华中科技大学

大数据技术与系统国家地方联合工程研究中心
服务计算技术与系统教育部重点实验室
集群与网格计算湖北省重点实验室
大数据安全湖北省工程研究中心

2025届研究生毕业留念 2025.5.12

我们毕业啦！

地 址：武汉市华中科技大学东五楼二楼
邮 编：430074
电 话：15387112483
E-mail：yanyf@hust.edu.cn
Homepage：http://grid.hust.edu.cn

大数据技术与系统国家地方联合工程研究中心
服务计算技术与系统教育部重点实验室
集群与网格计算湖北省重点实验室

并行与分布式计算通讯

BING XING YU FEN BU SHI JI SUAN TONG XUN

2025年第2期 总第61期 2025年06月

封面人物：谢子凡——批判性思维：科研路上的破雾利剑

2025届毕业生成果展示

首届计算机网络与系统前沿论坛顺利举办

探索与成长的科研之旅

http://grid.hust.edu.cn

2025届博士及硕士毕业生顺利完成答辩

首届计算机网络与系统前沿论坛顺利举办

风采

致实验室2025届毕业生

窗外的梧桐又添了一圈年轮，东五楼的灯光依旧在深夜里温柔地亮着。又是一年凤凰花开时，2025届的你们即将带着沉甸甸的收获，从这里启程，奔赴更广阔的天地。

回望这几年，实验室的服务器记录的不只是数据，更是你们从青涩到成熟的蜕变；展厅里的白板擦了又写，写满了灵感的迸发与思维的碰撞。从第一次组会的紧张忐忑，到如今能自信从容地分享研究成果；从最初代码的反复调试，到最终文章的顺利录用——这一路走来的每一步，都凝聚着你们的汗水与智慧。

在你们即将告别之际，我们特别整理了本届毕业生的论文摘要展示。这些成果或许只是漫漫科研路上的一小步，却见证了你们追逐真理的执着与勇气。愿你们带着这份科研精神的火种，在未来的道路上继续绽放光芒。

蔡敏志：愿一路乘风破浪，学以致用，回馈社会！**曹颖：**每一步都铿锵有力，未来的你前景可期！**陈豪：**祝你在未来的事业和生活中，芝麻开花节节高，早日实现自己的目标。**陈群锦明：**勤勉笃行前程广，勇攀人生新高峰！**陈瑞聪：**不必追赶每一场日出，愿你在追求理想时，不忘享受生活中的每一个小瞬间。**陈祎：**祝愿在博士阶段取得更多优秀成果！**陈雨欣：**愿接下来的日子，每个选择都有底气！**程传斌：**愿你在新单位，龙马精神，步步高升！**程健：**希望你能将学校所学在单位应用起来，发光发热！**储朝阳：**祝在未来的学术道路上大放异彩！**丁甲：**成长比成功更重要，努力比梦想更现实。**杜小虎：**愿你以不变的初心，迎接每一个未知的挑战，在未来工作与人生的道路上，始终坚定、自信、从容。**方正：**祝聪明乐观的你在新的人生征程中，取得更大的进步。**冯昊：**你的未来充满无限可能，愿每一个梦想都能开花结果！**伏子豪：**祝踏实勤奋的你在未来工作中前程似锦！**甘芮：**愿人生之路一路繁花，前程似景！**高文杰：**祝在今后的工作道路上一帆风顺！**郭超：**遇到困难别硬扛，工作生活中总有愿意搭把手的人。**郭海宏：**有想法、有韧性，勇于攻坚克难，你将更上层楼！**郭啸辰：**愿你在未来的道路上坚持热爱、保持思考，以坚韧与智慧迎接每一个新阶段的挑战。**郭正轩：**前路坦荡，愿你勇敢追光！**韩浩：**你的光彩大大的，记得常联系！**洪子骁：**祝工作顺利，健康平安！**侯宇翔：**虽总玩笑唤你'少爷'，但你骨子里的较真、执着与思维锋芒，早已深铭在心。愿前程锦绣，扶摇万里！**黄福龙：**愿此去繁花似锦，再相逢依然是少年。**黄晗翔：**感谢为serverless storage项目做出的重要贡献！祝宏图大展！**黄浩岩：**你在编程方面展现出的扎实功底令人钦佩，愿你在技术的道路上不断精进，勇攀高峰！**黄浚：**希望你能在新的岗位上勇攀高峰！**黄恺一：**FPGA大佬，未来继续带大家冲啊！**黄宗耀：**一帆风顺！**蒋晨昱：**愿你在工作单位开拓属于自己的辽阔人生！**靳晓忠：**愿你不忘初心，保持对未知的好奇心，勇攀学术生涯高峰！**李邦宇：**愿你心怀梦想，勇敢面对每一个新挑战，开创美好的未来！**李果：**愿你带着少年意气，一路乘风破浪，未来皆坦途！**李童天：**愿你始终相信自己的积累与勇气，大步向前！**李扬：**踏实认真、有想法，美好前程就在你脚下！**林立成：**祝早日成功申博，感受做研究的乐趣！**刘宝阳：**态度端正、踏实肯干，做好时间的主人，你将看到更好的风光！**刘存扬：**祝读博成果多多，早日成为刘博士！**刘董奇：**愿你永葆破局之勇，未来必见海阔天高！**刘威：**今朝有酒，何待明日，人生一场皆是空。**刘正涛：**方丈法力无边，前途光明！**卢浩宇：**愿你前程锦绣，扶摇万里，贡献社会！**陆思彤：**愿你带着勇气与热爱，奔赴灿烂远方，毕业快乐，未来闪闪发光！**陆筠潇：**愿前程似锦，所行皆坦途；盼一帆风顺，万事皆顺遂。**路琛泽：**愿你在工作单位闪闪发光！**罗陈亮：**以技为刃，步步踏实铺似锦前程；孜孜不倦，处处从容铸璀璨人生。**罗康：**愿你在新的单位开拓属于自己的璀璨人生！**马杰：**毕业，也是新的起点，祝你工作顺利，万事顺意，赢得属于自己的精彩人生！**马绍博：**祝你在未来新的岗位上成就自我，书写更美好的人生篇章！**聂龙宇：**新的旅程已至，愿你一路繁花，一路欢歌！**牛富平：**温柔细腻，

内心丰富，文学喜好继续保持啊！**潘晨高**：代码编织未来星辰，真诚沟通铸就通途！**潘力菘**：感谢为vkernel项目做出的重要贡献！祝前途似锦！**彭潇阳**：十年后再聚首，愿你眼里的光还像答辩通过那刻那么明亮！**齐豪**：以理想为顶点，以奋斗为边，编织无限可能，遍历星辰大海！**钱震宇**：风雨后有彩虹，加油！**乔辰奇**：未来一切顺利！**石煜庭**：带着少年勇气跃入人海，把每个“未知”都活成“值得”。**石月鑫**：踏实优秀，祝愿在今后的道路上成就更卓越的自己！**宋郅炜**：愿你的未来如同灿烂的花朵，绽放出五彩斑斓的色彩！**谈安东**：祝以后的工作生活都一帆风顺，前程似锦！**谭磊**：祝工作顺利、早日实现财富自由！**万伟**：广结善缘，左右逢源，前路漫漫便无忧！**汪啸宇**：祝你在未来的道路上，勇敢追梦，收获满满的幸福与成功。**汪易**：愿此去顺遂，顶峰再相见！**王冠雄**：毕业快乐，愿人生精彩继续！**王灏洲**：未来继续带着大家浪哈！**王虎**：愿你在未来道路上，始终这么脚踏实地，可靠有担当！**王可馨**：祝愿优秀的你一直以严谨为舟，以热忱作帆，前路自有星辰芬芳！**王坤明**：积极主动大胆闯，前程似锦万事顺！**王书林**：前程铺锦绣，智慧点星辰，愿人生之路常新！**王贤龙**：师夷长技，不忘初心，皇图霸业谈笑中。**王鑫蕾**：愿你在新的岗位上书写美好人生！**魏武才**：祝工作顺利、早日实现财富自由！**吴畅**：国产AI就靠你了！**吴浩**：祝学术生涯取得更多优秀成果！海外学成之后再回国发展！**吴漾**：祝在未来的学术道路上大放异彩！**吴羽飞**：愿你保持好奇，探索生活中更多未知精彩！**谢子凡**：道阻且长，行则将至；事虽难作，恒必成之。愿你在未来永葆赤子之心，以所学为炬火，照亮属于你的星辰大海。**熊淑怡**：愿你永远保有探索未知的勇气与创造美好的力量，温柔与果敢并存，智慧与坚韧共生！**徐成浩**：前程似锦，诸事如愿！**徐祯**：愿你以满腔的好奇与不懈的努力，开启新的征程。**杨奕驰**：你的未来充满无限可能，愿每一个梦想都能开花结果！**叶楚玥**：感谢你对项目的付出，祝你一帆风顺，前程似锦，生活美满！**易迎澳**：愿你在科技浪潮中开拓进取，闯出自己的天地。保持热爱，奔赴山海，未来皆坦途。**尹可汗**：以梦为马，不负韶华；前程似锦，未来可期！**尹伟行**：继续加油，在新的岗位大展宏图！**于跃**：感谢为serverless GPU项目做出的重要贡献！祝一帆风顺！**余辉**：愿你不忘初心，探索未知，坚持不懈，创造美好！**俞强**：春华结硕果，秋实启新程！**岳航**：祝将来工作生活一帆风顺！**张昊林**：AI系统大佬，就等你干趴下DeepSeek了！**张泷玲**：山高水长，后会有期，来日再续此深缘。**张茂荣**：携勇气出发，未知皆值得。**张梦颖**：愿你带着求知的热忱远行，在学术的星辰大海中探索未知，勇攀学术高峰事事顺。**张琪**：毕业了，即将踏入人生新的征程，希望你一切顺利，万事顺心，未来可期！**张世桀**：愿世桀开启精彩的学术之旅，闪闪发光！**张信民**：感谢为组里做出的杰出贡献！祝以后的工作生活都一帆风顺！**张业超**：水满则溢，月盈则亏，踏实前行才是真。**张熠**：盼守赤子之心，常怀青云之志！**章翔**：祝工作生活一帆风顺！**赵旭**：愿你在新的舞台上绽放才华，开启新精彩，既有职场人的干练，又不失内心的温暖。**周宇航**：愿如蓓蕾静绽光华，细腻初心伴你前行！**周卓然**：严谨细心，未来之路可期！**朱浩然**：愿你在星辰大海的征途上永葆初心，在未知的领域开拓属于自己的辽阔人生！**朱华**：鹏程万里，初心长明；山海坦荡，乘风自骋！**朱嘉豪**：愿你不忘初心，勇敢前行，把所学所悟融入实践，在人生与事业中书写更加精彩的篇章。**祝振宇**：天下断无易处之境，人间哪有空闲光阴。

亲爱的同学们，技术之路永无止境。愿你们始终保持对未知的好奇，在新的舞台上勇敢追梦；愿你们永远记得，在这个实验室里度过的日日夜夜，是你们职业生涯中最坚实的起点。

这里永远为你们亮着一盏灯，期待你们带着更精彩的故事回来分享。前路漫漫，勿忘初心，实验室永远是你们温暖的港湾！

实验室全体老师

二〇二五年六月

主　编：金　海
本期执行主编：戴小海
编　委：（按姓氏拼音排序）陈汉华、戴小海、丁晓锋、杜冰倩、段卓辉、耿　聪、顾　琳、何　强、胡胜山、华强胜、黄晨明、黄　航、黄　宏、黄　禹、黄　卓、蒋文斌、李钦宾、李婷婷、李　珍、李　志、廖小飞、刘海坤、刘英书、陆　枫、罗瑞坤、毛伏兵、邵志远、石宣化、陶　莉、万　瑶、王多强、王虹飞、王　雄、文　明、吴　松、吴月明、肖　江、徐　鹏、姚德中、姚鹏程、叶晨成、余　辰、余庚花、袁　斌、袁平鹏、张书豪、张　腾、张晓今、张　宇、赵　进、郑　龙、郑　然、邹德清
责任编辑：燕　燕

地　址：武汉市华中科技大学东五楼二楼
邮　编：430074
电　话：15387112483
E-mail：yany@hust.edu.cn
Homepage：http://grid.hust.edu.cn
（此刊仅供内部交流学习）

大模型赋能智能制造与产业升级

（史瑞泽 整理）

在工业界，大模型技术正加速与制造业深度融合，成为推动产业智能化升级的核心动力。大模型凭借其强大的知识理解与生成能力，正在重塑产品设计、生产控制、质量检测等多个关键环节。例如，华为发布的盘古大模型5.0在汽车造型设计领域实现突破，将传统设计周期从1至2年大幅缩短至数月，有效提升了产品迭代效率和企业市场响应速度。此外，华为还推出工业AI视觉质检平台，借助大模型对图像的精准识别能力，实现对生产线上产品缺陷的自动化检测，提升了整体制造质量的稳定性和一致性。

与此同时，科大讯飞的羚羊工业大模型2.0 能够在不同数据类型间进行高效联动与推理。这一模型覆盖从研发、生产、供应链到销售、服务和管理的“研产供销服管”全流程，广泛应用于工业文本生成、知识问答、工艺流程优化等场景。卡奥斯推出的COSMO-GPT工业大模型则强调机理融合与专家知识集成，具备较强的因果推理能力，已在多个工业场景中验证其对指标优化、数据生成、辅助决策等方面的高适应性与高精度，推理准确率达96%以上。

政策层面，国家也在积极布局。全国人大代表周云杰在两会中建议，推动工业大模型先行先试，探索标准化、可复制、成本可控的落地路径，特别是为中小制造企业提供普惠型人工智能能力，助力其数字化转型升级。

总体来看，大模型正成为新时代工业革命的重要推动力，不仅提升了制造环节的智能化水平，也为我国工业体系的高质量发展注入了新的活力与可能。

（参考链接：https://epaper.cena.com.cn/pc/content/202501/10/content_12634.html）

TEE辅助BFT共识协议——Achilles

（朱 景 整理）

在工业实践中，如区块链平台、金融交易系统或可信数据库中，BFT共识机制是实现系统高可用性的关键。然而，传统BFT协议通常面临通信复杂度高、延迟大、性能差的问题。近年来，通过TEE（如Intel SGX）增强节点的可信执行能力，成为提高BFT共识效率的主流方向。TEE能有效防止恶意节点在共识过程中进行消息“双重发送”，从而简化协议流程并提高性能。

但现实中，TEE存在“回滚攻击”风险，即攻击者可通过伪造旧状态数据让TEE回退到旧状态，导致节点行为不一致。现有方案使用昂贵的持久计数器（如TPM或ROTE）来防回滚，但这极大限制了协议的性能，尤其在每笔交易都需要访问计数器的情况下。

为打破“性能–容错”二元对立，Achilles提出一种“回滚弹性恢复”机制，将回滚防护逻辑移出交易提交的关键路径。该机制允许节点之间互助恢复状态，避免频繁访问低性能的持久计数器，从而显著减少性能开销。

在共识流程设计上，Achilles借鉴链式BFT（如HotStuff）的理念，引入定制化的链式提交规则，实现四步通信延迟和线性消息复杂度，首次将TEE辅助BFT的性能指标做到了与CFT（如Raft）协议持平。此外，Achilles的设计在网络抖动、节点故障后仍可通过高效恢复机制快速重建系统状态，兼具鲁棒性与高可用性。

实验评估表明，Achilles在局域网条件下可实现高达75KTPS的吞吐性能，远超现有代表性协议（如 Damysus-R、FlexiBFT和OneShot-R），且具备良好的扩展性与容错性。

对于以区块链基础设施、可信数据库和边缘计算为代表的工业系统，Achilles为TEE与

BFT共识的深度融合提供了切实可行、性能卓越的解决方案。

（参考链接：https://dl.acm.org/doi/10.1145/3689031.3717457）

CHERI 安全增强处理器原型公布：基于RISC V/AArch64的能力安全架构

（贾文朔 整理）

近期，多家学术与工业界合作发布了 CHERI 的最新进展，该架构在 ARM AArch64 与 RISC V 指令集上实现细粒度能力（capability）硬件扩展，旨在根本性提升系统的内存安全性与漏洞防御能力。

CHERI 在硬件层面为每个指针或能力引用附加元数据，限制访问权限与范围，从根本上杜绝越界访问与权限滥用，显著降低因 C/C++ 内存不安全导致的漏洞风险，而此类漏洞长期以来占据安全问题的主因。该架构的可移植性强，已在 MIPS、Arm、RISC V 等多个平台上实现。

系统软件方面，Linux、FreeBSD、OpenBSD 等主流操作系统纷纷开始支持 CHERI 能力指针模型，并调整内存模型与编译器（如 Clang 前端改进），以兼容 capability 语义。此外，C/C++ 等编译器在 CHERI 上启用了能力检查与转换工具链，使现有代码仅需少量修改便可受益于基于硬件的内存安全保障。

在工业层面，CHERI 架构已得到英国政府、美国 DARPA 等投资机构支持，多家公司与研究机构联合开发具生产潜力的 CHERI 增强处理器原型。2024–2025 年，多个早期样片已经性能评测完成，表明相较传统架构仅有小幅（≤ 10%）性能开销，却能有效防止许多内存漏洞攻击。此外，CHERI 的 compartmentalization 特性推动关键系统（如嵌入式设备、网络设备、云基础设施）成为了首批商业落地方向。

（参考链接：https://baijiahao.baidu.com/s?id=1830730885786016227&wfr=spider&for=pc）

热门Chrome插件因使用HTTP传输和硬编码凭证使得用户敏感信息泄露

（潘锶逵 整理）

Symantec网络安全研究人员近日发出警告：多个在谷歌Chrome网上应用商店中流行的浏览器扩展程序被发现存在严重设计缺陷，即便他们具备上万级别的下载量。这些插件将用户隐私信息通过HTTP明文传输。传输的信息包括浏览记录、机器ID、操作系统信息等敏感信息。这些信息很容易被中间人攻击所截获甚至篡改。被发现具有设计缺陷的插件包括SEMRush Rank、Browsec VPN、MSN New Tab 和DualSafe Password Manager & Digital Vault等。

此外，研究人员也发现了部分浏览器扩展直接将敏感信息嵌入到了JavaScript中。攻击者可以利用这些敏感信息构造恶意请求。这些插件包括Online Security & Privacy extension、Equatio – Math Made Digital、Awesome Screen Recorder & Screenshot、Antidote Connector等。获得这些敏感信息的攻击者可以应用在恶意攻击或者恶意脚本中。例如通过大量请求提高非公开API的访问量以提高开发者成本，获得存储桶写入权限以托管非法内容，向开发者发送欺骗性的遥测数据甚至伪造加密货币交易订单。其中一些攻击可能会导致被泄露敏感信息的开发人员的账号被停用。

值得注意的是，Antidote Connector依赖一个名为InboxSDK的第三方库，而InboxSDK向JavaScript中引入了敏感信息，这导致依赖InboxSDK的插件也会存在敏感信息硬编码的问题。现在有超过90个扩展利用了InboxSDK，引

发了供应链上的隐私泄露问题。

Symantec建议开发者不要在客户端存储凭证等敏感信息，并将传输手段从HTTP切换到HTTPS。对于用户，Symantec建议考虑将这些扩展程序移除，直到开发人员解决不安全HTTP传输或者凭证硬编码的问题。Symantec强调这些具备缺陷的扩展的风险不仅仅理论上存在，实际上未加密的流量很容易被捕获，用于分析、网络钓鱼或其他定向攻击。

最重要的教训是，即便一款插件可能拥有庞大的用户群或者由知名个体开发，这款插件并不一定能确保在加密方面采取最佳实践。用户需要仔细审查扩展程序使用的协议和共享的数据，以确保自己的敏感信息真正安全。

（参考链接：https://thehackernews.com/2025/06/popular-chrome-extensions-leak-api-keys.html）

中国联通发布全球首款抗量子安全手机 开启量子通信安全新时代

（郑　直　整理）

2025年5月16日，中国联通正式推出市场上首款抗量子安全手机，这是一款面向量子计算时代的高安全性通信设备，旨在应对量子计算机对传统加密算法的潜在威胁。该手机通过多重加密技术和量子安全解决方案，为政府、企业及行业用户提供高度安全的通信保障。

随着量子计算技术的发展，传统加密算法（如RSA、ECC）可能被量子计算机快速破解，因此抗量子密码算法（PQC）和量子密钥管理成为信息安全领域的战略方向。中国联通推出的抗量子安全手机正是为了应对这一挑战，确保通信数据在量子计算时代仍能保持安全。

该手机采用“终端+算法+应用”的三层防御体系：基于国产旗舰机型（如华为Mate 70系列）深度定制，确保硬件和软件供应链的安全可控；集成抗量子密码算法（PQC□和国密算法（SM系列），形成“双重防护机制”，既能抵御量子计算机攻击，又能防范传统网络威胁；内置量子随机数生成芯片，提供真随机密钥，增强通信加密的不可预测性。支持加密音视频通话、加密即时消息、文件加密传输及群组加密通信。配备强制水印、阅后即焚、防截屏/录屏/录音等功能，防止敏感信息泄露。提供在线管控策略、终端加固等服务，满足政府、司法、能源等行业的特殊需求。该手机已在多个行业试商用。

该产品由中国联通中讯设计院与本源量子合作，开创性地将抗量子密码算法与国密算法融合，取得了量子安全加密技术的革新突破。作为全球首款抗量子安全手机，该产品不仅填补了市场空白，还为量子计算时代的通信安全提供了可复制的解决方案，推动数字中国建设。

（参考链接：https://news.qq.com/rain/a/20250519A092AZ00）

存内计算的范式革新与技术突破

（郭建军　整理）

传统冯·诺依曼架构“计算–存储分离”陷入能效困局：数据在运算与存储单元间高频搬运，使AI任务能效不足0.1TOPS/W，访存能耗占比超90%。存内计算（IMC）以“存储介质原位运算”破局，将矩阵乘法等核心操作嵌入存储阵列，从硬件层重构算力供给逻辑，理论能效提升百倍，成为后摩尔时代算力突破的核心方向。

技术演进的三大核心突破

存内计算围绕硬件介质、算法适配、安全增强展开系统性攻关：

1. 硬件介质：非易失存储的“算存一体”RRAM（忆阻器）、MRAM（磁阻存储器）等

非易失存储是核心载体。北京大学基于HfO_2基RRAM阵列实现全原位矩阵运算，推理能效达257TOPS/W，CIFAR-10任务延迟降低72%；清华大学针对MRAM“写不对称性”难题，提出自旋轨道力矩（SOT）辅助策略，训练–推理兼容性提升3倍，成果登刊《Nature Electronics》。

2. 算法适配：从“推理加速”到“训练支撑”神经网络训练需反向传播的复杂操作，对存储阵列双向操作性要求严苛。浙江大学创新“可转置存算阵列”架构，首在28nm工艺芯片支持端到端训练，ResNet-18（CIFAR-10）训练精度达75.3%，打破“存内计算仅能推理”的技术桎梏。

3. 安全增强：能效与隐私的协同设计 存内计算“原位运算”易暴露中间态，面临侧信道攻击风险。上海交通大学提出“容偏混淆防御架构”，在RRAM阵列嵌入随机阻态扰动层，功耗侧信道攻击信息泄漏量降低89%；同时利用近似运算特性，将防御能效损失控制在15%以内，实现安全与能效平衡。

产业化落地的核心挑战

存内计算规模应用需跨越三道关卡：

1. 材料一致性：RRAM等新型存储的器件间阻态离散性导致计算精度损失，需构建“器件–电路–算法”三级容错机制；

2. 工具链缺失：缺乏非易失存储的存算协同编译、仿真平台（当前仅清华大学“UniNDP”覆盖DRAM近存计算）；

3. 跨域协同设计：需深度耦合类脑计算启发、量子材料创新与软件定义调度，推动“硬件定义算力”向“场景定义架构”升级。

存内计算本质是“存储即计算”的范式革命。随着非易失存储工艺成熟、算法–硬件协同理论完善，这一技术将重塑AI芯片、边缘计算的底层逻辑，成为后摩尔时代算力突破的关键变量。

（参考链接：
https://news.sohu.com/a/753580562_478183）

专家混合架构

（史瑞泽 整理）

截至目前，专家混合（Mixture of Experts，MoE）架构已成为大语言模型（Larger Language Model，LLM）研究的核心方向之一，因其在计算效率、模型可扩展性和任务适应性方面的优势而备受关注。

MoE的基本思想是通过引入多个“专家”子网络，并在每个输入上仅激活其中一部分，从而实现参数的稀疏激活。这种机制使得模型在保持高表达能力的同时，显著降低了训练和推理的计算成本。

近期的研究主要优化了MoE架构。例如，MoLE方法通过将专家操作映射到共享的低维潜在空间，并进行专家特定的变换，显著减少了参数数量和计算需求，同时保持了模型的表达能力。

此外，研究人员还开发了MoE–X模型，通过在每个专家内强制稀疏激活，并重新设计路由机制，提升了模型的可解释性，使得每个专家更专注于特定的语义特征。

在资源受限的环境下，MoE模型的内存效率也得到了验证。相应研究表明，MoE模型在相同的内存预算下，能够超越密集模型的性能，尤其在推理阶段表现出更低的KV缓存需求和更高的吞吐量。

为了进一步提升MoE模型的效率，微软提出了Pyramid–Residual MoE架构，通过在模型的不同层使用不同数量的专家，并结合残差连接，减少了模型的总体参数量，同时保持了性能。

尽管MoE架构在多个方面展现出优势，但

仍面临如训练稳定性、专家负载均衡和路由策略优化等挑战。未来的研究可能会集中在动态专家选择、跨任务迁移能力以及与新型硬件的集成等方向，以进一步提升MoE模型的实用性和效率。

（参考链接：https://arxiv.org/abs/2503.07639）

机器学习系统（MLSys）

（陶宇飞 整理）

机器学习系统（MLSys）领域近年来随着大模型技术的爆发式发展迎来了前所未有的关注度。这个交叉学科方向的核心目标是通过系统级优化来解决机器学习模型在训练、推理和部署过程中面临的效率瓶颈问题。从学术界的最新研究动态来看，大模型训练系统的优化无疑是当前最炙手可热的研究方向，这主要源于ChatGPT等大语言模型商业化落地过程中暴露出的显存墙、计算效率等关键技术挑战。

在大模型训练系统领域，分布式并行训练技术是最关键的突破口。传统的数据并行方式在应对千亿参数模型时已经捉襟见肘，研究者们开发出了更加精细的并行策略。以英伟达开源的Megatron-LM为代表的张量并行技术，通过将模型参数矩阵按维度切分到不同计算设备，实现了模型层面的分布式计算。微软开发的DeepSpeed框架则创新性地提出了Zero冗余优化器，通过智能划分优化器状态、梯度和参数，将显存占用降低了多达8倍。更复杂的混合并行方案如Alpa框架，可以自动为不同模型结构选择最优的并行组合，这些技术突破使得训练万亿参数模型成为可能。

显存优化是另一个重点研究方向。由于大模型的激活值会消耗大量显存，研究者提出了梯度检查点技术，通过牺牲部分计算时间来换取显存空间的释放。加州大学伯克利分校团队开发的FlashAttention创新性地从IO感知的角度重构了注意力计算过程，将训练速度提升了2-4倍。异构训练技术如PatrickStar则通过动态管理CPU和GPU内存，进一步突破了单卡显存限制。这些技术创新使得在有限硬件资源下训练超大模型成为现实。

在编译器优化层面，陈天奇团队开发的TVM框架引领了AI编译器的研究方向。通过引入自动调度生成和算子融合技术，TVM可以针对不同硬件后端生成高度优化的代码。阿里巴巴开源的BladeDISC专注于动态shape场景下的编译优化，解决了大模型输入长度可变带来的性能瓶颈问题。这些编译器技术显著提升了模型在推理阶段的执行效率。

值得关注的是，大模型训练系统的研究正在从单纯追求性能向易用性方向发展。ColossalAI 框架提供了统一的接口来配置各种并行策略和优化技术，大大降低了使用门槛。HuggingFace的accelerate库则进一步抽象了分布式训练的复杂性，使研究者可以专注于算法开发。这种工具链的完善对整个领域的健康发展至关重要。

当前该领域仍面临诸多挑战：如何设计更通用的并行策略以适应不同模型架构？如何降低分布式训练的通信开销？如何在保证训练效率的同时兼顾模型的收敛性？这些问题的解决需要算法和系统专家的深度协作。从ICML2024等顶会的最新论文来看，自动并行技术、通信压缩算法和新型硬件适配将成为未来的重点研究方向。

总的来说，大模型训练系统的研究正在推动整个机器学习系统领域的技术革新。这些突破不仅服务于当下的AI应用，也为未来更大规模的智能模型奠定了系统基础。随着技术的不断演进，我们有望看到更加高效、灵活和智能的机器学习系统支撑起下一代人工智能的发展。

（参考链接：https://blog.csdn.net/audyxiao001/article/details/141472839）

批判性思维：科研路上的破雾利剑

引言

时光飞逝，不知不觉已经进入了研究生的第五年生涯。作为硕转博的“实验室留守者”，见证同窗们陆续毕业时，我总会在工位前恍惚片刻，那些在论文返修中反复打磨的时光，此刻都化作键盘上斑驳的磨痕。相比于其他从本科就开接触科研的同学，我进入科研生活的时间相对更晚，虽然到如今也是有所成果，但也是更能体会到个中艰辛。正值此时收到实验室的季刊约稿，希望能将自己科研路上的感悟分析给大家，给大家一个继续勉励前行的动力。

初入科研

虽然已经过去了4年，到现在仍然清晰地记得那年暑期接到了文明老师的电话，知道了自己成功加入文老师团队的消息，不由得十分欣喜和惶恐。欣喜的是终于站在了科研的门口，惶恐的是自己毫无科研基础，连文献检索都磕磕绊绊，生怕跟不上团队的节奏，拖了大家的后腿。

在这种忐忑与期待交织的心态中，文老师给我了一个课题：通过统计和突变分析定位缺陷相关变量。对于我这样一个科研小白，进行这个项目需要从头开始学习所有前置知识，还要阅读大量文献。前置知识主要涉及程序分析技术，该技术是对程序行为进行建模的方法，在代码优化，漏洞检测等领域有大量应用。然而当时国内对于程序分析的课程和资料相对不足。幸运的是，南京大学的李樾老师首次在B站上开启了软件分析课程，而我也作为第一批学生在线上看完了所有的课程，课程系统地讲述了程序分析基础和进阶知识。对于近年来发表在软件工程、安全以及程序语言领域顶会的文章，它们所用到的程序分析的相关知识总体没有超出过课程的范围，这便说明了该课程的重要性。对于能看到本文的读者，如果你是从事系统安全相关的研究，我强烈推荐李樾老师和北京大学熊英飞老师的软件分析课程，从此海阔凭鱼跃，天高任鸟飞。

在学完了项目所需的前置知识之后，系统地完成一个科研项目并投稿论文，对我来说仍是一件困难的事情。这是因为一篇论文远不止是完成代码，更需要系统地阐释研究动机，框架设计等。我一度认为写作一篇英文论文的难度超过了写代码和做实验的难度。写论文的难点主要体现在两个方面，首先当时还没有ChatGPT能润色论文，导致论文出现大量语法或拼写错误。另一个更重要的难点在于对内容之间的逻辑梳理不足：论文需要句与句、段与段之间逻辑链条环环相扣，否则会显得牛头不对马嘴。这些难点导致我的第一稿论文几乎每一句都有问题，所幸在文老师的指导下，论文在几乎被重写一遍后，我能够在研一就投稿至A类会议。

