# 附件1

# 中国电子学会-天翼云科研专项（2025）申报主题说明

# 基于大语言模型跨架构代码迁移技术研究

随着全球信息技术快速发展，信息技术应用创新（信创）已成为推动国家信息化建设的核心力量，在这一背景下国产化信息技术软硬件的发展尤为关键。然而，现有大量软件基于x86架构开发，为实现其在国产ARM架构的高效运行，跨架构代码迁移技术的研究尤为迫切。本课题聚焦大模型技术，开发高效跨架构代码迁移解决方案。研究内容包括但不限于：

（1）模型架构设计与训练策略：如何构建适配多架构语义理解的大模型，通过预训练学习跨平台指令集、语法规则及代码优化模式，结合微调策略提升对目标架构的适配能力；

（2）语义保持与性能优化：探索代码功能等价性验证方法，结合程序分析、符号执行等技术确保迁移后代码逻辑的一致性，并融合硬件特性（如并行度、内存管理）实现性能优化；

（3）异构场景迁移泛化：构建涵盖x86/ARM等多架构的配对代码数据集，解决指令集差异、系统调用兼容性等挑战，通过强化学习动态优化跨平台迁移质量，提升工业场景实用性。

研发一款高效跨架构代码迁移工具, 突破传统手动迁移效率瓶颈，提升代码跨架构转换的精准性与规模化水平，提供技术报告并发表学术论文。

# 多维高速缓存与深度检测融合的实时Web防护系统

在实时Web防护场景中，虽然现有深度检测模型能够提供较高的检测精度，但其检测延迟难以满足实时响应的需求。融合多维缓存与深度检测技术以实现低延迟、高准确度的实时防护，是当前网络安全实时检测的难点问题。

本课题聚焦多维高速缓存与深度检测技术的融合，探索线性时间复杂度下的映射与泛化信息存储方案，以及调度执行模式、缓存同步策略等优化方法，将检测延迟控制在10ms以内，打破传统深度检测难以应用于实时防护的瓶颈。具体研究内容包括但不限于：

（1）设计场景适配的缓存线性映射技术，并支持泛化映射与流式映射，提升缓存的查找与信息存储效率；

（2）设计并实现深度检测算法与多维缓存系统的有机结合，通过模型融合与模块化部署，确保在通用场景下在线检测实现平均准确率≥99.99%、召回率≥99.99%；

（3）针对云上大规模集群部署环境，研究系统的可扩展性与实时性保障措施，确保在高流量情况下具备稳定的低延迟响应与高处理能力。

研发一款新型实时Web防护系统原型，为网络安全防护提供突破性的实时检测解决方案，完整交付核心算法实现、系统设计文档和性能测试报告。

# 复杂任务型对话系统大模型逻辑推理能力研究

随着大语言模型技术的发展，以大模型为基础的智能客服、智能助手等得到了广泛应用，但是在相关场景中，用户经常使用口语化表达，且描述模糊。造成用户与系统之间的无效沟通和反复确认，期望系统能够一次性准确理解并执行任务，提高交互效率。

本课题旨在探索多轮交互复杂任务对话场景下的大模型逻辑推理技术，从而提高对话意图的理解和任务规划的准确率。研究内容包括但不限于：

（1）提出模型优化方法或策略，提高在不规范表达对话中用户意图的理解和跟踪，例如口语化、隐喻或带有情绪色彩等描述情况下的意图理解；

（2）基于对话上下文的碎片信息关联整理，确保对话的逻辑性和连贯性，明确任务目标；

（3）通过知识推理对达成任务目标所需的条件和进行梳理，对缺失的关键内容进行合理推测和补全。

研发一套新型的多轮交互复杂任务对话系统及其相关模型推理能力技术方案，发表至少1篇CCF-B类及以上学术论文，并申请1项相关专利（含已受理专利）。

# 低碳编排调度原理与算法研究

面向2030碳达峰目标以及东数西算工程的推进，如何利用西部绿能算力是一个重要的研究方向。西部绿电资源池可部署非时延敏感型任务，通过跨域任务调度将计算任务与数据部署到绿电资源池为客户提供服务，从而降低算力碳排，但会引入额外的成本、带宽占用和网络设备碳耗等。基于这一背景，本课题提出如下研究内容：

