentralized flow monitoring and scheduling for tiny tasks. In IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications, pages 1–9. IEEE, 2016.

[204] Shouxi Luo, Hongfang Yu, Yangming Zhao, et al. Towards practical and near-optimal coflow scheduling for data center networks. IEEE Transactions on Parallel and Distributed Systems, 27(11):3366–3380, 2016.

[205] Mosharaf Chowdhury and Ion Stoica. Efficient coflow scheduling without prior knowledge. ACM SIGCOMM Computer Communication Review, 45(4):393–406, 2015.

[206] Pitch Patarasuk and Xin Yuan. Bandwidth optimal all-reduce algorithms for clusters of workstations. Journal of Parallel and Distributed Computing, 69(2):117–124, 2009.

[207] Zixian Cai, Zhengyang Liu, Saeed Maleki, et al. Synthesizing optimal collective algorithms. In Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, pages 62–75, 2021.

[208] Jiamin Cao, Shangfeng Shi, Jiaqi Gao, et al. Syccl: Exploiting symmetry for efficient collective communication scheduling. In Proceedings of the ACM SIGCOMM 2025 Conference, pages 645–662, 2025.

[209] Yanghua Peng, Yibo Zhu, Yangrui Chen, et al. A generic communication scheduler for distributed dnn training acceleration. In Proceedings of the 27th ACM Symposium on Operating Systems Principles, pages 16–29, 2019.

[210] Kshiteej Mahajan, Ching-Hsiang Chu, Srinivas Sridharan, and Aditya Akella. Better together: Jointly optimizing {ML} collective scheduling and execution planning using {SYNDICATE}. In 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23), pages 809–824, 2023.

[211] Chang Chen, Xiuhong Li, Qianchao Zhu, et al. Centauri: Enabling efficient scheduling for communication-computation overlap in large model training via communication partitioning. In Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3, pages 178–191, 2024.

[212] Yihao Zhao, Yuanqiang Liu, Yanghua Peng, et al. Multi-resource interleaving for deep learning training. In Proceedings of the ACM SIGCOMM 2022 Conference, pages 428–440, 2022.

[213] Sudarsanan Rajasekaran, Sanjoli Narang, Anton A Zabreyko, and Manya Ghobadi. Mltcp: A distributed technique to approximate centralized flow scheduling for machine learning. In Proceedings of the 23rd ACM Workshop on Hot Topics in Networks, pages 167–176, 2024.

[214] Manaf Bin-Yahya, Amir Shani, Hossein Shafieirad, et al. Symphony: Collective coordination in multi-tenant gpu clusters. In 2025 IEEE 33rd International Conference on Network Protocols (ICNP), pages 1–13. IEEE, 2025.

[215] Mario Minardi, Shree Krishna Sharma, Symeon Chatzinotas, and Thang Xuan Vu. A parallel link mapping for virtual network embedding with joint load-balancing and energy-saving. In CNSM Workshop on Network Infrastructure, 2021.

[216] Ziyang Wang and et al. Learning-based sla-aware online virtual network embedding. Computer Communications, 2024.

[217] Yingying Guan, Qingyang Song, Weijing Qi, et al. Multidimensional resource fragmentation-aware virtual network embedding in mec systems interconnected by metro optical networks. arXiv preprint arXiv:2303.15878, 2023. BiVNE: joint node-link embedding in MEC + optical networks.

[218] Farzad Habibi and Juncheng Fang. Devine: A decentralized virtual network embedding algorithm. arXiv preprint arXiv:2502.01807, 2025.

[219] Ziyang Duan et al. Towards learning-based energy-efficient online virtual network embedding. Computer Communications, 215:74–85, 2024.

[220] Shuai Wang, Kaihui Gao, and Kun et al. Qian. Predictable vfabric on informative data plane. In ACM SIGCOMM, 2022.

[221] Satyajeet Singh Ahuja, Varun Gupta, Vinayak Dangui, et al. Capacity-efficient and uncertainty-resilient backbone network planning with hose. In Proceedings of the 2021 ACM SIGCOMM 2021 Conference, pages 547–559, 2021.

[222] John P Eason, Xueqi He, Richard Cziva, et al. Hose-based cross-layer backbone network design with benders decomposition. In Proceedings of the ACM SIGCOMM 2023 Conference, pages 333–345, 2023.

[223] Jie Wu, Shuaibing Lu, and Huanyang Zheng. On maximum elastic scheduling of virtual machines for cloud-based data center networks. In IEEE ICC, 2018.

[224] Yusuf Qwareeq, Abdalaziz Sawwan, and Jie Wu. Maximum elastic scheduling of virtual machines in general graph cloud data center networks. Cyber-Physical Systems, 10(3):283–301, 2024.

[225] Yingling Mao, Xiaojun Shang, and Yuanyuan Yang. Joint resource management and flow scheduling for sfc deployment in hybrid edge-and-cloud network. In IEEE INFOCOM 2022-IEEE Conference on Computer Communications, pages 170–179. IEEE, 2022.

[226] 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.

[227] Yingling Mao, Xiaojun Shang, and Yuanyuan Yang. Ant colony based online learning algorithm for service function chain deployment. In IEEE INFOCOM 2023-IEEE Conference on Computer Communications, pages 1–10. IEEE, 2023.

[228] Rasoul Behravesh, David Breitgand, Dean H Lorenz, and Danny Raz. A practical near optimal deployment of service function chains in edge-to-cloud networks. In IEEE INFOCOM 2024-IEEE Conference on Computer Communications, pages 751–760. IEEE, 2024.

[229] Hongyi Huang, Wenfei Wu, Yongchao He, and Zehua Guo. Sfp: Service function chain provision on programmable switches for cloud tenants. In 2022 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pages 1239–1249. IEEE, 2022.

[230] Yicen Liu and Junning Zhang. Service function chain embedding meets machine learning: Deep reinforcement learning approach. IEEE Transactions on Network and Service Management, 21(3):3465–3481, 2024.

[231] Danyang Zheng and Xiaojun Cao. Provably efficient service function chain embedding and protection in edge networks. IEEE/ACM Transactions on Networking, 2024.

[232] schedmd. Slurm workload manager, 2025.

[233] The Linux Foundation. Kubernetes v1.19 scheduling framework, 2025.

[234] Ali Ghodsi, Matei Zaharia, Benjamin Hindman, et al. Dominant resource fairness: Fair allocation of multiple resource types. In 8th USENIX symposium on networked systems design and implementation (NSDI 11), 2011.

[235] Matei Zaharia, Dhruba Borthakur, Joydeep Sen Sarma, et al. Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling. In Proceedings of the 5th European Conference on Computer Systems, EuroSys '10, page 265–278, New York, NY, USA, 2010. Association for Computing Machinery.

[236] Zhen Xie, Murali Emani, Xiaodong Yu, et al. Centimani: Enabling fast AI accelerator selection for DNN training with a novel performance predictor. In 2024 USENIX Annual Technical Conference (USENIX ATC 24), pages 1203–1221, Santa Clara, CA, July 2024. USENIX Association.

[237] Hamid Arabnejad and Jorge G Barbosa. List scheduling algorithm for heterogeneous systems by an optimistic cost table. IEEE transactions on parallel and distributed systems, 25(3):682–694, 2013.

[238] Weitao Wang and T. S. Eugene Ng. Söze: One Network Telemetry Is All You Need for Per-flow Weighted Bandwidth Allocation at Scale. 2025.

[239] Yuanli Wang, Lei Huang, Zikun Wang, et al. Capsys: Contention-aware task placement for data stream processing. In Proceedings of the Twentieth European Conference on Computer Systems, EuroSys '25, page 654–670, New York, NY, USA, 2025. Association for Computing Machinery.

[240] Neil G Duffield, Pawan Goyal, Albert Greenberg, et al. A flexible model for resource management in virtual private networks. In Proceedings of the ACM SIGCOMM, pages 95–108, 1999.

[241] Matthias Rost, Carlo Fuerst, and Stefan Schmid. Beyond the stars: Revisiting virtual cluster embeddings. In ACM SIGCOMM, pages 14–27, 2015.

[242] Mohammad Bany Taha, Yousef Sanjalawe, Ahmad Al-Daraiseh, et al. Proactive auto-scaling for service function chains in cloud computing based on deep learning. IEEE Access, 12:38575–38593, 2024.

[243] Heng Liao, Bingyang Liu, Xianping Chen, et al. Ub-mesh: a hierarchically localized nd-fullmesh datacenter network architecture. arXiv preprint arXiv:2503.20377, 2025.

[244] Pengfei Zuo, Huimin Lin, Junbo Deng, et al. Serving large language models on huawei cloudmatrix384. arXiv preprint arXiv:2506.12708, 2025.

[245] Hong Liu, Ryohei Urata, Kevin Yasumura, et al. Lightwave fabrics: at-scale optical circuit switching for datacenter and machine learning systems. In Proceedings of the ACM SIGCOMM 2023 Conference, pages 499–515, 2023.

[246] Chenchen Shou, Guyue Liu, Hao Nie, et al. Infinitehbd: building datacenter-scale high-bandwidth domain for llm with optical circuit switching transceivers. In Proceedings of the ACM SIGCOMM 2025 Conference, pages 1–23, 2025.

[247] Xudong Liao, Yijun Sun, Han Tian, et al. Mixnet: A runtime reconfigurable optical-electrical fabric for distributed mixture-of-experts training. In Proceedings of the ACM SIGCOMM 2025 Conference, pages 554–574, 2025.

[248] Zijian Li, Zixuan Chen, Yiying Tang, et al. Muse: A runtime incrementally reconfigurable network adapting to hpc real-time traffic. In 2024 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pages 765–779. IEEE, 2024.

[249] Peirui Cao, Shizhen Zhao, Dai Zhang, et al. Threshold-based routing-topology co-design for optical data center. IEEE/ACM Transactions on Networking, 31(6):2870–2885, 2023.

[250] Youjie Li, Iou-Jen Liu, Yifan Yuan, et al. Accelerating distributed reinforcement learning with in-switch computing. In Proceedings of the 46th International Symposium on Computer Architecture, pages 279–291, 2019.

[251] Amedeo Sapio, Marco Canini, Chen-Yu Ho, et al. Scaling distributed machine learning with {In-Network} aggregation. In 18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21), pages 785–808, 2021.

[252] ChonLam Lao, Yanfang Le, Kshiteej Mahajan, et al. {ATP}: In-network aggregation for multi-tenant learning. In 18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21), pages 741–761, 2021.

[253] Daniele De Sensi, Salvatore Di Girolamo, Saleh Ashkboos, et al. Flare: Flexible in-network allreduce. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, pages 1–16, 2021.

[254] WANGQL LIUS et al. Netreduce: Rdma-compatiblein-networkreductionfordistributed dnn training acceleration. arXiv preprint arXiv:2009.09736, 2020.

[255] Jin Fang, Hongli Xu, Gongming Zhao, et al. Accelerating distributed training with collaborative in-network aggregation. IEEE/ACM Transactions on Networking, 32(4):3437–3452, 2024.

[256] Bohan Zhao, Chang Liu, Jianbo Dong, et al. Enabling switch memory management for distributed training with in-network aggregation. In IEEE INFOCOM 2023-IEEE conference on computer communications, pages 1–10. IEEE, 2023.

[257] Hao Wang, Yuxuan Qin, ChonLam Lao, et al. Preemptive switch memory usage to accelerate training jobs with shared in-network aggregation. In 2023 IEEE 31st International Conference on Network Protocols (ICNP), pages 1–12. IEEE, 2023.

[258] Arjun Singhvi, Nandita Dukkipati, Prashant Chandra, et al. Falcon: A reliable, low latency hardware transport. In Proceedings of the ACM SIGCOMM 2025 Conference, pages 248–263, 2025.

[259] Gautam Kumar, Nandita Dukkipati, Keon Jang, et al. Swift: Delay is simple and effective for congestion control in the datacenter. In Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication, pages 514–528, 2020.

[260] Wenxue Li, Xiangzhou Liu, Yunxuan Zhang, et al. Revisiting rdma reliability for lossy fabrics. In Proceedings of the ACM SIGCOMM 2025 Conference, pages 85–98, 2025.