虽然满怀期待，但论文仍在几个月后被拒。第一次看到审稿人的意见，才意识到一项工作从完成到完整的不易。审稿人指出了论文的各种缺点，例如我们方案的性能依赖于某些参数的设置，而我们未能充分地讨论这些参数的影响，这成为论文被拒的一个主要原因。随后我们总结了论文的各种问题，并进行了详尽的修改，投稿至另一个A刊。数月之后收到了大修的意见，令人震惊的是审稿意见比之前的会议还多，审稿人要求增加基线工具，在新的数据集上做实验等。我认为虽然期刊相比会议对论文的完备性要求更高，但投稿期刊的好处也显而易见，由于审稿人不会发生改变，只要按照审稿意见仔细地修改，论文大概率都能取得好的结果，于是在经历了大修、小修后，最终幸运地得到了录用。

渐入佳境

时间来到了研二下，在文老师的建议下，我申请了硕转博，继续攻读博士学位。与硕士阶段最大的不同是自主性：不是等着老师给课题，而是需要自己主动挖掘idea，并设计和实现方案。我首先开始了大量的论文阅读。如前文所述，程序分析的课程已经涵盖了软件工程和安全领域的大部分基础知识，并且该领域侧重于将已有的理论应用到实际问题中。例如对于最经典的漏洞定位问题，现有方法大多都会用到控制流分析、数据流分析、代码切片、值流分析等程序分析技术，再辅以神经网络模型/大模型来对提取的特征进行处理。

总的来说，阅读和理解本领域的顶会论文没有太大的难度。然而对于博士生，更重要的是带着批判性的思维给出独到的见解。我根据自己的科研习惯，总结找到新idea的有效方式是去做“we are the first”型的工作。这类工作可以是开拓新场景，例如解决被先前工作忽略的某种场景；也可以是将现有方法迁移到一个新的研究标的，例如将对Java项目的解决方案迁移到对Rust项目中。这类方式能够很好的体现新颖性，同时实验结果也通常能在特定场景中超越基线工具，从而使得论文更容易被录用。

有一个“we are the first”型工作的更具体的例子。研究背景是关于安卓应用组成成分分析，研究人员通过检测安卓应用所使用的第三方库版本，通过在公开漏洞数据库上搜索该第三方库版本是否已经被披露了安全漏洞，从而检测安卓应用是否被已有漏洞所影响。研究的主要难点在于安卓的代码在被发布时常常经历扰动（包括代码混淆和代码优化）。尽管相关的顶会顶刊论文有数十篇，我发现先前论文仅仅考虑代码混淆所带来的挑战，忽略了代码优化这一场景。于是我通过解决这个新的挑战，成功投稿了一篇A会并顺利得到了录用，后来还获得了ACM SIGSOFT杰出论文奖。

通过这几次的经历，我发现博士期间的科研需要更多批判性思维的。需要多下功夫，多思考先前工作的局限性，用批判性思维开拓认知的疆域，思考并解决新的挑战，找到适合自己的科研路线。

大模型时代的科研探索

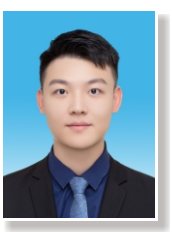
随后时间到了第四年下学期，我受到研究生院的资助前往新加坡南洋理工大学进行了为期半年的学术访问。在那里我看到了更广阔的科研世界。

与国内实验室单兵作战不同，这里的团队采用有组织的科研模式，合作导师向我们讲述了他的科研框架，即采用‘大模型智能体agent+代码理解’来解决软工和安全领域的问题。而大家需要做的，就是通过合作交流，思维的碰撞将这一框架复用到各种任务上，同时争取发表更多的论文。

通过这段经历，我意识到在大模型时代科研的范式已经发生了变化，AI for science已经成为主流。我们需要拥抱这一趋势，让自己的科研适应这一变化，才能更好的产出成果。

致 谢

回首五年学术征程，最应感恩的是文明教授的谆谆教诲。文老师以深厚的学术造诣为我搭建科研阶梯，从论文框架的构建到实验参数的推敲，让我在博士期间完成了多篇 A 类论文的突破。同时也感谢金老师给我的鼓励和支持，感谢实验室大家庭的温暖相伴，这些并肩作战的日夜终将成为记忆里的璀璨星辰。愿吾辈科研人既能如利剑劈开学术迷雾，亦能如春苗扎根研究沃土，在厚积薄发中收获真理的硕果。



谢子凡
2020级硕博连读生
研究方向：软件安全、软件测试与分析、代码大模型
Email：xzff@mail.hust.edu.cn

2025届毕业生成果展示

博士毕业生

面向漏洞检测大模型的对抗鲁棒性研究

姓　　名：杜小虎
研究方向：代码漏洞检测、对抗攻击
导　　师：文　明
指导老师：文　明
E-mail：xhdu2023@outlook.com
QQ：2571167
联系电话：18111632717
毕业去向：上海华为技术有限公司

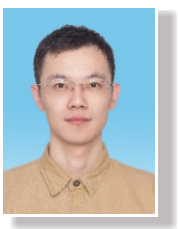
近年来，基于代码大模型的漏洞检测技术在对抗样本鲁棒性方面面临严峻挑战。现有对抗测试和模型增强方法存在三个主要局限性：标识符替换策略缺乏对代码上下文语义的考量，扰动生成过程未能针对模型决策边界进行优化，二分类训练模式导致模型过度依赖表层语法特征。为此，文章提出一种先测试再增强的解决方案：首先通过对抗测试揭示模型脆弱性，继而实施针对性的模型增强策略。

在对抗测试方面，首先提出标识符级测试框架BeamTest。通过分析代码上下文语义信息的影响，发现模型对if/for语句中标识符扰动表现脆弱。BeamTest采用三重机制：基于上下文的标识符替换策略、语义保持的候选标识符生成、束搜索算法迭代替换。实验表明其测试成功率提升了55.96%。其次提出语句级测试方法SLODA，创新性地设计两类分布外扰动策略：通用代码转换扰动和基于对向标签的代码片段插入。它们显著提升扰动针对性，并将测试成功率提升了77.21%。

在模型增强方面，提出VulLLM方法，它从两个维度进行优化。在数据层面，将对抗扰动纳入训练集增强多样性，防止模型依赖虚假特征；在训练层面，采用多任务指令微调，将二分类任务扩展为包含两个辅助任务的多任务学习，增强模型对漏洞上下文的理解。实验表明，该方法使整体F1值提升了8.58%。

存储系统中冗余数据高效检测机制研究

姓　　名：靳晓忠
研究方向：数据去重、分块、相似性检测
导　　师：刘海坤
指导老师：刘海坤
E-mail：3148438700@qq.com
QQ：3148438700
联系电话：13720244211
毕业去向：福建省莆田学院



冗余数据删除涉及三项关键技术：基于内容定义的数据分块、相同数据检索/去重、以及相似数据的差量压缩。输入数据先后被这三项技术处理并最终被写入存储系统。随着数据量的增长，数据分块的计算开销以及冗余数据检测过程带来的数据检索开销都急剧增加，造成存储系统的吞吐率大幅下降。针对这些问题，以提升冗余删除的计算效率为目标，通过缩减冗余数据搜索空间的方法，分别从以下三个方面展开研究，具体内容如下：（1）针对当前数据分块算法检测范围广、扫描速度慢、涉及的计算量大等问题，提出基于搜索空间跳跃来加速数据分块的机制，提升了数据分块的吞吐。

（2）针对当前数据块粒度的重复数据检索检测范围过广、检测效率过低的问题，提出了基于I/O因果关系的相同冗余数据检测机制。提升了冗余数据检测效率和去重率7倍左右。（3）对于相似数据的冗余检测，设计了基于近似匹配的相似性检测机制，在使用细粒度的特征增加检测精度的同时使用近似匹配缩小了相似数据块的搜索空间。在维持高去重率的前提下提升了吞吐。

冗余感知的图处理加速方法研究

姓　　名：牛富平
研究方向：图计算、近数据处理
导　　师：廖小飞
指导老师：岳建辉
E-mail：811073137@qq.com
QQ：811073137
联系电话：13627102510
毕业去向：小米科技（武汉）有限公司

图结构的稀疏性带来了数据结构不规则、存取粒度多样、时空分布不均匀等处理特征，对传统通用计算机体系结构在计算/访存等方面均提出了严峻挑战。研究发现图处理任务执行效率不佳的根本原因在于时空冗余难题，故开展了冗余感知的加速方法研究。

针对图随机游走算法，提出 SSD 内多层级并行处理加速方法以缩短数据移动路径和消除冗余访存。该方法感知图数据访存冗余规律，设计与固态硬盘访存模式相匹配的数据布局，采用在闪存芯片附近更新游走器的子图固定数据流，建立包含板级、通道级和芯片级的多层结构，以充分发掘闪存芯片的 I/O 并行性。

针对图神经网络训练，提出 SSD 内集中式异步处理加速方法以消除重复数据访问。该方法在运行时感知批次内数据依赖关系，提出避免重复访问相同数据块的最优调度模式，从而最大限度地利用闪存芯片聚合带宽，匹配计算和访存模块的延迟与吞吐率，实现高效的模型训练。

针对超图神经网络推理，提出基于细粒度数据依赖感知的加速方法，以精准捕捉推理请求间的数据依赖关系，进而消除冗余访存与计算。该方法采用流式处理体系，设计细粒度的子图重叠识别机制以实现快捷跨请求拓扑数据重用，并根据推理请求时间间隔决定调度策略以降低维护成本和处理开销。

面向边缘智能服务流程的时延优化研究

姓　　名：牛　鑫
研究方向：边缘智能、分布式计算
导　　师：余　辰
指导老师：余　辰
E-mail：956070265@qq.com
QQ：956070265
联系电话：18037596129

边缘智能将资源下沉至靠近数据源的边缘侧，支持实时服务。其任务流程包括分配、执行与完成，但低时延服务面临挑战：任务分配阶段服务器过载、执行阶段资源受限致模型选择与划分复杂、完成阶段冗余计算增加延迟。针对上述挑战提出优化策略。

针对边缘服务器过载问题，设计了基于终端设备时空位置的计算回载机制。通过构建马尔科夫链预测终端设备时空位置，边缘服务器结合预测位置与负载状态，将任务回载至空闲终端设备，缓解负载压力，提升执行效率。实验表明，该机制相较多种基准方法，最高可降低67.35%的时延。

为应对边缘设备计算资源受限与动态变化问题，设计了自适应模型选择与划分机制。该

机制基于时延预测框架，结合设备状态预测不同模型执行时延，并量化匹配任务需求，选择最优模型。引入适应度概念，动态划分模型计算任务。实验表明，该机制最高可降低5.92%任务时延，提升执行效率。

针对冗余计算增加任务完成时延的问题，设计了基于高排队时延与计算迁移的计算任务保障机制。首先将计算任务分为新到达的与部分完成的两类。新到达的计算任务通过迭代优化选择早退出点，部分完成的计算任务根据状态与时延要求动态调整退出点。实验表明，任务完成率超90%，时延最高降低61.81%。

边缘数据中心的高能效系统架构研究

姓　　名：裴强宇
研究方向：数据中心体系结构、ML系统
导　　师：余　辰
指导老师：余　辰
E-mail：492261796@qq.com
QQ：492261796
联系电话：15927465885
毕业去向：华为技术有限公司

随着计算逐渐从云端延伸至网络边缘，边缘数据中心近些年来蓬勃发展。相比于传统的大型云数据中心，边缘数据中心的电能利用效率（Power Usage Effectiveness，PUE）普遍偏低，显著加剧了其能源消耗压力。为了改善边缘数据中心的PUE，基于对边缘数据中心和边缘负载的特征和需求的深入挖掘，提出了一套精细化、自适应的高能效系统架构。为了满足多样化边缘数据中心对侵入式影响程度和能效提升幅度的不同要求，设计了具有不同技术特点和适用场景的多元化技术路线，具体包括以下三个方面：（1）从负载管理的角度，提出了一种热效应感知的推理任务重分配架构，包括一种准确、轻量的硬件功率消耗和温度表征估计方法，和一套具有多个级别的、硬件温度导向的应用部署与请求调度机制。（2）从制冷控制的角度，提出了一种硬件级的细粒度温水制冷架构，包括能够为每个硬件提供定制化制冷水的内外双循环水路，基于均温板的热点自适应感知的气液微循环水冷头，以及一种基于自然散热的定制化制冷调控机制。（3）从负载管理与制冷控制协同优化的角度，提出了一种基于分区的制冷调控和推理负载调度协同架构，包括一种基于冷水和温水制冷以利用“甜点”现象的分区水冷架构，以及一种基于深度强化学习的推理负载分区调度机制。

图结构感知的高性能图模式挖掘技术研究

姓　　名：齐　豪
研究方向：系统软件与体系结构、
导　　师：金　海
指导老师：张　宇
E-mail：2390631000@qq.com
QQ：2390631000
联系电话：16619714096

论文聚焦于构建面向图模式挖掘任务的高性能执行环境。探索并利用现实世界中图结构的特征，突破图模式挖掘的性能瓶颈，以高效挖掘海量图数据中蕴含的商业价值和科学洞察。

针对图模式挖掘任务中更新和计算性能难兼顾问题，提出差异性感知的图存储和更新机制，实现高效的图更新和图计算。核心思想是在确保数据局部性和有序性的基础上，构建低开销的差分索引，从而提高搜索效率，并规范数据移动，以提高数据局部性。

针对图模式挖掘任务中冗余计算开销大问题，提出相似性感知的计算共享范式。通过分析顶点的计算特征，揭示顶点在计算过程中的

相似行为，并提出相似性感知的执行模型，以高效处理共享相同计算的顶点。基于该模型，提出中间表示对冗余计算进行建模，并分析和优化中间表示，以生成执行计划来指导图模式挖掘的过程。

针对图模式挖掘任务中并行效率低问题，提出图结构感知的并行计算架构。提出不变性感知的增量执行模型，以避免计算全图所带来的不必要开销。通过分析图结构的相似性和差异性，揭示该模型运行时存在大量冗余计算和计算多样性的特点。设计冗余检测模块和混合计算机制，以提升并行效率。此外，提出结合数据并行和流水线的架构，以进一步提升计算并行性。

联邦学习鲁棒性评估与增强关键技术研究

姓　　名：万　伟

研究方向：人工智能安全

导　　师：胡胜山

指导老师：胡胜山

E-mail：wanwei_0303@hust.edu.cn

QQ：2411806418

联系电话：18162327140

毕业去向：澳门城市大学

联邦学习通过模型参数交换解决数据孤岛与隐私保护问题，但开放协同机制易受拜占庭与后门攻击，且数据异构增加防御难度。为此，本文提出集鲁棒性评估与增强于一体的技术路线。

鲁棒性评估方面，针对现有拜占庭攻击投毒强度难以精细控制的问题，提出了基于最小距离优化的精细化拜占庭攻击方法FPG。该方法以全局模型参数与攻击者期望次优解间距离最小化为目标，结合动态实时搜索策略与基于网络拓扑的误差反馈机制，实现对恶意模型更新的逐轮迭代调优，并通过严格的收敛性分析证明攻击有效性。针对数据缺失条件下后门植入难题，设计了无数据依赖的后门攻击 DarkFed。该方法利用影子数据集进行后门特征构建，并通过属性模仿技术仿真良性更新的模长、分布与一致性，增强攻击隐蔽性。

鲁棒性增强方面，针对数据异构场景下多元化攻击防御不足的问题，提出了多阶段协同防御算法 FPD。其第一阶段通过可靠客户端选择策略剔除可疑参与者；第二阶段基于更新相似度过滤高度相似的恶意更新；第三阶段运用谱分析剔除分布异常的模型参数；第四阶段引入自编码器对微小扰动进行去噪处理，从而实现对拜占庭攻击的多层次防护。针对复杂后门攻击防御难题，提出了恶意感知型后门防御算法 MARS。该算法以神经元后门能量为检测依据，结合层次化能量提取技术放大后门特征，最终通过 Wasserstein 距离聚类实现对后门模型的精准识别。

面向机器学习应用的服务器无感知计算系统优化研究

姓　　名：吴　浩

研究方向：服务器无感知计算

导　　师：吴　松

指导老师：吴　松

E-mail：598153737@qq.com

QQ：598153737

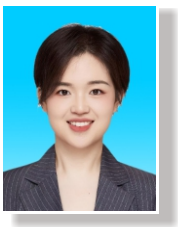
联系电话：15377528485

毕业去向：香港大学博士后

论文针对现有服务器无感知计算（Serverless）系统在部署机器学习应用时面临的函数资源分配不灵活、函数运行时启动开销大、函数间通信效率低的问题展开研究。为解决在基于Serverless架构的公有云平台上部署机

器学习训练任务时函数资源分配不灵活的问题，提出了基于资源需求感知的动态函数资源规划方法，从而有效降低训练任务的部署成本和执行延迟。实验结果表明，所提出的方法相比现有方案，在模型训练阶段，能够减少41%的延迟并降低38%的成本；为解决基于Serverless架构的机器学习推理系统中GPU函数运行时启动慢和空间占用大的问题，提出了基于GPU流机制的函数运行时轻量化方法，从而有效降低推理应用部署的运行延迟和GPU内存占用。实验结果表明，相较于现有方法，提出的方案能够减少Serverless推理系统中82%的GPU内存空间占用和98%的推理延迟，同时使系统吞吐量提高6.7倍；为解决基于Serverless架构的机器学习推理系统中GPU函数数据通信效率低的问题，提出了以GPU为中心的无状态函数高效通信方法，从而有效降低推理应用部署的端到端延迟。实验结果表明，相较于现有方案，所提出的方法能够将机器学习推理应用的端到端延迟降低90%，并使吞吐量提升12倍。三种方法结合机器学习应用的特点对Serverless系统中函数资源分配、函数运行时管理以及函数数据通信三个方面进行优化，有效降低Serverless系统中机器学习应用部署的成本和运行延迟。

面向边缘智能的资源调度优化机制研究



姓　　名：武　静

研究方向：边缘智能、资源优化

导　　师：余　辰

指导老师：余　辰

E-mail：wujinghust@hust.edu.cn

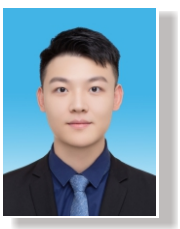
QQ：1932597204

联系电话：13315348169

毕业去向：中铝郑州有色金属研究院有限公司

边缘智能作为边缘计算与人工智能深度融合的新型计算范式，通过将智能任务下沉至网络边缘，来提升任务响应速度。然而，边缘智能任务的低时延与高算力需求，以及任务本身固有的动态性与关联性，对边缘数据中心资源的高效利用提出严峻挑战。现有资源优化机制侧重于静态优化和局部优化，极易引发资源分配不均衡。为此，文章从资源优化时间维度与空间维度出发，提出动态优化与全局优化：首先，延展时间跨度，提出动态资源优化机制，解决资源闲置问题。深入挖掘智能任务执行时延的变化性，充分利用任务内子任务间的时间依赖关系，根据上游子任务的执行时延，动态优化下游子任务的资源分配，显著降低资源占用。其次，细化空间粒度，提出资源复用机制，解决资源重复分配问题。针对智能任务间DNN层重叠现象，提出DNN“再对齐”概念，通过对DNN进行二次切分，合理提取公共层，实现跨任务的计算共享与资源复用，显著降低资源占用。最后，拓宽空间广度，提出联合资源优化机制，解决资源错置问题。针对边缘智能任务资源共享场景，提出将性能干扰、任务优先级需求和公平性约束纳入统一优化体系，实现跨任务的联合资源优化，有效提升系统吞吐。

面向代码扰动的软件组成分析与补丁存在性检测技术研究



姓　　名：谢子凡

研究方向：软件安全、程序分析

导　　师：文　明

指导老师：文　明

E-mail：xzf1244@gmail.com

QQ：1073200854

联系电话：15295525901

近年来，随着开源软件在各类应用中的广

泛部署，二进制软件中常嵌入大量开源组件。然而，这些开源组件往往存在大量漏洞，导致软件供应链安全问题日益突出。针对这一问题，目前主要采用软件组成分析和补丁存在性检测技术。论文提出：

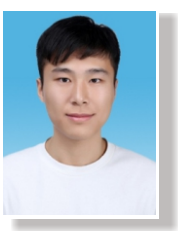
（1）抗优化扰动语义提取的软件组成分析。提出了一种抗优化扰动的软件组成分析方法，其核心思路在于通过细粒度静态分析模拟优化过程，利用调用图与数据流追踪对优化前后函数特征进行重构，并采用正则表达式增强的模糊签名机制，结合跨函数数据流分析，实现内联后的语义匹配。

（2）抗混淆扰动语义分析的补丁存在性检测。设计了一种抗混淆扰动语义分析的检测方案，其主要思路是在函数内部提取与补丁逻辑相关的执行路径、谓词约束及关键变量，并利用树状语义表征和动态规划匹配算法实现路径摘要的精确比对。

（3）抗演化扰动语义追踪的补丁存在性检测。构建了一种抗演化扰动的漏洞补丁追踪与检测框架，其主要思路在于首先通过程序依赖图和时间序列模型构建补丁演化四元组追踪模型；随后利用符号执行技术提取跨版本深层语义特征，从而生成具有跨版本鲁棒性的补丁签名，并在语义层面对特征等价性进行比较。

论文围绕现有图式区块链事务处理方法在应对交易负载的动态规模特征、数据访问倾斜特征以及复杂分支逻辑特征时弹性扩展能力不足的问题展开研究。针对图式区块链存储模型在动态负载规模下的扩展性受限问题，提出一种负载规模感知的弹性图式存储机制MorphDAG。该机制能够在当前负载规模下选择最优的存储并发度，有效适应负载规模的动态变化，实现系统性能与安全性的平衡；针对图式区块链并发控制机制在面向数据访问倾斜时并发处理效率低的问题，提出一种数据访问感知的图式自适应并发控制机制Nezha。该机制面向普通转账交易负载，通过双模式并发控制方法有效适应冷热账户在冲突特征上的差异，从而提升整体并发处理效率；针对图式区块链预执行机制在面向复杂分支逻辑时准确性低且低效的问题，提出一种分支逻辑感知的图式细粒度预执行机制Seer。该机制面向智能合约交易负载，引入分支预测的思想，通过预测所有与状态变量相关的分支方向提升了预执行结果的可复用性，从而增强预执行对实际执行的加速效用。上述三种机制分别面向图式区块链的存储与执行阶段进行优化设计，能够有效适应交易负载在不同阶段所呈现出的特征变化，综合提升了系统在存储与执行方面的性能表现与弹性扩展能力。

基于负载特征的图式区块链高效事务处理机制研究



姓　　名：张世荣
研究方向：区块链、分布式计算
导　　师：肖　江
指导老师：肖　江
E-mail：zsj19941203@126.com
QQ：454311804
联系电话：18362983295
毕业去向：南京邮电大学

服务器无感知计算系统垂直扩展机制研究

姓　　名：张信民
研究方向：服务器无感知计算、轻量级虚拟化、容器
导　　师：吴　松
指导老师：吴　松
E-mail：kingdo.m@foxmail.com
QQ：1440852110
联系电话：18571851973 / 13081661973

毕业去向：华为技术有限公司

为了现有基于水平扩展的服务器无感知计算平台面临的冷启动开销高、函数间通信效率低和调度延迟大等性能问题，提出了基于垂直扩展策略的服务器无感知计算系统设计。

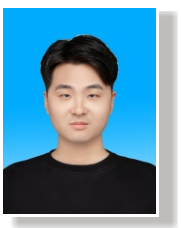
高效微虚拟机内存垂直扩展机制：针对现有微虚拟机内存扩展机制存在显著性能开销的问题，提出了面向函数的内存扩展机制。该机制通过块粒度的内存动态扩缩和基于页面交换的内存预填充技术，显著提升了微虚拟机的内存扩展性能，从而将垂直扩展策略应用于服务器无感知计算平台的理论构想转化为可行的实用技术。

面向垂直扩展的高效函数间通信机制：针对现有平台中函数间通信效率低的问题，提出了基于状态函数的内存共享机制。该机制在确保函数实例之间资源隔离的前提下，实现了低延迟的函数间数据交互，显著提升了垂直扩展策略下的数据通信效率，从而能够高效支撑复杂的数据交互密集型应用。

基于垂直扩展的可定制调度策略：针对现有平台采用统一的完全公平调度（CFS）策略，无法区分不同函数执行特点而导致调度延迟的问题，提出了基于垂直扩展的可定制调度框架。该框架综合考虑函数的初始化时间、调度等待时间和执行时间等特征，有效优化了整体平均完成时间，同时支持用户根据应用需求自定义调度策略，在调度公平性与执行效率之间进行权衡。

深度学习模型安全脆弱性迁移机制研究

姓　　名：张业超
研究方向：人工智能安全
导　　师：胡胜山
指导老师：胡胜山
E-mail：752860855@qq.com



QQ：752860855
联系电话：18362905788
毕业去向：新加坡南洋理工大学

深度学习模型在计算机视觉和自然语言处理等领域展现出优异性能，但在安全性上却脆弱，易受对抗攻击威胁。对抗攻击的“迁移性”特性使得攻击干扰能够在不同模型、数据或任务间传播。本文定义迁移性为对抗攻击干扰在不同环境中的有效性，包括跨数据迁移性、跨模型迁移性和跨任务迁移性。研究这些迁移机制有助于揭示深度学习模型的安全脆弱性，并为开发鲁棒性防御方法奠定基础。

为弥补当前研究不足，文章围绕三个核心主题展开研究：跨数据迁移的对抗扰动优化、跨模型迁移的对抗样本机理以及跨任务迁移的后门防御优化。具体贡献如下：

（1）针对跨数据迁移中的训练数据稀缺问题，提出基于动态最大最小优化的通用对抗扰动生成方法，利用目标模型的安全脆弱性降低对训练数据的要求。相较于现有对抗扰动生成方法，攻击性能提升12.11%。

（2）针对跨模型迁移缺乏解释的现状，建立基于模型光滑性和梯度相似性的对抗样本迁移性理论框架，设计稳定提升对抗样本跨模型迁移能力的代理模型训练方案，在20个不同目标模型中取得最佳迁移攻击效果。

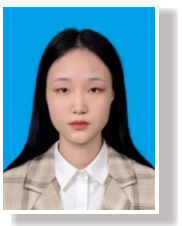
（3）针对跨任务迁移的后门防御缺乏通用方案的问题，分析现有防御在迁移学习中的可行性，发现现有反应式防御无法适应新的威胁。于是，提出基于可信核心引导的主动式干净模型训练方法，在涵盖5种编码器攻击、7种数据攻击及5个基准数据集的评估中，显著优于14种现有基线防御方案。

硕士毕业生

分布式组

面向联邦学习的通信与训练优化方法研究

姓　　名：陈雨欣
研究方向：联邦学习
导　　师：王　雄
指导老师：王　雄
E-mail：Allison_cyx@163.com
QQ：1057901453
联系电话：13940207442
毕业去向：中国农业银行股份有限公司



联邦学习是一种去中心化的分布式机器学习框架，允许多个客户端在不共享原始数据的前提下协作训练模型。为了加速训练，分别针对横向和纵向联邦学习提出了通信高效的训练优化框架。

在横向联邦学习场景下，提出了一种结合双动量更新和自适应客户端选择的联邦学习框架FedMoS。通过在客户端和参数服务器分别维护动量缓冲区，跟踪局部更新和全局更新方向，降低局部模型差异。同时基于动量结果，设计了一种自适应客户端选择方案，挑选有代表性的客户端子集参与训练，在保证无偏聚合的条件下减少采样方差，进一步加速训练过程。在纵向联邦学习场景下，提出了一种基于贡献度指导的客户选择框架P2VFL，采用截断Shapley值周期性地评估客户端的数据贡献，并结合硬件效率，优先选择既具有高质量数据又具备较强计算能力的客户端，从而减少不必要的数据交换，降低了训练过程中的通信开销，提升了模型的收敛速度。

通过多个数据集和模型下的实验表明，所提出的横向联邦学习框架有效缓解了客户端漂移问题，与现有方法相比，最高能够降低87%的通信开销，使训练效率实现了2.4-7.5倍的加速；所提出的纵向联邦学习框架将每轮训练的通信开销最高减少了40%，使训练时间加速了1.6-2.1倍，并且能够适应更大规模客户端的场景。

基于自适应默克尔树的高效区块链状态存储方法

姓　　名：伏子豪
研究方向：区块链
导　　师：肖　江
指导老师：肖　江
E-mail：1633898838@qq.com
QQ：1633898838
联系电话：18707189651
毕业去向：华为技术有限公司

区块链对底层状态存储机制提出更高的性能与可扩展性要求。然而，现有区块链存储方法主要依赖于默克尔树，受限于冗长的层级访问路径和分散的存储节点分布，导致频繁的I/O操作，严重制约系统的吞吐量。

为此，提出一种基于自适应默克尔树的高效区块链状态存储方法ZacChain，通过深入研究区块链状态数据的存储特性，以有效解决现有状态存储架构中访问延迟高、I/O开销大的问题。首先，设计了默克尔加树（Merkle Plus Trie，M+ Trie）结构，通过路径压缩和节点重构技术，显著减少状态存储的路径长度与冗余节点，从而降低状态访问过程中的I/O开销，提高读写操作的性能。其次，提出了一种自适应状态感知调整机制，基于区块链状态数据的访问频率，动态地对高频状态进行识别与迁移，将其存储到状态树的较高层，减少热状态数据的访问延迟，并有效避免冷数据的干扰，以进一步提升状态查询和更新效率。最后，设计了两阶段默克尔证明机制，通过规范化的分层验证方法，优化状态验证过程，降低默克尔证明规模，保障状态数据的一致性与完整性，增强系统的安全性。

适应混合网络的异步拜占庭共识机制设计

姓　　名：郭正轩
研究方向：区块链
导　　师：金　海
指导老师：戴小海
E-mail：stguoro@126.com
QQ：308195044
联系电话：18972115970
毕业去向：香港科技大学攻读博士

现有各类异步多值拜占庭共识协议普遍缺乏对良好网络的适应能力，在共识性能上与先进的半同步共识协议存在显著差距。针对该挑战，一类新的多值共识协议 Pako 被提出。为防止攻击者操纵领导者节点，现有多值共识协议必须在完成一致性广播后进行主节点选举。作为改进，Pako引入了提交预选主节点区块的乐观路径，节点直接提交携带聚合签名的该区块，减少了良好网络下的共识轮次。同时，Pako 利用异步二值共识（Asynchronous Binary Agreement，ABA）使所有节点一致选择乐观或悲观路径，保证协议的安全性及恶意攻击情况下的活性。通过权衡共识延迟与消息复杂度，为 Pako 实现了两种变体：Pako1 和 Pako2。其中 Pako1 在 O(n2) 的消息复杂度下实现了最优 3 轮的共识延迟，Pako2 则在最优情况下将消息复杂度降低为 O(n)，并实现了 5 轮的共识延迟。为评估 Pako 协议的性能，分别为 Pako1 和 Pako2 实现了原型系统，并与最先进的多值共识、半同步共识协议进行比较。实验结果表明，Pako1 和 Pako2 高效适应了兼具半同步与异步特征的混合网络。具体而言，与当前最先进的多值共识协议sMVBA 相比，Pako1 和 Pako2 可分别将延迟降低 51% 和 32%。

面向5G的视频流码率自适应算法研究

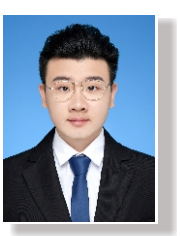
姓　　名：何逸卓
研究方向：自适应视频流
导　　师：余　辰
指导老师：余　辰
E-mail：854710227@qq.com
QQ：854710227
联系电话：15580995167
毕业去向：华为技术有限公司