（1）研究AI训推任务在计算设备上的电耗模型，探索数据传输时在网络设备上的电耗模型，建立电耗与碳耗的关系；

（2）研究低碳编排调度机制，结合提出的模型，建立跨区域算网调度协同机制；

（3）提出低碳编排算法，综合考虑算网成本、碳耗和资源占用等指标。

本课题拟解决的关键问题是通过算网调度为计算任务寻找绿电资源池和数据传输路径，降低算网碳排放量。需权衡碳排放量与成本间的矛盾，满足算网资源、碳排放量指标等要求，降低算网总成本，最大化营收利润。

通过上述研究，探索出电耗模型、低碳编排机制原理、低碳编排算法，可将相关成果落地在息壤平台。交付算法实现、机制设计文档、实验报告、1项专利(含已受理专利)、1篇CCF-B类及以上论文。

# 分布式系统一致性与正确性验证技术研究

随着云计算、边缘计算以及AI大模型的快速发展，超大规模分布式集群已成为新一代基础设施的核心支撑。在此背景下，分布式系统的正确性与一致性验证面临前所未有的挑战，如测试路径爆炸、场景复杂化、敏捷验证难等问题。常用的混沌测试通过随机注入网络、进程、硬件故障等事件验证程序的正确性与一致性，但存在效率低下、资源开销大以及缺乏系统性理论支撑等问题。本课题面向大规模分布式系统，聚焦于系统的正确性与一致性验证，旨在通过理论创新与工程实践相结合，提升大规模分布式系统验证理论和方法。研究内容可以包括但不限于以下方面：

（1）超大规模分布式集群的敏捷验证方法研究；

（2）故障快速复现与溯源技术研究；

（3）智能测试编排与设计研究；

（4）系统正确性、数据一致性的验证理论研究。

本课题期望的交付物可以是一种分布式系统的验证方法，需确保可行性与实用性。一篇高质量的论文，或一项具有创新性的专利，但不限于以上几种形式。此外，还可以包括详细的研究报告、完整的技术文档、以及用于验证的原型系统等。

# 基于高速总线网络的分离式内存池化技术在云基础设施中的研究和应用

在云计算场景下，云数据中心的规模正在快速增长。云厂商也不断推出大规格的计算实例，以满足高性能计算、人工智能训练和实时数据分析等场景的需求。但是云基础设施中单个计算实例（虚拟机、裸金属或容器）的规格增长速度（近5年增长2-4倍）不及数据中心规模的增长速度（近五年增长5-8倍），根本原因是单计算实例规格受计算节点的物理限制。本课题目标：一是通过高速总线网络的分离式内存突破物理服务器限制，满足云基础设施的超大内存实例规格的需求；二是通过跨物理计算节点的分离式共享内存池解决困扰云服务提供商内存资源碎片的问题；三是在保障云计算客户计算业务QoS的情况下解决通算、智算、超算集群的“内存墙”问题。但是在现有的计算机系统中引入分离式内存架构，面临无法有效解决内存资源高可用、计算性能劣化等问题。研究内容包括但不限于：

（1）面向云基础设施中分离式池化内存的黑盒负载性能优化与保障技术研究；

（2）分离式池化内存系统的爆炸半径扩大问题研究；

（3）分离式池化内存系统对大模型分布式推理性能优化研究。

提出一种能在云基础设施中应用的分离式内存的优化方法，能解决一个或者多个上述课题的问题，发表CCF-A类论文1篇，申请专利1项（含已受理专利）。

# 基于数据处理器的AI训练I/O加速技术研究

随着人工智能（AI）技术的快速发展，深度学习模型的规模和训练数据量呈现指数级增长。在AI训练过程中，数据读取和Checkpoint的频繁写入导致显著的I/O开销。传统基于CPU的数据预处理和存储管理方式难以满足高性能GPU的计算需求，导致GPU计算资源因等待数据而闲置，严重制约了训练效率。数据处理器（DPU, Data Processing Unit）作为一种新型智能网卡或专用加速器，具备高带宽I/O能力、可编程计算核心和硬件加速引擎，为优化AI训练I/O瓶颈提供了新的可能性。目前GPU与DPU的协同计算机制尚未成熟，文件系统在DPU上的卸载技术仍面临诸多挑战，亟需深入研究以提升AI训练的整体性能。本项目旨在探索基于DPU的I/O加速技术，优化AI训练中的数据访问和存储管理，系统性解决AI训练中的I/O墙问题，降低I/O延迟，提升GPU利用率，推动高效能AI训练系统的发展。研究内容包括但不限于：