[261] Adithya Gangidi, Rui Miao, Shengbao Zheng, et al. Rdma over ethernet for distributed training at meta scale. In Proceedings of the ACM SIGCOMM 2024 Conference, pages 57–70, 2024.

[262] Kun Qian, Yongqing Xi, Jiamin Cao, et al. Alibaba hpn: A data center network for large language model training. In Proceedings of the ACM SIGCOMM 2024 Conference, pages 691–706, 2024.

[263] Serhat Arslan, Yuliang Li, Gautam Kumar, and Nandita Dukkipati. Bolt:{Sub-RTT} congestion control for {Ultra-Low} latency. In 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23), pages 219–236, 2023.

[264] Goyal Prateesh, Shah Preey, Zhao Kevin, et al. Backpressure flow control. In NSDI, 2022.

[265] Wei Bai, Shanim Sainul Abdeen, Ankit Agrawal, et al. Empowering azure storage with {RDMA}. In 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23), pages 49–67, 2023.

[266] Yanqing Chen, Chen Tian, Jiaqing Dong, et al. Swing: Providing long-range lossless rdma via pfc-relay. IEEE Transactions on Parallel and Distributed Systems, 34(1):63–75, 2022.

[267] Zirui Wan, Jiao Zhang, Yuzhen Su, et al. Re-architecting traffic control in cross-datacenter rdma networks. IEEE Transactions on Networking, 2025.

[268] Rui Miao, Lingjun Zhu, Shu Ma, et al. From luna to solar: the evolutions of the compute-to-storage networks in alibaba cloud. In Proceedings of the ACM SIGCOMM 2022 Conference, pages 753–766, 2022.

[269] Shuai Wang, Kaihui Gao, Kun Qian, et al. Predictable vfabric on informative data plane. In Proceedings of the ACM SIGCOMM 2022 Conference, pages 615–632, 2022.

[270] 人民网. “算网一体·智惠未来”天翼云·算力大会在渝举行. http://cq.people.com.cn/n2/2025/0906/c365407-41343799.html, 9 2025. 人民网重庆频道.

[271] 上海市经济和信息化委员会. 【向新攀高当尖兵产业报国铸担当】上海电信：“智云上海”开启城市智能新时代. https://www.sheitc.sh.gov.cn/dsxxjyzl/20251022/28d28f2e3c6948f59d28cbac6733e62c.html, 5 2024.

[272] Weiyang Wang, Moein Khazraee, Zhizhen Zhong, et al. {TopoOpt}: Co-optimizing network topology and parallelization strategy for distributed training jobs. In 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23), pages 739–767, 2023.

[273] Zixuan Chen, Xuandong Liu, Minglin Li, et al. Rina: Enhancing ring-allreduce with in-network aggregation in distributed model training. In 2024 IEEE 32nd International Conference on Network Protocols (ICNP), pages 1–12. IEEE, 2024.

[274] Weitao Wang, Masoud Moshref, Yuliang Li, et al. Poseidon: efficient, robust, and practical datacenter {CC} via deployable {INT}. In 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23), pages 255–274, 2023.

[275] Yiran Zhang, Qingkai Meng, Chaolei Hu, and Fengyuan Ren. Revisiting congestion control for lossless ethernet. In 21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24), pages 131–148, 2024.

[276] Shiwei Tan et al. Edgeformer: A compact multi-modal transformer for edge devices. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022.

[277] Fang Zeng et al. Mobileclip: Lightweight image-text alignment for on-device intelligence. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

[278] Abraham Mhaidli et al. Data quality governance in crowd sensing systems. IEEE Transactions on Mobile Computing, 2022.

[279] Jie Hu et al. Privacylens: Controllable privacy-preserving vision for edge ai. In Proceedings of the ACM Conference on Computer and Communications Security (CCS), 2022.

[280] Wenbin Liu, Hao Du, et al. Spatio-temporal pyramid-based multi-scale data completion in sparse crowd-sensing. IEEE Transactions on Mobile Computing, 2025. To appear.

[281] Google Research. Federated analytics: Collaborative data analysis without raw data sharing. Google AI Blog, 2022.

[282] Yujun Lin, Song Han, Huizi Mao, et al. Deep gradient compression: Reducing the communication bandwidth for federated learning. In International Conference on Learning Representations (ICLR), 2018.

[283] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-iid data. In Advances in Neural Information Processing Systems (NeurIPS), 2019.

[284] Rohan Pathak et al. Fedsplit: An algorithmic framework for federated optimization. In International Conference on Machine Learning (ICML), 2022.

[285] Chien-Chun Xie, Oluwasanmi Koyejo, and Indranil Gupta. Asynchronous federated learning. In IEEE International Conference on Big Data (BigData), 2019.

[286] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In Advances in Neural Information Processing Systems (NeurIPS), 2019.

[287] Robin Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. In NIPS Workshop on Private Multi-Party Machine Learning, 2017.

[288] Virat Shejwalkar and Amir Houmansadr. Flame: Robust federated learning against malicious clients. In USENIX Security Symposium, 2022.

[289] Yixin Zhang et al. Visa: Verifiable and secure aggregation for federated learning. In Proceedings of the ACM Conference on Computer and Communications Security (CCS), 2023.

[290] Apple. Device-based learning for privacy-preserving personalization. Apple Machine Learning Research, 2022.

[291] Yuyi Mao, Changsheng You, Jun Zhang, et al. A survey on mobile edge computing: The communication perspective. IEEE Communications Surveys & Tutorials, 19(4):2322–2358, 2017.

[292] Xu Chen, Li Jiao, Wenzhong Li, and Xiaoming Fu. Efficient multi-user computation offloading for mobile-edge cloud computing. IEEE/ACM Transactions on Networking, 24(5):2795–2808, 2016.

[293] Juan Liu, Yuyi Mao, Jun Zhang, and Khaled B. Letaief. Delay-optimal computation task scheduling for mobile-edge computing systems. In 2016 IEEE International Symposium on Information Theory (ISIT), pages 1451–1455, 2016.

[294] Yuyi Mao, Jun Zhang, and Khaled B Letaief. A deep reinforcement learning approach for dynamic resource allocation in edge computing. IEEE Transactions on Wireless Communications, 19(7):4093–4107, 2020.

[295] Wenqiang Li, Jiajia Li, Fang Wang, et al. Privacy-preserving edge computing: Opportunities and challenges. IEEE Internet of Things Journal, 7(5):4312–4324, 2020.

[296] Yutao Lu, Ying Liu, Lei Zhang, et al. Blockchain and federated learning for privacy-preserved data sharing in industrial iot. IEEE Transactions on Industrial Informatics, 16(6):4177–4186, 2020.

[297] Antonio Brogi, Sara Forti, Abdelkarim Ibrahim, and Jacopo Soldani. Qos-aware deployment of iot applications through the fog. IEEE Internet of Things Journal, 6(2):3585–3594, 2019.

[298] Ankush Jindal, Agostino Poggi, and Maurizio Tomaiuolo. An extensible framework for workflow orchestration in edge computing environments. Future Generation Computer Systems, 115:430–445, 2021.

[299] Zhiyuan Gao, Weihua Liang, Weiming Xu, et al. Optimal service function chain orchestration in edge computing and cloud computing hybrid system. IEEE Transactions on Network Science and Engineering, 7(4):2923–2937, 2020.

[300] Juncheng Li, Dong Lin, Chao Xu, et al. Reinforcement learning-based adaptive service orchestration for edge computing: A survey. ACM Computing Surveys, 55(10):1–36, 2023.

[301] Peng Wang, Xue Chen, Yong Wang, et al. Blockchain-based service function chain management in edge computing environments. IEEE Access, 8:113924–113936, 2020.

[302] Lei Xu, Hong Wang, Qi Li, et al. Blockchain-based secure and trustworthy service composition in edge computing. IEEE Transactions on Services Computing, 14(6):1824–1837, 2021.

[303] Yumeng Liang, Jianhui Chang, Mingyuan Zang, and Jie Wu. Rcnet: Resilient collaborative dnn inference for wireless networks with high packet loss. IEEE Transactions on Network Science and Engineering, 12(5):3694–3708, 2025.

[304] Luyao Gao, Jianchun Liu, Hongli Xu, et al. Accelerating end-cloud collaborative inference via near bubble-free pipeline optimization. In IEEE INFOCOM 2025 - IEEE Conference on Computer Communications, pages 1–10, 2025.

[305] Yubin Duan and Jie Wu. Optimizing job offloading schedule for collaborative dnn inference. IEEE Transactions on Mobile Computing, 23(4):3436–3451, 2023.

[306] Guanyu Xu, Zhiwei Hao, Yong Luo, et al. Devit: Decomposing vision transformers for collaborative inference in edge devices. IEEE Transactions on Mobile Computing, 23(5):5917–5932, 2023.

[307] Mingyue Zhao, Jiayi Shi, Zhengyuan Zhang, et al. C2f: Enabling context-aware edge-cloud collaborative inference for foundation models. In IEEE INFOCOM 2025-IEEE Conference on Computer Communications. IEEE, 2025.

[308] Zewei Xin, Qinya Li, Chaoyue Niu, et al. Adaptive routing of text-to-image generation requests between large cloud model and light-weight edge model. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 19482–19491, 2025.

[309] Google. Federated learning in the android ecosystem. Google AI Blog, 2022.

[310] Chao Liu, Yuxuan Chen, Xuefeng Zhang, and Min Zhao. Edge computing for autonomous driving: Opportunities and challenges. IEEE Network, 33(5):54–61, 2019.

[311] Weisong Shi, Jie Cao, Quan Zhang, et al. Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5):637–646, 2016.

[312] Mahadev Satyanarayanan. The emergence of edge computing. Computer, 50(1):30–39, 2017.

[313] Jun Zhang, Jun Wang, and Wen Yao. Joint task offloading and resource allocation in mobile edge computing networks with deep reinforcement learning. IEEE Internet of Things Journal, 8(13):10632–10644, 2021.

[314] Jin Du, Lin Zhao, Ji Feng, and Shuo Zhang. Computation offloading and resource allocation in mixed fog/cloud computing systems with min-max fairness guarantee. IEEE Transactions on Communications, 69(6):3861–3874, 2021.

[315] Jing Ren, Guanding Yu, Yuan He, and Guoqiang Li. Collaborative cloud and edge computing for latency minimization. IEEE Transactions on Vehicular Technology, 68(5):5031–5044, 2019.

[316] Hwan Jeong, Osvaldo Simeone, and Jihong Kang. Mobile edge computing via a uav-mounted cloudlet: Optimization of bit allocation and path planning. In IEEE Transactions on Vehicular Technology, volume 67, pages 2049–2063. IEEE, 2018.

[317] Yang Liu, Xiaojie Chen, and Zhikun Wu. Multi-agent deep reinforcement learning for computation offloading and resource allocation in multi-access edge computing. IEEE Access, 8:117313–117327, 2020.

[318] Yu Sun, Sheng Zhou, and Jie Xu. Emm: Energy-aware mobility management for mobile edge computing in ultra dense networks. IEEE Journal on Selected Areas in Communications, 38(2):486–501, 2020.

[319] Nicola Dragoni, Ivan Lanese, Søren Larsen, et al. Microservices: Yesterday, today, and tomorrow. Present and ulterior software engineering, pages 195–216, 2017.

[320] Jose Santos, David Macedo, Victor Silva, et al. Microservices orchestration in cloud environments: Survey and research challenges. Journal of Network and Computer Applications, 135:62–74, 2019.

[321] Chia-Mu Chang, Chia-Yu Lin, and Jiun-Hung Yu. Deep reinforcement learning-based service function chain orchestration in edge computing. IEEE Transactions on Network and Service Management, 19(2):1948–1960, 2022.

[322] 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.

[323] 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.

[324] Jianhui Chang. Generative image coding with diffusion prior. arXiv preprint arXiv:2509.13768, 2025.

[325] Jing Liu, Bohan Zhuang, Zhuangwei Zhuang, et al. Discrimination-aware network pruning for deep model compression. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(8):4035–4051, 2022.

[326] Linyi Jiang, Silvery D. Fu, Yifei Zhu, and Bo Li. Janus: Collaborative vision transformer under dynamic network environment. In IEEE INFOCOM 2025 - IEEE Conference on Computer Communications, pages 1–10, 2025.