现有自适应视频流算法难以感知5G和QUIC技术加持的新兴网络环境下的吞吐量变化，导致性能下降，无法充分发挥这些技术在提高视频流用户体验质量方面的优势。

针对以上问题，以5G和QUIC协同场景下的点播视频流为对象，以目标场景下网络吞吐量的实际性能为关注点，提出了一种吞吐量变化感知的码率自适应算法。算法通过分层强化学习以及跨层协作，直接获取并分析传输层实时状态信息来学习并感知网络吞吐量变化特性。其首先对用户当前正经历的网络进行分类，并切换至适配的具体码率自适应策略。随后算法利用视频实时下载状态等应用层信息，交付于码率自适应策略以决策当前视频块的码率。通过以上方式使得算法能够适应目标场景下复杂的网络条件。

以全面的码率自适应算法作为比较对象，通过模拟目标场景下的网络以开展实验。结果显示，与基准算法相比，所提出的算法将视频流的用户体验质量提高了至少14.44%。

基于服务器无感知计算的图神经网络推理资源优化研究



姓　　名：胡海川
研究方向：云计算Serverless
导　　师：余　辰
指导老师：余　辰
E-mail：huhc@hust.edu.cn
QQ：931333253

联系电话：18296157515
毕业去向：华为技术有限公司

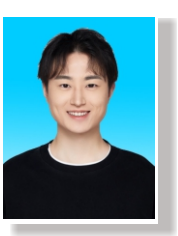
服务器无感知计算（Serverless）凭借弹性伸缩和按需付费的优势，为图神经网络（GNN）推理提供了降本增效的机会。然而，以请求为中心的通用Serverless方案在任务调度方面存在大量的计算与内存冗余，在资源管理方面难以应对GNN操作的高度耦合性，导致资源效率低下。

针对上述问题，在任务调度方面，分析发现不同GNN推理请求的计算图之间存在显著的数据局部性。在资源管理方面，观察到GNN具体由两类资源敏感性相反的子操作组合而成。基于此，设计了一套基于Serverless架构的资源高效型GNN推理服务系统λ Grapher，其设计包括：（1）以图为中心的任务调度策略：通过设置多缓冲区，将具有显著数据局部性的请求计算图合并调度，并利用图共享机制重新规划计算图，实现高效批处理，以最小化推理时的计算与内存冗余；（2）以资源为中心的函数管理机制：将云函数实例按计算与内存资源组的形式进行分类管理，分别负责处理解耦后资源敏感性不同的GNN子操作，并通过精细地编排函数协同推理流水线，以最大化系统资源效率。

基于真实的在线应用数据集，对λ Grapher系统进行性能评估。实验结果表明，在满足GNN推理时延需求的条件下，λ Grapher与现有的先进方案相比，平均可节省61.5%的内存资源和47.2%的计算资源，能够为云服务提供商带来可观的经济效益。

面向服务器无感知计算的混合存储系统研究

姓　　名：黄晗翔
研究方向：服务器无感知计算
导　　师：吴　松
指导老师：吴　松
E-mail：hndsnhuang@gmail.com



QQ：965357884
联系电话：17729014600
毕业去向：支付宝（杭州）信息技术有限公司

服务器无感知计算（Serverless Computing）是一种新兴的云计算执行模式，开发者无需管理底层基础设施。然而，云函数实例之间的通信依赖于外部存储（如远端对象存储），如果存储访问开销较大，会显著影响通信和计算效率。因此，如何优化中间数据存储系统，在降低数据访问延迟的同时控制存储成本，已成为该领域亟待解决的核心问题。

为解决此问题，提出一种高效的混合存储系统 HeStore。HeStore 充分利用不同存储介质的性能与成本特性，将内存、持久内存和固态硬盘进行分层管理，构建了一种多层次、高效能的数据存储架构。在此基础上，HeStore 结合服务器无感知计算环境中应用的访问特征，创新性地设计了动态预测引擎和强化学习决策引擎，动态预测引擎利用神经网络等技术，实时预测数据访问模式、热度和数据间的相关性；强化学习决策引擎则根据预测结果和系统实时状态，如延迟、吞吐量、存储利用率，智能决策数据的最佳放置层级、迁移时机以及缓存管理策略。通过这两个核心引擎的协同工作，HeStore能够实时调整数据在不同存储介质间的分布，使热数据优先驻留在高速介质以降低访问延迟，而冷数据或生命周期较长的数据则被迁移至低速介质中以优化存储成本。此外，系统还优化了混合存储执行层，利用PMEM直接访问、批量数据降级流水线等技术加速数据操作。通过这一系列策略，HeStore 在保证高性能数据访问的同时，有效控制存储资源的开销，显著提升了服务器无感知计算环境下的整体数据管理效率。

实验表明，相较于纯内存的远端存储系统Jiffy，HeStore的成本仅占其30%，在最好情况

下延迟仅比Jiffy增加16%，吞吐量最高可达Jiffy的87%。与基于DUO缓存策略的远端存储系统相比，HeStore的成本为其65%，吞吐量最高可达其1.83倍。

基于可信硬件的区块链高效可验证分析性查询方法研究

姓　　名：林立成
研究方向：区块链、可验证查询、可信硬件
导　　师：肖　江
指导老师：肖　江
E-mail：546140049@qq.com
QQ：546140049
联系电话：18757751525

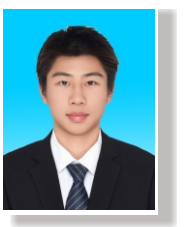
随着Web3.0生态的快速发展，去中心化应用的涌现使得区块链分析性查询逐渐成为研究热点。然而，由于区块链网络运行于拜占庭环境，分析性查询需确保查询结果的可验证性，即同时保证数据完整性和计算健全性。现有可验证查询方案大多侧重于数据完整性，而对计算健全性缺乏有效保障，导致分析性查询的可验证性难以得到充分满足。此外，现有方法未能优化查询计算过程，使得查询效率低下，难以适应大规模区块链数据分析需求。

针对上述问题，提出了一种基于可信硬件的区块链高效可验证分析性查询方法（Efficient and Verifiable Analytical query, EVA），提升查询效率，同时保证数据完整性和计算健全性。具体而言，在查询效率方面，设计了基于两阶段处理的区块链分析性查询框架，将分析性查询划分为预处理阶段和可信硬件内的查询执行处理阶段。在预处理阶段，设计了可验证数据结构的变体，完备性约束索引，采用键链接组织区块链索引数据，仅存储单一交易属性，从而提升数据检索效率。在查询执行阶段，利用可信硬件Intel SGX的明文安全计算能力以执行可验证查询计算，并通过读写一致性内存虚拟地扩大硬件安全空间，以容纳更大规模数据。同时，分别提出了索引评分机制和批量读写验证的优化方法，进一步提升查询性能。在可验证性方面，提出了基于混合证明的区块链可验证查询机制，通过完备性索引和读写一致性内存共同验证数据完整性，同时利用Intel SGX提供的远程证明服务确保计算的健全性。

实验结果表明，EVA相较于当前国际领先的区块链可验证分析性查询方法IQUERY，在服务提供商查询性能上提升了2.5倍，在客户端查询性能上提升了3倍，同时网络传输的可验证对象大小降低了1.2倍，有效提升了区块链分析性查询的查询效率和可验证性。

面向云原生的私有内核视图构建方法研究

姓　　名：潘力松
研究方向：操作系统、虚拟化
导　　师：吴　松
指导老师：吴　松
E-mail：panlisong2010@qq.com
QQ：940464190
联系电话：13397656786
毕业去向：深圳市腾讯计算机系统有限公司



云原生场景下，云服务厂商会根据用户对安全性的需求提供不同的运行时，有以KVM、Xen为代表的传统虚拟机，有以gVisor、Kata为代表的轻量级虚拟机，同时还有以docker为代表的容器。这些技术在一定程度上增强了内核隔离性，但是不同的运行时之间依然共享宿主机操作系统的内核视图（内核地址空间中的代码段和数据段）。内核视图的共享和云原生环境中租户互不信任的假设形成了矛盾，同时它也导致难以裁剪内核代码以减小攻击面、隔离不同进程的内核数据以保证数据隐

私性。

针对共享内核视图在云原生场景中带来的问题，提出了面向云原生的私有内核视图构建方法。在保证共享内核性能的前提下，支持定制私有内核视图，提高隐私数据的安全，并针对性裁剪应用的内核视图。（1）针对进程共享内核视图的这一问题，设计了基于私有内核视图页表的内核视图隔离框架，以几乎零成本的方式解耦了不同进程之间的共享内核视图；（2）针对进程隐私数据在共享内核视图中隔离性偏弱的缺陷，基于私有内核视图设计了隐私数据区，保证隐私数据区仅仅只出现在对应进程和特权进程的内核视图中，在内核态中保证了隐私数据的安全性。针对在共享内核视图下无法充分裁剪内核的问题，基于私有内核视图进行内核裁剪，进而排除了其他进程和内核线程的干扰；（3）针对不同场景下应用安全策略的定制化需求，基于私有内核视图构建方法，设计并实现了异步安全策略执行框架，为管理员提供了保护内核、分析监控进程的工具。

面向模型异构的一次性联邦学习研究

姓　　名：石月鑫
研究方向：联邦学习、分布式学习
导　　师：姚德中
指导老师：姚德中
E-mail：shiyuexin2000@163.com
QQ：450825297
联系电话：15652351218
毕业去向：北京达佳互联信息技术有限公司

模型异构的联邦学习允许具备不同模型结构的客户端协同训练。一次性联邦学习通过单轮通信缓解了高昂的计算和通信开销。然而，现有方法在处理模型异构的一次性联邦学习场景时，仍存在诸多局限性。基于模型压缩的方法难以适应一次性训练的特殊场景；基于知识蒸馏的方法过度依赖与本地数据分布相似的公共数据集；表达能力受限的小型模型在模型聚合中给全局模型引入错误知识。这些方法均会造成全局模型准确率降低。

为应对上述挑战，设计了一种面向模型异构的一次性联邦学习算法FedMHO。该算法在联邦分类任务中将深度分类模型部署在计算资源充足的客户端，而在资源受限的客户端部署轻量级生成式模型。FedMHO算法实现了数据生成、数据优化与知识聚合三个模块。在数据生成阶段，生成式模型的解码器生成合成样本。在数据优化阶段，采用一种无监督数据优化算法提升样本质量。在知识聚合阶段，基于分类模型初始化全局模型参数，再利用合成样本有监督训练。为了缓解知识聚合阶段的知识遗忘问题，提出了两种改进机制FedMHO-MD和FedMHO-SD。其中，FedMHO-MD采用多教师蒸馏，FedMHO-SD则基于自蒸馏。此外，FedMHO的泛化界限被理论证明。与最优基线方法相比，FedMHO、FedMHO-MD和FedMHO-SD的平均准确率分别提高了5.17%、8.35%和8.25%。

基于轻量化广播的高效图式拜占庭共识协议

姓　　名：王冠雄
研究方向：区块链、拜占庭共识
导　　师：金　海
指导老师：戴小海
E-mail：13613041957@163.com
QQ：757081859
联系电话：13613041957
毕业去向：武汉金山云信息技术有限公司

区块链技术凭借其去中心化架构和分布式账本特性，在数据透明性、系统安全性及记录不可篡改性方面展现出显著优势。作为区块链技术的核心机制，拜占庭容错（Byzantine Fault

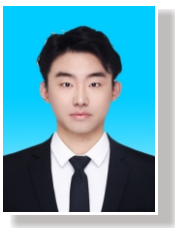
Tolerant，BFT）共识协议近年来持续受到学术界关注。为突破传统区块链链式结构的吞吐量瓶颈，一系列工作将有向无环图（Directed Acyclic Graph，DAG）拓扑结构引入拜占庭容错共识协议设计中，称为图式拜占庭共识协议。然而，现有图式共识方案普遍依赖可靠广播协议（Reliable Broadcast，RBC）进行区块广播。由于每个RBC包含3轮网络通信，这导致了显著的高延迟问题，极大制约了图式共识协议的应用潜力。

针对上述问题，提出了一种轻量化的图式共识协议LightDAG。LightDAG采用一致性广播协议（Consistent Broadcast, CBC）和普通广播协议（Plain Broadcast, PBC）替代RBC协议。该协议的两个变体LightDAG1和LightDAG2旨在在协议的最佳延迟和预期最差延迟之间进行权衡。具体而言，LightDAG1使用CBC代替RBC，在实现最佳延迟为5轮网络通信的同时保证了14轮网络通信的预期最差延迟。同时LightDAG1设计了区块检索机制来解决CBC缺乏整体性的问题。LightDAG2则使用PBC与CBC的组合代替RBC，进一步将最佳延迟降低至4轮网络通信。考虑到拜占庭节点可能利用PBC发起混淆攻击，LightDAG2中禁止引用矛盾区块。此外，LightDAG2还设计了一套高效的拜占庭节点检测机制，该机制能够在检测到恶意行为时识别并排除拜占庭节点，从而保证协议的活性并进一步提升共识效率。

为验证LightDAG协议的性能优势，分别对LightDAG1和LightDAG2进行了系统原型实现，并与现有图式共识进行实验对比。实验结果表明，LightDAG1和LightDAG2具有低延迟、高吞吐及强可扩展性等优点。相较于Tusk，LightDAG1和LightDAG2可分别将吞吐量提升至1.69倍与1.91倍。相较于BullShark，LightDAG1 和LightDAG2可分别将延迟降低10.2%与16.2%。

分布式图神经网络训练加速机制研究

姓　　名：王书林
研究方向：分布式图神经网络、机器学习系统
导　　师：王　雄
指导老师：王　雄
E-mail：953550366@qq.com
QQ：953550366
联系电话：13356209416
毕业去向：北京三快在线科技有限公司



在多 GPU、分布式图神经网络训练过程中，计算资源得不到充分利用，且往往因跨GPU、跨计算节点的数据传输延迟等问题，致使训练效率低下。

为了提升训练效率，针对静态图和动态图的不同场景，分别设计了基于异步并行和高效缓存的分布式图神经网络训练架构PSC-GNN，以及基于流水线并行的时序图神经网络训练架构Pipe-TGL。PSC-GNN将分布式图神经网络训练中同步的通信与计算过程异步并行化，消除因通信时间过长而带来的计算延迟。PSC-GNN还设计了高效的缓存模块，充分利用空闲的GPU显存空间，并合理设置缓存策略，保证缓存具有较高命中率，减少重复数据的传输，进一步提升图神经网络分布式训练效率。Pipe-TGL通过模块化设计时序图神经网络的训练架构，以适应连续时间动态图的训练特性。Pipe-TGL充分利用CPU与GPU计算资源的计算特点并合理分配计算任务，最大化发挥异构计算资源的性能优势。为了充分提升训练效率，Pipe-TGL将不同操作流水线并行化，避免了数据处理与传输时间延迟，确保每个GPU都能持续高效地参与训练过程，提升整体训练速度。

实验结果表明，PSC-GNN架构能够将分布式图神经网络训练的收敛速度加快3.6~5.0倍；Pipe-TGL架构相较于现有的时序图神经网络训练架构有约11%~45%的综合训练速度提升。

边缘计算场景中干扰感知的容器化服务放置优化方法研究

姓　　名：王源博
研究方向：边缘计算、容器技术
导　　师：陈汉华
指导老师：顾　琳
E-mail：1012079827@qq.com
QQ：1012079827
联系电话：18771778813
毕业去向：中国移动通信集团湖北有限公司

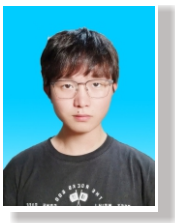
由于容器的弱隔离性，同一边缘服务器上部署的多个容器在运行时极易产生性能干扰，导致高延迟波动。这种干扰效应在资源受限的边缘节点中尤为显著，严重影响了边缘服务的质量。因此，亟需研究干扰感知的容器化服务放置优化方法来实现干扰缓解与延迟优化。

针对容器弱隔离性引发的边缘服务性能干扰问题，文章提出基于深度确定性策略梯度（DDPG）的干扰感知容器化边缘服务动态放置优化算法（BC-DDPG）。该方法首先建立了多维资源约束下的干扰感知优化模型，证明其属于NP难问题。接着通过实验测试构建容器性能干扰数据集，采用集成回归树建立干扰预测模型。为解决模型难显式表达的问题，利用无模型DDPG算法通过与边缘环境自主交互感知干扰并优化放置决策。进一步提出基于模仿学习的BC-DDPG加速算法，通过行为克隆预训练网络提升初始质量，大幅加快收敛速度。

实验表明，BC-DDPG相较于国际先进算法OLAIA和iPlace平均降低延迟18.19%和11.64%，比基线DDPG算法收敛速度提升28.49%，有效保障了边缘服务质量。

算网融合下的路由协议研究

姓　　名：徐家豪
研究方向：边缘计算、算力网络
导　　师：金　海
指导老师：余　辰
E-mail：jh157709@163.com
QQ：1364470533
联系电话：15770962200
毕业去向：腾讯科技（深圳）有限公司



对于云边端异构的计算资源、动态的网络负载和多样化的用户需求，如何有效的进行资源的调度和路由成为了算力网络面临的核心挑战之一。针对这些问题，围绕算网状态感知和算力资源路由两个问题，设计了一套用于解决算力网络中资源的路由和调度的系统——多层次的计算资源系统（Computing Resource System，CRS），主要做出如下工作：

首先，为了降低广域网络中算网状态的感知开销以及减少算力资源路由时的搜索空间，依据计算节点的算力资源和共享性的差异，对算力网络进行了水平和垂直层面的划分。其次，为实现计算节点之间算网感知，设计并实现了算网感知策略，引入辖区系统的概念，定义了辖区系统内部的域内感知规则和不同辖区之间的域间感知规则。同时，为了实现计算任务的卸载，提出了一种基于贪心的资源路由算法（Greedy-Based Resource Routing Algorithm，GBRA），以最大化任务调度数和辖区系统的负载均衡，能为每个用户任务生成独特的感知搜索树。最后，融合了算网状态感知和算力资源的路由，设计了协议的报文结构，协议通过CRS请求报文、授权通告报文、通告确认报文和CRS响应报文来完成资源的申请与调配工作，同时可以捎带节点的算网状态信息。

面向深度学习推理任务的服务器无感知计算显存管理机制研究

姓　　名：于　跃
研究方向：服务器无感知计算、深度学习、显存管理

导　　师：吴　松
指导老师：吴　松
E-mail：2091568941@qq.com
QQ：2091568941
联系电话：17838271656
毕业去向：支付宝（杭州）信息技术有限公司

深度学习推理应用广泛普及，其密集型矩阵运算高度依赖图形处理器（Graphics Processing Unit，GPU）加速。为最大化资源利用率，GPU并发处理多推理请求成为常态。然而，在日益主流的服务器无感知计算（Serverless computing）范式下，为并发推理任务进行高效的GPU显存管理已构成关键瓶颈，成为亟待解决的学术与工程难题。

现有服务器无感知计算系统显存管理机制存在三个问题。（1）GPU运行时冗余显存占用。现有系统为各请求分配独立GPU运行时，忽视了请求间共享GPU运行时的可能，显著加剧了并发请求间的显存竞争。（2）中间变量冗余显存占用。作为各计算层输出数据，中间变量只在相临几层被访问。然而，现有系统在计算前后为所有中间变量统一分配、释放显存，导致显存资源浪费。（3）显存池化方案存在缺陷。现有系统使用显存池化方案降低显存分配开销。然而，一方面，现有显存池化方案缺乏空闲显存回收能力，导致显存池资源利用率低；另一方面，缺乏主动扩张能力，显存需求超出池规模时显存池被动扩张，导致请求延迟增加。针对上述问题，设计能够实现细粒度显存管理的服务器无感知计算系统MemFlexSl。（1）针对GPU运行时冗余显存占用，提出GPU运行时共享方案。通过对CUDA Stream机制的运用及对传输行为的细粒度切分，实现同一GPU运行时下各请求间的算力隔离与传输带宽共享。（2）针对中间变量冗余显存占用，提出层粒度的显存分配与释放方案。通过延迟未被访问的中间变量的显存的分配，并在其不再被访问后主动释放其显存空间，降低请求实时显存占用。（3）针对现有显存池化方案缺陷，提出动态扩缩的显存池化方案。通过周期性的显存需求预测及池规模预扩缩，提升了显存池利用率并避免了临时显存池扩张导致的请求延迟增加。

实验结果表明，相比企业界及学术界最新相关系统，MemFlexSl的吞吐量提升分别达到7.45倍和1.14倍，P99延迟缩减分别达到72.55%和17.13%。

分布式图卷积网络采样优化研究

姓　　名：俞　强
研究方向：图卷积网络采样
导　　师：王　雄
指导老师：王　雄
E-mail：916333023@qq.com
QQ：916333023
联系电话：18705078703
毕业去向：北京三快在线科技有限公司

随着图卷积网络（Graph Convolutional Networks, GCN）在社交网络分析、推荐系统、生物医药等领域的广泛应用，处理大规模图数据的需求日益增长。然而，分布式GCN训练面临着严峻的通信瓶颈问题。由于跨工作节点之间频繁的数据交互，导致训练时间显著延长。尽管已有多种优化方法尝试提升训练效率，但这些方法往往以牺牲模型性能为代价，且在降低通信开销方面效果有限。为此，文章提出了一种面向分布式GCN训练的框架EVRS-GCN，通过高效的采样策略与基于采样的缓存机制，从根本上缓解通信瓶颈并提升模型整体性能。

具体而言，EVRS-GCN融合了可解释性驱动的采样与最小化方差采样。通过分析节点在模型决策中的重要性，优先采样对模型贡献较

大的节点，相较于传统随机采样方法，能够更精准地捕捉关键结构信息。同时，进一步对采样方差进行建模，建立其与节点采样概率的数学关系，并从优化目标出发推导出最小化采样方差的闭式解。这一优化策略不仅显著降低了冗余计算，还有效加速了模型收敛过程。

在此基础上，EVRS-GCN引入了联合采样驱动的特征缓存模块。该模块针对高采样概率的热点节点进行本地缓存，确保频繁访问的节点特征能够无需跨节点通信即可快速获取，从而进一步缓解网络负载，提升通信效率。在多个真实世界图数据集上的实验结果表明，EVRS-GCN在训练效率和模型性能方面均优于现有主流方法，在提升模型精度1.12%~4.34%的同时，实现了高达56%的训练时间缩减，展示了其在大规模图数据分布式训练中的卓越潜力。

基于大语言模型的增强型容器镜像构建指令生成方法

姓　　名：岳　航

研究方向：云原生、容器镜像

导　　师：吴　松

指导老师：樊　浩

E-mail：1772503531@qq.com

QQ：1772503531

联系电话：15671942185

毕业去向：华为技术有限公司

容器镜像构建指令的自动生成与优化面临效率低、冗余多和执行错误等问题，大语言模型虽提供解决思路，但因缺乏依赖关系感知和指令结构优化能力，其生成效果仍有待提升。

为了解决上述问题，提出了一种基于大语言模型的增强型容器镜像构建指令生成方法。首先，针对大语言模型对容器领域知识匮乏的问题，提出了一种系统化的知识图谱构建方法。然后，利用该知识图谱，设计了基于大语言模型的容器镜像构建指令检索增强生成系统。最后，针对大语言模型无法优化指令结构以及现有镜像指令存在的问题，设计了基于指令静态依赖的指令重构策略，提升镜像构建效率，优化镜像大小，提升构建指令的可维护性。

实验全面测试了知识图谱的有效性以及镜像构建指令生成方法的性能。该知识图谱识别到的指令语义关系有39条，相较于Binnacle多了16条，且提取速度提升了97.24%。镜像构建指令检索增强生成系统相比原生大语言模型，可以将生成准确率提升10%~24%。相较于其他指令优化工具，指令重构策略可以减少20%以上的镜像构建时间，同时降低14%~29%的存储开销。

纵向联邦学习的计算与通信优化研究

姓　　名：张　熠

研究方向：纵向联邦学习、并行分布式计算

导　　师：金　海

指导老师：王　雄

E-mail：892405310@qq.com

QQ：892405310

联系电话：18770254225

毕业去向：中国工商银行湖北省分行

联邦学习作为一种新兴的分布式机器学习框架，能够让多个数据参与方在不共享原始数据的前提下协同训练一个全局模型，为隐私保护和数据孤岛问题提供了一个有效的解决方案。然而，联邦学习在实际应用中仍面临诸多挑战，尤其是在纵向联邦学习场景下，由于广域网上频繁的跨参与方数据交换，通信开销成为制约其训练效率的主要瓶颈。当前的纵向联邦学习框架通常面临着的通信开销过大的问题，而现有的一些优化技术在加速训练的同时可能会削弱学习的准确性。为了在保障模型性能的前提下减少通信开销，提高纵向联邦学习的训练效率，提出了两种优化策略。

首先提出了一种基于有界模型陈旧度的异步纵向联邦学习框架BS-VFL（Vertical Federated Learning with Bounded Model Staleness）。BS-VFL通过引入有界模型陈旧度约束，将数据交换与模型计算流水线化处理，减少了通信开销，同时确保了良好的模型性能。通过分析收敛误差，证明BS-VFL能达到与同步纵向联邦学习框架相当的结果。此外，开发了一个通用框架来推导BS-VFL的闭式训练时间，提供了一个衡量其运行效率的方法，并突出了其显著的通信缩减。同时，利用这种收敛和时间分析，优化学习参数以最小化收敛误差，从而在不影响训练效率的情况下优化了BS-VFL的性能。其次设计了一种自适应参数冻结机制APF（Adaptive Parameter Freezing），能够在本地模型参数稳定时暂停其更新，从而减少不必要的通信开销和本地计算。此外，提出了基于统计量的阈值策略，以及基于动态冻结期的训练恢复策略，在提升训练效率的同时保障了模型的准确性。

实验结果表明，相对于基准方法，BS-VFL不仅显著减少了通信轮次，还保持了模型的收敛性和准确性，训练时间减少了48%-90%，收敛速度提升了1.9-10.1倍。APF相对于FedBCD实现了1.37-2.45倍的训练加速，并减少了最高35.8%的数据传输量，同时保持了相当甚至更优的模型性能。

系　统　组

面向动态图处理的高效数据访存机制研究

姓　　名：蔡敏志

研究方向：计算机体系结构、图计算

导　　师：肖　江

指导老师：张　宇

E-mail：910958809@qq.com

QQ：910958809

联系电话：17866553135

毕业去向：海思技术有限公司

随着社交网络、金融交易监测和实时推荐系统等应用的快速发展，动态图处理需求日益增长。动态图的存储和计算需要同时满足高效的图更新和低延迟的图分析，而传统的静态图存储结构难以兼顾这两方面的性能。现有动态图系统虽在提高搜索效率和减少数据移动方面做了很多工作，但在高度动态数据场景下，仍存在数据更新过程中的大量冗余搜索和数据移动；同时为了提升在实际场景下的通用性，也需要支持数据压缩和多版本管理。

针对现有系统的局限性，首先在其基础上设计了以局部为中心的图数据批量更新机制，通过局部性感知的任务划分和调度策略和基于细粒度的多阶段批量更新策略，减少了更新和计算过程中的冗余搜索和数据移动开销。其次设计了增量感知的多版本动态图数据存储策略，引入稀疏快照管理版本，同时结合增量日志降低存储开销并支持高效的历史数据查询。此外，实现了对混合索引树的规则压缩，通过检测和替换邻接边中的重复模式，进一步提升存储密度并降低冗余存储成本。最后结合上述设计实现了高效数据访存的动态图处理引擎LUGraph。

在多个真实数据集和合成数据集上的实验评估表明，LUGraph在动态图处理性能上显著优于现有方案。相较于高性能动态图引擎LSGraph，LUGraph在图更新吞吐量方面提升1.15至2.61倍、在图计算性能方面提升1.12至1.47倍、在存储效率上提升2.29至4.02倍，为动态图存储与计算提供了一种高效、可扩展的解决方案。

面向非易失性内存的轻量级磨损均衡机制研究

姓　　名：陈瑞聪

研究方向：计算机体系结构

导　　师：刘海坤

指导老师：刘海坤

E-mail：1065313246@qq.com

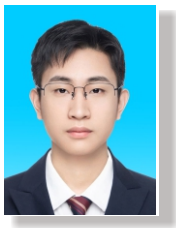
QQ：1065313246

联系电话：15958185055
毕业去向：海光云芯集成电路设计（上海）有限公司

随着人工智能和大数据的发展，非易失性内存（NVM）因其高密度、低功耗等优势成为解决DRAM容量瓶颈的有效补充。但其有限的写入耐久性导致热点区域磨损加剧，严重影响系统寿命。为此，文章提出一种面向NVM内存扩展系统的轻量级磨损均衡机制。该机制包括三项核心设计：首先，设计双层级写入流量均衡策略，在页内通过状态位驱动的细粒度偏移映射实现局部均衡，当页内触发次数达到阈值时启动跨页写入均衡，分散写入压力。其次，提出基于位翻转检测的分级阈值触发机制，通过检测局部位翻转程度动态判断是否触发均衡操作，替代传统写入计数器方案，显著降低元数据开销。最后，构建结合状态位与LRU信息的元数据缓存优化机制，引入权重函数提升缓存命中率，降低访问延迟。实验结果表明，该机制在多种负载下有效延长NVM寿命，最高可达传统方法的14倍，执行时间减少约13.9%，适用于资源受限的实际系统环境，具备良好的实用价值与推广前景。

异构内存带宽限制的页面迁移机制研究

姓　　名：郭　超
研究方向：异构内存、页面迁移
导　　师：余　辰
指导老师：刘海坤
E-mail：1508717884@qq.com
QQ：1508717884
联系电话：18771005987
毕业去向：武汉达梦数据库股份有限公司



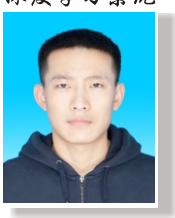
异构内存系统中，需要充分考虑不同内存介质的特性，如Dynamic Random Access Memory（DRAM）、Non-Volatile Memory（NVM）和Compute Express Link（CXL）扩展内存设备，将冷热页面分别迁移至慢存和快存，来提升程序的性能。

基于平行化异构内存架构，提出了带宽限制的页面分类算法和基于共享区的页面迁移机制，实现了异构内存的带宽平衡。首先，采用硬件采样的方式实现了页面热度监测与收集，并设计了采样周期动态调整策略。其次，在传统依赖快存容量的分类策略中引入了带宽因素，提出了带宽限制的页面分类模型。通过限制快存的带宽，确保其延迟保持在较低水平，从而更合理地发挥快存的低延迟优势。最后，设计了共享区机制，共享区中的页面能够在快存和慢存同时存在，简化了部分页面迁移过程，仅需删除旧的页表项并创建新的页表项；对于其他页面，优化了页面迁移的方法，确保程序在迁移过程中仍能正常访问页面，显著减少了页面迁移对程序执行的阻塞时间。并将上述模块集成为BW系统。此外，该系统同时支持非易失存储器NVM和CXL扩展内存设备，具有较强的可扩展性。

实验结果显示，在Redis和NAS Parallel Benchmarks（NPB）应用测试中，在基于CXL设备的异构内存架构下，采用BW系统相比于不进行迁移的策略性能分别提升了32.4%和34.39%；对比TPP系统，性能分别提升了21.43%和22.19%。同时测试了访问模式、数据量以及快存和慢存的比例对系统的影响，结果显示在不同的场景下BW系统相比于不进行迁移的策略性能提升20%~30%；并且在快存容量只占总容量的20%时，BW系统相比于不进行迁移的策略的提升达到37.23%。

基于行窗口采样的图神经网络训练加速方法研究

姓　　名：郭海宏
研究方向：大规模图神经网络训练，深度学习系统
导　　师：蒋文斌
指导老师：蒋文斌
E-mail：1941409241@qq.com
QQ：1941409241



联系电话：18734768083
毕业去向：北京小米移动软件有限公司

图神经网络因其在社交网络分析、推荐系统等领域的表现而受到广泛关注。然而，大规模图训练中常出现“邻域爆炸”问题，导致计算开销剧增。现有优化方法如图采样算法难以兼顾模型性能与资源消耗，而消息传递加速仅支持全图训练，难以高效支持小批次训练。