（1）研究GPU-DPU协同的文件访存机制，实现高并发、低延迟的数据访问，解决AI训练数据访问I/O路径延迟高的问题；

（2）研究面向DPU的文件系统卸载技术，降低数据管理开销，解决AI训练数据访存元数据操作开销大，管理复杂的问题。
 研究成果应包含但不限于原理、方法、技术、原型系统演示，能降低20%的读延迟，10%的写延迟，GPU利用率提升10%以上。申请专利，发表论文等。

# 基于云电脑场景的屏幕内容图像增强算法研究

云电脑是一种面向用户实时交互的软件产品。对于算力、时延、带宽、终端硬件都有极高的要求。矢量图像在云电脑的传输过程中占据主导地位（90%以上），如何能用尽可能少的云端算力实现高质量低时延传输是关键难点。目前云电脑传输过程中主要受限于终端解码能力的限制(终端硬解码能力仅支持420格式)，采用 yuv420 编码，相对于 yuv444 的编码格式色度上存在降采样，用户体验上存在不足。

本课题聚焦云电脑内主流的矢量图像，研究一种端侧的色彩增强方法，在解码器输出yuv420格式的基础上，进行针对性的增强，生成yuv444图像，还原色彩分辨率。在还原色彩的同时，保证生成的实时性要求。

交付一种图像增强算法,将YUV420增强生成为YUV444图像，平台为intel N100、N97低功耗云终端平台，性能要求30fps及以上实时处理2K（2560x1440）, 平均占用不超过1.5个核心，图像质量尽量与原图直接转换的yuv444一致。

# 大模型训练中慢节点检测、定位及恢复技术研究

慢节点问题是大模型分布式训练中最常见的性能瓶颈之一。在大模型训练场景下，任何单节点性能的下降都可能影响全局训练效率，造成算力成本浪费。目前智算网络难以实现对底层设备信息和数据的轻量化实时精细采集，同时，智算网络信息的复杂性使得在大模型训练过程中快速发现和准确定位慢节点极为困难，进而难以及时采取有效的隔离恢复措施。本课题的目的是快速发现并定位智算网络中的慢节点，并在保证AI模型训练效率的情况下，完成慢节点的隔离和对应任务的无损恢复。预期从大模型训练集合通信的角度出发，寻找模型参数同步节奏差异，拟支持以下几个方面的研究，包括但不限于：

（1） 研究智算网络大模型训练过程中关键数据性能的精细化采集，包括但不限于分布式GPU集合通信时延、网卡带宽、交换机队列长度等，在兼顾高精度的同时满足轻量化需求；

（2） 研究慢节点的分析和定位算法，根据采集到的信息进行跨物理机全局分析，诊断训练慢或hang住的原因并实现秒级90%以上定位精度；

（3） 研究慢节点隔离和模型任务无损恢复策略，在保证大模型训练效率的情况下对慢节点进行隔离，并实现分钟级模型任务的无损恢复。

设计并研制出一款智算网络慢节点检测系统原型，完整交付算法实现，提供在至少两个典型慢节点场景下的测试用例以及测试报告。

# 大规模智算网络仿真技术研究

大模型训练与推理是智算网络的核心应用，然而基于实际AI服务器集群的性能优化和新技术验证成本过高。本课题聚焦高保真的AI集群仿真技术，综合考虑计算侧和网络侧各因素的影响，提高仿真的准确度，并保证仿真速度和稳定性。应用场景包括但不限于：AI集群训练与仿真。研究课题包括但不限于：

（1）高精度计算侧与网络侧建模：实现细粒度的内核模拟，支持天翼云AI集群国产GPU，并具备可拓展性；支持网卡、交换机核心功能及性能建模，支持常见集合通信模式，支持GPU总线仿真、RDMA传输仿真，网络模型可拓展支持第三方通信协议；

（2）针对大模型特有的计算-通信交替依赖特性，构建网络侧与计算侧的联合仿真框架，支持多样分布式训练模式，支持并行仿真框架，满足差异化的仿真精度、时间要求；

（3）构建面向训练/推理的业务模型，设计兼具物理真实性与可配置性的负载生成框架。

研制一款智算网络AI训练和推理仿真软件，准确度误差在5%以内，支持72小时长稳测试，实现国产GPU万卡级规模仿真，支持验证网络新技术。