[327] Xing Liu, Wei Yu, Fan Liang, et al. Toward deep transfer learning in industrial internet of things. IEEE Internet of Things Journal, 8(15):12163–12175, 2021.

[328] 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.

[329] Bufang Yang, Lixing He, Neiwen Ling, et al. Edgefm: Leveraging foundation model for open-set learning on the edge. In Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems, pages 111–124, 2023.

[330] Edward J Hu, Yelong Shen, Phillip Wallis, et al. Lora: Low-rank adaptation of large language models. ICLR, 1(2):3, 2022.

[331] Zhiqiang Cao, Yun Cheng, Zimu Zhou, et al. Edge-cloud collaborated object detection via bandwidth adaptive difficult-case discriminator. IEEE Transactions on Mobile Computing, 24(2):1181–1196, 2025.

[332] Gartner. Gartner2026 年十大战略技术趋势. 2025.

[333] 中华人民共和国国务院. 新一代人工智能发展规划. http://www.gov.cn/zhengce/content/2017-07/20/content$_5$211996.htm, 7 2017.

[334] Reinhard Diestel. Graph theory. Springer Nature, 2025.

[335] Cliff Click and Keith D Cooper. Combining analyses, combining optimizations. ACM Transactions on Programming Languages and Systems (TOPLAS), 17(2):181–196, 1995.

[336] 陈宝林. 最优化理论与算法. 清华大学出版社有限公司, 2005.

[337] Natallia Kokash. An introduction to heuristic algorithms. Department of Informatics and Telecommunications, 1:1–7, 2005.

[338] Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. ACM computing surveys, 56(9):1–36, 2024.

[339] Sudharshan Suresh, Haozhi Qi, Tingfan Wu, et al. Neuralfeels with neural fields: Visuotactile perception for in-hand manipulation. Science Robotics, 9(96):eadl0628, 2024.

[340] Yunwei Zhang, Jing Tian, and Qiaochu Xiong. A review of embodied intelligence systems: a three-layer framework integrating multimodal perception, world modeling, and structured strategies. Frontiers in Robotics and AI, 12:1668910, 2025.

[341] Jiajun Xi, Yinong He, Jianing Yang, et al. Teaching embodied reinforcement learning agents: Informativeness and diversity of language use. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4097–4114, 2024.

[342] Ye Liu, Peishan Huang, Fan Yang, et al. Quasyncfl: Asynchronous federated learning with quantization for cloud–edge–terminal collaboration enabled aiot. IEEE Internet of Things Journal, 11(1):59–69, 2023.

[343] John McCarthy, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon. A proposal for the dartmouth summer research project on artificial intelligence. AI magazine, 27(4):12–12, 1955.

[344] Stuart Russell, Peter Norvig, and Artificial Intelligence. A modern approach. Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs, 25(27):79–80, 1995.

[345] Edward A Feigenbaum et al. The art of artificial intelligence: Themes and case studies of knowledge engineering. 1977.

[346] Stephan R Sain. The nature of statistical learning theory, 1996.

[347] Christopher M Bishop and Nasser M Nasrabadi. Pattern recognition and machine learning, volume 4. Springer, 2006.

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

[349] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[350] Werner Vogels. Web services at amazon. com. In 2006 IEEE International Conference on Services Computing (SCC'06). IEEE Computer Society, 2006.

[351] 中华人民共和国国务院. 新一代人工智能发展规划. https://www.gov.cn/zhengce/content/2017-07/20/content$_5$211996.htm, 7 2017. 国发〔2017〕35 号.

[352] 工业和信息化部, 中央网络安全和信息化委员会办公室, 教育部, et al. 算力基础设施高质量发展行动计划. https://www.gov.cn/zhengce/zhengceku/202310/content$_6$907900.htm, 10 2023. 工信部联通信〔2023〕180 号.

[353] 中华人民共和国中央网络安全和信息化委员会办公室. 全球人工智能治理倡议. https://www.cac.gov.cn/2023-10/18/c$_1$699291032884978.htm, 10 2023. 中国网信网发布.

[354] Jaime Sevilla, Lennart Heim, Anson Ho, et al. Compute trends across three eras of machine learning. In 2022 international joint conference on neural networks (IJCNN), pages 1–8. IEEE, 2022.

[355] Mario Villamizar, Oscar Garcés, Harold Castro, et al. Evaluating the monolithic and the microservice architecture pattern to deploy web applications in the cloud. In 2015 10th computing colombian conference (10CCC), pages 583–590. IEEE, 2015.

[356] Marc Brooker, Tao Chen, and Fan Ping. Millions of tiny databases. In 17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20), pages 463–478, 2020.

[357] Yingnong Dang, Qingwei Lin, and Peng Huang. Aiops: real-world challenges and research innovations. In 2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), pages 4–5. IEEE, 2019.

[358] Qian Cheng, Doyen Sahoo, Amrita Saha, et al. Ai for it operations (aiops) on cloud platforms: Reviews, opportunities and challenges. arXiv preprint arXiv:2304.04661, 2023.

[359] Haoran Qiu, Weichao Mao, Archit Patke, et al. Reinforcement learning for resource management in multi-tenant serverless platforms. In Proceedings of the 2nd European Workshop on Machine Learning and Systems, pages 20–28, 2022.

[360] Guangyao Zhou, Wenhong Tian, Rajkumar Buyya, et al. Deep reinforcement learning-based methods for resource scheduling in cloud computing: A review and future directions. Artificial Intelligence Review, 57(5):124, 2024.

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

[362] Laurence A Wolsey. Integer programming. John Wiley & Sons, 2020.

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

[364] TN Kipf. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.

[365] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[366] Richard S Sutton, Andrew G Barto, et al. Reinforcement learning: An introduction, volume 1. MIT press Cambridge, 1998.

[367] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529–533, 2015.

[368] Zihan Yan, Dan Li, Li Chen, et al. From atop to zcube: Automated topology optimization pipeline and a highly cost-effective network topology for large model training. In Proceedings of the ACM SIGCOMM 2025 Conference, pages 861–881, 2025.

[369] Liangyu Zhao, Siddharth Pal, Tapan Chugh, et al. Efficient {Direct-Connect} topologies for collective communications. In 22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25), pages 705–737, 2025.

[370] Prithwish Basu, Liangyu Zhao, Jason Fantl, et al. Efficient all-to-all collective communication schedules for direct-connect topologies. In Proceedings of the 33rd International Symposium on High-Performance Parallel and Distributed Computing, pages 28–41, 2024.

[371] Sushant Jain, Alok Kumar, Subhasree Mandal, et al. B4: Experience with a globally-deployed software defined wan. ACM SIGCOMM Computer Communication Review, 43(4):3–14, 2013.

[372] Harsha Vardhan Simhadri, George Williams, Martin Aumüller, et al. Results of the neurips' 21 challenge on billion-scale approximate nearest neighbor search. In NeurIPS 2021 Competitions and Demonstrations Track, pages 177–189. PMLR, 2022.

[373] Wei Gao, Zhisheng Ye, Peng Sun, et al. Unisched: A unified scheduler for deep learning training jobs with different user demands. IEEE Transactions on Computers, 73(6):1500–1515, 2024.

[374] Haitao Yuan, Jing Bi, and MengChu Zhou. Geography-aware task scheduling for profit maximization in distributed green data centers. IEEE Transactions on Cloud Computing, 10(3):1864–1874, 2020.

[375] Qing Li, Shangguang Wang, Ao Zhou, et al. Qos driven task offloading with statistical guarantee in mobile edge computing. IEEE Transactions on Mobile Computing, 21(1):278–290, 2020.

[376] Haitao Yuan, Jing Bi, and MengChu Zhou. Spatiotemporal task scheduling for heterogeneous delay-tolerant applications in distributed green data centers. IEEE Transactions on Automation Science and Engineering, 16(4):1686–1697, 2019.

[377] Hongjian Shi, Weichu Zheng, Zifei Liu, et al. Automatic pipeline parallelism: A parallel inference framework for deep learning applications in 6g mobile communication systems. IEEE Journal on Selected Areas in Communications, 41(7):2041–2056, 2023.

[378] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[379] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078, 2014.

[380] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.

[381] Jitendra Kumar, Rimsha Goomer, and Ashutosh Kumar Singh. Long short term memory recurrent neural network (lstm-rnn) based workload forecasting model for cloud datacenters. Procedia computer science, 125:676–682, 2018.

[382] Bei Zhu, Jing Li, Rongbin Gu, and Liang Wang. An approach to cloud platform log anomaly detection based on natural language processing and lstm. In Proceedings of the 2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence, pages 1–7, 2020.

[383] Hang Su, Qian He, and Biao Guo. Kpi anomaly detection method for data center aiops based on gru-gan. In 2021 10th International Conference on Internet Computing for Science and Engineering, pages 23–29, 2021.

[384] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, pages 4171–4186, 2019.

[385] Alec Radford, Jeffrey Wu, Rewon Child, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

[386] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI conference on artificial intelligence, volume 35, pages 11106–11115, 2021.

[387] Peijun Zheng, Heng Zhou, Jiang Liu, and Yosuke Nakanishi. Interpretable building energy consumption forecasting using spectral clustering algorithm and temporal fusion transformers architecture. Applied Energy, 349:121607, 2023.

[388] Betsy Beyer, Chris Jones, Jennifer Petoff, and Niall Richard Murphy. Site reliability engineering: how Google runs production systems. " O'Reilly Media, Inc.", 2016.

[389] Honghua Chen, Xinyuan Qiu, Kelan Ren, and Xiaolong Cui. An aiops approach to data cloud based on large language models. In Proceedings of the 2024 4th International Conference on Artificial Intelligence, Big Data and Algorithms, pages 634–641, 2024.

[390] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, et al. Neural message passing for quantum chemistry. In International conference on machine learning, pages 1263–1272. Pmlr, 2017.

[391] Zhongxia Yan, Jingguo Ge, Yulei Wu, et al. Automatic virtual network embedding: A deep reinforcement learning approach with graph convolutional networks. IEEE Journal on Selected Areas in Communications, 38(6):1040–1057, 2020.

[392] Ahsan Shehzad, Feng Xia, Shagufta Abid, et al. Graph transformers: A survey. arXiv preprint arXiv:2407.09777, 2024.

[393] Farzad Habibi, Mahdi Dolati, Ahmad Khonsari, and Majid Ghaderi. Accelerating virtual network embedding with graph neural networks. In 2020 16th International Conference on Network and Service Management (CNSM), pages 1–9. IEEE, 2020.

[394] Yanghao Xie, Lin Huang, Yuyang Kong, et al. Virtualized network function forwarding graph placing in sdn and nfv-enabled iot networks: A graph neural network assisted deep reinforcement learning method. IEEE Transactions on Network and Service Management, 19(1):524–537, 2021.

[395] Jinwoo Park, Byungkwon Choi, Chunghan Lee, and Dongsu Han. Graf: A graph neural network based proactive resource allocation framework for slo-oriented microservices. In Proceedings of the 17th International Conference on emerging Networking EXperiments and Technologies, pages 154–167, 2021.

[396] Hanzhang Wang, Zhengkai Wu, Huai Jiang, et al. Groot: An event-graph-based approach for root cause analysis in industrial settings. In 2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE), pages 419–429. IEEE, 2021.

[397] Ruibo Chen, Jian Ren, Lingfeng Wang, et al. Microegrcl: An edge-attention-based graph neural network approach for root cause localization in microservice systems. In International Conference on Service-Oriented Computing, pages 264–272. Springer, 2022.

[398] Jiangbo Wang, Stéphane Zuckerman, and Juan Angel Lorenzo del Castillo. Evaluation of multi-armed bandit algorithms for efficient resource allocation in edge platforms. In Euro-Par 2024: Parallel Processing Workshops: Euro-Par 2024 International Workshops, Madrid, Spain, August 26–30, 2024, Proceedings, Part II, page 46–56, Berlin, Heidelberg, 2025. Springer-Verlag.

[399] Nikhil Shalikram Mondhe. Kubernetes Proactive Resources Scheduling using Multi-armed bandit Algorithm. PhD thesis, Dublin, National College of Ireland, 2024.

[400] Francesco Orabona. A modern introduction to online learning. arXiv preprint arXiv:1912.13213, 2019.