为此，提出了RWS-GNN，一种面向大规模图的小批次训练加速框架，通过协同设计图采样算法与消息传递加速内核，提升训练效率并降低内存占用。该框架包含两项关键创新：（1）行窗口采样算法，将源节点划分为多个行窗口，在每个窗口内基于列密集性执行层采样，有效缓解邻域爆炸并保持模型性能，同时结合基于特征相似性的图扩散预处理，增强关键连接信息捕获；（2）定制化加速内核，针对采样后稀疏结构设计新型存储格式和计算流程，使Tensor Core可加速前后向传播，并结合数据预取策略进一步减少传输开销。

基于PyTorch与CUDA实现的RWS-GNN在多个大规模图数据集上验证了其优势。在Reddit和ogbn-products上，RWS-GNN综合性能分别较基准提升1.39倍和1.57倍，训练耗时平均缩短14.96%，在模型精度、收敛速度与资源效率方面均取得显著改进。

CPU–FPGA协同的局部敏感哈希加速方法研究

姓　　名：黄福龙
研究方向：系统结构
导　　师：金　海
指导老师：叶晨成
E-mail：2835345719@qq.com
QQ：2835345719
联系电话：18813126051
毕业去向：腾讯科技（深圳）有限公司

传统的基于CPU的局部敏感哈希算法在处理高维数据时受限于CPU核数并行计算能力较差，难以应对高维向量的密集计算需求，而现有的基于图形处理器（GPU）和可编程门阵列（FPGA）的加速方案在系统架构协同和资源利用等方面也存在一些缺点和不足，导致加速效果并不理想。

CPU-FPGA协同的局部敏感哈希加速方案可以通过算法–硬件协同设计实现高效的近似最近邻搜索。针对局部敏感哈希算法中计算复杂度较高的哈希编码计算和相似向量选择（TopK）任务，设计并实现了两个关键的硬件加速内核：1）LSH内核采用动态分层存储策略，结合数组划分和循环展开等优化技术，实现高维向量哈希编码的并行计算；2）TopK内核构建四级流水线架构，通过任务级并行与多路归并排序，实现高效的相似向量筛选。在CPU-FPGA协同系统中，CPU负责轻量级任务调度与哈希表管理，FPGA通过硬件架构加速计算密集型任务。该系统能够充分利用CPU和FPGA 各自的硬件优势，通过灵活的内存管理机制应对不同规模的数据，有效提高资源利用率，并利用FPGA的任务并行和数据并行设计加速计算。同时系统采用基于动态资源调度的混合计算架构，通过智能任务分发机制与分片处理策略，提高系统处理不同规模数据任务的能力。

实验结果表明，与传统CPU实现的局部敏感哈希方法相比，基于CPU-FPGA的协同加速系统对哈希编码计算和TopK筛选步骤提速分别可以达到32倍和2.7倍以上，对向量插入和查询操作分别提速5.5倍和2.2倍以上，有效加速了近似最近邻搜索过程。

以数据为中心的高性能自适应基数树研究

姓　　名：岳　航

研究方向：图计算
导　　师：张　宇
指导老师：张　宇
E-mail：623793872@qq.com
QQ：623793872
联系电话：13459307227
毕业去向：北京快手科技有限公司

近年来，出现了许多自适应基数树优化变体，试图全方位提高自适应基数树的整体性能，以满足日益增长的数据处理需求。然而，在实际执行自适应基数树的各类操作的过程中，仍然存在大量的冗余树节点遍历开销和巨额的同步成本。而大多数真实世界的工作负载的操作往往倾向于自适应基数树的一小部分节点，且这部分节点会在短时间内被不同的操作频繁访问，这些操作之间存在很强的时间相似性和空间相似性。

针对上述两个问题，为了利用自适应基数树上的不同操作之间的时间和空间相似性，提高自适应基数树的处理性能，提出了一种以数据为中心的高性能自适应基数树处理方法，通过这种方式，能够显著减少冗余树节点遍历开销，并降低不同处理单元之间的同步成本。同时，设计了价值感知的内存子系统，减少不同类型数据的缓存冲突，避免高价值树节点频繁发生缓存抖动，提高缓存利用率。

实验结果显示，和现有最先进的自适应基数树变体SMART和CuART相比，DCART的锁开销仅为3.2%~15.9%和7.1%~19.7%，部分键匹配次数仅为6.5%~14.3%和8.8%~15.9%，从而实现了21.1~44.2倍加速和71.1~148.9倍节能。

面向多GPU平台的超图神经网络训练框架

姓　　名：蒋晨昱
研究方向：计算机体系结构、图神经网络
导　　师：廖小飞
指导老师：张　宇
E-mail：517300883@qq.com
QQ：517300883
联系电话：18593932923
毕业去向：国家能源集团广西公司

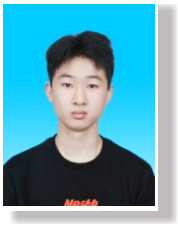
超图神经网络因其能够灵活建模现实世界中复杂高阶交互关系而受到广泛关注。然而其在实际应用中面临计算稀疏性强、计算复杂度高等多重挑战，使得现有框架在超图神经网络多卡并行计算中面临着严重的负载不均衡和高昂的通信开销问题，使得多GPU并行训练的实际加速效果远低于理论预期。

为充分利用多GPU计算和互联资源加速超图神经网络训练，提升系统扩展性，提出了面向多GPU平台的超图神经网络训练框架，从数据划分、任务调度和数据通信三个维度展开系统优化。针对超图神经网络训练中存在的的高阶依赖关系，通过计算负载分割的二级超图划分方法，将超图拓扑结构与超图神经网络计算负载进行协同分割，以生成高质量、高平衡性的子图，显著减少冗余的数据存储和跨设备通信的邻域副本数量。根据子图划分的结果，实现基于平衡数据块的协同调度方式，通过块级任务的计算分配方式，确保训练语义的正确执行，提升多卡处理能力，同时最大化全局并行度。而针对通信瓶颈，通过去重特征数据的通信策略，实现训练过程中的通信优化，通过冗余通信识别与全局映射，利用设备间高速互联带宽减少主机与设备之间的冗余数据传输，显著降低系统的通信延迟。

实验表明，在4块V100 GPU的服务器上，该框架可以高效支持千万级别规模的超图数据集训练，减少43%~61%的主机与设备的通信开销。对比主流的超图神经网络框架DistDGL和HyperGEF显著提升了超图神经网络的模型处理能力，并实现了3.1~5.2倍的性能提升。同时还展现出了良好的系统扩展性。

面向图神经网络的众核处理器异构内核融合研究

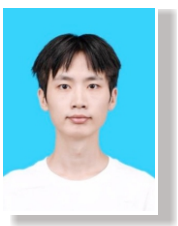
姓　　名：李　扬
研究方向：图神经网络加速，稀疏矩阵乘法
导　　师：蒋文斌
指导老师：蒋文斌
E-mail：3032842800@qq.com
QQ：3032842800
联系电话：15907990939
毕业去向：阿里巴巴（中国）有限公司



图神经网络凭借对图结构数据的高效表征能力，广泛应用于社交网络等领域。然而，图结构的不规则性导致图神经网络模型在图节点聚合阶段的数据访问模式效率较低，且聚合算法对硬件资源的调度存在不足。论文引入基于图结构相似性的节点聚类思想，结合GPU异构计算核心的协同执行模式，提出HCF-GNN——一种图结构与聚合计算协同优化的图神经网络加速系统。系统首先基于节点邻居列索引执行初步聚类，将相似节点归入同一行块，再根据行块的非重复邻居列索引实施二次聚类，以提高GPU缓存命中率，并在整个过程中持续更新聚类后的索引状态。在此基础上，采用“分核映射”方式划分图节点聚合负载，实现GPU异构核心（Tensor Core和CUDA Core）并行计算，同时设计定制化的加速内核，协同提升图神经网络的聚合计算效率。

面向数据流架构的稀疏矩阵高效处理内核研究

姓　　名：刘宝阳
研究方向：高性能计算
导　　师：蒋文斌
指导老师：蒋文斌
E-mail：1599607211@qq.com
QQ：1599607211
联系电话：15342753271
毕业去向：海思技术有限公司



高通量数据流众核处理器（DFU）作为面向多应用的通用数据流加速器，虽然其在规则计算上已展现出优异的能效比和计算资源利用率，但稀疏计算频繁的数据流图重构以及DFU不支持稀疏计算内核编译等问题使SpMM内核难以利用加速器的指令复用和数据流动两个核心特性。论文主要从DFU的两个核心特性入手：（1）提出面向数据流指令复用的分层稀疏矩阵处理方法。该方法首先基于行上非零元的个数进行粗粒度的聚类，产生不同的稀疏块，再对稀疏块内的行进行基于二的次幂的细粒度划分及二次聚类，确保稀疏块具有规则的计算模式，以减少SpMM内核的数据流图重传次数。同时，对于具有不同特性的稀疏块，使用不同的策略对其进行处理，以最大限度地利用系统资源。（2）提出面向数据流图优化的自适应汇编代码生成方法。引入观察者–执行者的思想，获取稀疏矩阵的结构信息并划分片上存储资源，在此基础上生成汇编代码，解决DFU不支持SpMM内核编译的问题。此外，从任务分配和寄存器重用两个角度对代码生成过程中的数据流图构造进行优化，挖掘DFU上SpMM内核的数据重用，确保生成内核的高效性。

面向高吞吐并发点对点查询的共享机制研究

姓　　名：卢浩宇
研究方向：图计算系统、高性能计算
导　　师：张　宇
指导老师：张　宇
E-mail：2729903242@qq.com
QQ：2729903242
联系电话：13223041056
毕业去向：北京沃东天骏信息技术有限公司

点对点查询是一种专注于分析图中指定图顶点对之间拓扑关系的算法，广泛应用于路径规划和社交网络分析等领域。然而，随着图规

模呈指数级增长，现有系统在支持高并发点对点查询时要么没法利用点对点查询特性有效进行计算结果共享和剪枝从而面临大量冗余计算开销，要么难以应对高并发场景下不规则数据访问和缓存抖动导致的数据局部性差和冗余数据访问问题。

为了解决上述问题，提出了一种冗余感知的并发点对点查询共享方法，充分利用并发点对点查询之间的遍历相似性提高其系统吞吐量。具体而言，该方法首先提出一种基于热路径的计算共享策略。该策略将图遍历过程分解为图遍历路径的集合并高效识别出其中的热路径，然后基于热路径高效预估未知路径查询结果，通过对不满足收敛条件的路径进行高效剪枝以减少遍历开销，同时通过构建由热路径组成的状态传播网络，实现图顶点状态的快速传播以减少冗余计算。此外，为了进一步减少并发查询的冗余数据访问开销，提出了以数据局部性为中心的数据访问共享策略，通过对图结构数据和查询任务的细粒度调度，优化并发点对点查询的数据访问局部性，从而提高缓存命中率。最后，基于上述方法，构建了一个冗余感知的并发图查询系统GraphCPP，通过高效消除并发点对点查询之间冗余数据访问和计算开销并提高数据局部性，高效支持并发点对点查询。

实验结果表明，与现有最好的点对点查询系统SGraph相比，GraphCPP能够减少并发点对点查询中20.3%~52.2%的计算开销和29.7%~35.6%的数据访问开销，从而在支持并发点对点查询时综合性能提升了2.9~3.5倍。

面向FPGA的非阻塞式缓存的研究与设计

姓　　名：陆思彤
研究方向：FPGA硬件加速
导　　师：邵志远
指导老师：邵志远
E-mail：sitong-lu@qq.com
QQ：614149243
联系电话：18177133465
毕业去向：中国移动通信集团广西有限公司

在现场可编程门阵列（FPGA）上加速如图计算等应用时，通常会因不规则内存访问而导致内存传输效率较低。对此，现有研究面向FPGA提出了优化的非阻塞式缓存架构，通过配置大量的未命中状态保持寄存器（MSHR）来减少缓存未命中产生的系统停顿。现有研究采用块随机访问存储器（BRAM）来实现容量固定的MSHR阵列，由于实际应用的MSHR需求具有高度的不确定性和可变性，这会造成潜在的存储资源利用率问题，并且使缓存系统配置复杂化。

通过对非阻塞式缓存中的缓存行与MSHR在功能、资源以及需求等方面的相关性进行分析，并设置实验测量系统运行时的缓存命中率与MSHR利用率变化曲线，对所提出的分析加以验证，从而揭示了采用固定数量MSHR所带来的配置合理性问题。根据理论分析与实验观察，设计了一种缓存行与MSHR共享存储空间的非阻塞式缓存架构方案，允许存储空间根据缓存命中情况在缓存行与MSHR之间进行动态切换，能够在运行时根据应用的实际访存模式进行实时自适应调整，从而提高存储资源的利用效率。此外，为进一步提升缓存系统的处理吞吐量，设计了一种并行双流水线架构，能够同时分别处理应用访存请求和内存返回数据。

在Xilinx Alveo U280加速卡上进行实验测试的结果表明，在缓存容量与MSHR数量均为等效的配置下，相较于现有面向FPGA的非阻塞式缓存设计，所提出设计能够带来总体相近的性能表现，并且最高可以节省17%的BRAM存储资源消耗。在消耗相近数量的BRAM存储资源，且为现有设计针对数据集进行最优化配置后的情况下进行对比，所提出设计最高可以带来1.50倍的性能加速，且所提出设计无需进行配置调整就能动态自适应具有不同访存局部性表现的场景。

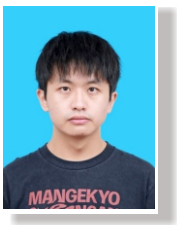
数据驱动的图神经网络训练优化技术研究

姓　　名：路琛洋
研究方向：图神经网络
导　　师：张　宇
指导老师：张　宇
E-mail：2210439120@qq.com
QQ：2210439120
联系电话：18234610796
毕业去向：中国证券登记结算有限责任公司深圳分公司

随着图规模的不断扩大，利用当前系统进行图神经网络的训练任务仍面临着资源需求大与网络数据传输量大等挑战，造成图神经网络的训练低效。为了解决上述问题，提出了数据驱动的图神经网络执行模型，其充分利用图特征数据的维度信息，通过分层的任务编排方法重构图神经网络模型的执行策略，以减少模型训练所需原始特征传输量，并基于该模型构建了一个图神经网络训练系统LCGNN。为了满足不同环境部署的需求，LCGNN不仅支持基于单机的图神经网络训练，还高效支持分布式的图神经网络训练。为充分利用单机环境下的训练资源，LCGNN构建了两级缓存的策略将图数据中热数据的计算任务卸载到主机，并实现热嵌入的重用，将网络中的原始特征数据传输转换为热嵌入的传输，减少不规则内存访问与网络中特征传输量。此外，为实现分布式扩展并保持节点间的负载均衡，实现了基于哈希的图结构分区与基于维度的特征数据分区两级独立分区方法，将网络通信延迟到模型底层计算之后，消除了智能图分区的高额开销并大幅减少分布式环境中的网络通信量。同时，实现了统一的双阈值策略管理的历史缓存，能方便地集成到单机与分布式环境下，减少模型计算量。

实验结果表明，LCGNN与目前最先进的系统DUCATI、DGL相比，性能分别提升了1.19~1.89倍、1.19~2.59倍。

基于GPU的动态图模式匹配的研究



姓　　名：罗　康
研究方向：图模式匹配、图计算
导　　师：张　宇
指导老师：张　宇
E-mail：1143550127@qq.com
QQ：1143550127
联系电话：18540138420
毕业去向：腾讯科技（深圳）有限公司

动态图模式匹配旨在识别因动态图结构变化而新增或删除的与给定模式图同构的子图，广泛应用于社交网络分析、网络对比等领域。然而，随着图数据规模的指数级增长和实时更新需求的增加，基于CPU的图模式匹配系统在计算吞吐量上面临着严重的性能瓶颈。而将现有的基于增量计算的动态图模式匹配模型移植到GPU上时，仍然面临着冗余计算、负载不均衡与资源利用率低等挑战。

为此，提出了冗余消除的GPU动态图模式匹配模型。具体而言，该模型通过分析模式图的拓扑结构，为同构的模式图边生成共享执行计划，在动态图模式匹配过程中减少冗余计算和对称性检查。基于该模型，构建了基于负载感知的GPU动态图模式匹配系统PGMiner，以在GPU上高效执行动态图模式匹配任务。PGMiner基于负载感知策略从不同层次感知任务负载以充分利用GPU的计算资源：在多GPU环境下，使用任务开销预测模型指导任务分配，以缓解多GPU设备间的负载不均衡；在GPU内，采用负载感知的两级负载均衡策略，通过自适应任务拆分方法感知并拆分高负载任务，以及动态工作窃取方法感知高负载线程束并窃取其任务，以此缓解GPU内不同线程束间的负载不均衡；在线程束内，通过感知候选集顶点的负载情况，实施负载融合的动态循环展开机制，并行执行多个交集计算，从而提升线程利用率。此外，PGMiner提出动态内存分配策略，通过运

行时按需为线程束分配存储候选集所需的内存，提升GPU的内存利用率。

实验结果表明，相比于最先进的动态图模式匹配系统GraphSet-P和G2Miner-P，PGMiner分别取得了2.18~7.81倍和3.85~9.21倍的性能加速，并且在多GPU环境下具有良好的扩展性。

面向向量数据的图查询加速器研究

姓　　名：马铭博
研究方向：图计算
导　　师：金　海
指导老师：姚鹏程
E-mail：353902592@qq.com
QQ：353902592
联系电话：13838308051
毕业去向：中国移动通信集团河南有限公司

向量检索因其强大的抽象能力被广泛地应用于人工智能、推荐系统和知识图谱等领域。通过采用定制化的硬件架构，图查询加速器可以取得通用处理器数十倍的性能，开展面向图查询的领域专用加速器研究有着重要的现实意义。

然而，现有图查询加速器忽略了不同向量维度对检索结果的贡献差异性，导致严峻的冗余访存问题。针对该问题，提出基于提前终止的图查询加速器架构，通过提前终止已识别的冗余访存请求，显著降低图查询的访存总量。基于该执行模式，进一步提出基于数据流调度的向量检索流水线。该设计通过基于数据流模型的任务调度机制，在保证整体并行执行效率的同时，实现向量串行的执行效果，避免加速器提前预取被跳过的负载数据。此外，通过挖掘向量检索过程中的幂律遍历特征，提出基于差分管理的片上缓存架构。该架构通过优先缓存频繁被访问的向量数据，进一步降低片外访存总量。

实验结果表明，在保证90%以上查询正确率的约束条件下，基于提前终止思想的图查询加速器相比与中央处理器和最先进的图查询加速器分别取得17.7~30.5倍和1.4~2.0倍的吞吐量提升，在高维数据集上降低约20%的搜索延迟，实现图查询低延迟，高吞吐量，高精度的现实应用需求。

面向大规模分布式图计算的轻量级容错机制研究

姓　　名：聂龙宇
研究方向：图计算、容错
导　　师：毛伏兵
指导老师：毛伏兵
E-mail：1466410548@qq.com
QQ：1466410548
联系电话：15964182061
毕业去向：华为技术（杭州）有限公司

在分布式图计算系统中，计算节点的增加会导致故障发生频率上升，进而影响整个系统的稳定性和性能。因此，容错机制对分布式图计算系统至关重要。

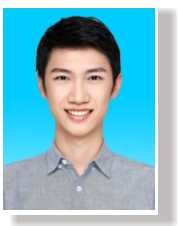
提出了一种兼顾容错成本与恢复效率的容错机制LDC-RAR，包括两项关键技术：轻量级动态检查点保存方案（Lightweight Dynamic Checkpointing，LDC）和基于冗余感知的失效恢复策略（Redundancy-Aware Recovery Strategy，RAR）。LDC首先构建了基于增量拓扑保存和消息在线生成的轻量级检查点保存方案，然后在此基础上设计了成本敏感的动态检查点保存方案，在运行时自适应地调整检查点写入时机以降低开销。RAR对日志中导致冗余计算的顶点状态进行感知和过滤，并基于过滤后的消息日志进一步实现了一套高效的失效恢复策略，通过减少失效恢复期间冗余消息的传播和失效顶点的计算次数，显著提高了系统从失效中恢复的速度。

对提出的容错机制进行了性能测试。实验结果表明，轻量级检查点保存方案的时间开销仅为传统方案的2.1%~18.6%，动态检查点方案在此

基础上进一步将时间开销降低21.9%~72.9%。基于冗余感知的失效恢复策略比基于分区的恢复方法减少了15.4%~64.8%的恢复时间。此外，在综合性能方面，当执行期间有失效发生时，LDC-RAR的总执行时间总是优于最先进的无状态冗余容错机制。

基于高带宽内存的图计算众核加速器研究

姓　　名：潘晨高
研究方向：图计算加速器、片上互联网络
导　　师：郑　龙
指导老师：姚鹏程
E-mail：987374783@qq.com
QQ：987374783
联系电话：15024423659
毕业去向：浙江省政府



随着高带宽存储器等新型存储技术发展，内存带宽实现数量级的提升，同时也对加速器的扩展能力提出更高的要求。现有的加速器扩展能力受到片上互联网络制约，每周期可并发处理的计算任务较少，在部署高带宽内存时内存带宽利用率通常低于20%。设计基于高带宽内存的图计算加速器的核心是实现高吞吐的片上互联网络，通过高效的众核扩展提升内存级并行度，从而高效利用高带宽存储器带宽。针对该问题，设计基于高带宽内存的众核图计算加速器GraphNexus，通过面向高带宽内存的高并发执行模型降低访存数量与通讯总量，提高带宽利用率。提出低延迟易扩展的片上互联网络，通过基于分层互联的高带宽片上互联模型和基于请求合并的片上网络结构，兼顾架构扩展性与通信效率，在核心数量超过32时相比于传统结构有显著性能优势，显著提升图计算加速器的可扩展性。提出负载均衡的片上通讯机制，通过拥塞感知的众核路由优化策略与维度感知的任务调度映射机制，满足图计算多样化通信需求，保持加速器负载均衡，高效利用存储器带宽。

面向键值存储的地址翻译硬件加速方法研究

姓　　名：石煜庭
研究方向：系统结构
导　　师：金　海
指导老师：叶晨成
E-mail：1362101324@qq.com
QQ：1362101324
联系电话：16605221638

现有键值存储加速技术通常根据访问局部性缓存频繁使用或近期访问的值以提升系统的整体性能。然而此技术对于地址转换过程中的内存访问缺乏足够关注且通常需要设计新的指令，这限制了该类方案在更广泛应用场景中的实用性。

针对现有方法中存在的问题，提出了不引入新指令的加速方法，从地址缓存机制、预取功能、运行时监测三个方面进行了优化设计。首先，通过深入分析键值存储系统的查找过程，提出了用于缓存固定大小的物理地址的硬件结构IPT，用于存储后续可能频繁访问的页表地址和数据地址。其次，对传统预取机制进行修改，使其与IPT协同工作实现特定功能的预取，这一过程未引入新的指令，有效降低了额外开销。此外，设计了运行时监测模块，可在IPT缺失相关表项时捕获程序的访存行为，将后续的访问地址记录至IPT相应表项中，实现IPT的更新，提升了自适应能力。最后，通过将加速技术进行优化并应用于哈希聚合操作中，实现了哈希聚合的性能提升并验证了加速系统的灵活性。

实验表明，与常规的键值查找过程相比，键值存储加速系统在不同结构上的加速比达到1.19~2.26；与常规的哈希聚合操作相比，平均加速比达到1.13。加速技术显著提升了查找过程的执行效率，体现出较低的系统复杂性，具备良好的灵活性与可移植性。

面向分离式内存池的页面访问加速机制

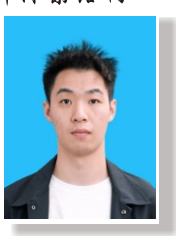
姓　　名：宋邵炜
研究方向：计算机体系结构、分离式内存
导　　师：廖小飞
指导老师：刘海坤
E-mail：1072860370@qq.com
QQ：1072860370
联系电话：18942928945
毕业去向：中兴通讯股份有限公司

分离式内存架构受限于内核慢速交换路径和有限的本地缓存资源，同时不同应用共享交换路径和资源，易发生资源竞争和干扰。针对上述问题，提出了面向分离式内存池的页面访问加速机制ACC。

首先，在本地引入由压缩缓存和快速缓存组成的二级缓存架构，通过压缩提升存储效率，并结合解压延迟与CPU负载动态调整访问路径和缓存比例。其次，设计基于负反馈的自适应预取机制，整合流式、历史和多数投票三种策略，实时评估预取准确率并动态选择最优策略，同时调节预取窗口以控制缓存污染和避免错失。最后，面向多应用场景，引入应用隔离换页机制，依据延迟敏感性和带宽需求划分缓存空间，结合公平调度减少应用间干扰。实验结果显示，ACC可将缓存命中率提升11.7%~57.1%，延迟降低53.2%，带宽提升1.3倍，在多应用场景下性能提升达30.7%~57.9%。

面向动态图神经网络的通信与计算优化机制研究

姓　　名：谈安东
研究方向：高性能计算，系统软件体系结构
导　　师：张　宇
指导老师：张　宇
E-mail：2296204523@qq.com
QQ：2296204523
联系电话：18771926899
毕业去向：杭州阿里巴巴网络技术有限公司



动态图神经网络已经在许多领域广泛应用，以处理现实生活中随着时间维度不断发生变化的图数据并从中获取有用信息。为了高效支持动态图神经网络，一些方法已经被提出。然而由于需要访问和重新计算每个图快照中所有图数据，现有方法在支持动态图神经网络时仍然存在大量冗余计算和高额通信开销。此外，在动态图神经网络模型中，广泛应用的循环神经网络计算部分具有计算复杂度高的特点，因此现有系统在支持动态图神经网络时存在庞大的计算开销。

为解决上述问题，提出了局部性感知的动态图神经网络加速方法，充分利用了动态图神经网络的时间和空间局部性以提高执行效率。具体而言，该方法首先提出一种依赖驱动的冗余消除策略，通过高效计算图顶点之间的依赖关系识别未修改顶点状态以感知冗余计算，然后计算未修改顶点的聚合依赖，根据这些依赖关系驱动图神经网络复用上一快照的状态信息到当前快照计算中实现当前快照的快速增量计算，正确地避免对相同状态的重新计算与访问，有效减少冗余计算和不必要的数据通信。其次，为解决动态图神经网络计算复杂度高的问题，该方法提出一种面向长短期记忆网络的近似计算策略，通过时间步跳过与状态最小化实现对循环神经网络进行近似计算，降低计算复杂度以提高动态图神经网络计算效率。最后，为了高效支持该方法，构建了依赖驱动的动态图神经网络加速系统RAKP，通过减少冗余计算与数据通信以及提供高效近似计算，支持动态图神经网络模型的高效执行。

时序依赖感知的高性能时序图处理系统研究

姓　　名：王鑫蕾
研究方向：图计算、时序图
导　　师：廖小飞
指导老师：张　宇
E-mail：2232797012@qq.com

QQ：2232797012
联系电话：18561011619
毕业去向：中国南方电网有限责任公司

现有时序图处理系统普遍面临冗余计算开销大、数据时空局部性差等问题。因此，如何设计高性能时序图处理系统成为时序图领域亟待突破的问题。

为解决上述问题，提出时序依赖感知的高性能时序图处理模型，突破传统快照独立处理范式，提高时序图处理的效率和系统扩展性。该模型首先利用时序依赖感知的共同遍历路径挖掘方法解析多快照间的结构相似性，并构建跨快照统一遍历路径，从而有效减少冗余数据加载。通过基于时序依赖的数据调度策略和时序感知的缓存替换策略，融合顶点结构信息与时间特征，实现缓存优先级的动态优化，在保证数据局部性的同时降低不必要的存储访问开销。通过实例感知的并行计算方法，将不同快照中对同一边的计算任务进行合并，提高系统整体吞吐率。此外，为了使所提出的高性能时序图处理模型能够适应大规模数据场景，进一步构建大规模时序图处理外部存储优化框架，该框架利用固态硬盘外部存储，结合动态分块压缩与异步流水线I/O技术，有效解决了外存访问延迟和I/O瓶颈等问题，从而显著提高数据访问效率和计算吞吐率。在此基础上，构建了一个时序依赖感知的高性能时序图处理系统TDA，以减少处理过程中冗余数据开销并支持大规模时序图数据场景，保证时序图处理的高效性。

图Transformer模型训练优化技术研究

姓　　名：叶楚玥
研究方向：图神经网络、系统软件
导　　师：张　宇
指导老师：张　宇
E-mail：308591412@qq.com

QQ：308591412
联系电话：18811573858
毕业去向：华为技术有限公司

图Transformer是一种在图学习领域超越传统图神经网络的新的网络架构。目前图Transformer模型的应用大都局限于小图，论文设计了一个基于稀疏注意力机制的图Transformer训练系统，能够高效地支持大规模图上的图 Transformer 模型训练，提升计算效率和内存效率的同时保证模型的收敛。首先，为了优化标准注意力的平方级计算复杂度，设计了基于全局树的图拓扑诱导注意力计算机制，是一种稀疏的注意力计算机制，通过构建全局树实现对全连接注意力的近似，同时结合通信感知的采样策略来利用图拓扑指导注意力计算，引入稀疏性的同时兼顾了对全局信息和图结构信息的学习。其次，在此基础上设计并实现了一个高效、可扩展、保证模型精度的分布式图Transformer训练系统，系统基于上述稀疏注意力机制来做注意力计算，使用树结构感知的图划分策略实现高效图分区且减少全局树构建引入的跨分区边，结合冗余消除的消息传递执行机制和基于历史嵌入缓存的依赖解耦合策略来加速消息传递并隐藏通信延迟，通过关联性感知的图数据访问方法实现访存高效的注意力计算。

实验表明，上述图Transformer训练系统能够有效地加速模型训练，和现有方法相比能实现2~4倍的整体性能提升，且系统可扩展性良好，支持的最大序列长度比标准注意力高6~36倍。

基于图索引的近似最近邻搜索GPU加速机制

姓　　名：尹伟行
研究方向：图计算、近似最近邻搜索
导　　师：张　宇
指导老师：张　宇
E-mail：hannyin@163.com

QQ：645184252
联系电话：15527129500
毕业去向：网易（杭州）网络有限公司

最近邻搜索是一种在给定数据集中查找与查询点最相似或者距离最近的目标点的过程，广泛应用于分类、推荐、图像检索等领域。

针对搜索过程中的并行不充分问题以及内存瓶颈，提出了一种基于优先级的数据管理策略以及状态感知的数据预取策略和内存合并访问机制。具体而言，基于优先级的数据管理策略以顶点之间的距离作为优先级，提前获取优先级最高的顶点，将排序与邻居列表的获取解耦，使其能够并行执行，更加充分地利用GPU的多线程优势。状态感知的数据预取策略和内存合并访问机制通过预测下次迭代所需顶点，将其对应的邻居列表和向量数据提前加载到共享内存中，并将多次内存访问合并执行，掩盖访存延迟，提高带宽利用率。最后还提出了多查询任务并发执行机制，使用一个线程块同时处理多个查询，有效提高查询吞吐量。通过集成上述策略，在GPU上实现了高效的近似最近邻搜索系统PGANN，能够充分利用GPU的硬件资源。

实验结果表明，在大批次查询中，PGANN与目前最好的CPU和GPU近似最近邻搜索系统HNSW和CAGRA相比，搜索吞吐量分别能够提升22.7~64.3和1.6~3.7倍。在小批次查询中，搜索延迟能够分别降低1.9~4.6和1.2~1.4倍。

基于FPGA的大规模动态图计算加速器

姓　　名：周宇航
研究方向：图计算加速器、图划分
导　　师：郑　龙
指导老师：郑　龙、姚鹏程
E-mail：944676203@qq.com
QQ：944676203
联系电话：18913937686
毕业去向：南京欧珀软件科技有限公司

现有的动态图加速器在大规模图场景下会出现严重的异构通信延迟。由于加速器板载内存容量有限，现有工作通常采用主机与加速器异构协同架构，通过利用主机端的内存资源实现动态大图处理。受限于外围组件快速互连总线较低的传输带宽，这类设计通常会导致加速器被高延迟的远端访存频繁阻塞，停滞时间占总体计算时间的70%，最终显著降低整体执行效率。针对该问题发现图划分数据的迭代收敛特征，并基于该特征提出未来值计算的设计思想。围绕该思想，首先提出局部性感知的动态图结构管理机制，采用轻量级的启发式划分提高动态图计算的局部性，加速未来值收敛并减少增量计算时的异构通信次数。然后，提出软硬协同的动态图计算优化策略，通过零通信的标记机制和冗余感知的硬件架构，消除删除边标记阶段的异构通信并减少加速器端的冗余计算与访存开销。

基于上述设计实现了基于现场可编程逻辑门阵列的大规模动态图加速器NeoStream，协调主机与加速器的数据调度，大幅减少两者之间的无效通信，提升图数据处理的效率。实验结果表明，相比于当前最先进的动态图加速器工作JetStream的异构版本JetStream-H，NeoStream取得了平均17.99倍的性能加速比。

面向基因组图分析的存内计算加速器研究

姓　　名：周卓然
研究方向：存内计算、存算一体
导　　师：廖小飞
指导老师：黄　禹
E-mail：939921357@qq.com
QQ：939921357
联系电话：15907220138
毕业去向：云南省政府

基因组分析是现代生物信息学与精准医学领域中的核心研究方向。基因组图通过引入图

结构整合多个个体的遗传信息，从而显著提升比对的覆盖率与准确性。基因组图分析主要面临两方面挑战：播种阶段需执行大规模索引查询，从而带来较高的存储访问延迟；对齐阶段则涉及到序列比对的高度密集计算，对算力资源提出严苛要求。

为应对上述挑战，针对基因组图分析流程进行系统优化设计并提出一种轻量级、兼容主流动态随机存取存储器的存内计算架构，以软硬件协同设计高效加速基因组图分析流程。该架构结合近存计算与位级原位计算的优势，从存储访问模式优化与计算并行性提升两个维度开展系统设计。首先，在播种阶段引入基于索引的卸载机制，借助近存计算的低访问延迟特性，有效缓解不规则内存访问引发的性能瓶颈；其次，在对齐阶段通过行级并行计算与距离感知调度机制，提升资源利用率，挖掘任务间潜在并行性，从而降低整体计算延迟。

在标准人类基因组数据集上的实验评估表明，所提出的架构在短（长）读长比对任务中相较于现有CPU、GPU与ASIC方案，分别实现了最高502倍（30.2倍）、272倍（15.1倍）和5.5倍（8.3倍）的性能提升，同时能耗分别降低1628倍（85.6倍）、1443倍（77.1倍）与7.8倍（11.7倍）。

大　数　据

面向多重标记图的高效子图匹配系统设计与优化

姓　　名：曹　颖

研究方向：数据库、数据挖掘、子图匹配

导　　师：丁晓锋

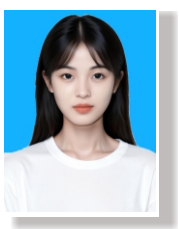
指导老师：袁平鹏

E-mail：2661254767@qq.com

QQ：2661254767

联系电话：15527815802

毕业去向：阿里云计算有限公司



子图匹配作为图挖掘中的核心任务，对模式发现、结构化查询和推理具有重要意义。随着图数据规模与结构复杂度的不断增长，传统算法在多标记场景下难以兼顾匹配精度与计算效率。

针对以上问题，提出了一套解决方案。所设计的LMiner系统采用图结构信息与标记信息解耦处理的方法：结构部分采用改进的压缩稀疏行格式，以提升结构存储与访问效率；标记部分引入质数乘积映射策略，对顶点与边的多重标记进行编码，并结合质数复用机制，有效压缩标记信息。在此基础上，构建顶点的双重索引机制：标记域索引通过构建顶点邻接边的标记特征描述体系，并以域面积为依据进行起始点筛选与候选集过滤；邻域索引记录顶点的局部拓扑结构，为匹配过程提供结构约束支持。匹配过程中，算法依据标记约束生成初始匹配计划，并结合顶点的最大后代数动态优化匹配顺序，最终采用多线程并行执行策略，有效提升整体匹配效率。

实验结果表明，LMiner系统在多种图类型的匹配任务中展现出较为优异的性能表现。

算力差异感知的微批流处理系统动态数据分区优化研究

姓　　名：陈　豪

研究方向：大数据系统

导　　师：陈汉华

指导老师：陈汉华

E-mail：517888189@qq.com

QQ：517888189

联系电话：18672361173

毕业去向：国家知识产权局专利局专利审查协作湖北中心

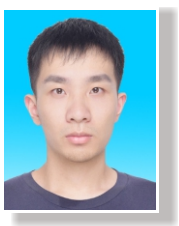
微批流处理系统结合了批处理和流处理的优点，通过将连续数据流离散化为微批实现近实时高吞吐数据处理。其技术优势体现为：

基于批次缓冲机制降低流式传输开销，利用缓存局部性优化提升计算效率，并通过分布式数据并行架构扩展至大规模集群。然而，现有的微批流处理系统普遍基于同构环境设计，依赖均匀数据分区维持节点间负载均衡，难以适配当前数据中心广泛存在的异构算力环境，由于计算节点间存在算力差异，对相同任务的处理能力不同，均等划分的数据块将引发处理时间不均衡，最终导致系统吞吐率显著下降。

针对异构算力环境下微批流处理系统负载不均衡和任务特性与节点算力不匹配的问题，提出了算力差异感知的动态分区与节点分配调度联合优化框架。在数据分区层面，构建基于XGBoost回归的异构算力模型，通过量化分析节点硬件指标（如CPU算力、内存容量）、数据特征（如批次大小、键值分布）与任务逻辑间的耦合关系，预测各节点处理时延；基于此设计贪心动态分区算法，以处理时间均衡为目标，实现异构集群的数据分区。在节点分配调度层面，提出任务特性驱动的差异化节点分配策略，结合流水线并行调度机制，依据任务计算密集度调整节点分配，最大化异构资源利用率。

基于全同态加密的低延迟安全外包矩阵乘法

姓　　名：陈　祎
研究方向：分布式算法，大数据处理
导　　师：华强胜
指导老师：华强胜
E-mail：860081469@qq.com
QQ：860081469
联系电话：18270609619
毕业去向：北京大学攻读博士



全同态加密（Fully Homomorphic Encryption, FHE）允许直接在加密的敏感数据上做任意计算，是目前实现无交互式安全外包计算的最有前景技术。基于现有FHE方案的安全外包计算相较于明文计算，在服务器上有5个数量级的性能差距，使其难以落地应用。

主要工作如下：（1）针对任意矩阵维度矩阵乘法，构建基于超立方体 编码的密文–密文矩阵乘算法，将同态旋转操作次数降低至与矩阵维度平方根成正比，同态乘法次数降低至矩阵最小维度值；（2）引入以减少同态操作为目标的“块布局”中间表示，有效规避单个RLWE密文无法容纳全部输入数据的问题，同时适配不同FHE参数配置，并提升同态计算任务在分布式环境下的可调度性；（3）针对传统分布式矩阵乘法在同态加密场景下存在冗余计算及密钥存储开销的痛点，设计四维矩阵乘法（FHE4DMM算法），通过联合优化大规模密文–密文矩阵乘法的通信和计算开销，显著降低总延迟。

相较于现有最优分布式方案，FHE4DMM在所有支持的矩阵类型上加速比最高达16.62。在真实安全外包推理任务中，基于密态模型和密态MNIST、CIFAR-10数据集的实验结果表明，FHE4DMM将推理整体加速提升至3.54倍和4.22倍，其中矩阵乘模块分别加速4.09倍和4.87倍，验证了该方法在实际应用中显著的性能优势，有助于推动FHE实用化落地。

类型信息增强的知识图谱表示学习方法

姓　　名：程传斌
研究方向：知识图谱表示学习、知识图谱嵌入
导　　师：袁平鹏
指导老师：袁平鹏
E-mail：765050834@qq.com
QQ：765050834
联系电话：18870320805
毕业去向：浙江省衢州市政府

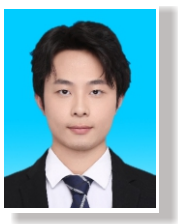
知识图谱作为结构化语义网络，在智能搜索、推荐系统等领域具有重要应用价值。本文提出类型增强的层次化注意力网络模型TEHAN

(Type-Enhanced Hierarchical Attention Network)，通过构建类型模式与类型交互模式的语义约束体系，实现知识图谱的层次化建模。

TEHAN通过Transformer编码器学习实体多维度类型标签的语义关联，生成可解释的类型嵌入。在此基础，定义类型模式划分三元组语义场景，结合二维卷积提取类型组合的泛化环境信息；进一步引入类型交互模式，通过关系约束下的卷积策略捕获细粒度语义特征，动态筛选有效类型模式。层次化注意力网络分为两阶段：局部网络聚焦类型模式内四元组（实体、关系、尾实体、语义信息）的差异化语义建模，通过注意力机制量化贡献权重；全局网络融合跨类型模式的环境信息与局部表征，通过多头注意力和融入初始嵌入提高稳定性。

实验表明，TEHAN在FB15k-237和WN18RR数据集上的链路预测任务中表现优异。在FB15k-237数据集，实验指标Hits@10较主流基线模型最高提升13.7%，MRR分别达到0.482和0.510，较最优基线提升9.05%。在WN18RR数据集，实验指标Hits@3较主流基线模型最高提升23.0%，MRR达到0.510，较最优基线提升18.6%。此外，信息可视化验证了语义信息和环境信息的作用，消融实验与敏感性分析验证类型增强机制和层次化注意力网络的有效性，揭示参数对性能的影响。

基于图神经网络的代码漏洞检测模型可解释性研究



姓　　名：储朝阳
研究方向：软件智能化
导　　师：万　瑶
指导老师：万　瑶
E-mail：zychu418@gmail.com
QQ：3199266412
联系电话：15907163884
毕业去向：伦敦大学学院读博

近年来，基于图神经网络的漏洞检测技术凭借其高效捕捉代码结构化语义信息的能力，获得了广泛关注。然而，当前研究普遍关注检测精度的提升，忽视了对模型可解释性的探究。这种局限性在实际应用中导致安全人员既难以验证漏洞判定结果的可靠性，也难以根据预测结果追溯漏洞触发根源。

针对上述问题，基于假设分析范式提出首个面向基于图神经网络的漏洞检测模型的反事实推理方法CFExplainer。该方法旨在识别能够翻转漏洞判定结果的最小代码图扰动，从而揭示模型决策的边界条件与核心依据。具体而言，首先将离散的反事实扰动搜索问题转化为对连续边掩码的优化任务，通过可微分掩码表示扰动强度；然后设计联合优化目标，在最小化代码图扰动规模与最大化模型预测翻转之间实现动态平衡；最终利用优化得到的扰动模式生成反事实解释，明确漏洞触发的根本原因并为修复提供可操作指引。

为验证CFExplainer的有效性，在四种代表性的图神经网络架构（图卷积网络、门控图神经网络、图同构网络、k维图神经网络）上开展系统性实验。定量评估结果表明，相较于基于事实推理的基线解释器，CFExplainer在面向漏洞定位精度的评估指标——准确率、精确率、召回率及F1分数方面，分别实现了19.72%、7.36%、24.11%和13.05%的平均性能提升。

面向代码大模型的高效增量学习方法

姓　　名：丁　甲
研究方向：增量学习，大模型
导　　师：金　海
指导老师：张　腾
E-mail：1027796877@qq.com
QQ：1027796877
联系电话：15334050031
毕业去向：华为技术有限公司

随着机器学习训练数据量的爆发式增长，批量学习范式因依赖全量数据进行训练，对计算存储资源消耗巨大。另一方面，增量学习因其能利用新增数据进行局部更新的特点，在如今模型参数量动辄百亿千亿的大模型时代，无论是通用任务上的预训练还是下游垂类任务上的微调，都得到广泛运用。大模型增量学习方法普遍采用混合专家结合低秩适配的方式，但其面临两个问题，一是收敛速度缓慢，二是知识迁移效率低。对此，提出一种加权专家低秩适配方法（Weighted Expert Low-Rank Adaptation, WELoRA），旨在实现大模型增量学习的参数高效更新和抗灾难性遗忘。

该方法首先将历史任务的低秩适配矩阵构建为专家库，每个专家对应特定的任务。其次，设计基于缩略核均值嵌入的任务相似性度量方法，该方法通过核映射将任务数据分布映射到再生核希尔伯特空间，计算任务间的分布距离，从而动态筛选与当前任务最相关的历史专家。最后，引入负权重机制与稀疏过滤策略，通过优化相似性权重矩阵，抑制冲突任务的干扰，同时保留关键任务的知识迁移。这一系列设计使得模型能够自适应地组合历史专家知识，在保证参数效率的同时实现跨任务的稳健知识迁移。

所提方法在代码垂类领域进行了实验验证，CodeXGLUE和XLCoST数据集上的实验结果表明，结合了WELoRA的代码大模型在代码生成、翻译等任务中显著优于基准增量学习方法，平均遗忘率仅为4.3%，且时间开销减少40%。此外，通过消融实验验证了WELoRA各个组件的有效性，进一步证明了该方法在训练效率与减少模型遗忘方面的优势。

代码语言模型的词汇表构建与迁移研究

姓　　名：杜荣明

研究方向：代码智能、代码语言模型、词汇表

导　　师：万　瑶

指导老师：万　瑶

E-mail：reecedoo@163.com

QQ：2283677214

联系电话：15327402190

毕业去向：联通数据智能有限公司

代码语言模型作为人工智能技术的重要应用，正逐渐展现出其强大的潜力和广阔的应用前景。然而，尽管代码语言模型的应用备受关注，其核心技术之一——代码的有效表示与理解——却未获得足够的重视。词汇表作为构建代码表示的基石，对模型的整体性能有着至关重要的影响。由于代码与自然语言在语义和结构上存在显著差异，直接应用自然语言的词汇表构建策略可能导致语义混淆、结构信息丢失等问题，从而影响模型性能。

针对上述现象，深入探讨了代码语言模型的词汇表构建方式及其迁移策略。在词汇表构建方面，聚焦于自然语言与编程语言的语义模糊性问题，系统分析了不同词汇表构建方式对模型性能的影响机制。同时，研究深入探讨了词元词缀对词汇表冗余和模型语义学习效率的影响，进一步阐明了其对模型性能的影响。在词汇表迁移方面，提出了创新的“词元使用率”评估指标，用于精准衡量代码语言模型词汇表的词元利用效率。并基于该指标，设计了一种优于平均嵌入迁移方式的词汇表迁移策略，为代码语言模型的优化提供了新的思路和方法。

实验结果表明：1）自然语言与编程语言的语义差异对代码语言模型的词元语义学习具有显著影响，但统一分词器可以有效利用两者的语义重叠性，提升模型性能。2）词缀信息对嵌入向量的贡献有限，去除词缀信息并不会导致模型性能下降。3）“词元使用率”评估指标能够有效衡量词元利用效率，并为词汇表优化提供了评估方式。4）基于词元使用率的词汇表迁

移策略在下游任务中表现出优异性能，为跨语言模型的词汇表适配提供了创新解决方案。这些成果为优化代码语言模型的分词策略和提升性能提供了理论基础和实践指导，推动了其在软件工程领域的应用。

面向频繁激增负载的实时流处理系统高效伸缩技术研究

姓　　名：方　正

研究方向：大数据流处理、伸缩性、频繁激增负载

导　　师：陈汉华

指导老师：陈汉华

E-mail：278668539@qq.com

QQ：278668539

联系电话：18371217805

毕业去向：长江存储有限公司

现实实时流式应用的运行特征表明，流负载往往呈现出频繁激增的数据模式，这些负载激增具有瞬时触发、短期持续、高频发生的特点，给系统的资源管理和动态伸缩带来了极大挑战。现有最新工作利用快速启动的无服务器函数实例处理突发负载峰值。然而此方法的周期性实例回收策略难以适配频繁激增的输入负载，导致系统资源利用率降低。此外，同等规格下无服务器函数的价格成本远高于虚拟机，系统利用其处理频繁出现的负载峰值会带来严重的成本开销。

针对频繁激增负载所导致的资源利用率降低及成本开销过高的问题，提出一种面向频繁激增负载的峰值预测系统ForMa。设计一种新颖的峰值感知负载预测模型，融合双态峰值识别以及双域周期性特征增强机制，深入挖掘频繁激增负载中的峰值时刻以及周期性信息，从而有效预测峰值负载。进一步，设计分段式滑动预测窗口结构，对预测结果窗口划分不同处理段，利用前后段的滑动覆盖机制，消除了负载峰值非均匀分布导致的处理优先级差异。最后，提出基于峰值负载预测失效的处理机制，细分预测失效场景以及差异化失效处理策略来确保系统在预测失效时的高效运行。

基于实际系统大规模数据集的实验表明，ForMa既有效应对了频繁激增的数据负载，又避免了频繁使用无服务器函数实例带来的资源低效利用和成本过高的问题。相较于现有最新工作，ForMa在保证处理时延的同时，将资源利用率提升了50.2%，吞吐率提升了57%。

分布式多重标记图查询系统

姓　　名：甘　芮

研究方向：分布式图数据库查询

导　　师：顾　琳

指导老师：袁平鹏

E-mail：749977062@qq.com

QQ：749977062

联系电话：18782866706

毕业去向：四川省政府

随着大数据的快速发展，图数据展现出规模呈指数级增长、结构复杂度提升以及语义内涵更加丰富的特点。作为处理海量数据的关键技术，分布式图查询已成为研究的热门领域。然而，现有系统存在明显的性能瓶颈：基于子图连接的查询方式产生的冗余中间结果不仅占用内存资源，还增加了通信开销；而基于搜索的查询方法提高了剪枝效率，但并行度低且难以高效利用计算资源。因此，如何充分利用集群资源提升查询效率成为亟待解决的问题。

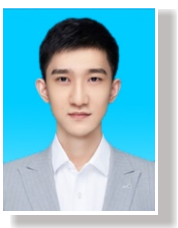
为了解决上述问题，本研究设计并实现了分布式多重标记图查询系统（Distributed Multi-labeled Graph Query，DMGQ）。首先，系统采用多重标记图表示形式，通过在顶点和边上附加多重标记，实现对实体及其关系更丰富的语义表征。基于此设计了一种基于双向邻接表和索引表的分布式存储方案，以优化数据访问效率，并结合贪心图划分方法，保留数据局部性以减少

通信次数。其次，提出了一种混合搜索策略，将查询过程划分为细粒度的任务，通过广度优先搜索生成新任务，再由深度优先搜索筛选数据，结合异步通信机制，实现任务流水化执行，从而优化查询效率。最后，提出了一种节点内和跨节点的两级负载均衡策略，通过动态调度任务平衡工作负载，提升资源的利用率。

实验结果表明，DMGQ系统在性能和可扩展性方面均表现出显著优势，尤其适用于处理复杂的查询任务。在查询性能方面，DMGQ系统较对比系统的平均性能提升了1.44至3.95倍。在可扩展性方面，DMGQ系统在水平扩展和垂直扩展上的加速比接近线性，充分证明了其在大规模图数据处理中的高效性和可扩展性。

基于全同态加密的分布式隐私推理系统加速技术研究

姓　　名：洪子晓
研究方向：分布式算法、大数据处理
导　　师：华强胜
指导老师：华强胜
E-mail：513464215@qq.com
QQ：513464215
联系电话：18561932088
毕业去向：华为技术有限公司



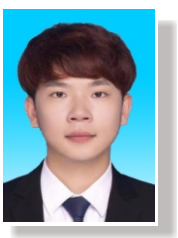
在数据安全与隐私保护需求日益增长的背景下，全同态加密成为关键技术。但由于全同态加密存在密文膨胀、计算效率低等问题，限制了其广泛应用。目前学术界研究大部分集中于硬件加速，少部分分布式相关研究未能针对全同态加密本身特点专门优化。

针对以上问题，设计多种基于全同态加密的分布式加速技术，并以此构建分布式全同态加密卷积神经网络推理系统。首先，提出分布式场景下的全同态加密矩阵连乘算法；同时，提出分布式批量自举算法，支持多密文自举，将自举操作内部密文数据合理划分，以减少旋转密钥数量、降低全同态加密参数，最终提升自举吞吐量和自举效率。基于矩阵连乘算法构建高效全连接层计算框架并实现基于层次全同态加密的分布式卷积神经网络推理系统，并对网络层算法和数据排布进行优化。在基于分布式批量自举的非层次全同态加密ResNet卷积神经网络推理中，结合最小自举策略等多种优化技术，降低整体推理时延。

实验结果显示，在不同数据集上，基于层次全同态加密的分布式卷积神经网络推理系统在64节点下较单机推理性能提升160倍，相较于现有最优分布式卷积神经网络推理系统端到端推理时延降低4倍左右。此外，提出的分布式批量自举算法相较于平凡分布式自举实现了1.39倍整体自举性能提升。以分布式批量自举为核心的基于非层次全同态加密的分布式ResNet-20网络推理系统在与目前最优全同态加密实现对比中，推理速度提升1.56倍。对上层应用分布式加速算法的探索，为全同态加密加速提供新思路，对推动全同态加密技术的实际应用具有重要意义。

面向乳腺钼靶分类的对抗性四视图深度学习模型

姓　　名：侯宇翔
研究方向：医学影像，计算机视觉
导　　师：陆　枫
指导老师：陆　枫
E-mail：1792188973@qq.com
QQ：1792188973
联系电话：17355139910
毕业去向：华为技术有限公司



乳腺癌是全球女性健康面临的重大公共卫生挑战，在世界范围内位居女性恶性肿瘤发病率和死亡率前列。在乳腺癌防治过程中，通过影像筛查实现早期发现并进行及时治疗非常关键。近年来，利用神经网络技术提高乳腺癌筛

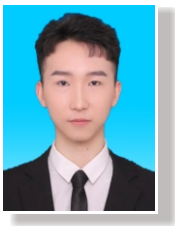
查准确性的研究日益受到关注。然而，现有方法主要局限于分析患侧乳腺视图，未能充分发掘非患侧乳腺视图的潜在临床价值。此外，现有对抗样本生成方法往往缺乏临床合理性，制约了其在真实医疗场景中的应用。

针对上述问题，基于放射科医生阅片经验，提出了一种对抗性四视图分类网络（NaFV-Net），通过同时分析患侧与非患侧乳腺影像来提升分类性能。NaFV-Net由三个核心部分组成：乳腺肿块定位部分、对抗样本生成部分和四视图分类部分。首先，基于增强型乳腺钼靶影像的特性，采用不对称敏感算法和混合注意力机制，精准定位并分割患侧减影图像中的肿块区域，并将位置信息映射至低能图像。随后，通过类mixup算法将患侧低能图中的肿块区域移植到非患侧低能图的对称位置并增加肿块位置信息的扰动，生成具有临床意义的对抗样本。这种位置扰动策略能有效抑制模型过拟合，增强分类鲁棒性。最后，四视图分类网络整合多视角乳腺X光影像信息，结合卷积神经网络局部特征提取与Transformer的全局建模能力，实现病灶局部细节与全局关联特征的多层次挖掘，有效提高了乳腺癌分类的准确性。

实验结果表明，NaFV-Net在乳腺癌良恶性分类任务中表现出优秀的性能。与现有国内外同类方法相比，模型在多项关键评价指标上均取得显著提升。在内部验证集上，NaFV-Net的AUC值、准确率和召回率分别提高了11.1%、10.2%和10.0%。在外部公开数据集的验证中，模型同样表现出良好的泛化能力，上述指标分别提升了8.2%、8.6%和6.8%。这些改进增强了模型在临床乳腺癌早期诊断中的应用价值。

基于FPGA的椭圆曲线多标量乘法加速器研究

姓　　名：黄恺一
研究方向：零知识证明、硬件加速
导　　师：石宣化
指导老师：石宣化
E-mail：ky_huang_work@outlook.com
QQ：932229607
联系电话：15827051262
毕业去向：快手网络科技有限公司



零知识证明作为隐私计算与区块链领域的核心技术，其大规模应用受制于较低的证明生成效率。在非交互式零知识证明协议中，椭圆曲线多标量乘法占据证明者模块约70%的计算开销，是制约系统性能的关键因素。现有分布式CPU方案因网络延迟严重制约性能，GPU方案针对有限域大整数运算能效低下，而目前的FPGA方案缺乏对计算负载与资源分配的系统性剖析，未充分探索优化资源分配后的性能潜力。

针对该问题，提出聚焦资源的FPGA椭圆曲线多标量乘法加速器架构设计。从计算资源角度出发，创新性地引入共享负载策略与分治剪枝优化算法，通过降低单计算单元资源消耗实现多计算单元并行部署，在有限硬件资源约束下达成系统吞吐量的线性扩展。在存储资源层面，通过预计算机制将部分计算负载转化为存储开销，并引入动态窗口缩放策略实现主机端与设备端的负载均衡调度，从而在异步计算框架下获得更优的资源利用率和更低的计算延迟。

基于Xilinx U250 FPGA开发平台完成了原型验证与性能评估。针对BLS12-381椭圆曲线在220~225点规模下综合实验结果表明：与CycloneMSM、Hardcaml-MSM和BSTMSM等主流FPGA椭圆曲线多标量乘法硬件加速器设计方案相比，在单计算单元的资源占用上最大可以节省30.3%的寄存器资源，同时计算性能最高提升5.4倍。

基于用户画像与子图关联的推荐方法研究

姓　　名：黄宗耀

研究方向：大语言模型
导　　师：石宣化
指导老师：石宣化
E-mail：1500923230@qq.com
QQ：1500923230
联系电话：13320155849
毕业去向：上海寻梦信息技术有限公司

当前推荐场景文本属性与图结构并存，存在用户文本信息不可见场景下用户画像表征难问题和传统推荐模型图结构获取效率低问题。为解决问题，提出用户画像表征学习技术与轻量级图结构信息获取技术。用户画像表征学习技术通过利用用户历史交互和相关项目文本属性，设计针对性指令对大语言模型进行微调。基于微调之后的大语言模型，引入文本属性信息，实现用户画像的个性化构建；并利用训练得到的嵌入表示模型，生成用户嵌入表示，有效解决用户文本信息未知场景下的用户画像表征难题。轻量级图结构获取技术首先设计子图采样算法，高效地获取子图信息，生成节点多跳嵌入；同时，设计子图关联算法，训练多层感知机模型，完成轻量级的用户–项目相关性计算，从而实现图结构信息的高效精准获取。在此基础上，进一步构建基于用户画像–项目子图关联的推荐方法，通过基于用户画像表征的嵌入生成模块和基于用户画像–项目子图关联模块的结合，达成用户画像表征以及轻量级图结构获取。实验证明了方法的有效性。

标记组合约束的可达查询处理方法研究

姓　　名：刘津源
研究方向：图查询、数据库
导　　师：袁平鹏
指导老师：袁平鹏
E-mail：614903495@qq.com
QQ：614903495
联系电话：17640444467

图数据中的可达性查询作为基础性操作，在社交网络关系挖掘、知识图谱路径推理等场景中具有关键作用。在实际场景中，图的边常带有标记信息，标记图上的可达性查询通常带有标签限制，其中标记约束可达性（Label-Constrained Reachability, LCR）查询得到了广泛研究，但现有的LCR查询在功能性和性能方面也有所不足。在功能层面，LCR查询仅支持路径标记均在给定标记集内的基础查询场景，在导航规划等复杂查询场景中无法得到结果。在性能层面，当前最优的P2H+方法在十亿级图数据上索引构建超24小时，且存储开销达PB级，严重制约实际应用。

针对现有LCR查询的局限性，提出了一种标记组合约束可达性查询解决方案。该方案通过引入IN、NOT、OR和AND四种标记组合约束，显著扩展了传统LCR查询的功能边界，使其能够支持更复杂的查询场景。在技术实现层面，设计了一种基于拓扑感知的标记索引（Topology- aware Label Index, TLI）架构，该架构通过整合标记索引、顶点索引和合并表结构，实现了顶点到叶子顶点路径标记信息的高效记录与查询。

实验结果表明，与P2H+相比，TLI索引构建时间减少95%以上，存储空间降低20%至85%。此外，在NOT、AND、OR标记组合约束下，TLI的可达查询处理速度较深度优先搜索提速1至4个数量级，验证了其高效性与扩展性。本研究拓展了LCR查询的语义，为智能导航、社交网络分析和供应链优化等实际应用提供了高效可靠的技术支持，对推动图数据管理领域的发展具有重要意义。

知识图谱时序知识抽取与补全技术研究

姓　　名：刘正涛
研究方向：自然语言处理、知识图谱

导　　师：石宣化
指导老师：石宣化
E-mail：1317666188@qq.com
QQ：1317666188
联系电话：15705901295
毕业去向：华为技术有限公司

时序知识是知识图谱中相关联的事件按时间先后顺序产生的一组排列，抽取出事件对间的时序关系并补全缺失的时序节点有助于揭示事件的演化规律。现有的时序关系抽取算法存在事件到关系的误差传播以及对不同的时序关系区分度不足的问题，而主流的时序节点补全算法对于未见事件的补全精度不佳，模型的训练效率低下且泛化能力较差。

针对上述问题，提出了基于多任务联合学习的时序关系抽取算法和基于相似性模式挖掘的时序节点补全算法。基于多任务联合学习的时序关系抽取算法将对比学习与联合抽取相结合。联合抽取通过共享参数的方式，同时进行事件识别和时序关系抽取，增强了子任务间的交互性，减少了事件到关系的误差传播。对比学习模块通过实现合理的正负样本构造策略，学习更深层次的语义特征，以捕捉不同时序关系之间的差异，从而提高抽取结果的稳定性和可靠性。提出的基于相似性模式挖掘的时序节点补全算法包含相似性匹配模块和历史频率学习模块两部分。在相似性匹配模块中设计了一个高效的时序位置编码策略，结合线性堆叠的多层感知机，对实体间的结构依赖性进行建模，提升了未见事件的补全精度。历史频率学习模块利用图谱序列的过往频率信息，学习实体间交互的长期行为模式。提出的时序节点补全算法简洁有效，在提升节点补全准确率的同时显著降低了模型训练时的资源消耗。

在真实数据集TB-Dense、MATRES和ICEWS上分别对提出的时序关系抽取算法和时序节点补全算法进行测试，实验结果表明提出的算法在通用指标上均优于现有的国际先进算法。时序关系抽取算法在准确率和F1指标上分别实现了最高17.8%和9.3%的性能提升，时序节点补全算法在MRR指标上实现了最高5.2%的性能提升，且模型的训练效率平均提升了10倍。提出的算法通过提升时序关系抽取的准确性和节点补全的泛化能力，为复杂事件链的因果推理和趋势预测提供了可靠的支撑。

安全可验证的跨模态kNN查询处理方法研究

姓　　名：马　杰
研究方向：安全索引、跨模态检索
导　　师：丁晓锋
指导老师：丁晓锋
E-mail：2276518549@qq.com
QQ：2276518549
联系电话：15880370787
毕业去向：华为技术有限公司