[401] Tor Lattimore and Csaba Szepesvári. Bandit algorithms. Cambridge University Press, 2020.

[402] Kevin Murphy. Reinforcement learning: an overview. arXiv preprint arXiv:2412.05265, 2024.

[403] Chen Tessler, Yuval Shpigelman, Gal Dalal, et al. Reinforcement learning for datacenter congestion control. ACM SIGMETRICS Performance Evaluation Review, 49(2):43–46, 2022.

[404] Zhaoyang Du, Chunrong Peng, Tsutomu Yoshinaga, and Celimuge Wu. A q-learning-based load balancing method for real-time task processing in edge-cloud networks. Electronics, 12(15), 2023.

[405] Yan Gu, Zhaoze Liu, Shuhong Dai, et al. Deep reinforcement learning for job scheduling and resource management in cloud computing: An algorithm-level review. arXiv preprint arXiv:2501.01007, 2025.

[406] Reyhane Ghafari and Najme Mansouri. Reinforcement learning-based solution for resource management in fog computing: A comprehensive survey. Expert Systems with Applications, page 127214, 2025.

[407] Deep Bodra and Sushil Khairnar. Machine learning-based cloud resource allocation algorithms: a comprehensive comparative review. Frontiers in Computer Science, 7:1678976, 2025.

[408] Ajay Kattepur, Sushanth David, and Swarup Kumar Mohalik. Model-based reinforcement learning for router port queue configurations. Intelligent and Converged Networks, 2(3):177–197, 2021.

[409] Omid Jafarzadeh, Mehdi Dehghan, Hadi Sargolzaey, and Mohammad Mehdi Esnaashari. A model-based reinforcement learning protocol for routing in vehicular ad hoc network. Wireless Personal Communications, 123(1):975–1001, 2022.

[410] Farzan Karami and Babak Hossein Khalaj. Model-based reinforcement learning approach for federated learning resource allocation and parameter optimization. Computer Communications, 228:107957, 2024.

[411] Armando Ordonez, Oscar Mauricio Caicedo, William Villota, et al. Model-based reinforcement learning with automated planning for network management. Sensors, 22(16):6301, 2022.

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

[413] Ravindra K Ahuja, Thomas L Magnanti, and James B Orlin. Network Flows: Theory, Algorithms, and Applications. Prentice Hall, 1993.

[414] Jon Kleinberg and Éva Tardos. Algorithm Design. Pearson, 2005.

[415] Christos H Papadimitriou and Kenneth Steiglitz. Combinatorial Optimization: Algorithms and Complexity. Dover, 1998.

[416] Laurence A. Wolsey. Integer Programming. Wiley, 1998.

[417] Bernhard Korte and Jens Vygen. Combinatorial Optimization: Theory and Algorithms. Springer, 2012.

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

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

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

[421] Dong Hock et al. Fat-tree: A scalable and efficient data center network. In Proceedings of IEEE INFOCOM Workshops, 2010.

[422] Xuting Liu, Behnaz Arzani, Siva Kesava Reddy Kakarla, et al. Rethinking machine learning collective communication as a multi-commodity flow problem. In Proceedings of the ACM SIGCOMM 2024 Conference, pages 16–37, 2024.

[423] Ruixing Zong, Jiapeng Zhang, Zhuo Tang, et al. Topology-aware interleaved all-reduce communication for dragonfly network. IEEE Transactions on Networking, 2025.

[424] Pablo Adasme, Ali Dehghan Firoozabadi, and Enrique San Juan. Bridging classic operations research and artificial intelligence for network optimization in the 6g era: A review. Symmetry, 17(8):1279, 2025.

[425] Martin Aumüller, Erik Bernhardsson, and Alexander Faithfull. Ann-benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. Information Systems, 87:101374, 2020.

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

[427] Junhua Zhang, Wentao Li, Long Yuan, et al. Shortest-path queries on complex networks: experiments, analyses, and improvement. Proceedings of the VLDB Endowment, 15(11):2640–2652, 2022.

[428] 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.

[429] Di Jin, Zhizhi Yu, Pengfei Jiao, et al. A survey of community detection approaches: From statistical modeling to deep learning. IEEE Transactions on Knowledge and Data Engineering, 35(2):1149–1170, 2021.

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

[431] Chunlin Li, Yihan Zhang, Zhiqiang Hao, and Youlong Luo. An effective scheduling strategy based on hypergraph partition in geographically distributed datacenters. Computer Networks, 170:107096, 2020.

[432] Luyao Luo, Gongming Zhao, Hongli Xu, et al. Achieving cost optimization for tenant task placement in geo-distributed clouds. IEEE/ACM Transactions on Networking, 32(2):1391–1406, 2023.

[433] Xiaobo Hao, Pengcheng Liu, and Yanhui Deng. Joint optimization of operational cost and carbon emission in multiple data center micro-grids. Frontiers in Energy Research, 12:1344837, 2024.

[434] Arezoo Jahani, Marco Lattuada, Michele Ciavotta, et al. Optimizing on-demand gpus in the cloud for deep learning applications training. In 2019 4th International Conference on Computing, Communications and Security (ICCCS), pages 1–8. IEEE, 2019.

[435] Abbas Kiani and Nirwan Ansari. Profit maximization for geographically dispersed green data centers. IEEE Transactions on Smart Grid, 9(2):703–711, 2016.

[436] Wenxin Li, Sheng Chen, Keqiu Li, et al. Efficient online scheduling for coflow-aware machine learning clusters. IEEE Transactions on Cloud Computing, 10(4):2564–2579, 2020.

[437] 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.

[438] 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.

[439] Wei Wang, Sheng Wang, Jinyang Gao, et al. Rafiki: Machine learning as an analytics service system. arXiv preprint arXiv:1804.06087, 2018.

[440] Pan Zhou, Xinshu He, Shouxi Luo, et al. Jpas: Job-progress-aware flow scheduling for deep learning clusters. Journal of Network and Computer Applications, 158:102590, 2020.

[441] Fei Xu, Jianian Xu, Jiabin Chen, et al. igniter: Interference-aware gpu resource provisioning for predictable dnn inference in the cloud. IEEE Transactions on Parallel and Distributed Systems, 34(3):812–827, 2022.

[442] Jürgen Schmidhuber. Deep learning in neural networks: An overview. Neural networks, 61:85–117, 2015.

[443] Jie Zhou, Ganqu Cui, Shengding Hu, et al. Graph neural networks: A review of methods and applications. AI open, 1:57–81, 2020.

[444] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. arXiv preprint arXiv:2012.09699, 2020.

[445] Sepideh Goodarzy, Maziyar Nazari, Richard Han, et al. Resource management in cloud computing using machine learning: A survey. In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 811–816. IEEE, 2020.

[446] Amer Al-Mazrawe and Bahaa Al-Musawi. Anomaly detection in cloud network: A review. In BIO Web of Conferences, volume 97, page 00019. EDP Sciences, 2024.

[447] Dexian Yang, Jiong Yu, Xusheng Du, et al. Energy saving strategy of cloud data computing based on convolutional neural network and policy gradient algorithm. Plos one, 17(12):e0279649, 2022.

[448] Soumyendu Sarkar, Antonio Guillen, Zachariah Carmichael, et al. Enhancing data center sustainability with a 3d cnn-based cfd surrogate model. In NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning, 2023.

[449] Thirumalai Selvan Chenni Chetty, Vadim Bolshev, Siva Shankar Subramanian, et al. Optimized hierarchical tree deep convolutional neural network of a tree-based workload prediction scheme for enhancing power efficiency in cloud computing. Energies, 16(6):2900, 2023.

[450] Shaojie Bai. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271, 2018.

[451] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks, 5(2):157–166, 1994.

[452] Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. International journal of forecasting, 37(4):1748–1764, 2021.

[453] Josh Achiam, Steven Adler, Sandhini Agarwal, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[454] Berend JD Gort, Godfrey M Kibalya, Maria A Serrano, and Angelos Antonopoulos. Forecasting trends in cloud-edge computing: Unleashing the power of attention mechanisms. IEEE Communications Magazine, 2024.

[455] Junjie Zha, Xinwen Shan, Jiaxin Lu, et al. Leveraging large language models for efficient alert aggregation in aiops. Electronics, 13(22):4425, 2024.

[456] Lingzhe Zhang, Tong Jia, Mengxi Jia, et al. A survey of aiops in the era of large language models. ACM Computing Surveys, 2025.

[457] Chengxuan Ying, Tianle Cai, Shengjie Luo, et al. Do transformers really perform badly for graph representation? Advances in neural information processing systems, 34:28877–28888, 2021.

[458] Guangyu Huo, Yong Zhang, Boyue Wang, et al. Hierarchical spatio–temporal graph convolutional networks and transformer network for traffic flow forecasting. IEEE Transactions on Intelligent Transportation Systems, 24(4):3855–3867, 2023.

[459] Prohim Tam, Inseok Song, Seungwoo Kang, et al. Graph neural networks for intelligent modelling in network management and orchestration: a survey on communications. Electronics, 11(20):3371, 2022.

[460] Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. In Proceedings of the AAAI conference on artificial intelligence, volume 34, pages 1234–1241, 2020.

[461] Ruijia Wang, Shuai Mou, Xiao Wang, et al. Graph structure estimation neural networks. In Proceedings of the web conference 2021, pages 342–353, 2021.

[462] Petar Veličković, William Fedus, William L Hamilton, et al. Deep graph infomax. arXiv preprint arXiv:1809.10341, 2018.

[463] Yu Rong, Yatao Bian, Tingyang Xu, et al. Self-supervised graph transformer on large-scale molecular data. Advances in neural information processing systems, 33:12559–12571, 2020.

[464] Jiawei Liu, Cheng Yang, Zhiyuan Lu, et al. Graph foundation models: Concepts, opportunities and challenges. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

[465] Shijie Wang, Jiani Huang, Zhikai Chen, et al. Graph machine learning in the era of large language models (llms). ACM Transactions on Intelligent Systems and Technology, 16(5):1–40, 2025.

[466] Bowen Jin, Gang Liu, Chi Han, et al. Large language models on graphs: A comprehensive survey. IEEE Transactions on Knowledge and Data Engineering, 2024.

[467] Huachi Zhou, Jiahe Du, Chuang Zhou, et al. Each graph is a new language: graph learning with llms. arXiv preprint arXiv:2501.11478, 2025.

[468] Antonio Galli, Vincenzo Moscato, Simon Pietro Romano, and Giancarlo Sperlí. Playing with a multi armed bandit to optimize resource allocation in satellite-enabled 5g networks. IEEE Trans. on Netw. and Serv. Manag., 21(1):341–354, February 2024.

[469] Hui Xia, Ning Huang, Xuecai Feng, et al. Starlet: Network defense resource allocation with multi-armed bandits for cloud-edge crowd sensing in iot. Digital Communications and Networks, 10(3):586–596, 2024.

[470] Brijen Thananjeyan, Kirthevasan Kandasamy, Ion Stoica, et al. Resource allocation in multi-armed bandit exploration: Overcoming sublinear scaling with adaptive parallelism. In International Conference on Machine Learning, 2021.

[471] Yongxin Xu, Shangshang Wang, Hengquan Guo, et al. Learning to schedule online tasks with bandit feedback. arXiv preprint arXiv:2402.16463, 2024.

[472] Qingsong Liu, Weihang Xu, Siwei Wang, and Zhixuan Fang. Combinatorial bandits with linear constraints: Beyond knapsacks and fairness. Advances in Neural Information Processing Systems, 35:2997–3010, 2022.

[473] Samarth Gupta, Jinhang Zuo, Carlee Joe-Wong, et al. Correlated combinatorial bandits for online resource allocation. In Proceedings of the Twenty-Third International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing, pages 91–100, 2022.

[474] Abhineet Agarwal, Anish Agarwal, Lorenzo Masoero, and Justin Whitehouse. Mutli-armed bandits with network interference. Advances in Neural Information Processing Systems, 37:36414–36437, 2024.

[475] Qingsong Liu and Zhixuan Fang. Decentralized scheduling with qos constraints: Achieving o (1) qos regret of multi-player bandits. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 13981–13989, 2024.

[476] Mengfan Xu and Diego Klabjan. Decentralized randomly distributed multi-agent multi-armed bandit with heterogeneous rewards. Advances in Neural Information Processing Systems, 36:74799–74855, 2023.

[477] S.K. Tesfatsion, E. Wadbro, and J. Tordsson. A combined frequency scaling and application elasticity approach for energy-efficient cloud computing. Sustainable Computing: Informatics and Systems, 4(4):205–214, 2014. Special Issue on Energy Aware Resource Management and Scheduling (EARMS).

[478] Yongchao Xiang, Zhen Liu, and Guoqiang Zhang. Cloud data centers energy-saving scheduling algorithm based on cpu frequency scaling.

[479] Arvindhan Muthusamy and Rajesh Kumar Dhanaraj. Dynamic q-learning-based optimized load balancing technique in cloud. Mobile Information Systems, 2023(1):7250267, 2023.

[480] Hussain Kahil, Shiva Sharma, Petri Välisuo, and Mohammed Elmusrati. Reinforcement learning for data center energy efficiency optimization: A systematic literature review and research roadmap. Applied Energy, 389:125734, 2025.

[481] Wenbing Yang, Mingqiang Zhao, Jingbo Li, and Xingjun Zhang. Energy-efficient dag scheduling with dvfs for cloud data centers. J. Supercomput., 80(10):14799–14823, March 2024.

[482] Shashank Swarup, Elhadi M. Shakshuki, and Ansar Yasar. Task scheduling in cloud using deep reinforcement learning. Procedia Computer Science, 184:42–51, 2021. The 12th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 4th International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops.

[483] Li Chen, Justinas Lingys, Kai Chen, and Xudong Liao. Datacenter traffic optimization with deep reinforcement learning. Communication Networks and Service Management in the Era of Artificial Intelligence and Machine Learning, pages 223–259, 2021.

[484] Yasar Sinan Nasir and Dongning Guo. Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks. IEEE Journal on selected areas in communications, 37(10):2239–2250, 2019.

[485] Nathan Jay, Noga Rotman, Brighten Godfrey, et al. A deep reinforcement learning perspective on internet congestion control. In International Conference on Machine Learning, pages 3050–3059. PMLR, 2019.

[486] Hao Ye, Geoffrey Ye Li, and Biing-Hwang Fred Juang. Deep reinforcement learning based resource allocation for v2v communications. IEEE Transactions on Vehicular Technology, 68(4):3163–3173, 2019.

[487] Juan Chen, Peng Chen, Xianhua Niu, et al. Task offloading in hybrid-decision-based multi-cloud computing network: a cooperative multi-agent deep reinforcement learning. Journal of Cloud Computing, 11(1):90, 2022.

[488] Akito Suzuki, Masahiro Kobayashi, and Eiji Oki. Multi-agent deep reinforcement learning for cooperative computing offloading and route optimization in multi cloud-edge networks. IEEE Transactions on Network and Service Management, 20(4):4416–4434, 2023.

[489] Andrzej Małota, Paweł Koperek, and Włodzimierz Funika. Towards understanding of deep reinforcement learning agents used in cloud resource management. In International Conference on Computational Science, pages 561–575. Springer, 2023.

[490] Lilian Weng. Llm-powered autonomous agents. lilianweng.github.io, Jun 2023.

[491] John McCarthy, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. AI magazine, 27(4):12–12, 2006.

[492] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, et al. Practices for governing agentic ai systems. Research Paper, OpenAI, 2023.

[493] Andrew Ng. The rise of agentic workflow. DeepLearning.AI, The Batch Newsletter, 2024. Accessed: 2025-01-28.

[494] Shunyu Yao, Jeffrey Zhao, Dian Yu, et al. React: Synergizing reasoning and acting in language models. In The eleventh international conference on learning representations, 2022.

[495] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, et al. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th annual acm symposium on user interface software and technology, pages 1–22, 2023.

[496] Alan Chan, Rebecca Salganik, Alva Markelius, et al. Harms from increasingly agentic algorithmic systems. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pages 651–666, 2023.

[497] Guohao Li, Hasan Hammoud, Hani Itani, et al. Camel: Communicative agents for" mind" exploration of large language model society. Advances in Neural Information Processing Systems, 36:51991–52008, 2023.

[498] Nuo Chen, Yan Wang, Haiyun Jiang, et al. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 8506–8520, Singapore, December 2023. Association for Computational Linguistics.

[499] Maciej Besta, Nils Blach, Ales Kubicek, et al. Graph of thoughts: Solving elaborate problems with large language models. In Proceedings of the AAAI conference on artificial intelligence, volume 38, pages 17682–17690, 2024.

[500] Nuo Chen, Hongguang Li, Jianhui Chang, et al. Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations. In Owen Rambow, Leo Wanner, Marianna Apidianaki, et al., editors, Proceedings of the 31st International Conference on Computational Linguistics, pages 755–773, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[501] Xinyan Guan, Jiali Zeng, Fandong Meng, et al. Deeprag: Thinking to retrieve step by step for large language models. arXiv preprint arXiv:2502.01142, 2025.

[502] Jae-Woo Choi, Hyungmin Kim, Hyobin Ong, et al. Reactree: Hierarchical task planning with dynamic tree expansion using llm agent nodes. 2025.

[503] Noah Shinn, Federico Cassano, Ashwin Gopinath, et al. Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems, 36:8634–8652, 2023.

[504] Nuo Chen, Hongguang Li, Baoyuan Wang, and Jia Li. From good to great: Improving math reasoning with tool-augmented interleaf prompting. In Bhavana Dalvi Mishra, Greg Durrett, Peter Jansen, et al., editors, Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2024), pages 64–79, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[505] Cheng Qian, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. Toolink: Linking toolkit creation and using through chain-of-solving on open-source model. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 831–854, 2024.

[506] Sirui Hong, Mingchen Zhuge, Jonathan Chen, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In The Twelfth International Conference on Learning Representations, 2023.

[507] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. Nature, 624(7992):570–578, 2023.

[508] Xiangru Tang, Anni Zou, Zhuosheng Zhang, et al. Medagents: Large language models as collaborators for zero-shot medical reasoning. In Findings of the Association for Computational Linguistics: ACL 2024, pages 599–621, 2024.

[509] Yuqi Zhu, Shuofei Qiao, Yixin Ou, et al. Knowagent: Knowledge-augmented planning for llm-based agents. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 3709–3732, 2025.

[510] Aman Madaan, Niket Tandon, Prakhar Gupta, et al. Self-refine: Iterative refinement with self-feedback. Advances in Neural Information Processing Systems, 36:46534–46594, 2023.

[511] Nuo Chen, Ning Wu, Jianhui Chang, et al. ControlMath: Controllable data generation promotes math generalist models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12201–12217, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[512] Ceyao Zhang, Kaijie Yang, Siyi Hu, et al. Proagent: building proactive cooperative agents with large language models. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 17591–17599, 2024.

[513] Zhibin Gou, Zhihong Shao, Yeyun Gong, et al. Critic: Large language models can self-correct with tool-interactive critiquing. In The Twelfth International Conference on Learning Representations.

[514] Shuyang Jiang, Yuhao Wang, and Yu Wang. Selfevolve: A code evolution framework via large language models. arXiv preprint arXiv:2306.02907, 2023.

[515] Xiao Liu, Hao Yu, Hanchen Zhang, et al. Agentbench: Evaluating llms as agents. In The Twelfth International Conference on Learning Representations.

[516] Yixing Jiang, Kameron C Black, Gloria Geng, et al. Medagentbench: Dataset for benchmarking llms as agents in medical applications. arXiv e-prints, pages arXiv–2501, 2025.

[517] Rafael Barbarroxa, Luis Gomes, and Zita Vale. Benchmarking large language models for multi-agent systems: A comparative analysis of autogen, crewai, and taskweaver. In International Conference on Practical Applications of Agents and Multi-Agent Systems, pages 39–48. Springer, 2024.

[518] Chen Qian, Wei Liu, Hongzhang Liu, et al. Chatdev: Communicative agents for software development. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15174–15186, 2024.

[519] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, et al. Voyager: An open-ended embodied agent with large language models. Transactions on Machine Learning Research, 2023.

[520] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33:9459–9474, 2020.

[521] Darren Edge, Ha Trinh, Newman Cheng, et al. From local to global: A graph rag approach to query-focused summarization. arXiv preprint arXiv:2404.16130, 2024.

[522] Mingxu Tao, Dongyan Zhao, and Yansong Feng. Chain-of-discussion: A multi-model framework for complex evidence-based question answering. In Proceedings of the 31st International Conference on Computational Linguistics, pages 11070–11085, 2025.

[523] Lei Wang, Wanyu Xu, Yihuai Lan, et al. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2609–2634, Toronto, Canada, July 2023. Association for Computational Linguistics.

[524] Mengkang Hu, Yao Mu, Xinmiao Yu, et al. Tree-planner: Efficient close-loop task planning with large language models. In ICLR, 2024.

[525] Rui Yang, Lin Song, Yanwei Li, et al. Gpt4tools: Teaching large language model to use tools via self-instruction. Advances in Neural Information Processing Systems, 36:71995–72007, 2023.

[526] Thibault Le Sellier de Chezelles, Maxime Gasse, Alexandre Lacoste, et al. The browsergym ecosystem for web agent research. Transactions on Machine Learning Research.

[527] Yidong Huang, Jacob Sansom, Ziqiao Ma, et al. Drivlme: Enhancing llm-based autonomous driving agents with embodied and social experiences. In 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 3153–3160. IEEE, 2024.

[528] Qingyun Wu, Gagan Bansal, Jieyu Zhang, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In First Conference on Language Modeling, 2024.

[529] Yilun Du, Shuang Li, Antonio Torralba, et al. Improving factuality and reasoning in language models through multiagent debate. In Forty-first International Conference on Machine Learning, 2023.

[530] Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, et al. Aflow: Automating agentic workflow generation. In The Thirteenth International Conference on Learning Representations.

[531] Zijun Liu, Yanzhe Zhang, Peng Li, et al. A dynamic llm-powered agent network for task-oriented agent collaboration. In First Conference on Language Modeling, 2024.

[532] Yixuan Weng, Minjun Zhu, Fei Xia, et al. Large language models are better reasoners with self-verification. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 2550–2575, 2023.

[533] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, et al. Self-rewarding language models. In Forty-first International Conference on Machine Learning.

[534] Chengdong Ma, Ziran Yang, Hai Ci, et al. Evolving diverse red-team language models in multi-round multi-agent games. arXiv preprint arXiv:2310.00322, 2023.

[535] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, et al. Webgpt: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332, 2021.

[536] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, et al. Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems, 36:68539–68551, 2023.

[537] Yifan Song, Weimin Xiong, Dawei Zhu, et al. Restgpt: Connecting large language models with real-world restful apis. arXiv preprint arXiv:2306.06624, 2023.

[538] Lifan Yuan, Yangyi Chen, Xingyao Wang, et al. Craft: Customizing llms by creating and retrieving from specialized toolsets. In 12th International Conference on Learning Representations, ICLR 2024, 2024.

[539] Vasilios Mavroudis. Langchain v0.3. 2024.

[540] Jerry Liu. LlamaIndex. 2022.

[541] Alex Leatherwood and Vic Matta. Building ai applications with dify. ai: A hands-on workshop. 2025.

[542] Abul Ehtesham, Aditi Singh, Gaurav Kumar Gupta, and Saket Kumar. A survey of agent interoperability protocols: Model context protocol (mcp), agent communication protocol (acp), agent-to-agent protocol (a2a), and agent network protocol (anp). arXiv preprint arXiv:2505.02279, 2025.

[543] Xiang Deng, Yu Gu, Boyuan Zheng, et al. Mind2web: Towards a generalist agent for the web. Advances in Neural Information Processing Systems, 36:28091–28114, 2023.

[544] Tianbao Xie, Danyang Zhang, Jixuan Chen, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. Advances in Neural Information Processing Systems, 37:52040–52094, 2024.

[545] Xiao Liu, Tianjie Zhang, Yu Gu, et al. Visualagentbench: Towards large multimodal models as visual foundation agents. In The Thirteenth International Conference on Learning Representations.

[546] Yunsheng Ma, Can Cui, Xu Cao, et al. Lampilot: An open benchmark dataset for autonomous driving with language model programs. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 15141–15151, 2024.