随着多媒体互联技术的推广，跨模态kNN查询技术被应用于许多检索场景中。在数据上云的方案中，涉及多模态数据量化与查询处理。数据量化关注如何保留数据间的语义关联，需要综合考量模态不变性损失、模态内/模态间判别损失等多种指标。云服务器上的查询处理则是带来了数据安全与结果可验证的担忧。明文域上的查询存在数据隐私、结果隐私等泄露的风险，数据加密能够避免未授权的访问。结果可验证要求重现异构数据的计算过程，而目前大多数方法都基于单模态场景设计。

针对以上问题，提出了基于跨模态学习的量化方法、安全可验证的树形索引SVB-tree（Secure and Verifiable Ball Tree）以及kNN查询方法。首先，在量化方法设计上，采用多个子网学习不同模态的表示，并用权重共享保留原数据间的语义关联。接着，在SVB-tree设计上，采用数据分区对向量数据进行划分，基于此构建的多叉树的节点存储分区信息，并引入Merkle摘要

实现结果验证功能，最后对索引参数用Paillier加密。查询方法分为kNN搜索与验证对象生成两步骤进行。kNN搜索每次设定一个递增的搜索半径，根据节点相交情况决定是否向下搜索，当在叶子层搜寻到足够的数据点时该过程终止。验证对象生成阶段再次遍历SVB-tree，保留可用于重构根节点摘要的最小子树。为了应对高维空间中查询方法所面临的高延迟和通信开销的问题，提出了基于数据压缩的优化方案，并分析了查询过程的复杂度与安全性。

最后，采用Python实现了以上方法，在Wikipedia 等四个真实数据集上进行了实验评估。与同领域工作的对比说明了量化方法的优异性能，在多参数设定下的索引构建过程的时间与空间开销分析说明了SVB-tree的合理性。随后，在多种情境下对查询延迟进行了对比。相比同类方法，所提查询方案的性能提升近10倍，而基于数据压缩的优化方案更是将性能提升至60倍。安全性分析与实验结果充分表明所提方案的有效性与可靠性。

基于大语言模型的知识图谱补全方法研究

姓　　名：钱震宇
研究方向：知识图谱
导　　师：余　辰
指导老师：谢　夏
E-mail：634193463@qq.com
QQ：634193463
联系电话：15852583178
毕业去向：百度国际科技(深圳)有限公司

知识图谱在拓展知识库有效性、支撑智能问答和知识推荐系统方面发挥着关键作用，而知识图谱补全技术则是维持其完整性和适应新增知识的核心环节。

设计了一种基于大语言模型的知识图谱补全框架并进行了系统实现。首先，为解决小样本场景下的邻域表示建模问题，所设计的邻居编码器融合实体邻域的拓扑结构与自然语言描述，构建统一的语义–结构联合表示空间；其次，为缓解大语言模型知识遗忘问题并降低训练资源开销，二阶段背景知识图谱嵌入配合低秩适应训练策略，将参数更新限定于低秩子空间；然后，为弥补现有方法在置信控制方面的缺失，提出基于蒙特卡洛扰动采样的不确定性量化方法，通过对输入进行随机微扰动并结合核密度估计，刻画模型输出的可信度，并实现置信度的概率化校准。

在NELL-One、Wiki-One和FB15K-237三个小样本知识图谱补全基准数据集上，基于设计框架的系统分别达到了99.2%、83.1%和89.9%的Hits@1准确率，较现有主流模型平均提升3.5个百分点。在Tox21等7个图分类基准数据集上，系统实现了平均13.7%的性能提升，证明了其对复杂图结构的优秀建模能力。不确定性估计框架在Wiki-One数据集上获得0.82的AUC评分。

金融知识图谱超关系的语义增强抽取与图结构感知补全技术

姓　　名：乔辰奇
研究方向：深度学习、关系抽取
导　　师：石宣化
指导老师：石宣化
E-mail：1561196285@qq.com
QQ：1561196285
联系电话：15623755820
毕业去向：深圳市腾讯计算机系统有限公司

构建金融知识图谱超关系作为信息的结构化表示方法能够为自然建模高维数据提供基础。抽取和补全是图谱超关系构建的关键子任务。

设计了基于层叠指针框架的语义增强超关系抽取模型和基于条件信息传递的超关系补全方法。超关系抽取模型中设计了基于短语的实

体识别模型解决嵌套实体分类问题；提出了实体类别信息的隐式融入方法增强文本中的语义信息；采用基于优先级指针标注的层叠指针框架用于一对多等复杂关系的预测。超关系补全方法设计并实例化超关系知识图谱的条件信息传递框架，增强模型归纳性，通过分析证明了相比于传统超图神经网络，所提出模型表达能力更强。此外，收集并标注数据，提出了一个中文金融领域关系抽取任务数据集。

为验证所提出框架和模型的有效性，在两个任务多个数据集上进行了实验。对于超关系抽取任务，分别在公开关系抽取任务数据集和中文金融关系抽取任务数据集上进行了一系对比试验。在公开数据集DUIE上较基线模型在F1值衡量标准上提高了16.98%，在中文金融数据集上F1值提高了26.66%。对于超关系补全任务，在多个数据集上进行对比测试，所提出模型在所有数据集上均达到最优结果。此外在两个任务上进行了一系列消融实验，验证了所提出模块对于模型性能的提升作用。

基于全局依赖图的动态检索增强代码补全研究

姓　　名：谭　磊

研究方向：代码智能、大语言模型

导　　师：万　瑶

指导老师：万　瑶

E-mail：545811257@qq.com

QQ：545811257

联系电话：18973594585

毕业去向：华为技术有限公司

近年来，尽管基于大语言模型的代码补全技术取得了显著进展，但其难以有效利用全局上下文信息。因此，如何有效结合外部知识检索与大模型生成能力，实现精准高效的仓库级代码补全，成为亟待解决的重要问题。

针对上述问题，研究了仓库级代码补全任务中的全局依赖关系建模与检索增强生成技术，提出了一种基于全局依赖图的动态检索增强代码补全方法GloDynCoder。通过构建覆盖整个代码仓库的全局依赖图，精准捕获跨文件的语义关联与依赖关系。在此基础上，设计了基于结构相关性和语义相似性的混合代码上下文检索及融合策略，为模型提供充分的跨文件语义信息。并通过基于实时信息需求检测的动态检索触发机制，动态评估代码补全过程中跨文件上下文的实时需求，仅在必要时触发检索，以减少冗余计算、提高模型补全速度。

为验证提出方法的有效性，在CrossCodeEval与RepoBench基准数据集上进行了全面实验评估。实验结果表明，相比于现有的先进基线方法，代码精确匹配率提高了10.8%~20.4%。同时基于动态触发的检索机制显著降低了平均补全耗时，耗时较固定检索策略降低达31.8%。此外，还深入探讨了不同相似度检索方法、上下文窗口长度、动态触发检索阈值等因素对性能的影响，为实际应用中的参数选择提供了有益参考。

基于GPU的蒙哥马利模乘加速器设计与实现

姓　　名：王溆洲

研究方向：并行计算、加密计算

导　　师：石宣化

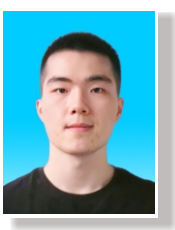
指导老师：石宣化

E-mail：whz_tec03@163.com

QQ：2247805411

联系电话：19967080758

毕业去向：汉口银行股份有限公司



随着数据规模的增长和安全问题的日益严重，隐私保护计算技术变得越来越重要，而蒙哥马利模乘作为RSA等加密系统的核心运算已经成为计算效率瓶颈。 GPU凭借其高效的并行计算架构成为突破性能瓶颈的关键，但现有基

于GPU的蒙哥马利模乘加速方案对输入位宽适应性差、采用的并行策略效率不足且未充分利用新型计算单元。针对这些问题，提出一种基于GPU的蒙哥马利模乘加速器TCSMM，旨在实现高效、通用的模乘加速。

以提升多精度模乘运算吞吐量与动态位宽支持为目标，首先提出分段整数乘法算法，将多精度整数划分为固定长度段，通过封装的单元乘法实现段间并行计算，支持256位至8192位输入且性能衰减可控。其次，设计二维并行化策略，基于计算图分析线程负载与通信开销，合理划分线程工作负载并采用共享内存和寄存器存储中间结果，降低线程通信开销，提升并行度与硬件资源利用率。进一步地，提出Tensor Core协同加速机制，将多精度整数乘法映射为矩阵运算得以充分利用现代GPU强大算力，同结合寄存器即时压缩技术降低格式转换开销，充分提升计算吞吐量。

在NVIDIA A100与RTX 4090 GPU平台上的实验结果表明，TCSMM在密钥长度为2048位的RSA、ElGamal和Paillier加密系统中，加密解密吞吐量较目前最优方案CGBN提升2.08至2.54倍。同时，面对256位至8192位的不同输入位宽均保持领先性能，相较于CGBN实现了平均2.11倍的加速比。对比实验验证了分段整数乘法、二维并行策略与Tensor Core协同加速的有效性。

动态超图桁架维护并行算法研究

姓　　名：王　猛
研究方向：动态图
导　　师：华强胜
指导老师：华强胜
E-mail：1622417059@qq.com
QQ：1622417059
联系电话：18995805296

K-桁架（K-truss）是检测图中凝聚子图结构的重要工具，其根据图中各边构成三角形数量对边进行分层，相比其他凝聚结构兼具计算高效性与凝聚紧密性的优点。但目前大部分对于桁架的研究局限于传统图，超图上相关研究较为欠缺。

文章对超图桁架值维护的理论建立与算法设计进行了研究，将超图中不同类型的三角形分为内、外三角形，并分别用内、外桁架值对两者进行度量。对于两种桁架值的增减情况，提出了相关桁架值维护理论并分别给出单个结点更新后两者的变化上界。内桁架值在单个结点更新后至多变化1，外桁架值变化上界与原图中点对和更新点对间形成的新三角形数相关。对于变化点对的范围，给出了单个结点更新时桁架值变化的点对分布。内桁架值的变化只发生在更新超边中，外桁架值的变化只发生在特定指标满足条件的区间内。

基于上述理论，文章提出作为基线的静态分解算法，并结合动态图的变化规律提出基于h指数的收敛维护算法。同时利用work-depth模型对时间复杂度进行分析，揭示影响算法性能的关键要素。在真实世界数据集上进行广泛实验。实验结果表明维护算法在插入与删除场景下均具有良好的稳定性、拓展性、并行性与普适性。相比于作为基线的分解算法，维护算法最高可快100倍。

面向代码大语言模型的指令微调研究

姓　　名：魏武才
研究方向：代码大模型
导　　师：万　瑶
指导老师：万　瑶
E-mail：2951562858@qq.com
QQ：2951562858
联系电话：15216269307
毕业去向：华为科技有限公司

代码大语言模型凭借强大的语言理解与生成能力，成为实现代码智能的核心驱动力。其中指令微调作为提升模型任务对齐能力的关键技术，能够促使模型更好地理解和遵循各种任务指令。

研究聚焦代码生成场景，剖析影响指令微调性能的关键要素，系统研究指令数据表征与模型微调效果的关系。具体地，从多方面围绕指令微调展开研究：（1）探究注释信息对微调效果以及其在不同规模模型的影响规律；（2）观察思维链信息对模型性能的影响及其与基座模型推理能力的关联；（3）分析不同复杂度思维链信息在不同基座模型上的微调效果差异及规律；（4）对比指令微调场景下不同参数高效微调方法的差异。并构造了相应表征的高质量指令数据，在HumanEval、MBPP上开展实验，量化分析相关要素对CodeLlama、Qwen2.5-Coder等系列模型的微调性能影响。

实验表明：（1）注释缺失对模型代码生成能力存在影响，尤其小模型受影响显著；（2）合理地引入思维链作为输出过渡信息，有利于提升模型的推理、代码生成等能力；（3）思维链信息的复杂度对微调效果的影响与模型规模存在关联，模型越大，越能适配更复杂的思维链；（4）不同参数高效微调方法存在显著优劣差异，如P-tuning和Prompt Tuning等仅在嵌入层添加可训练层的方法，不适用于指令微调。这些研究成果丰富了指令数据表征与代码大模型适配微调的相关实践经验，为面向代码生成的指令数据构建和微调策略优化提供了重要参考依据。

面向MindSpore的细粒度自动并行技术研究

姓　　名：吴　畅
研究方向：深度学习、分布式训练
导　　师：石宣化
指导老师：石宣化
E-mail：1813735859@qq.com
QQ：1813735859
联系电话：18371302451
毕业去向：华为技术有限公司

张量并行作为当前分布式训练领域主流的并行方法，在深度学习框架昇思（MindSpore）中依旧存在两个问题：首先，由于张量并行为每个算子指定一个切分策略，基于所有算子切分策略的组合会形成巨大的并行策略搜索空间，现有张量方法只能通过简单的代价模型和快速搜索算法来寻找并行策略，但策略性能不佳。其次，张量并行将计算分配到多个设备上并行执行，导致设备间存在大量的数据通信，现有通信优化方法并未考虑算子特性以及硬件属性，难以有效隐藏巨大的通信开销。

针对MindSpore中现有张量并行方法并行策略搜索空间大、搜索效率低以及代价模型不准确的问题，提出算子依赖感知的自动张量并行方法ODA（Operator Dependency-Aware）。ODA通过分析算子间细粒度数据依赖关系构建粗粒度算子簇OperatorCluster，并基于算子簇为基本单元重构并行空间；同时结合子图同构性特征提取独特子图，并利用子图开销组合构建准确代价模型，以优化并行策略选择，获得最佳训练性能。针对MindSpore中现有张量并行通信开销大、设备利用率低的问题，提出细粒度通信重叠的通信优化方法FGCO（Fine-Grained Communication Overlap）。FGCO基于数据依赖分析对共享张量进行分解，并结合昇腾NPU硬件特性对计算内核和通信内核进行细粒度切分并融合，同时在融合内核内部实现流水线式调度，从而有效地隐藏了张量并行训练中的通信开销。

基于大语言模型的自然语言驱动数据可视化实证研究

姓　　名：吴　瀑

研究方向：数据可视化，代码智能
导　　师：万　瑶
指导老师：万　瑶
E-mail：wuyang_emily@hust.edu.cn
QQ：1047772929
联系电话：18016571865
毕业去向：苏黎世联邦理工大学攻读博士

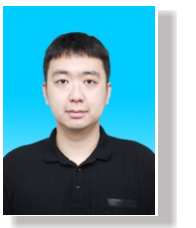
自然语言驱动的自动数据可视化任务，旨在基于自然语言描述生成结构化表格数据的可视化表示，以帮助用户从海量数据中获取洞察。近年来，尽管多种基于深度学习的方法在此任务取得了进展，但它们在处理未见过的数据库或跨多表数据时仍面临挑战。受大语言模型强大代码生成能力的启发，开展了一项实证分析研究，旨在挖掘其在生成可视化方面的潜力，并探索上下文学习提示策略在该任务中的有效性。

围绕以下核心问题展开：（1）如何将结构化表格数据高效编码为顺序文本提示，以最大化保留数据语义。主要探索了表格序列化、表格摘要、表格标记和表格编码四种表示方法。（2）如何通过上下文学习策略提升大语言模型在自动数据可视化任务中的表现，能否突破现有模型的性能瓶颈。对比了两种类型的大语言模型——微调模型（如T5-Small和T5-Base）和仅推理模型（如GPT-3.5系列、GPT-4和DeepSeek-V3）——与现有方法在基准数据集 nvBench 上的表现。（3）如何设计优化策略进一步提升模型的能力。实验分析了模型生成错误的类型，并且设计了一系列的优化策略。

实验验证了：（1）将结构化表格用编程语言进行编码表示，能够有效保留数据的语义信息，从而显著提升模型的性能。表模式（表名与列名）是提示词构建的核心信息，适当增加表关系信息（如外键）有助于提高模型在跨域任务中的准确度。（2）大语言模型在跨领域和域内任务中相比传统基线模型分别提升了26%和16%匹配率，尤其在提供足够上下文学习示例时，仅推理模型的性能可以逐步超越微调模型。（3）链式思考、角色扮演、自我修复和代码解释器等优化策略，分别提升了模型9.3%、12.8%、13.3%和50.3%的准确率。

深度学习分布式训练流水线并行性能优化研究

姓　　名：张昊林
研究方向：深度学习、流水线并行
导　　师：石宣化
指导老师：石宣化
E-mail：374462522@qq.com
QQ：374462522
联系电话：13127253129
毕业去向：阿里云飞天信息技术有限公司

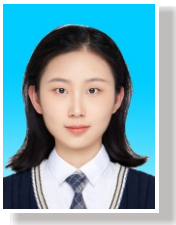


在模型参数规模持续扩张的大模型时代，流水线并行已成为多设备协同训练的关键技术。针对流水线并行训练中的显存瓶颈，提出了基于细粒度静态计算图的流水线并行训练框架 FGPipe。基于编译的思想，设计了细粒度剖析器对模型进行剖析并对训练过程中的模型状态信息实时监控，生成静态计算图并通过理论计算和实时模型信息协同更新计算图中的节点信息，使 FGPipe 可以精细且准确地感知模型状态信息。在此基础之上，FGPipe 通过多层次运行时显存优化技术，对不同的训练场景选择不同显存需求的调度策略以及细粒度应用不同内存优化技术，大幅度优化了流水线并行训练中的显存占用，大大了提升可训练模型规模。同时 FGPipe 通过非连续层移动技术，高灵活度地平衡流水线的计算负载，提升模型训练的吞吐量。

面向隐私保护的多模态近似最近邻查询方法研究

姓　　名：张　琪
研究方向：安全索引、隐私保护、多模态
导　　师：丁晓锋

指导老师：丁晓锋
E-mail：1070925249@qq.com
QQ：1070925249
联系电话：17396127967
毕业去向：交通银行股份有限公司



随着人工智能和物联网技术的快速发展，多模态数据（如文本、图像、视频、传感器信号等）呈现爆发式增长。由于其能够提供更加全面和准确的查询结果，多模态数据查询已成为智能推荐、医疗诊断、智慧城市等应用场景中的核心需求。然而，随着来自不同模态的数据的融合，个人敏感信息的推断和泄露风险也显著增加。现有的隐私保护方法通常是针对单模态数据设计的，难以同时兼顾多种模态数据的特性，同时也无法有效应对模态间关联性引发的隐私泄露问题，因此在多模态场景下的适用性存在一定的局限性。

针对上述问题和挑战，提出了一种面向隐私保护的高效多模态近似最近邻查询方法，包含多模态统一向量化模型，基于差分隐私的高维向量索引DP-HNSW，以及基于该索引的安全查询算法。首先，多模态统一向量化模型利用现有的预训练编码器，将文本、图像和音频数据分别转换为特征向量，并结合交叉注意力机制与对比学习，将不同模态的数据对齐到统一的高维向量空间，从而增强不同模态数据之间的语义一致性。接着，基于差分隐私的高维图索引通过在向量集合中随机插入伪节点，并基于这些伪节点构建索引，实现了节点级的隐私保护。该方法有效隐藏了真实节点及其邻接关系，从而降低隐私泄露风险。最后，安全查询算法采用“完全拒绝，选择回溯”的策略，有效解决了查询结果因伪节点引起的精度下降问题，确保了查询的准确性。

最后，实现了多模态统一向量化模型、高维向量索引和安全查询算法。在通过多模态统一向量化模型自行构造的向量数据集，以及多个现有的大型向量数据集上，评估了索引的构建时间和空间开销，验证了其可行性；并进一步评估了查询算法的检索性能和隐私保护能力。大量实验结果表明，在合理参数配置下，该方法能够在有效保护隐私的同时，将检索效率保持在接近原始索引水平的90%左右。

面向事件序列的因果发现算法研究

姓　　名：朱　华
研究方向：事件序列建模、因果发现
导　　师：黄　宏
指导老师：黄　宏
E-mail：1780776761@qq.com
QQ：1780776761
联系电话：15979710081
毕业去向：腾讯科技（深圳）有限公司

面向事件序列的因果发现算法是分析事件序列的重要方法，旨在学习不同类型事件之间的因果关系，以揭示真实世界的运行机制，进而为决策任务提供科学依据。然而，现有方法常面临事件序列建模的灵活性与可解释性难以兼顾、因果图优化效率不足以及非独立同分布事件序列处理受限等挑战。

为此，文章提出了一种显式交互感知注意力网络EIAN，通过类型独立的自注意力机制与从实例到类型的交叉注意力机制，实现事件交互的显式建模，兼顾了模型的灵活性与可解释性。然后，在EIAN的基础上，本文设计了一种基于变分推断和格兰杰因果的因果变分推理框架CausalVI，通过因果图显式地干预事件交互过程和建模事件因果关系，并以概率图的方式在无环性约束下高效地优化因果图。最后，面向拓扑网络中的非独立同分布事件序列，本文提出了一种拓扑感知因果注意力网络CausalNET，通过融合拓扑图与因果图来建模节点间的复杂事件依赖，摆脱了独立同分布假设的限制。

在一系列来自不同领域的真实数据集和合

成数据集上的广泛实验表明，本文提出的算法具备优越的事件序列建模能力和可解释性，在多个场景下的因果发现任务中均表现出卓越的性能、灵活性、可拓展性以及数据利用效率，为复杂场景下的事件因果分析提供了新的思路。

基于最优传输的主动学习算法研究

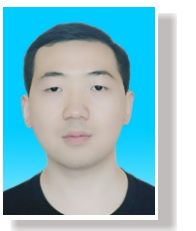
姓　　名：祝振宇
研究方向：机器学习、主动学习
导　　师：张　腾
指导老师：张　腾
E-mail：3109656543@qq.com
QQ：3109656543
联系电话：18871580661
毕业去向：阿里云计算有限公司

提出了一种基于最优传输的主动查询（Active Query by Optimal Transport，AQOT）策略。用OT系数分布的信息熵衡量未标记样本和正负样本的相似度，定义样本的最优传输置信度，与基于模型的置信度结合，从分布层面对样本信息量进行评估。提出了基于未标记样本OT系数的距离衡量方式，结合基于Softmax的预算分配方式为各聚类分配样本查询预算，选取具有代表性的核心集样本，从核心集样本中选取信息量较大的样本进行查询。提出了一种基于最优传输置信度的样本加权策略，对未标记样本进行加权，权值缩放策略随主动学习进度变化，并提出了一种半监督主动学习框架。将该框架与支持向量机（Support Vector Machine，SVM）结合实现了AQOT-SVM方法，并基于Rademacher复杂度分析了AQOT-SVM的泛化性能。还将该框架与梯度提升决策树、多层感知机结合，验证了策略的通用性。

为了评估AQOT的性能，在20个表格数据集上与基线模型进行对比，与现有最先进的主动学习策略相比，在查询不超过未标记样本总数10%的前提下，AQOT-SVM的F1分数平均提升了3.58%，AUC平均提升了3.22%，准确率平均提升了3.56%。三种AQOT模型与去除半监督模块、去除信息量模型、去除代表性模块相比，F1分数平均分别提升了1.95%、3.96%、2.61%。通过实验深入分析了各超参数对AQOT 的影响。2个图像数据集上的实验进一步表明了AQOT对数据模态和深度学习模型的广泛普适性。

安　全　组

面向TLS协议不同实现版本的指纹识别研究



姓　　名：陈群锦明
研究方向：网络安全
导　　师：邹德清
指导老师：袁　斌
E-mail：2493739087@qq.com
QQ：2493739087
联系电话：17872554187
毕业去向：娄底市供电公司

尽管TLS协议规范为协议实现提供了标准化的框架，但不同厂商的协议实现却表现出显著的多样性。这些实现之间的细节差异不断累加，形成了TLS协议的指纹，为TLS协议的安全性分析提供了新的视角。协议指纹的研究为应对现代互联网复杂化带来的多重挑战提供了重要解决方案。

因此，提出了一种创新性的TLS协议实现服务器的指纹识别方法。该方法借助状态机推演技术，能够精准地捕捉不同TLS协议实现服务器之间的细微差异，并通过独特的状态转移路径提取出每个协议实现的指纹，显著提高了指纹识别的粒度和准确性。基于上述方法，设计并实现了自动化工具TSI。TSI不仅能够高效执行指纹提取流程，还可以面向现实世界网站进行探测。仅依赖少量探针，TSI即可准确识别网站部署的TLS协议实现服务器的类型和版本。此外，TSI还能够为每个网站提供安全性分

析所需的重要信息，在TLS协议的安全评估和提升网络安全性方面发挥着关键作用。

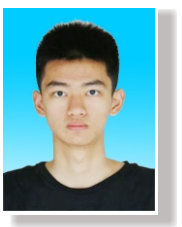
实验结果表明，相比于现有的其他指纹识别工具，TSI在指纹识别的粒度和准确性方面表现出了显著优势，能够捕捉到TLS协议实现之间的细微差异。同时，在针对现实世界747051个热门网站的探测分析结果中，TSI发现了8095个存在显著安全隐患的风险网站，展示了TSI在现实网络环境的广泛适用性。

基于检索增强种子生成的Python解释器测试技术

姓　　名：郭啸辰
研究方向：模糊测试、静态分析
导　　师：文　明
指导老师：文　明
E-mail：1059750328@qq.com
QQ：1059750328
联系电话：15161763296
毕业去向：杭州阿里云飞天信息技术有限公司

Python解释器作为Python语言的核心执行环境，负责解析、编译并运行 Python代码。考虑到Python在人工智能、数据科学等领域的广泛应用，其解释器的稳定性直接影响众多应用的正常运行。然而，传统的模糊测试方法存在缺乏高质量的初始种子池，测试范围局限以及种子变异存活率低等问题，难以有效发现潜在缺陷。为了设计高效的缺陷检测方法，本研究首先收集了Python解释器的历史缺陷数据并对缺陷分布、缺陷成因以及缺陷测试代码的结构特征进行分析。基于分析结果，本研究提出了PyFuzz，一种基于检索增强种子生成的Python解释器模糊测试技术。PyFuzz从代码库、语言文档等相关语料中提取信息构建索引，驱动大模型生成针对性的测试代码，从而构建高质量初始种子池。此外，PyFuzz拓展了一系列针对Python语言特性的变异算子，更有效地探索Python解释器中特有的缺陷。在变异的过程中结合高效的类型推断和语义检查机制，确保程序变体的语义正确性，提高测试的执行成功率，避免额外开销。最后在差分测试的流程中，PyFuzz融入了高效的错误种子 修复方法，提升其在Python解释器缺陷检测的有效性。

基于补丁分析的软件漏洞关联检测技术研究



姓　　名：李童天
研究方向：软件代码安全
导　　师：邹德清
指导老师：李　珍
E-mail：3207671754@qq.com
QQ：3207671754
联系电话：13554696939
毕业去向：中国移动湖北公司

当前研究中存在补丁识别和过滤工作缺乏持续化手段，补丁差异分析粒度粗，关联检测中程序分析方法不完善且易受噪声干扰等问题。

基于补丁代码差异信息提取的思想，提出了一种关联漏洞检测方法，通过已有补丁检查软件代码中未修补完全的1-day漏洞。该方法分为补丁分析与差异提取和基于程序分析的关联漏洞检测两部分。在补丁差异信息提取中通过对补丁残缺语句构建语法分析方法，并设计语法树比较和残缺语句比较方案，全面高效解析补丁文件差异信息。关联漏洞检测中根据补丁差异信息在待测软件构建的数据流图上定位节点，并通过不同漏洞类型的漏洞触发点分析方案，有效定位到漏洞触发点。判断流向漏洞触发点的潜在数据分支并将不满足漏洞特征的分支标记为潜在漏洞并输出。同样针对控制流语句设计了基于基本块遍历的控制流变量识别方法和遗漏修补的检测方案。

为了验证两个部分和方法整体的有效性，实现框架利用来自4个软件的120个真实补丁文件

全面评估了补丁分析中词法分析、语法分析、语法树比较和残句比较功能，证明了其高效的语法分析和特征提取能力。在该软件仓库上的228个补丁文件上评估了关联漏洞检测效果，达到了49.5%的安全覆盖率和5.8%的误报率，并发现了4个关联漏洞。

基于知识检索增强和多代理架构的智能漏洞修复研究

姓　　名：刘董奇
研究方向：大语言模型应用、软件安全
导　　师：徐　鹏
指导老师：李　珍
E-mail：louisliunova@outlook.com
QQ：542408455
联系电话：18207136039
毕业去向：南京信息技术研究院

基于开源软件代码仓库难以长期进行安全保障的问题，针对自动修复任务，结合大模型提出一种基于多代理技术和检索–增强–生成技术的自动智能漏洞修复系统，通过加入项目内外的知识库增强大模型的修复智能并减缓大模型的幻觉造成的错误补丁生成的问题。基于真实开源项目的测试表明优于仅适用大模型的简单修复，相比于基于学习的自动修复方法，具有无需依赖高质量数据集进行模型训练的优势，且更加适合中大型开源代码仓库安全保障的应用场景。

基于三维点云干扰分析的后门样本检测技术研究

姓　　名：刘　威
研究方向：三维点云、人工智能安全
导　　师：胡胜山
指导老师：胡胜山
E-mail：2410830772@qq.com
QQ：2410830772
联系电话：13618661704
毕业去向：腾讯科技（北京）有限公司

自动驾驶、虚拟现实等应用的兴起，极大地推动了以深度学习为核心的三维点云技术发展。三维点云技术的广泛应用不仅吸引了密切关注，也让研究者逐渐重视深度学习模型的安全与隐私问题。其中，后门攻击以高隐蔽性和可控性给人工智能安全构成严重威胁。由于点云数据的几何不变性和离散性，使得研究者提出的新型针对三维点云的后门攻击带来独特挑战。尤其是以方向为后门触发器的后门攻击不会改变原始点云的几何形状与语义信息，让现有的后门样本检测算法难以检测。

针对上述问题，采取从数据集整体到样本个体的分析方式进行。首先，针对不同的点云属性特征构建了点云干扰集，并提出了一种通用评估框架，预测稳定性测试。该评估技术仅通过查询模型输出的标签变化，使用一系列统计学指标来评估模型预测输入点云在不同干扰条件下，模型预测标签结果的稳定性，从而以数据集视角为正常样本和后门样本构建了统一评估标准。在此基础上，为了检测三维点云后门样本，提出了一种基于标签预测一致性的后门样本检测算法PointCRT。该算法适用于完全黑盒的模型推理阶段，旨在不依赖任何假设判断输入是否为后门样本。为了在个体层面上定量评估样本对干扰的敏感度，引入了干扰临界强度作为评估指标。该指标定义为使模型输出标签发生变化的最小干扰强度作为该干扰下样本的标签预测稳定性的临界值。同时，为了提高对未知后门样本的泛化性，引入基于分类模型的后门样本检测模块进行最终的推断。该模块显著地提升了检测后门样本的能力。

在多种实验场景下进行了检测后门样本的实验，结果表明PointCRT在多个实验场景下都能保持优异的后门检测性能，在基准数据

集上比现有的后门样本检测基线算法的平均性能超过18%~28%。而PointCRT在面对未知后门样本和真实数据集场景下，使用干扰临界强度能展现出良好的泛化性，同时显示出了基于非线性分类模型检测模块的实用性和可靠性。

基于可靠性的物理不可克隆函数机器学习攻击方法研究

姓　　名：陆箫潇
研究方向：物理不可克隆函数、机器学习、硬件安全
导　　师：王虹飞
指导老师：王虹飞

E-mail：897695505@qq.com
QQ：897695505
联系电话：13739113037
毕业去向：华为技术有限公司

随着物理不可克隆函数（Physical Unclonable Function，PUF）在轻量级硬件安全领域的广泛应用，其面临的机器学习建模攻击问题愈发严峻。现有的建模攻击方法存在数据需求量大、攻击时间长以及预测准确率低等问题，这限制了对PUF安全性的有效评估。因此，需要探索更高效的攻击策略，以准确评估PUF的安全性。