[547] Yuge Zhang, Qiyang Jiang, XingyuHan XingyuHan, et al. Benchmarking data science agents. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5677–5700, 2024.

[548] Frank F Xu, Yufan Song, Boxuan Li, et al. Theagentcompany: benchmarking llm agents on consequential real world tasks. arXiv preprint arXiv:2412.14161, 2024.

[549] Matthew Kenney. Ml research benchmark. arXiv preprint arXiv:2410.22553, 2024.

[550] Junlin Xie, Zhihong Chen, Ruifei Zhang, et al. Large multimodal agents: A survey. arXiv preprint arXiv:2402.15116, 2024.

[551] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. View in Article, 2(5):1, 2023.

[552] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In International Conference on Machine Learning, pages 9118–9147. PMLR, 2022.

[553] Jacky Liang, Wenlong Huang, Fei Xia, et al. Code as policies: Language model programs for embodied control. In IEEE International Conference on Robotics and Automation. IEEE, 2023.

[554] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35:23716–23736, 2022.

[555] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In The International Conference on Learning Representations, pages 19730–19742. PMLR, 2023.

[556] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3372–3382, 2021.

[557] Wenliang Dai, Junnan Li, Dongxu Li, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems, 36:49250–49267, 2023.

[558] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft Coco: Common objects in context. In European Conference on Computer Vision, pages 740–755. Springer, 2014.

[559] Rowan Zellers, Jiasen Lu, Ximing Lu, et al. Merlot reserve: Neural script knowledge through vision and language and sound. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 16375–16387, 2022.

[560] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1728–1738, 2021.

[561] Vishnu Sashank Dorbala, Gunnar A Sigurdsson, Jesse Thomason, et al. Clip-nav: Using clip for zero-shot vision-and-language navigation. In Workshop on Language and Robotics at CoRL 2022.

[562] Vishnu Sashank Dorbala, James F Mullen, and Dinesh Manocha. Can an embodied agent find your "cat-shaped mug" ? llm-based zero-shot object navigation. IEEE Robotics and Automation Letters, 9(5):4083–4090, 2023.

[563] Kiana Ehsani, Winson Han, Alvaro Herrasti, et al. Manipulathor: A framework for visual object manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4497–4506, 2021.

[564] Mayank Mittal, Calvin Yu, Qinxi Yu, et al. Orbit: A unified simulation framework for interactive robot learning environments. IEEE Robotics and Automation Letters, 8(6):3740–3747, 2023.

[565] C Karen Liu and Dan Negrut. The role of physics-based simulators in robotics. Annual Review of Control, Robotics, and Autonomous Systems, 4(1):35–58, 2021.

[566] Tuomas Haarnoja, Ben Moran, Guy Lever, et al. Learning agile soccer skills for a bipedal robot with deep reinforcement learning. Science Robotics, 9(89):eadi8022, 2024.

[567] Daniel Ho, Kanishka Rao, Zhuo Xu, et al. Retinagan: An object-aware approach to sim-to-real transfer. In IEEE International Conference on Robotics and Automation, pages 10920–10926. IEEE, 2021.

[568] Tongzhou Mu, Zhan Ling, Fanbo Xiang, et al. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

[569] Yu-Zhu Sun, He-Li Sun, Jian-Cong Ma, et al. Multimodal agent ai: A survey of recent advances and future directions. Journal of Computer Science and Technology, 40(4):1046–1063, 2025.

[570] Elias Dritsas, Maria Trigka, Christos Troussas, and Phivos Mylonas. Multimodal interaction, interfaces, and communication: a survey. Multimodal Technologies and Interaction, 9(1):6, 2025.

[571] Caelan Reed Garrett, Rohan Chitnis, Rachel Holladay, et al. Integrated task and motion planning. Annual review of control, robotics, and autonomous systems, 4(1):265–293, 2021.

[572] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Bev-guided multi-modality fusion for driving perception. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21960–21969, 2023.

[573] Jacob Krantz, Shurjo Banerjee, Wang Zhu, et al. Iterative vision-and-language navigation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14921–14930, 2023.

[574] Yuli Wang, Yuwei Dai, Craig Jones, et al. Enhancing vision-language models for medical imaging: bridging the 3d gap with innovative slice selection. volume 37, pages 99947–99964, 2024.

[575] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. IEEE transactions on pattern analysis and machine intelligence, 46(8):5625–5644, 2024.

[576] QuantumBlack, AI by McKinsey. The state of ai in 2025: Agents, innovation, and transformation, November 2025. Forthcoming. Available upon release at the QuantumBlack insights page.

[577] Sam Altman. Reflections. https://blog.samaltman.com/reflections, 1 2025.

[578] Microsoft AI. Towards Humanist Superintelligence. https://microsoft.ai/news/towards-humanist-superintelligence/, 11 2025.

[579] Di Cooke, Abigail Edwards, Sophia Barkoff, and Kathryn Kelly. As good as a coin toss: Human detection of ai-generated content. Communications of the ACM, 68(10):100–109, 2025.

[580] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, et al. Managing extreme ai risks amid rapid progress. Science, 384(6698):842–845, 2024.

[581] C114 通信网. 5g 赋能柳钢热轧厂智能升级：中国电信工业互联网平台助力传统制造业转型. 新闻网站, 1 月 2025.

[582] Microsoft. Microsoft named a leader in the 2025 gartner® magic quadrant for global industrial iot platforms. Azure 博客, 9 月 2025.

[583] yicaiai.com. 【科技前沿】亚马逊云科技 2025 年 re:Invent 大会首日亮点全景解析. 澎湃新闻, 12 月 2 日 2025. 访问日期: 2025-12-04.

[584] 李礼 and 贾川民. 智能图像视频编码的未来发展之路. 计算, 1(7):69–77, 2025.

[585] Andreas Haas, Andreas Rossberg, Derek L Schuff, et al. Bringing the web up to speed with webassembly. In Proceedings of the 38th ACM SIGPLAN conference on programming language design and implementation, pages 185–200, 2017.

[586] 中国电信翼支付. 甜橙区块链服务：密流安全计算平台. 中国电信翼支付平台官方, 2025. 可访问: https://blockchain.bestpay.com.cn/fmpc.html.

[587] 阿里云团队. 阿里云机密计算能力. 阿里云官方文档, 2025. 可访问: https://www.alibabacloud.com/help/zh/ecs/user-guide/trusted-computing-and-confidential-computing/.

[588] Microsoft Azure. Azure confidential computing). 微软官方文档. 可访问: https://learn.microsoft.com/zh-cn/azure/confidential-computing/.

[589] Google Cloud. Google cloud confidential computing 概览. Google Cloud 官方文档, 2025. 可访问: https://docs.cloud.google.com/confidential-computing/docs/confidential-computing-overview.

[590] 张晓兰 and 黄伟熔. 低空经济发展的全球态势、我国现状及促进策略. 经济纵横, (8):53–62, 2024.

[591] 中共中央关于制定国民经济和社会发展第十五个五年规划的建议. 中国共产党中央委员会. https://www.gov.cn/zhengce/202510/content$_7$046050.htm, 10 2025. 新华社发布.

[592] 中国航空学会. 2024 低空经济场景白皮书. https://www.stdaily.com/web/gdxw/2024-10/23/content$_2$47065.html, 10 2024. 在第七届中国航空科学技术大会上发布，科技日报报道.

[593] Nokia. Nokia to revolutionize mobile networks with cloud ran and ai-powered by nvidia. https://www.nokia.com/newsroom/nokia-partners-with-nvidia/, 2025.

[594] 中国电信. 全球首创！中国电信实现百公里级量子-经典空芯共纤传输重大突破. 中国电信官网新闻, 11 2025.

[595] 工业和信息化部教育部科学技术部交通运输部文化和旅游部 国务院国有资产监督管理委员会中国科学院. 关于推动未来产业创新发展的实施意见, 01 2024.

[596] 工业和信息化部办公厅教育部办公厅文化和旅游部办公厅 国务院国资委办公厅国家广播电视总局办公厅. 元宇宙产业创新发展三年行动计划（2023—2025 年）, 08 2023.

[597] 工业和信息化部科技部 国家能源局国家标准化管理委员会. 新产业标准化领航工程实施方案（2023—2035 年）, 08 2023.

[598] Hongyu Li, Liwei Guo, Yexuan Yang, et al. An empirical study of {Rust-for-Linux}: The success, dissatisfaction, and compromise. In 2024 USENIX Annual Technical Conference (USENIX ATC 24), pages 425–443, 2024.

[599] Ruolin Xing, Mengwei Xu, Ao Zhou, et al. Deciphering the enigma of satellite computing with cots devices: Measurement and analysis. In Proceedings of the 30th Annual International Conference on Mobile Computing and Networking, pages 420–435, 2024.

[600] Qing Li, Shangguang Wang, Chenren Xu, et al. Exploring real-time satellite computing: From energy and thermal perspectives. In 2024 IEEE Real-Time Systems Symposium (RTSS), pages 161–173. IEEE, 2024.

[601] Qing Li, Shangguang Wang, Xiao Ma, et al. Battery-aware energy optimization for satellite edge computing. IEEE Transactions on Services Computing, 17(2):437–451, 2024.

[602] Qiyang Zhang, Xin Yuan, Ruolin Xing, et al. Resource-efficient in-orbit detection of earth objects. In IEEE INFOCOM 2024-IEEE Conference on Computer Communications, pages 551–560. IEEE, 2024.

[603] Abdelkader Mekrache, Adlen Ksentini, and Christos Verikoukis. Next-generation 6g network management with oss-gpt. In Proceedings of the ACM SIGCOMM 2025 Posters and Demos, pages 158–160. 2025.

[604] Weijun Wang, Liang Mi, Shaowei Cen, et al. Region-based content enhancement for {Efficient} video analytics at the edge. In 22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25), pages 613–633, 2025.

[605] Oscar Adamuz-Hinojosa, Lanfranco Zanzi, Vincenzo Sciancalepore, et al. Oranus: Latency-tailored orchestration via stochastic network calculus in 6g o-ran. In IEEE INFOCOM 2024-IEEE Conference on Computer Communications, pages 61–70. IEEE, 2024.

[606] Mary Hogan, Gerry Wan, Yiming Qiu, et al. Efficient {Multi-WAN} transport for 5g with {OTTER}. In 22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25), pages 1243–1267, 2025.

[607] Mattia Bevilacqua, Marco Mezzavilla, Eugenio Moro, et al. Ecoran: A novel programmable framework for dynamic and energy-efficient resource optimization in multi-tenant, neutral host o-ran systems. In Proceedings of the 1st Workshop on Open Research Infrastructures and Toolkits for 6G, pages 27–32, 2025.

[608] Chang Liu, TD Khoa Le, Rahul Saini, et al. Vota: Parallelizing 6g-ran experimentation with virtualized over-the-air workloads. In Proceedings of the 1st Workshop on Open Research Infrastructures and Toolkits for 6G, pages 15–20, 2025.

[609] Akihiro Nakao. In-band context signaling for cross-layer qos in software-defined local 6g. In Proceedings of the 1st Workshop on Open Research Infrastructures and Toolkits for 6G, pages 1–6, 2025.

[610] Junwei Zhao, Qianchun Luo, Shiliang Zhang, et al. HDCFN: Haze distribution-aware cross-modal fusion network for infrared-guided dense haze removal in uavs. In Proceedings of the 33rd ACM International Conference on Multimedia, pages 10867–10875, 2025.

[611] Wang Zhang, Chunsheng Liu, Faliang Chang, and Ye Song. Multi-scale and occlusion aware network for vehicle detection and segmentation on uav aerial images. Remote Sensing, 12(11):1760, 2020.

[612] Fan Liu, Delong Chen, Zhangqingyun Guan, et al. Remoteclip: A vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 62:1–16, 2024.

[613] Oleg Sautenkov, Yasheerah Yaqoot, Artem Lykov, et al. Uav-vla: Vision-language-action system for large scale aerial mission generation. In 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pages 1588–1592. IEEE, 2025.

[614] Jiazhi Chen, Xianbin Wang, and Xuemin Shen. Goal-driven trusted collaborator selection and task offloading in dynamic collaborative systems. IEEE Internet of Things Journal, 2024.

[615] Chao Wang, Yiran Zhang, Qing Li, et al. Satellite computing: A case study of cloud-native satellites. In 2023 IEEE International Conference on Edge Computing and Communications (EDGE), pages 262–270. IEEE, 2023.

[616] Shangguang Wang and Qing Li. Satellite computing: Vision and challenges. IEEE Internet of Things Journal, 10(24):22514–22529, 2023.

[617] NASA Jet Propulsion Laboratory. High-performance spaceflight computing (hpsc) project overview. Technical report, NASA, 2023.

[618] A. Ahmed and B. Rinner. Distributed computing in space: A survey of leo constellation architectures. Ad Hoc Networks, 130:102795, 2022.

[619] Daniel P. Mandl, Steve Miller, et al. Sensorweb and event-driven onboard processing for earth-observation satellites. Acta Astronautica, 161:66–78, 2019.

[620] HSBC. Hsbc demonstrates world’s first-known quantum-enabled algorithmic trading with ibm, 2025. Accessed: 2025-11-27.

[621] Fujitsu. Fujitsu’s quantum simulator challenge 2024, 2024. Accessed: 2025-11-27.

[622] ITU-R WP5D. Imt-2030 (6g) framework and overall objectives. https://www.itu.int, 2023. International Telecommunication Union Recommendation.

[623] 3GPP. 3gpp tr 38.821: Study on non-terrestrial networks (ntn) in nr. https://www.3gpp.org, 2024.

[624] China Telecom Research Institute. 6g network architecture white paper. https://www.cttl.org.cn, 2023. 中国电信 6G 网络架构白皮书.

[625] IMT-2030 (6G) Promotion Group. 6g overall development report. https://www.caict.ac.cn, 2023. 工信部 IMT-2030 推进组白皮书.

[626] Z. Zhang, Y. Xiao, Z. Ma, and M. Xiao. 6g wireless networks: Vision, requirements, architecture, and key technologies. IEEE Communications Surveys & Tutorials, 23(3):1341–1397, 2021.

[627] 3GPP. Study on artificial intelligence (ai)/machine learning (ml) for nr air interface (tr 38.843). Technical report, 3GPP, 2022.

[628] O-RAN Alliance. Research report on cross-domain ai. https://www.o-ran.org/research-reports/research-report-on-cross-domain-ai-rr-2024-02, 2024.

[629] ETSI ZSM Group. Gr zsm 015 —deployed ai model assessment and model management (gr_zsm015 v1.1.1). Technical report, ETSI, 2024.

[630] Next G Alliance. Sustainable ai in telecom: Promises and challenges in 6g. https://nextgalliance.org/white$_p$apers/sustainable-ai-in-telecompromises-and-challenges-in-6g/, 2024.

[631] Hexa-X II Consortium. Hexa-x ii deliverable d1.4 —6g value, requirements and ecosystem. Technical report, Hexa-X II, 2025.

[632] R. K. Singh, E. Hossain, F. Murshed, and M. B. I. Reaz. Energy consumption analysis of lpwan technologies and lifetime estimation for iot devices. Sensors, 20(17):4794, 2020.

[633] V. Bonilla, R. Forno, L. Muñoz, et al. A systematic review of lorawan: Sensors, deployments and applications. Sensors (Open Access) / PubMed Central, 2023.

[634] M. Jouhari, A. Ksentini, and A. Ben Dhaou. A survey on scalable lorawan for massive iot. arXiv preprint, 2022.

[635] U. Raza, P. Kulkarni, and M. Sooriyabandara. Low power wide area networks: An overview. In IEEE Communications Surveys & Tutorials (overview paper / arXiv/tech report versions available), 2017. Overview of LPWAN design objectives, key technologies (LoRa, Sigfox, NB-IoT) and open challenges.

[636] H. Rajab, A. Al-Dubai, et al. Evaluation of energy consumption of lpwan technologies: A lorawan case study and model. EURASIP Journal on Wireless Communications and Networking, 2023.

[637] 数字低空工作组. 低空经济场景应用与通信需求白皮书, 4 2025. 全球 6G 技术与产业生态大会上发布.

[638] Sijie Wang, Siqi Li, Yawei Zhang, et al. Uavscenes: A multi-modal dataset for uavs. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 28946–28958, 2025.

[639] Jinyuan Liu, Bowei Zhang, Qingyun Mei, et al. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 2226–2235, 2025.

[640] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 779–788, 2016.

[641] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems, volume 28, pages 91–99, 2015.

[642] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2021.

[643] Kaiming He, Xinlei Chen, Saining Xie, et al. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15979–15988, 2022.

[644] Yezhen Cong, Samar Khanna, Chenlin Meng, et al. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. In Advances in Neural Information Processing Systems, volume 35, 2022.

[645] Fan Liu, Delong Chen, Zhangqingyun Guan, et al. Remoteclip: A vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 62:1–16, 2024.

[646] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pages 8748–8763. PMLR, 2021.

[647] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In Advances in Neural Information Processing Systems, volume 36, 2023.

[648] Peter E Hart, Nils J Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. IEEE transactions on Systems Science and Cybernetics, 4(2):100–107, 1968.

[649] Xin Zhou, Zhepei Wang, Hongkai Ye, et al. Ego-planner: An esdf-free gradient-based local planner for quadrotors. IEEE Robotics and Automation Letters, 6(2):478–485, 2021.

[650] Jesús Tordesillas, Brett T. Lopez, Michael Everett, and Jonathan P. How. Faster: Fast and safe trajectory planner for navigation in unknown environments. IEEE Transactions on Robotics, 38(2):922–938, 2022.

[651] Angel Romero, Sihao Sun, Philipp Foehn, and Davide Scaramuzza. Model predictive contouring control for time-optimal quadrotor flight. IEEE Transactions on Robotics, 38(6):3340–3356, 2022.

[652] Angel Romero, Sihao Sun, Philipp Foehn, and Davide Scaramuzza. Model predictive contouring control for time-optimal quadrotor flight. IEEE Transactions on Robotics, 38(6):3340–3356, 2022.

[653] Jesús Tordesillas and Jonathan P. How. Mader: Trajectory planner in multiagent and dynamic environments. IEEE Transactions on Robotics, 38(1):463–476, 2022.

[654] Baskın Şenbaşlar and Gaurav S Sukhatme. Dream: Decentralized real-time asynchronous probabilistic trajectory planning for collision-free multi-robot navigation in cluttered environments. IEEE Transactions on Robotics, 2024.

[655] Iván Maza, Jesús Capitán, Luis Merino, and Anıbal Ollero. Multi-uav cooperation. Unmanned Aircraft Systems, page 347, 2016.

[656] Changle Li, Mengqiu Tian, Yilong Hui, et al. On-demand environment perception and resource allocation for task offloading in vehicular networks. IEEE Transactions on Wireless Communications, 2024.

[657] Yingchao Wang, Chen Yang, Shulin Lan, et al. End-edge-cloud collaborative computing for deep learning: A comprehensive survey. IEEE Communications Surveys & Tutorials, 26(4):2647–2683, 2024.

[658] James Flemings, Bo Jiang, Wanrong Zhang, et al. Estimating privacy leakage of augmented contextual knowledge in language models. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25092–25108, 2025.

[659] Zhan Li, Yongtao Wu, Yihang Chen, et al. Membership inference attacks against large vision-language models. Advances in Neural Information Processing Systems, 37:98645–98674, 2024.

[660] Wen-jie Lu, Zhicong Huang, Qizhi Zhang, et al. Squirrel: a scalable secure {Two-Party} computation framework for training gradient boosting decision tree. In 32nd USENIX Security Symposium (USENIX Security 23), pages 6435–6451, 2023.

[661] Martin Abadi, Andy Chu, Ian Goodfellow, et al. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pages 308–318, 2016.

[662] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, et al. Machine unlearning. In 2021 IEEE symposium on security and privacy (SP), pages 141–159. IEEE, 2021.

[663] Junlin Liu, Xinchen Lyu, Qimei Cui, and Xiaofeng Tao. Similarity-based label inference attack against training and inference of split learning. IEEE Transactions on Information Forensics and Security, 19:2881–2895, 2024.

[664] Yuzheng Hu, Fan Wu, Qinbin Li, et al. Sok: Privacy-preserving data synthesis. In 2024 IEEE Symposium on Security and Privacy (SP), pages 4696–4713. IEEE, 2024.

[665] Xinyu He, Dongqi Fu, Hanghang Tong, et al. Temporal heterogeneous graph generation with privacy, utility, and efficiency. In The Thirteenth International Conference on Learning Representations, 2025.

[666] James Liyuan Wang, Ran Li, Junfeng Yang, and Chengzhi Mao. Raft: Realistic attacks to fool text detectors. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 16923–16936, 2024.

[667] Iuri Frosio and Jan Kautz. The best defense is a good offense: Adversarial augmentation against adversarial attacks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4067–4076, 2023.

[668] Yichen Gong, Delong Ran, Jinyuan Liu, et al. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 23951–23959, 2025.

[669] Yilei Jiang, Xinyan Gao, Tianshuo Peng, et al. HiddenDetect: Detecting jailbreak attacks against multimodal large language models via monitoring hidden states. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14880–14893, Vienna, Austria, July 2025. Association for Computational Linguistics.

[670] Sheng Yang, Jiawang Bai, Kuofeng Gao, et al. Not all prompts are secure: A switchable backdoor attack against pre-trained vision transfomers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24431–24441, 2024.

[671] Haoran Li, Yulin Chen, Zihao Zheng, et al. Simulate and eliminate: Revoke backdoors for generative large language models. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 397–405, 2025.

[672] Xianjun Yang, Wei Cheng, Yue Wu, et al. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. arXiv preprint arXiv:2305.17359, 2023.

[673] Yuhui Shi, Qiang Sheng, Juan Cao, et al. Ten words only still help: Improving black-box ai-generated text detection via proxy-guided efficient re-sampling. arXiv preprint arXiv:2402.09199, 2024.

[674] Guangyu Nie, Changhoon Kim, Yezhou Yang, and Yi Ren. Attributing image generative models using latent fingerprints. In International Conference on Machine Learning, pages 26150–26165. PMLR, 2023.

[675] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24480–24489, 2023.

[676] Yang Wu, Ruijia Wang, and Jie Wu. Non-existent relationship: Fact-aware multi-level machine-generated text detection. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 3757–3768, Suzhou, China, November 2025. Association for Computational Linguistics.

[677] Andrea Continella, Mario Polino, Marcello Pogliani, and Stefano Zanero. There's a hole in that bucket! a large-scale analysis of misconfigured s3 buckets. In Proceedings of the 34th Annual Computer Security Applications Conference, pages 702–711, 2018.

[678] Michael Meli, Matthew R McNiece, and Bradley Reaves. How bad can it git? characterizing secret leakage in public github repositories. In NDSS, 2019.

[679] Xueqiang Wang, Yuqiong Sun, Susanta Nanda, and XiaoFeng Wang. Credit karma: Understanding security implications of exposed cloud services through automated capability inference. In 32nd USENIX Security Symposium (USENIX Security 23), pages 6007–6024, 2023.

[680] Zakir Durumeric, Zane Ma, Drew Springall, et al. The security impact of https interception. In NDSS, 2017.

[681] Henry Birge-Lee, Yixin Sun, Anne Edmundson, et al. Bamboozling certificate authorities with {BGP}. In 27th USENIX Security Symposium (USENIX Security 18), pages 833–849, 2018.

[682] Nicholas Carlini, Steve Chien, Milad Nasr, et al. Membership inference attacks from first principles. In 2022 IEEE symposium on security and privacy (SP), pages 1897–1914. IEEE, 2022.

[683] Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. Did the neurons read your book? document-level membership inference for large language models. In 33rd USENIX Security Symposium (USENIX Security 24), pages 2369–2385, 2024.

[684] Nikhil Kandpal, Krishna Pillutla, Alina Oprea, et al. User inference atta cks on large language models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 18238–18265, 2024.

[685] Zhikun Zhang, Min Chen, Michael Backes, et al. Inference attacks against graph neural networks. In 31st USENIX Security Symposium (USENIX Security 22), pages 4543–4560, 2022.

[686] Xiuling Wang and Wendy Hui Wang. Group property inference attacks against graph neural networks. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, pages 2871–2884, 2022.