基于可靠性的物理不可克隆函数机器学习方法在现有的使用协方差矩阵自适应进化策略的可靠性建模攻击方法的基础上，提出了新改进思路：一是减少算法运行时间如仅使用协方差矩阵对角线元素对原矩阵进行近似减少矩阵更新的计算，并同时使用每代生成的最优和最差候选解的信息来使种群分布能更快逼近最优区域。二是提升预测准确率如使用更能表示PUF特性的斯皮尔曼相关系数代替原来使用的皮尔逊相关系数。在这些优化方法的基础之上，针对多种强PUF结构使用了不同的可靠性建模攻击方法进行了高效建模攻击。除上述方法外还提出了一种基于可靠性的神经网络建模攻击方法，使用“激励–可靠性”进行训练并对激励对应的响应进行预测。同时为了减少训练所需数据量，针对协方差矩阵自适应进化策略和神经网络分别使用了不同的迁移学习方法。

实验结果表明，在使用加速方法时，对于异或PUF、多路选择器PUF和插入型PUF可分别减少约10.27%、11.69%和13.74%的时间；使用提升准确率的方法时，准确率则分别提升了约1.84%、1.87%和1.20%。当使用基于可靠性的神经网络建模攻击方法时，对异或数量为6的异或PUF成功建模仅需使用十万级别的数据集，是传统机器学习建模攻击成功所需数据集大小的十分之一。当加入迁移学习后，使用协方差矩阵自适应进化策略和神经网络进行攻击所需数据量分别减少至原来的70%和53.78%。

数字集成电路多故障可诊断性和分辨率提升研究

姓　　名：罗陈亮
研究方向：逻辑诊断
导　　师：王虹飞
指导老师：王虹飞

E-mail：315656939@qq.com
QQ：315656939
联系电话：13618658850
毕业去向：华为技术有限公司

芯片制造过程中的缺陷不可避免。识别缺陷的根本原因首先需要逻辑诊断确定缺陷的位置。随着现代芯片的复杂度上升，芯片倾向于同时存在多个故障，它们之间的互相作用和巨大的搜索空间使得其中的任一故障都难以被识别。因此亟需高质量的多故障逻辑诊断以提高其可诊断性和分辨率。

研究提出了一种具有两个阶段的多故障逻辑诊断方法。在第一阶段中首先使用一种保守的路径追踪算法以获得初始的候选故障。对于每一个候选故障基于固定型故障模型和X故障模

型设计并提取共计36个不同特征。依靠直方图梯度提升算法构建机器学习模型来过滤假故障。为了在容忍少量真故障损失的前提下尽可能多得排除假故障，通过在训练过程中引入代价矩阵以及使用召回率调整预测的阈值来构建保守的模型。在第二阶段中，通过将多故障逻辑诊断问题视作组合优化问题，使用一种二进制差分进化算法的变体在剩余的候选故障中寻找真故障，为了进一步提高算法找到全局最优解的概率，研究改进了差分进化算法的初始化过程和交叉过程。

实验结果表明，直方图梯度提升算法能够去除98.21%的候选故障并保留95.22%的真故障，较随机森林算法性能更好。在预过滤候选故障的基础上，改进的二进制差分进化方法的平均可诊断性为94.45%，平均分辨率为82.78%，显著优于粒子群优化算法和商业诊断工具。

面向漏洞代码数据集构建的安全补丁识别与分类方法研究

姓　　名：王　虎

研究方向：漏洞代码数据集构建

导　　师：徐　鹏

指导老师：李　珍

E-mail：1984141285@qq.com

QQ：1984141285

联系电话：15623738228

毕业去向：浙江网商银行股份有限公司

大规模高质量的漏洞数据集是支撑开源软件漏洞检测和漏洞修复等研究的基础。通过真实软件代码仓库中的安全补丁可以获取大量真实软件的漏洞代码，因此研究安全补丁的识别与安全补丁漏洞类型的分类方法具有重要意义。

针对现有方法的问题，提出了基于多模态融合分类的安全补丁识别方法和基于样本优化与伪标签学习的安全补丁漏洞类型分类方法。在安全补丁的识别方面，综合补丁的文本信息和代码修改两方面的特征进行多模态融合分类。对于安全补丁漏洞类型分类，通过提取关键语义信息来优化样本质量，并运用伪标签学习策略扩充训练样本规模。最终实现了一个基于安全补丁识别和安全补丁漏洞类型分类的漏洞代码数据集的构建方法。

提出的安全补丁识别方法充分利用了补丁的文本信息和代码修改方面的特征，在真实软件安全补丁数据集上的F1分数达到66.99%，相较于现有最优的开源安全补丁的识别方法，F1分数提高了16.62%。提出的安全补丁漏洞类型分类方法在训练样本的质量和数量方面均优于现有方法，在安全补丁数据集上的加权F1分数达到了63.37%，提高了17.77%。提出的基于上述方法的漏洞代码数据集构建方法在真实场景中具有较强的实用性，漏洞代码样本识别的平均准确率为96.75%，其漏洞类型识别的平均准确率为87.34%。

基于专家知识优化样本生成的软件智能漏洞检测

姓　　名：王可馨

研究方向：软件漏洞检测、深度学习

导　　师：徐　鹏

指导老师：李　珍

E-mail：761077821@qq.com

QQ：761077821

联系电话：18292727890

毕业去向：保密单位

基于深度学习的漏洞检测方法在真实软件应用中存在不足：代码复杂性导致切片样本难以包含完整漏洞信息，传统模型对代码逻辑理解有限。为了解决上述问题，实现了用漏洞领域专家知识辅助的深度学习模型的C/C++源代码漏洞检测方法。首先，构建了基于软件源代码漏洞特征的知识，通过人工分析现有的漏洞代

码，总结了六种C/C++常见漏洞类型的漏洞根本原因和漏洞触发代码特征。其次，利用专家知识和大模型优化了漏洞模型的切片样本。通过大模型的代码理解能力，能够更精确地锁定疑似漏洞代码，减少切片数量。使用漏洞代码特征的专家知识对切片样本进行截断能够保证切片中漏洞信息完整性，同时尽可能减少漏洞无关代码。最后，在漏洞检测模型方面，构建漏洞指令微调数据集，用该数据微调现有代码大模型，使其更好地适应漏洞检测的需求。然后利用提示工程让模型聚焦漏洞检测任务于污点数据的跟踪和净化操作的检测，并且给出详细的漏洞原因说明，增强模型漏洞检测的准确性和可解释性。

在CVEFixes数据集上进行了实验，实验结果表明，与现有的深度学习漏洞检测模型相比，本系统的F1分数指标达到了78.0%，比现有最先进的方法平均提升了10.8%，并在openEuler真实软件上发现了三个漏洞。

基于动态测试的车联网协议漏洞挖掘

姓　　名：王坤明

研究方向：车联网协议、模糊测试

导　　师：袁　斌

指导老师：袁　斌

E-mail：3291138501@qq.com

QQ：3291138501

联系电话：15936934239

毕业去向：中证股转科技有限公司

车联网协议模糊测试受限于驱动程序编写，并且无法识别协议状态，导致模糊测试效率低下。针对上述问题，提出基于数据流的函数依赖识别方式，通过分析车联网协议库函数公开的说明文档中函数签名，依据参数类型推断的函数之间存在的依赖关系图，根据得到的依赖关系自动化构建模糊测试驱动程序。同时，提出基于报文格式的协议状态感知方式，使用同一种报文结构的函数被锁定为同一种协议状态，从每个协议状态下的报文结构出发，在函数依赖关系图中提取完整的调用链，从而生成对应协议状态下的模糊测试驱动程序。最后，根据每个模糊测试驱动程序中函数的参数类型和使用类型为不同函数参数设定不同的初始值和变异策略，减少模糊测试探索空间。

通过在Android Automotive真实库中的应用，评估了所提方法的有效性，并与FuzzGen 和GraphFuzz进行对比。实验结果表明，基于协议状态感知的模糊测试驱动程序生成方法可以有效学习协议之间的约束关系，并成功构建模糊测试驱动程序。该方案在代码覆盖率和错误发现方面都优于其他模糊测试程序，平均函数覆盖率方面达到93.95%，相较FuzzGen和GraphFuzz分别提高了625.48%、106.53%，并发现了Android Automotive 通信协议中的漏洞。

面向深度神经网络数据中毒的防御技术研究

姓　　名：王贤龙

研究方向：人工智能安全

导　　师：胡胜山

指导老师：胡胜山

E-mail：767249213@qq.com

QQ：767249213

联系电话：17764060053

毕业去向：香港城市大学攻读博士

近年来，许多研究表明深度神经网络训练过程容易受到数据中毒的安全威胁。为此，一系列防御技术相继被提出，但仍存在：缺乏广泛有效性、假设不够实际，且对干净数据损伤过大等局限性。

为克服现有防御的局限性，分别针对上述两类投毒场景的两种全新的防御方案被提

出。在加性范数可用性投毒攻击场景中，提出了一种基于稀疏扩散模型与轻量补偿模块相结合的防御方案。该方案通过稀疏同分布数据训练，使防御假设更为合理，并通过结合净化去噪和轻量图像转换技术，实现了广泛有效的防御，并且补偿模块的轻量化设计有效减小了对干净数据的损伤。同时，在乘性卷积可用性投毒场景中，另一种基于双线性插值的随机乘性图像转换防御技术被提出，其通过随机乘性转换打破中毒样本和标签的错误映射关系，有效防御了乘性卷积可用性投毒攻击，弥补了现有乘性卷积投毒场景防御技术的不足。

不同基准图像数据集和模型上的实验结果表明，所提出的稀疏净化防御方案在加性范数可用性投毒攻击场景下，相较于现有10种投毒防御方案，平均测试准确率提高了2.75%-65.13%；在乘性卷积可用性投毒攻击场景下，所提出的随机乘性转换防御方案在不同数据集上的测试准确率均表现出显著优势，在不同数据集上提高25.37%-49.17%。

面向容器环境的隔离机制协同方法研究

姓　　名：徐　祯
研究方向：Linux内核安全、云安全
导　　师：袁　斌
指导老师：李　志
E-mail：1369858268@qq.com
QQ：1369858268
联系电话：15086729272
毕业去向：印第安纳大学攻读博士

在无服务器计算等新兴范式的驱动下，跨命名空间资源共享需求导致容器隔离边界扩展，而控制组的生命周期管理未能同步演进，从而产生了命名空间与控制组群异步化（Namespace-Cgroup Desynchronization，简称NCD）这一新型安全风险。这种异步化使得攻击者可利用残留资源恶意占用主机资源，引发拒绝服务攻击乃至系统崩溃等严重后果。

为评估风险影响范围，研发了一套自动化漏洞检测框架，系统性验证了Kubernetes、Docker、Podman等主流容器平台中存在的由NCD风险引入的漏洞。相关成果已获得厂商确认，累计收录4个CVE及3个CNVD漏洞编号，证实了漏洞的广泛性和严重性。此外，对真实场景下的应用进行了调研，结果表明128,935个仓库中，占比15.64%的仓库使用到了命名空间的共享配置，揭示了共享命名空间配置使用的普遍性。

针对该安全隐患，提出了一个内核层面的解决方案——命名空间关联控制组（Cgroup Associating with Namespaces，简称CANs）。该方案使命名空间与控制组群可相互感知，实现了命名空间资源管理与控制组限制的动态耦合。通过构建气球控制组架构，有效地弥合了两类机制的生命周期差异。实验评估表明，CANs方案在Linux内核、容器运行时及典型的容器应用负载中仅引入可忽略的性能损耗，以近乎零开销的代价实现了用户无感知的安全加固。

个人信息可验证删除研究

姓　　名：易迎澳
研究方向：数据安全
导　　师：徐　鹏
指导老师：徐　鹏
E-mail：641122748@qq.com
QQ：641122748
联系电话：17373346170
毕业去向：华为技术有限公司

随着互联网产业的迅猛发展，个人信息从最初的私有数据逐渐转变为具有经济价值的数字资产，随之出现的一个现象就是数据的所有者与控制者分离。因此，相关法律与时俱进，

要求保护用户的数据安全。其中，针对用户数据的“删除权”（又称“被遗忘权”），现有法律法规明确要求，若收到用户删除请求或者数据存储期限已过，就必须删除所有用户数据副本或将数据处理为不可使用的形式。为了保护用户数据的“删除权”，当前研究人员提出了许多方案以实现对于用户数据的确定性删除。然而，现存的方案大多关注于实现确定性删除本身，却忽视了对于删除结果的验证问题，而少数实现了验证功能的方案也面临恶意服务器临时计算删除证明的风险。

针对现有方案的局限性，设计了一种个人信息可验证确定性删除方案。删除方案借助特别构建的可验证延迟函数生成伪随机数据，并使用这些数据对待删除用户数据进行覆写。在后续验证阶段，用户要求服务器在一定时限内返回指定覆写数据片段，并验证片段是否由可验证延迟函数正确生成。构建的可验证延迟函数具备计算不可并行性，杜绝了恶意服务器临时生成覆写数据的可能，同时函数的验证算法相对高效，即使是普通用户也能轻松验证结果。

在系统测试环节，实验构造了一定规模的仿真数据集，涉及6000万个不同的用户，涵盖20个不同的应用场景，包括结构化数据与不同类型的文件数据，同时涉及明文与密文形式的用户数据。实验表明，用户能够使得服务器必须经过一定的延迟时长才能生成覆写数据，通过合理地设置删除参数，用户能够确保最小延迟时长的取值在1.2 秒至482.2秒之间浮动。这一计算延迟时长使得用户能够区分按时删除的诚实服务器与临时计算覆写数据的恶意服务器。与此同时，相同环境下的验证流程最长不会超过0.6秒，即使是用户所持有的普通机器也能快速完成验证流程。实验结果表明，系统能够阻止临时生成覆写数据的恶意服务器欺骗用户，同时支持多种模态的用户数据可验证删除，具备可行性与实用性。

核电站分布式控制系统网络安全加固研究

姓　　名：余啸海
研究方向：网络安全、零信任
导　　师：徐　鹏
指导老师：徐　鹏
E-mail：374365920@qq.com
QQ：374365920
联系电话：15549672019
毕业去向：中国联合网络通信有限公司湖北省分公司

核电站分布式控制系统（Distributed Control System，DCS）在早期设计中对网络安全的防护设计不足，如今面临着越发严峻的网络安全威胁。

为应对核电站DCS当前面临的网络安全挑战，调研并分析了真实核电站DCS的网络特性、网络安全问题和网络安全需求，基于零信任的安全思想，通过融合软件定义边界和软件定义网络的核心技术，设计了一种核电站DCS网络安全加固方案。针对认证机制脆弱问题，方案改进了通用的单包授权方法，显著提升了身份认证的可靠性。针对权限管理不当和访问控制缺陷问题，设计了多指标结合的信任评估方法，对异常访问行为实施快速精准阻断。针对网络稳定性和生存能力的严格要求，设计了集中化的网络监管功能，实现了网络故障识别和自动快速恢复。

根据提出的核电站DCS网络安全加固方案，进一步设计并实现了核电站DCS网络安全加固系统，基于现有的通用开源组件，在Linux环境中设计了一款基于Python语言的控制台程序，并搭建虚拟测试平台验证方案的正确性和稳定性。实验测试证明核电站DCS网络安全加固方案能够提供稳定精确的身份认证和动态授权功能，以及高效可靠的网络管理和故障恢复功能。在性能测试中显示了方案在测试环境中

的延迟保持在毫秒级，为未来的实际应用的可行性提供了有力的支持。

面向联邦学习公平性评估的贡献估计技术研究

姓　　名：张泷玲
研究方向：联邦学习
导　　师：胡胜山
指导老师：胡胜山
E-mail：1590914475@qq.com
QQ：1590914475
联系电话：13549304343
毕业去向：京东集团股份有限公司

随着人工智能发展，联邦学习成为处理多源数据隐私问题的关键技术，但实际应用中存在协作与性能不公平性。针对该问题，本文提出两种全新贡献估计方案。在协作公平性场景，文章提出基于梯度更新长期贡献评估与自适应集成的公平性方案。该方案通过多轮训练累积客户端梯度更新差异，利用滑动窗口平均与二次归一化策略，突破同质数据和强隐私保护假设限制，在保障协作公平性的同时提升整体训练性能。在性能公平性场景，提出基于梯度更新长期贡献评估与动态偏差矫正的公平性方案，通过累计去除客户端数据前后的损失差异构建长期贡献值，结合动态偏差矫正模块，在数据异质条件下实现各客户端性能均衡，兼顾性能公平性与整体训练性能提升。实验在不同基准数据集和模型架构上展开。结果显示，协作公平性方案在分类任务中平均准确率提升0.53%-2.40%，基尼系数降低0.52%-4.66%；分割任务中平均DICE系数提高1.47%-5.24%，基尼系数降低1.15%-5.21%。性能公平性方案在分类任务平均准确率提升1.09%-2.96%，准确性差值降低1.28%-5.39%；分割任务平均DICE系数提升1.47%-5.24%，DICE系数差值降低1.39%-8.34%。两种方案有效增强联邦学习公平性，性能优于现有技术，为联邦学习应用提供有力支撑。

面向联邦学习的细粒度研究

姓　　名：张梦颖
研究方向：数据删除、联邦学习、遗忘学习
导　　师：徐　鹏
指导老师：徐　鹏
E-mail：2865120877@qq.com
QQ：2865120877
联系电话：18916419251
毕业去向：皇家墨尔本理工大学攻读博士

随着《通用数据保护条例》等全球数据保护法规的实施，企业在客户请求下必须删除个人数据，以满足日益严格的隐私保护要求。现有的联邦学习下的遗忘方法通常只能删除整个客户端的数据，而无法实现更精细的细粒度遗忘，且大部分方法需要大量未发起遗忘请求的客户端的参与。在实际应用中，任何用户均可随时发起遗忘请求，要求无关客户端为每个请求分配大量计算资源会带来巨大的开销。为了解决这一问题，提出了一种新型联邦遗忘方法，旨在实现细粒度数据遗忘，同时在遗忘期间仅要求相关客户端参与。该方法受影响力函数启发，扩展其思想到联邦学习场景中，设计面向联邦学习的影响计算方法，以估计任意细粒度待遗忘数据对全局模型参数的影响，并通过从全局模型参数中去除目标影响来实现遗忘。在计算样本级的影响时，通过计算块对角经验费歇耳信息矩阵来近似完整的海森矩阵，从而显著降低计算、存储、通信开销。此外，当多个客户端需要遗忘同一类数据时，通过斜对角线近似法来计算类级影响，该方法允许持有部分目标类别训练数据的一个客户端代表全部客户端执行类级遗忘操作，最大限度减少了类级遗忘需要的客户端数量。

Ext4文件系统的不可恢复性删除研究

姓　　名：赵　旭
研究方向：数据安全、不可恢复性删除
导　　师：徐　鹏
指导老师：徐　鹏

E-mail：2505693205@qq.com
QQ：2505693205
联系电话：15799032500
毕业去向：工业和信息化部电子五所

《中华人民共和国个人信息保护法》等相关法律法规明确要求个人信息处理者在个人信息不再使用时要及时执行删除操作。然而，现有文件系统在执行数据删除过程中却往往无法彻底清除数据，增加了数据泄露的风险。

为应对这一挑战，在主流的Ext4文件系统上开展了深入研究，提出了一种不可恢复性删除方法。该方法通过在现有的删除流程中集成安全覆写功能，对Ext4文件系统的底层操作函数，如ext4_mb_clear_bb()和ext4_free_inode()进行重写和优化，不仅实现了文件本身的彻底删除，还清除了在文件处理过程中产生的所有关联内容，包括文件的历史副本、隐藏副本、文件元数据和文件关联日志中的相关信息。

在磁盘和闪存存储介质上开展了广泛实验，验证了不可恢复性删除方法的有效性。实验结果表明，在执行删除操作时，优化后的内核能够正确地覆写文件及其关联内容，确保这些内容不可恢复。该方法对整体输入输出（I/O）性能的影响并不明显，优化后内核的I/O 性能约为原生内核I/O性能的98%。此外，不可恢复性删除方法的实施大大缩小了潜在的攻击窗口，有效降低了数据泄露的风险。这些结果表明，提出的不可恢复性删除方法为Ext4文件系统提供了一种安全可靠的数据删除解决方案，在保障用户隐私和数据安全方面具有重要的应用价值。

基于代码多模态表示对齐的漏洞检测异构模型构建技术

姓　　名：朱嘉豪
研究方向：软件安全、程序漏洞检测
导　　师：文　明
指导老师：文　明

E-mail：2629904428@qq.com
QQ：2629904428
联系电话：18868578336
毕业去向：蚂蚁科技集团有限股份公司

当前漏洞检测模型可能通过虚假特征（如代码风格差异）实现高预测准确率，而非真正理解漏洞语义。为突破这一局限，提出以“成对预测”（即模型同时正确分类漏洞样本及其修复版本）为核心评估标准，旨在构建更严格的漏洞检测模型评估框架，并探索提升模型语义理解能力的新方法。

为实现上述目标，首先基于主流的漏洞检测数据集构建了一个包含漏洞 – 修复样本严格配对的评估数据集 VulFixed，并定义了全新的漏洞检测评估指标：成对预测率（PP）。通过系统评估主流的文本结构模型与图结构模型，研究发现：现有模型在成对预测任务中表现显著不足（PP 均低于 25%），且文本模型注意力分散于漏洞语义无关的代码 tokens，图模型则缺乏对全局语义的把握。为此，提出多模态融合的漏洞检测异构模型 VulGraphPTM，其通过融合代码文本模态与代码图模态实现特征互补，从而强化漏洞语义建模。

实验表明，VulGraphPTM 在成对预测任务中 PP 值达到了 36.86%，较图模型（17.12%）和文本模型（21.37%）分别提升 115.30% 与 72.48%。在常规漏洞检测任务中，漏洞检测能力最高超越基线模型 17.06%，跨数据集检测能力最高提升 74.52%。上述结果验证了多模态异构模型对抑制虚假特征的有效性。

全场景AI推理中的WebAssembly技术研究

（彭俊辉 https://blog.csdn.net/hellojunbo/article/details/148686152?spm=1001.2014.3001.5502）

一、前言：边缘智能时代的技术变革需求

随着全球数字化转型进入深水区，算力资源分布正经历从“中心化云平台”向“云–边–端三级架构”的历史性转变。据IDC 2024年预测，边缘设备数量将于2027年突破280亿台，同时欧盟《人工智能法案》、中国《数据安全法》等法规对隐私保护提出严苛要求，这使得本地化AI推理成为产业刚需。传统云计算模式面临三重困境：首先，敏感数据上传云端导致合规风险，医疗、金融等领域的数据跨境限制使云端模型训练难以实施；其次，网络带宽与时延制约实时性，工业质检等场景要求毫秒级响应；最后，异构硬件生态碎片化，ARM、x86、RISC-V架构差异迫使开发者维护多套代码库。

早期浏览器内推理方案如TensorFlow.js与ONNX.js尝试通过JavaScript重写框架实现跨平台，但受限于语言本质缺陷：解释执行机制使ResNet-50推理时延高达850毫秒（Google V8团队2023年测试），垃圾回收引发内存抖动导致20-60毫秒卡顿，动态类型漏洞占据浏览器攻击事件的43%（Mozilla 2024安全报告）。更严峻的是，JS缺乏硬件隔离能力，模型参数在内存中以明文存在，不符合HIPAA、GDPR等隐私法规要求。

WebAssembly（WASM）作为W3C主导的开放标准，以三重革新突破困局：在性能维度，其二进制指令集与堆栈式虚拟机架构，使代码执行效率较JS提升2-10倍，Unity引擎实测显示3D渲染帧率从45fps跃升至63fps；在安全维度，线性内存隔离模型配合能力限制机制（Capability-based），构建端到端可信执行环境，阻断99%的内存攻击向量；在生态维度，.wasm格式实现真正的“一次编译，处处运行”，覆盖从浏览器到边缘网关的异构硬件，运行时仅4MB体积（Docker容器通常>100MB）。这些特性使WASM成为连接算力网络（CFN）中云、边、端节点的新型算力载体，为全场景AI推理提供基础设施级支撑。

二、WASM与AI推理融合的三级技术路径演进

（一）LLM APP代理模式：效率优先的折衷方案

该模式代表为LlamaEdge，核心思想是通过WASI-NN接口将外部推理引擎（如PyTorch、ONNX Runtime）与WASM胶水代码绑定。其技术实现分为四步：JavaScript输入数据序列化为FlatBuffer格式；WASM调用wasi_nn_load()接口传输数据；宿主系统执行实际计算（支持GPU/NPU加速）；结果反序列化返回浏览器。该架构最大优势在于硬件兼容性，由于实际计算在宿主系统上进行，宿主系统上所有的硬件加速都能使用，其本质缺陷在于安全性妥协——模型权重明文存储于宿主内存，序列化过程消耗35%推理时间，且移植性受本地驱动制约。此类方案适用于需调用专用硬件的视觉检测场景，但无法满足医疗等高敏

感需求。

（二）算子编译模式：平衡性能与安全的实践路径

以WebLLM为代表的方案创新性地结合TVM编译栈与WASM内存模型。核心技术突破包含三层：首先，TVM编译器将Relay IR降阶为WASM字节码，对GEMM算子采用分块策略（Blocking Factor=64），卷积层应用Winograd算法减少40%乘加操作；其次，通过共享内存（SharedArrayBuffer）实现张量数据零拷贝传递，华为昇腾团队实测内存复制开销降低90%；最后，设计分层线程调度模型——浏览器主线程负责任务分派与结果聚合，8个Worker线程基于信号量机制并行执行算子。在RK3588边缘网关的实测中，ResNet-50推理时延压缩至210毫秒（较JS方案提速3.16倍），内存碎片减少85%。然而该架构仍存在安全短板：计算图构建暴露于JavaScript环境，存在控制流劫持风险；动态内存分配可能引发侧信道攻击。此类方案适用于智慧城市等对实时性要求较高但隐私敏感度中等的场景。

（三）全栈沙箱模式：隐私优先的终极架构

此路径致力于构建端到端安全推理闭环，技术突破体现在三大层面：计算流程重构方面，从提示词Tokenization到结果Detokenization的完整链路在沙箱内完成，采用WASM-GC管理计算图内存生命周期；硬件加速集成方面，通过WASI-WebGPU接口将MatMul等算子卸载至GPU，利用WGSL着色器语言重构计算内核，微软测试显示矩阵乘法速度提升8倍；安全增强机制方面，模型参数分页加密加载（每页4MB），解密过程仅在Intel SGX Enclave执行，输出结果注入Laplace噪声（ε =0.1）满足差分隐私。代价是性能损失：Llama-7B推理时延达6800毫秒，较原生方案慢8倍。这表明全栈架构适用于医疗诊断、金融风控等最高隐私等级场景，但需持续优化性能瓶颈。

三、关键技术突破的系统性突破

（一）性能优化机制的深度创新

面对WASM与原生代码的效率鸿沟，研究者从指令集、内存管理、异构计算三维度突破。指令集层面，128位SIMD向量指令实现单周期处理4个FP32运算，结合Winograd算法优化，卷积层计算速度提升3.8倍；原子指令（Atomic.wait/notify）支持多线程同步，Attention层任务拆解至8线程后时延降低58%。内存管理领域，线性内存（Linear Memory）提供连续地址空间支持直接偏移访问，消除传统JS引擎的哈希表查询开销；内存复用池（Memory Pool）技术预分配模型参数存储区，Llama-7B推理中动态分配次数减少92%。异构计算融合成为最新趋势，WASI-WebGPU 1.0草案允许直接调用浏览器GPU API，通过WGSL实现计算内核跨平台部署，异步流水线设计使计算与数据传输重叠执行，端到端时延降低34%。

（二）安全架构的范式重构

传统浏览器安全模型依赖沙箱隔离，而WASM创新构建五级纵深防御：编译阶段LLVM前端插入边界检查指令，阻断缓冲区溢出攻击；加载过程采用AES-256-GCM算法动态解密模型参数，密钥通过TPM 2.0芯片保护；执行环境部署Intel SGX Enclave隔离敏感计算，确保即使宿主系统被攻破也无法窃取模型；输出层注入差分噪声，使攻击者无法通过输出反推原始数据；审计系统实现指令级日志跟踪，实时匹配MITRE ATT&CK攻击模式库。该架构已通过ISO/SAE 21434车规认证，工业控制系统漏洞数量减少76%，医疗数据泄露风险下降82%。

四、未来挑战与发展路径

（一）性能瓶颈的攻坚方向

当前最紧迫的挑战是缩小与原生代码的性能差距，需在三大方向突破：SIMD 256位指令集扩展（支持单周期8个FP32运算）预计今年纳入标准，可将卷积计算速度再提升2.1倍；即时编译优化技术（Profile-Guided Optimization）通过运行时热点分析动态生成最优机器码；算子量化加速利用INT8精度替代FP32，在保持90%精度前提下降低75%计算负载。

（二）安全机制的持续进化

面对量子计算威胁，后量子密码（PQC）迁移成为重点，CRYSTALS-Kyber密钥交换方案已在测试中实现每秒千次密钥交换；侧信道防护需集成Cache访问混淆技术，通过随机化内存访问模式抵御时序攻击；零信任模型要求建立TPM远程认证机制，确保只有可信环境可加载敏感模型。

（三）产业生态的协同建设

标准化进程加速推进：wasi-nn 2.0接口将支持Rotary Position Embedding等大模型特有算子；Kubernetes通过Krustlet项目原生支持WASM工作负载调度；ISO/ECC联合工作组正在制定边缘AI能效比评价体系（FPS/W、安全覆盖率等12项指标）。产业协作将成为突破“性能–安全–移植性”不可能三角的关键力量。

五、总结

理论上，WASM技术通过架构革新（三级融合路径）、机制创新（五级安全防御）、生态协同（标准化与工具链）三维合力，将重塑全场景AI推理基础设施。但就现状来看，WASM技术虽可以为AI推理带来极佳的移植性和安全沙箱防护，但代价是慢于原生数倍的推理速度（主要源于WASM缺少对应的硬件加速），这对于大型模型推理几乎是不可接受的，这有赖于WASM社区的共同努力。但是在个人PC、智能手机及其他边缘设备上做一些中小型模型的推理如Llama-7B则完全可行，性能损失可以接受，WASM这个载体完美的解决了移植和隐私保护问题，实现了本地AI推理的极简部署。

参考文献

[1] W3C. WebAssembly Core Specification v2.0. 2023

[2] Google V8 Team. JavaScript vs WASM Performance Benchmark. 2023

[3] Mozilla Security Lab. WebAssembly Security Audit Report. 2024

[4] WebLLM Team. TVM-based Compilation for Browser Inference. MLSys 2023

[5] Intel. SGX-Protected WASM in Edge Computing. IEEE IoT Journal Vol.19, 2024

精 彩 言 论（转载）

学海常存启智的晨光与求索的星火。唯有燃起心中的热望，方能守得纯粹的初心，不负青春的。（刘欣鹏 https://b23.tv/jsrBP1K）

把握住今天，胜过两个明天。（马啸埃 https://zhuanlan.zhihu.com/p/349997385）

不完美是一种美，疯狂是一种天分，不靠谱总好过无聊至极。（Imperfection is beauty, madness is genius and it's better to be absolutely ridiculous than absolutely boring.）——玛丽莲·梦露（Marilyn Monroe）（潘锶逵 https://language.chinadaily.com.cn/easyEnglish/2014-07/31/content_18223493.htm）