[687] Ehsanul Kabir, Lucas Craig, and Shagufta Mehnaz. Disparate privacy vulnerability: Targeted attribute inference attacks and defenses. In 34th USENIX Security Symposium, pages 5445–5463, 2025.

[688] Francesco Diana, Othmane Marfoq, Chuan Xu, et al. Attribute inference attacks for federated regression tasks. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 16271–16279, 2025.

[689] Junjie Chu, Zeyang Sha, Michael Backes, and Yang Zhang. Reconstruct your previous conversations! comprehensively investigating privacy leakage risks in conversations with gpt models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6584–6600, 2024.

[690] Jacob Choi, Shuying Cao, Xingjian Dong, and Sai Praneeth Karimireddy. Contextleak: Auditing leakage in private in-context learning methods. In The Impact of Memorization on Trustworthy Foundation Models: ICML 2025 Workshop.

[691] Shenglai Zeng, Jiankun Zhang, Pengfei He, et al. The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag). In Findings of the Association for Computational Linguistics: ACL 2024, pages 4505–4524, 2024.

[692] Mingrui Liu, Sixiao Zhang, and Cheng Long. Mask-based membership inference attacks for retrieval-augmented generation. In Proceedings of the ACM on Web Conference 2025, pages 2894–2907, 2025.

[693] Alexander Viand, Patrick Jattke, Miro Haller, and Anwar Hithnawi. {HECO}: Fully homomorphic encryption compiler. In 32nd USENIX Security Symposium (USENIX Security 23), pages 4715–4732, 2023.

[694] Yijia Chang and Songze Li. Arbitrary-threshold fully homomorphic encryption with lower complexity. In 34th USENIX Security Symposium (USENIX Security 25), pages 8403–8422, 2025.

[695] Muhammad Faisal, Jerry Zhang, John Liagouris, et al. {TVA}: A multi-party computation system for secure and expressive time series analytics. In 32nd USENIX Security Symposium (USENIX Security 23), pages 5395–5412, 2023.

[696] Cynthia Dwork. Differential privacy. In International colloquium on automata, languages, and programming, pages 1–12. Springer, 2006.

[697] Lijie Hu, Ivan Habernal, Lei Shen, and Di Wang. Differentially private natural language models: Recent advances and future directions. Findings of the Association for Computational Linguistics: EACL 2024, pages 478–499, 2024.

[698] Chengkun Wei, Weixian Li, Gong Chen, and Wenzhi Chen. Dc-sgd: Differentially private sgd with dynamic clipping through gradient norm distribution estimation. IEEE Transactions on Information Forensics and Security, 2025.

[699] Hilal Asi, Vitaly Feldman, and Kunal Talwar. Optimal algorithms for mean estimation under local differential privacy. In International Conference on Machine Learning, pages 1046–1056. PMLR, 2022.

[700] Yuhan Liu, Tianhao Wang, Yixuan Liu, et al. Edge-protected triangle count estimation under relationship local differential privacy. IEEE Transactions on Knowledge and Data Engineering, 36(10):5138–5152, 2024.

[701] Zihang Xiang, Tianhao Wang, and Di Wang. Preserving node-level privacy in graph neural networks. In 2024 IEEE Symposium on Security and Privacy (SP), pages 4714–4732. IEEE, 2024.

[702] Qiuchen Zhang, Hong kyu Lee, Jing Ma, et al. Dpar: Decoupled graph neural networks with node-level differential privacy. In Proceedings of the ACM Web Conference 2024, pages 1170–1181, 2024.

[703] Dian Lei, Zijun Song, Yanli Yuan, et al. Achieving personalized privacy-preserving graph neural network via topology awareness. In Proceedings of the ACM on Web Conference 2025, pages 3552–3560, 2025.

[704] Yuxin Qi, Jun Wu, Xi Lin, et al. Differentially private graph neural network with importance-grained noise adaption. IEEE Transactions on Dependable and Secure Computing, pages 1–14, 2025.

[705] Tingting Tang, Yue Niu, Salman Avestimehr, and Murali Annavaram. Edge private graph neural networks with singular value perturbation. Proceedings on Privacy Enhancing Technologies, 2024.

[706] Yuxin Qi, Xi Lin, Ziyao Liu, et al. Linkguard: Link locally privacy-preserving graph neural networks with integrated denoising and private learning. In Companion Proceedings of the ACM Web Conference 2024, pages 593–596, 2024.

[707] Tamara T Mueller, Johannes C Paetzold, Chinmay Prabhakar, et al. Differentially private graph neural networks for whole-graph classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(6):7308–7318, 2022.

[708] Ningxin Su and Baochun Li. Asynchronous federated unlearning. In IEEE INFOCOM 2023-IEEE conference on computer communications, pages 1–10. IEEE, 2023.

[709] Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. In 30th Annual Network and Distributed System Security Symposium, NDSS 2023, 2023.

[710] Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, et al. Mixed-privacy forgetting in deep networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 792–801, 2021.

[711] Xiaoyu Cao, Jinyuan Jia, Zaixi Zhang, and Neil Zhenqiang Gong. Fedrecover: Recovering from poisoning attacks in federated learning using historical information. In 2023 IEEE Symposium on Security and Privacy (SP), pages 1366–1383. IEEE, 2023.

[712] Ming Hu, Yue Cao, Anran Li, et al. Fedmut: Generalized federated learning via stochastic mutation. In Proceedings of the AAAI conference on artificial intelligence, volume 38, pages 12528–12537, 2024.

[713] Jingxue Chen, Hang Yan, Zhiyuan Liu, et al. When federated learning meets privacy-preserving computation. ACM Computing Surveys, 56(12):1–36, 2024.

[714] Zheng Lin, Guanqiao Qu, Xianhao Chen, and Kaibin Huang. Split learning in 6g edge networks. IEEE Wireless Communications, 31(4):170–176, 2024.

[715] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, et al. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

[716] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, et al. Improving transferability of adversarial examples with input diversity. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2730–2739, 2019.

[717] Yu Guo, Weiquan Liu, Qingshan Xu, et al. Boosting adversarial transferability through augmentation in hypothesis space. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 19175–19185, 2025.

[718] Fengfan Zhou, Bangjie Yin, Hefei Ling, et al. Improving the transferability of adversarial attacks on face recognition with diverse parameters augmentation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 3516–3527, 2025.

[719] Ziqi Zhou, Shengshan Hu, Minghui Li, et al. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In Proceedings of the 31st ACM International Conference on Multimedia, pages 6311–6320, 2023.

[720] Jin Li, Ziqiang He, Anwei Luo, et al. Advad: Exploring non-parametric diffusion for imperceptible adversarial attacks. Advances in Neural Information Processing Systems, 37:52323–52353, 2024.

[721] Bin Liang, Hongcheng Li, Miaoqiang Su, et al. Deep text classification can be fooled. arXiv preprint arXiv:1704.08006, 2017.

[722] Ziyi Yin, Muchao Ye, Tianrong Zhang, et al. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. Advances in Neural Information Processing Systems, 36:52936–52956, 2023.

[723] Haodi Wang, Kai Dong, Zhilei Zhu, et al. Transferable multimodal attack on vision-language pre-training models. In 2024 IEEE Symposium on Security and Privacy (SP), pages 1722–1740. IEEE, 2024.

[724] Hao Fang, Jiawei Kong, Wenbo Yu, et al. One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4090–4100, 2025.

[725] Zhaoyu Chen, Bo Li, Jianghe Xu, et al. Towards practical certifiable patch defense with vision transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15148–15158, 2022.

[726] Edoardo Mosca, Shreyash Agarwal, Javier Rando Ramírez, and Georg Groh. "that is a suspicious reaction!": Interpreting logits variation to detect NLP adversarial attacks. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7806–7816, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[727] Ashim Gupta, Carter Wood Blum, Temma Choji, et al. Don't retrain, just rewrite: Countering adversarial perturbations by rewriting text. arXiv preprint arXiv:2305.16444, 2023.

[728] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. arXiv preprint arXiv:2402.12336, 2024.

[729] Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24625–24634, 2024.

[730] Songlong Xing, Zhengyu Zhao, and Nicu Sebe. Clip is strong enough to fight back: Test-time counterattacks towards zero-shot adversarial robustness of clip. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 15172–15182, 2025.

[731] Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, et al. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. arXiv preprint arXiv:2308.06463, 2023.

[732] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? Advances in Neural Information Processing Systems, 36:80079–80110, 2023.

[733] Andy Zou, Zifan Wang, Nicholas Carlini, et al. Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043, 2023.

[734] Xiaojun Jia, Tianyu Pang, Chao Du, et al. Improved techniques for optimization-based jailbreaking on large language models. arXiv preprint arXiv:2405.21018, 2024.

[735] Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, et al. Tree of attacks: Jailbreaking black-box llms automatically. Advances in Neural Information Processing Systems, 37:61065–61105, 2024.

[736] Hao Wang, Hao Li, Junda Zhu, et al. Diffusionattacker: Diffusion-driven prompt manipulation for llm jailbreak. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22193–22205, 2025.

[737] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, et al. Are aligned neural networks adversarially aligned? Advances in Neural Information Processing Systems, 36:61478–61500, 2023.

[738] Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. arXiv preprint arXiv:2307.14539, 2023.

[739] Ruofan Wang, Juncheng Li, Yixu Wang, et al. Ideator: Jailbreaking and benchmarking large vision-language models using themselves. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8875–8884, 2025.

[740] Jiabao Ji, Bairu Hou, Alexander Robey, et al. Defending large language models against jailbreak attacks via semantic smoothing. arXiv preprint arXiv:2402.16192, 2024.

[741] Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Gradient cuff: Detecting jailbreak attacks on large language models by exploring refusal loss landscapes. Advances in Neural Information Processing Systems, 37:126265–126296, 2024.

[742] Reshabh K Sharma, Vinayak Gupta, and Dan Grossman. Defending language models against image-based prompt attacks via user-provided specifications. In 2024 IEEE Security and Privacy Workshops (SPW), pages 112–131. IEEE, 2024.

[743] OpenAI. Gpt-4 technical report, 2024.

[744] Nisan Stiennon, Long Ouyang, Jeffrey Wu, et al. Learning to summarize with human feedback. Advances in neural information processing systems, 33:3008–3021, 2020.

[745] Rafael Rafailov, Archit Sharma, Eric Mitchell, et al. Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems, 36:53728–53741, 2023.

[746] Yunhao Gou, Kai Chen, Zhili Liu, et al. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. In European Conference on Computer Vision, pages 388–404. Springer, 2024.

[747] Zenghui Yuan, Pan Zhou, Kai Zou, and Yu Cheng. You are catching my attention: Are vision transformers bad learners under backdoor attacks? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24605–24615, 2023.

[748] Micah Goldblum, Dimitris Tsipras, Chulin Xie, et al. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2):1563–1580, 2022.

[749] Naibin Gu, Peng Fu, Xiyu Liu, et al. A gradient control method for backdoor attacks on parameter-efficient tuning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3508–3520, 2023.

[750] Yanzhou Li, Tianlin Li, Kangjie Chen, et al. Badedit: Backdooring large language models by model editing. arXiv preprint arXiv:2403.13355, 2024.

[751] Zhenyang Ni, Rui Ye, Yuxi Wei, et al. Physical backdoor attack can jeopardize driving with vision-large-language models. arXiv preprint arXiv:2404.12916, 2024.

[752] Khoa D Doan, Yingjie Lao, Peng Yang, and Ping Li. Defending backdoor attacks on vision transformer via patch processing. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pages 506–515, 2023.

[753] Akshayvarun Subramanya, Aniruddha Saha, Soroush Abbasi Koohpayegani, et al. Backdoor attacks on vision transformers. arXiv preprint arXiv:2206.08477, 2022.

[754] Xuanli He, Jun Wang, Benjamin Rubinstein, and Trevor Cohn. Imbert: Making bert immune to insertion-based backdoor attacks. arXiv preprint arXiv:2305.16503, 2023.

[755] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 364, 2019.

[756] Tianyun Yang, Ziyao Huang, Juan Cao, et al. Deepfake network architecture attribution. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 4662–4670, 2022.

[757] Chu Luo and Vassilis Kostakos. Toward ubiquitous operating systems: Lessons from the field. Communications of the ACM, 68(10):24–27, 2025.