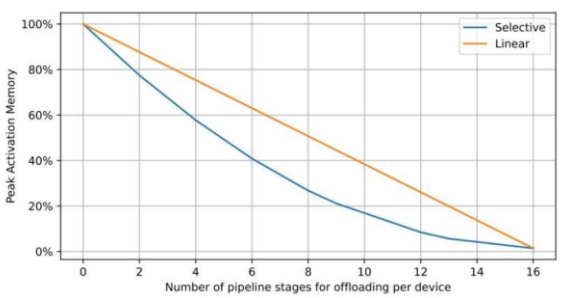
我的成功要归功于好运气、努力工作以及朋友和导师的支持和建议。但最重要的是，这取决于我在失败后继续努力。（‘My success was due to good luck, hard work, and support and advice from friends and mentors. But most importantly, it depended on me to keep trying after I had failed.’）——马克·华纳（Mark Warner）（郑 直 https://au.indeed.com/career-advice/career-development/hard-worker-quotes）

PipeOffload：通过内存优化提升流水线并行的可扩展性

（王世杰 https://weibo.com/8004429662/PvUonyNN7）

一、引言

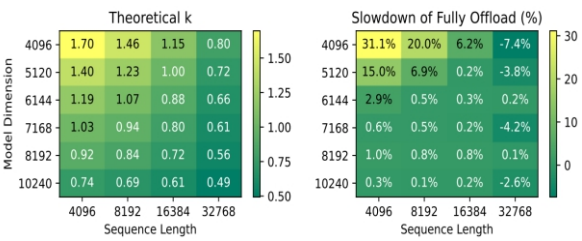
流水线并行（PP）在训练大型语言模型（LLMs）中应用广泛，然而随着流水线并行度的提高，训练过程中的微批次数量增加，过高的激活内存消耗常常限制了其可扩展性。SeaAI Lab提出的PipeOffload聚焦于利用流水线并行中尚未充分探索的内存卸载策略来应对这一挑战。通过实证研究发现，在大多数标准配置下，至少一半甚至全部的激活都可以在开销可忽略的情况下进行卸载。当无法完全卸载时，其引入了一种新颖的选择性卸载策略，该策略能以优于线性的方式降低峰值激活内存。此外，PipeOffload还将内存卸载与其他技术相结合，综合考虑整体吞吐量和内存限制。实验证明，每设备的激活内存可以随着阶段总数的增加而有效减少，使流水线并行成为比张量并行更优的选择，甚至能在更低的内存消耗下实现高达19%的加速。



二、流水线并行的内存挑战与内存卸载的潜力

现代大型Transformer模型的参数规模朝着万亿级别扩展，模型并行对于在多个设备上分布模型参数变得至关重要。与ZeRO和张量并行相比，流水线并行（PP）具有更低的通信量和更高的计算强度。尽管流水线并行通过在设备间划分层来减少参数内存需求，但其可扩展性仍然受到激活内存的限制。增加流水线阶段的数量会减少每个设备上的层数，但需要更多的微批次来最小化流水线气泡，这种权衡使得整体激活内存需求保持不变。

在这项工作中通过将内存卸载到主机来解决流水线并行的内存限制问题。虽然内存卸载在数据并行（DP）中已被广泛采用，但其在流水线并行中的潜力在很大程度上尚未被探索。流水线并行特别适合内存卸载，因为前向传播和反向传播之间的时间间隔为卸载和重新加载激活内存创造了一个自然的窗口，而不会干扰其他计算。这与激活重计算形成了鲜明对比，因为后者会引入显著的重新计算开销。如果安排得当并与其他计算重叠，内存卸载可以说是一种“免费的午餐”。



三、核心技术

1. 选择性卸载策略

当无法完全卸载激活内存时，PipeOffload采用部分卸载策略，优先选择生命周期长（前

向与反向传播间隔久）的激活，因其对峰值内存贡献更大。通过在同一设备放置多阶段，利用不同阶段生命周期差异，可实现优于线性的内存减少。例如，对比交错1F1B，PipeOffload卸载阶段0可使峰值内存减少3/4，而前者最多仅能减少1/2。

2. 内存与吞吐量的权衡

流水线并行中，激活内存与流水线气泡存在权衡：

· 均匀重复策略内存减少效率高，但气泡较多；交错策略吞吐量更优，适合内存压力小的场景。

· 零气泡优化拆分反向传播为激活梯度（B）和权重梯度（W）计算，在不引入额外气泡的前提下，将预热阶段前向次数减少，峰值内存从(dv+d-1)降至(dv)。

· 缩短生命周期调度（GIS-H）通过调整微批次重复模式，在内存减少与气泡增加间平衡，极端情况下峰值内存可降至原始交错1F1B的50%。

3. 卸载实现技术 –

· 效率优化：对GeLU等轻计算层重计算，减少40%激活内存；利用硬件拓扑感知策略稳定PCI-E带宽；通过连续缓冲区降低主机内存开销。

· 单流调度：采用单流处理卸载与重新加载，避免多流延迟波动。基于均匀重复策略，按“先卸载后加载”顺序分配槽位，并通过设备间同步避免PCI-E冲突。

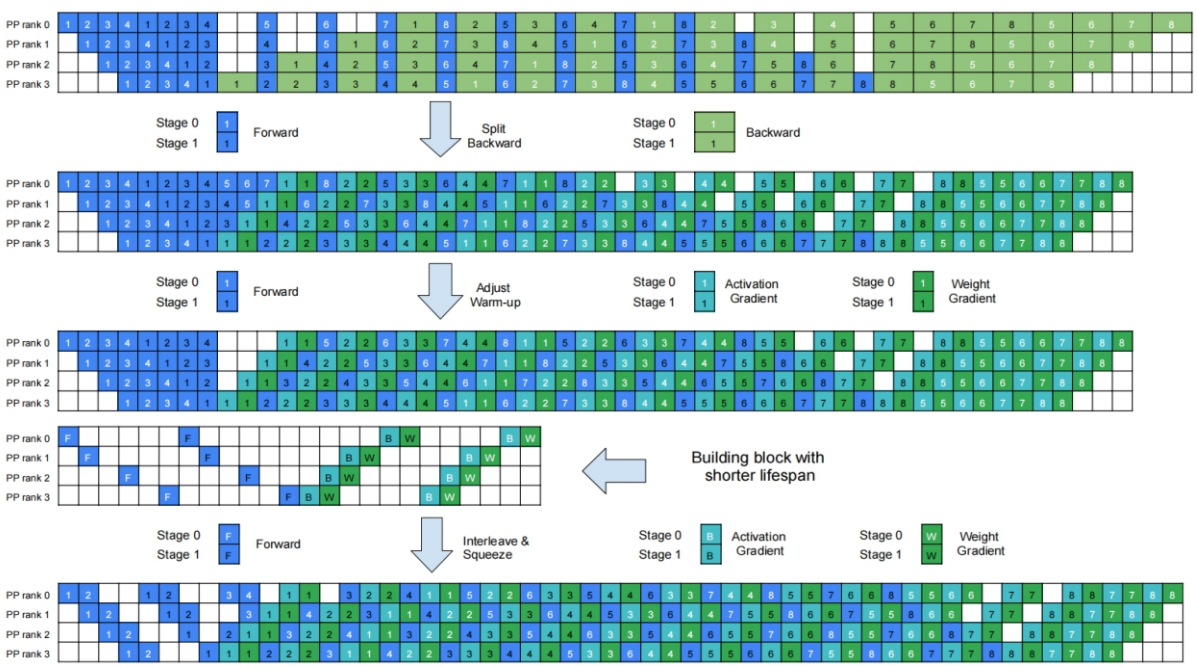
四、未来展望

总的来说，PipeOffload作为一项极具创新性的流水线并行技术，具有巨大的潜力和应用价值。尽管该技术并不能在各种训练环境下保持最佳效果，但它已经引起了人们的广泛兴趣和关注，未来的发展前景非常值得期待。希望未来能够充分利用海量数据驱动的优势，实现更加大规模的大型语言模型训练。

参考链接

· https://arxiv.org/abs/2503.01328v1

· https://zhuanlan.zhihu.com/p/1885394078486733999



首届计算机网络与系统前沿论坛顺利举办

燕 燕

2025年4月27日，首届计算机网络与系统前沿论坛在江城武汉盛大举办。本届论坛由Frontiers of Computer Science期刊编委会主办，华中科技大学计算机科学与技术学院、服务计算技术与系统教育部重点实验室、湖北省计算机学会共同承办。论坛汇集了来自全国各地高校100多位计算机网络与系统领域专家、学者，共同探讨人工智能时代网络与系统创新发展的前沿话题。

大会由华中科技大学计算机科学与技术学院院长廖小飞担任大会主席，华中科技大学教授刘海坤、中国地质大学（武汉）教授曾德泽、华中科技大学教授郑龙担任大会程序委员会主席。大会采用“主旨论坛+青年论坛+现场访谈”相结合的形式，设置了2个论坛主旨报告，14个青年专家报告和1场网络与系统领域青年学者发展Panel。

开幕式由华中科技大学计算机科学与技术学院院长廖小飞主持，Frontiers of Computer Science期刊主编周志华教授、华中科技大学副校长冯丹发表致辞，他们分别向与会者表示感谢和欢迎，Frontiers of Computer Science期刊张德发主任向大家介绍了期刊的历史和发展现状，涉及的研究领域，以及进一步提升期刊影响力的一系列举措。

华中科技大学教授刘海坤、中国地质大学（武汉）教授曾德泽主持了主旨报告环节，2位专家带来精彩分享。中国科学院计算技术研究所副所长、处理器芯片全国重点实验室主任、中国计算机学会体系结构专委会主任陈云霁教授以“从人工智能到处理器芯片”为主题进行报告，他认为人工智能和处理器芯片都是信息产业的基石技术，都是大国博弈的焦点，二者之间相互交叉、彼此推动，着重介绍了二者交叉研究的核心思想和研究现状，并探讨了未来的发展趋势。清华大学全球创新学院院长刘云浩教授发表“万物相联万物生：具身智能与数字先行”的主题报告，指出物联网的发展经历了从“设想”到“现实”的重要转变。不断提升的智能化、更广泛的互联互通和更深刻的感知能力，使物联网推动了工业互联网的变革，引领全球迈向万物互联数字先行的具身智能时代。

现场访谈环节由南京大学田臣教授主持，山东大学成秀珍教授、中国科学院计算技术研究所陈云霁研究员、上海交通大学过敏意教授、上海交通大学管海兵教授、国防科技大学苏金树教授、香港科技大学张黔教授围绕网络与系统领域青年学者成长过程中面临的诸多问题进行了深入讨论，各位嘉宾针对本领域青年学者发展面临的困难深刻地剖析了原因，并从职业生涯学术方向规划、科研评价体系优化、平台和团队建设、产学研合作等多个维度给出了举措和建议，现场气氛活跃，互动频繁。嘉宾们一致认为，青年学者的成长需要个人努力、团队支持和制度保障的协同，并呼吁学界共同营造更健康的科研生态。

14位青年学者就各自领域研究也做了精彩分享。中国科学技术大学副教授李诚作题为“AI+BigData驱动的跨尺度系统创新”的报告，西北工业大学孙卓教授作题为 “感通策协同的多具身智能体网络初探” 的报告，南京大学副教授郑嘉琦作题为“区域IP任播:全球部署、性能和潜力”的报告，浙江工业大学教授姚信威作题为“智能物联网中感算存一体化”的报告，重庆大学教授李秀华作题为“云边协同环境下高效模型推理优化”的报告，蚂蚁技术研究院高级研究员张明喆作题为“对GPU价值的再思考”的报告，中国人民大学谢旻晖博士作题为“面向推荐大模型的参数存储系统研究”的报告，中国科学技术大学特任副研究员田晗作题为“基于深度强化学习的智能网络控制”的报告，南开大学副教授张圣林作题为“大模型时代的网络智能运维”的报告，华中科技大学副教授文明作题为“大模型赋能的软件安全测试与分析”的报告，浙江大学研究员汪睿作题为“时序感知的大规模动态图智能计算系统”的报告，山东大学教授吴思作题为“数据中心系统可靠高效编码研究”的报告，上海交通大学助理教授刘方鑫作题为“面向人工智能模型的自适应精度加速研究”的报告，华中科技大学副教授赵进作题为“高性能软硬协同图计算技术研究”的报告。

本次大会以“智驱质变：人工智能时代的网络与系统创新发展”为主题，主旨论坛聚焦前沿趋势，专家学者汇聚一堂探讨青年学者发展举措，青年论坛探讨分享各自细分领域，探索交流迸发思想的火花。

燕　燕
负责事务：实验室宣传、科研项目管理等
Email：yany@hust.edu.cn

Systematic CXL Memory Characterization and Performance Analysis at Scale

刘万奇 推荐

“Systematic CXL Memory Characterization and Performance Analysis at Scale”这篇论文是收录在计算机体系结构领域的国际顶级会议ASPLOS 2025的一篇文章，该会议于2025年3月30日至4月3日在荷兰鹿特丹(Rotterdam, Netherlands)举行。

在内存密集型应用需求不断增长的推动下，对内存容量的需求正在迅速上升。新兴的互连技术，如Compute Express Link（CXL），有望在服务器/机架级别实现内存扩展。各种内存供应商已经推出了CXL内存扩展器，有助于访问比以前更大数量的DRAM。但与传统的套接字本地DRAM相比，CXL内存扩展引入了更高的延迟。

考虑到PCIe上CXL协议栈的复杂性增加以及第三方存储控制器优化带来的可变性，仅仅将CXL内存视为较慢的DRAM是不够的。虽然之前的研究为CXL性能对一些HPC工作负载的影响提供了宝贵的见解，但它们主要关注粗粒度分析，忽略了几个关键方面：（1）CXL性能稳定性（即尾部延迟）；（2）CPU对各种工作负载中长时间CXL延迟的容忍度，以及CXL的架构影响；（3）缺乏系统的方法来剖析CXL下的工作负载性能和CPU效率低下。

本文提出了Melody，一个用于深入表征CXL性能的综合框架。Melody首次公开、量化并分析了CXL尾延迟，通过微基准测试与真实应用提供了深入洞见。Melody在4个CXL设备、7种内存延迟配置（140–410ns）和5个处理器平台上评估了265个工作负载，系统性地分析了CPU和工作负载对CXL延迟的容忍性。Melody引入了一种新颖的方法来诊断CXL性能问题，仅使用9个CPU性能计数器即可进行轻量、准确、细粒度的工作负载行为分析。

首先论文对延迟进行分析。发现并非所有CXL设备的延迟都是相同的；每个设备不仅具有独特的平均延迟和带宽特性，还具有独特的延迟稳定性以及延迟–带宽关系特性。更重要的是，与本地/NUMA内存相比，CXL设备表现出不稳定且更高的尾部延迟。平均延迟和带宽并不能完全反映CXL设备的性能影响。CXL MC无法在轻负载下缓解尾部延迟，而高内存利用率将进一步加剧CXL尾部延迟，而本地内存和NUMA则保持延迟稳定。并发读写对CXL延迟的影响不同，在混合工作负载下尾部延迟会恶化。虽然CPU硬件预取器可以改善平均内存访问延迟，但它们无法消除尾部延迟。CXL尾部延迟对应用程序性能产生负面影响。基于FPGA的CXL设备无法充分利用CXL的双向数据传输能力，导致其在读写混合工作负载下的性能特征与基于ASIC的设备存在显著差异。

然后论文对实际的工作负载进行分析。发现工作负载性能随着CXL延迟的增加而超线性恶化；更重要的是，相对降速超过了延迟增加的速度。尾部延迟更差的CXL设备在所有评估

的工作负载中经历更显著的降速，并且在CXL+NUMA配置下的尾部延迟可能导致工作负载出现惊人的高降速。从积极的一面看，许多工作负载能够容忍较长的CXL延迟（高达410ns），经历的降速小，这表明CXL可能对某些实际应用有用。在相似的带宽容量下，CXL和NUMA之间的性能差距缩小，使CXL内存成为本地/NUMA内存的可行替代方案。然而，对于延迟敏感的工作负载，延迟差距仍然是一个挑战。

最后论文提出了一种新的性能分析方法SPA，它仅使用9个CPU计数器就能以高精度（95%）和最小误差（5%）对所有工作负载的CXL性能进行建模。现有方法无法可靠地将应用性能下降准确映射到系统/架构事件，如Intel的TMA方法，借助CPU计数器/事件提供的丰富信息来分析代码效率。然而，TMA在CXL性能下降分析方面存在不足，TMA无法提供差异性分析来解释由于不同后端内存（即CXL与本地DRAM）导致的管道差异。TMA不能精准关联架构层面指标与工作负载性能下降。其指标旨在捕获特定硬件组件的性能或争用情况，而非整体工作负载行为。SPA的关键优势在于能够将CXL性能下降归因于特定的“源头”，例如CPU缓存层次结构和CXL内存，从而实现对CXL引发性能瓶颈更细致精确的分析。通过分离这些组件的贡献，SPA能全面剖析工作负载下降原因，并量化它们对整体性能退化各自的影响。

SPA采用了自下而上的分析方法，其核心思路是深入研究CXL与本地内存之间在CPU停顿周期上的差异。这种方法能够为CXL性能下降提供精确的分析结果，而单独分析本地内存或CXL环境下的性能表现则无法达到同样的效果。SPA的目标是精准定位那些导致CXL引发性能下降的具体停顿源头，从而有效弥合架构层面与工作负载层面性能表现之间的差距。SPA的实现基于9个CPU计数器，通过对这些计数器数据的分析，SPA能够深入检查涉及内存请求处理的CPU管道组件，并剖析CXL在指令执行过程中对这些组件造成的影响。

通过分析，CXL的性能下降主要因为内存子系统(ΔsMemory)的额外停顿周期，论文提出了估算CXL性能下降的公式：

$$S = \frac{\Delta c}{c} \approx \frac{\Delta s}{c} \approx \frac{\Delta s\ \text{Backend}}{c} \approx \frac{\Delta s\ \text{Memory}}{c}$$

并对公式的准确性进行了验证，通过SPA计数器估算工作负载的性能下降，将估算结果与应用层面指标（如执行时间和吞吐量）对比，发现两者之间一致性非常高。论文还对基于SPA的性能下降进行剖析，将性能下降分为Sstore + SL1 + SL2 + SL3 + SDRAM，每一部分上的性能下降。论文发现缓存性能下降的根源是预期效率下降，CXL延迟增加，预取数据抵达变慢，预取覆盖范围缩小。即使数据在缓存中，因预取抵达慢，加载操作仍被延迟。

刘万奇
2023级硕士研究生
研究方向：CXL分离式内存系统
Email：2464387631@qq.com

HouseFuzz: Service-Aware Grey-Box Fuzzing for Vulnerability Detection in Linux-Based Firmware

江宗泽 推荐

随着物联网（IoT）设备的普及，基于Linux的固件已成为关键基础设施，为全球超过43%的IoT设备提供动力。然而，这些设备普遍存在的安全漏洞使其极易受到网络攻击，可能引发远程代码执行等严重后果，对供应商和用户构成巨大威胁。为了应对这一挑战，论文“HouseFuzz: Service-Aware Grey-Box Fuzzing for Vulnerability Detection in Linux-Based Firmware”应运而生，并被网络与系统安全领域的国际顶级会议IEEE S&P 2025收录。该研究指出，现有的固件模糊测试技术因未能充分理解固件服务的内在特性而效果不彰。为此，作者提出了一种新颖的、服务感知的灰盒模糊测试工具HouseFuzz，通过三大核心技术系统性地解决了固件测试的瓶颈问题：整体性服务识别，多进程协同模糊测试，以及服务协议指导的测试用例生成。

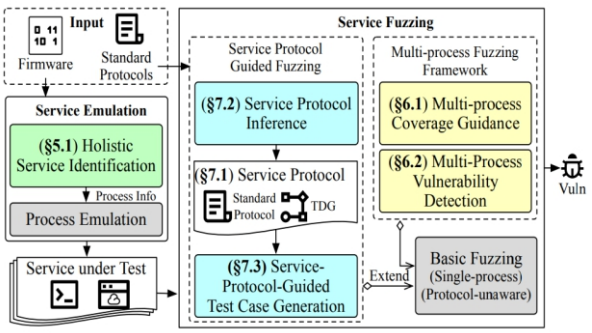


图1 HouseFuzz架构图

灰盒模糊测试是检测固件漏洞的关键技术，但现有方法在模拟、反馈和测试用例生成三个核心环节存在严重缺陷，这些缺陷主要源于对固件服务特性的认知不足，导致测试范围受限、效率低下。比如服务模拟的局限性：Linux固件服务通常由多个进程协同工作，包括直接面向网络的进程 和通过进程间通信 (IPC) 触发的守护进程。现有方法难以完整地模拟这种多进程环境。又比如基于白名单的方法：这类方法仅当进程名出现在预设的列表中时才会进行模拟。这导致大量名称特殊但功能关键的进程被直接忽略。还有基于全系统模拟的方法：这类方法尝试模拟整个操作系统，理论上更完整，但常因硬件依赖、NVRAM配置错误等问题而过早中断。此外还有测试用例生成的局限性，固件服务常在标准协议（如HTTP）之上实现大量定制化协议，引入了丰富的语义约束，例如特定的键值对、严格的数据格式或参数间的依赖关系。现有模糊测试工具对此缺乏感知，仅对标准协议模板进行随机、盲目的字节变异。这些生成的测试用例绝大多数因为无法通过服务前几层的语义校验而被直接拒绝，无法触及深层的、可能存在漏洞的业务逻辑，造成了大量的无效测试。

为解决上述挑战，HouseFuzz设计了如图1所示的创新流程，包含三大协同工作的核心模块：整体性服务识别 (Holistic Service Identification)，多进程模糊测试框架 (Multi-Process Fuzzing

Framework) 和服务协议指导的模糊测试 (Service-Protocol-Guided Fuzzing)。

为实现全面的服务模拟，HouseFuzz创造性地从分析和修复系统初始化过程入手。具体而言，HouseFuzz利用QEMU用户模式执行固件的INIT进程，并同步追踪其创建的所有子进程，以监控完整的系统启动序列。在此过程中，当检测到因“模拟路障”导致的进程崩溃或挂起时，系统会自动分析该进程执行轨迹的末端以精确定位异常代码点，并应用补丁。例如，用NOP指令替换一个函数调用以绕过该异常。值得注意的是，该模块还集成了一套鲁棒性增强机制以应对潜在的误报。如果一个补丁导致后续模拟中识别出的服务通道数量减少，系统会判定该补丁具有破坏性并将其自动撤销。通过这种细致的分析与修复循环，HouseFuzz 能够确保系统初始化流程的完整性，从而成功启动并识别出比传统方法多得多的可测试服务。

Algorithm 1 INIT Emulation

Input: Img - Firmware image,
N - Max attempt times
TO - Timeout of each emulation run
Output: Img - Patched firmware image,
C - Identified network or IPC channels

$P \leftarrow IdentifyInitProgram(Img)$
$C \leftarrow Nil$
repeat
  $T \leftarrow TraceEmulation(Img, P, TO)$
  $PrevC \leftarrow C; C \leftarrow IdentifyChannels(T)$
  $Ex \leftarrow IdentifyException(T)$
  if $Ex = Nil$ **then**
    return Img, C
  end if
  if $PrevC.size > C.size$ **then**
    return $PrevImg, PrevC$
  end if
  $PrevImg \leftarrow Img; Img \leftarrow PatchException(Img, Ex)$
  $N \leftarrow N - 1$
until $N = 0$

图2 固件INIT进程模拟算法

其次，为克服传统单进程分析模型的局限性，HouseFuzz设计并实现了一个创新的多进程模糊测试框架，其核心思想是将构成一个完整服务的所有相关进程——包括网络前端进程、守护进程及工具进程——作为一个整体进行并发监控与分析。其关键设计主要体现在三个方面：第一，在测试引导方面，该框架通过合并所有受监控进程的代码覆盖率信息，构建了一个全局性的覆盖率视图，从而为模糊测试引擎提供更全面的反馈，使其能够探索以往被忽略的多进程交互逻辑，更深入地覆盖服务状态空间。第二，为确保多进程环境下覆盖率收集的准确性与稳定性，该框架引入了测试完成事件（TCE）检测机制，它为不同类型的进程定义了不同的完成标志，例如工具进程的终止、网络前端进程网络资源的释放，或守护进程重新进入I/O监听的系统调用。第三，该框架集成了多进程漏洞预言机 (Multi-Process Vulnerability Oracle)，能够精确识别并报告在任意服务进程中触发的内存损坏及命令注入等高危漏洞。

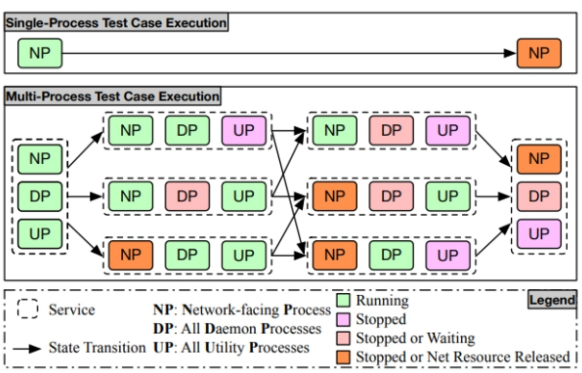


图3 多进程模糊测试模块概览

为解决传统模糊测试在处理定制化服务协议时的局限性，HouseFuzz设计了服务协议指导的模糊测试模块。该模块的核心是引入了令牌依赖图 (Token Dependency Graph, TDG) 这一概念，旨在对定制化协议中蕴含的复杂语义约束进行形式化建模。在该图中，节点 (Node) 代表

协议中的关键字符串（即令牌），并被赋予如路径 (Path)、键 (Key)、值 (Value) 等语义类型；而边 (Edge) 则表示令牌之间存在的控制流或数据流依赖关系。为确保TDG的全面性与准确性，HouseFuzz采用了一种离线与在线相结合的互补策略来自动构建该图谱：离线推断：在模糊测试开始前，系统通过对固件二进制文件进行静态的控制流与数据流分析，预先提取潜在的令牌及其依赖关系。在线推断：在模糊测试过程中，系统利用动态插桩技术（例如，监控strcmp等字符串比较函数），实时观测服务处理输入数据的方式，从中动态地推断出新的令牌与依赖关系。

在测试用例生成环节，HouseFuzz将TDG与描述标准协议的上下文无关文法 (CFG) 相结合，以TDG指导变异过程。这种方法确保了生成的测试用例不仅在语法上符合标准（由CFG保证），更在语义上满足了定制化协议的特定约束。最终，这种策略能够生成高质量的输入，有效绕过服务前端的协议校验逻辑，从而深入探索以往难以触及的代码路径，显著提升了发现深层漏洞的概率。

该论文在包含60个真实固件的数据集上对HouseFuzz进行了全面、多维度的评估，并与SoTA工具 GREENHOUSE 进行了深入对比。在识别可测试服务方面，HouseFuzz展现出压倒性优势。它总共识别了 311 个网络服务，而 FirmAE 和 GREENHOUSE 分别只识别了 128 和 44 个。

Protocol	# Service	GREENHOUSE * Recall	FirmAE* Recall	HOUSEFUZZ* Recall
HTTP	98	44.9%	43.9%	95.9%
Telnet	40	0	30.0%	100.0%
UPNP	28	0	39.3%	100.0%
NetBIOS	24	0	33.3%	91.7%
DHCP	23	0	34.8%	91.3%
(DNS)	22	0	18.2%	90.9%
NCI	18	0	11.1%	100.0%
mDNS	13	0	61.5%	76.9%
LLMNR	11	0	63.6%	63.6%
SSH	7	0	28.6%	100.0%
RIP	4	0	0	100.0%
AFP	4	0	0	100.0%
TFTP	3	0	33.3%	100.0%
STP	1	0	0	100.0%
Unknown	142	0	64.1%	47.2%
Summary	438	10.0%	45.0%	79.0%

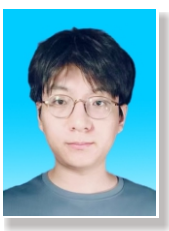
图4 识别可测试服务实验结果

TABLE 3: General fuzzing performance comparison (**RQ1**).

Vendor	Series	# Services	Avg. Edge Code Coverage[1]		Number of Detected Vulnerabilities	
			GREENHOUSE	HOUSEFUZZ	GREENHOUSE	HOUSEFUZZ
D-Link	DAP	12	20704	28924	1	14
	DIR	3	4480	5678	0	0
	DSP	1	2449	2969	4	5
	GO	1	1442	1841	0	0
Netgear[2]	WN/WNCE	2	1522	2625	6	19
	WN*AP/WAC	7	15715	16408	0	0
	WN(D/DR/R)	8	6886	11209	20	57
	WPN	1	766	994	2	2
	X(AV/W)N	2	3620	4745	5	22
Trendnet	TEW	4	3012	4068	8	9
Summary		41	11072	14767	46	128

图5 覆盖率以及检出漏洞数目比较

在双方都能测试的41个服务上，HouseFuzz在代码覆盖率上平均高出 24.8%，且在统计上显著优于基线。在漏洞发现方面，HouseFuzz检测到 128 个漏洞（其中110个为0-day），而 GREENHOUSE 仅发现 46 个（其中40个为0-day），0-day漏洞发现数量提升了 175%。



江宗泽
2024级博士研究生
研究方向：软件安全
Email：jiangzongze@outlook.com

探索与成长的科研之旅

黄浩琴

三年的研究生生活如白驹过隙，一眨眼已成过去。在实验室度过的时光，既有辛勤付出，也有满载收获的喜悦。我不仅收获了丰富的学术知识和技能，更磨练了自己的意志力和解决问题的能力。总的来说，我非常感谢实验室，希望我的经历能对大家有所帮助。

勤于阅读，善于思考

刚进入实验室时，我感到茫然无措，不知道该从何下手。虽然明白阅读文献是科研的基础，但面对浩如烟海的文献，我一度感到无从下手。在这种情况下，等待和犹豫只会浪费时间。于是，我决定主动出击，积极向导师和同学请教。在组会中，我实事求是地汇报自己的科研进展，暴露不足，寻求帮助。正是这种主动的态度，让我在科研的起步阶段少走了许多弯路。

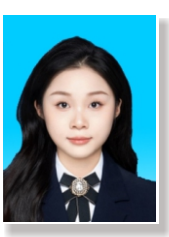
科研的起点始于阅读文献，而阅读文献的关键在于批判性思维。每当阅读文献时，我都会仔细分析文章的结构，研究作者是如何提出问题、设计解决方案并解决问题的，并思考其中的方法和逻辑是否合理，是否存在改进的空间。通过这种深入的阅读和思考，我逐渐形成了自己的知识体系。对于重要的文献，我会做详细的笔记，记录关键点和自己的思考。这些笔记不仅帮助我加深理解，也为后续的科研工作提供了宝贵的参考。

发现问题，解决问题

在大量阅读文献的基础上，我逐渐明确了自己的研究方向。寻找科研问题是一个从理论到实践的过程。我结合文献中的研究空白，寻找具有创新性和可操作性的研究问题。这个过程中，导师和同学的意见非常重要。他们的经验和不同视角往往能够帮助我更全面地认识问题，并发现其中的潜在挑战和机遇。这种交流和碰撞，不仅激发了我的科研灵感，也提升了我解决问题的能力。

在发现问题后，解决问题便成为科研的核心任务。通过查阅文献、请教导师和同学，我尝试了多种解决方案，并不断进行实验验证和优化。在这个过程中，我深刻体会到科研是一项反复试验、不断修正的工作。每一次实验失败都是一次宝贵的学习机会，每一次成功都离不开前期的努力和积累。这个过程不仅让我学会了如何在挫折中坚持，更让我懂得了科研需要持之以恒的探索精神和严谨细致的工作态度。

最后，科研是一项充满挑战和乐趣的工作，需要不断地学习和探索。我深知，未来的科研道路上还有许多未知的困难和挑战，但我相信，只要坚持不懈、勇于探索，就一定能够在科研的道路上取得更大的进步和收获。愿大家在科研的道路上勇往直前，不断创新，产出高质量的论文成果。



黄浩琴

2024届硕士毕业生

研究方向：图计算、稀疏矩阵乘法

Email：hqhuang@hust.edu.cn