東南大學
SOUTHEAST UNIVERSITY

边缘计算与边缘智能研究小组

指导老师：东方 教授

提纲

1 导师介绍

2 小组概况

3 方向简介

4 分组安排

5 小组制度

导师介绍

研究方向
边缘计算与边缘智能

东方，博士，东南大学**青年首席教授**，博士生导师，入选**国家级青年人才**计划，现任**东南大学大数据计算中心主任**。同时担任ACM中国理事会常务理事、ACM南京分会主席、江苏省计算机学会高性能计算专委会副主任&云计算专委会委员。

作为项目负责人主持科技创新2030重大专项课题、国家重点研发计划项目子课题、国家自然科学基金等多项国家级项目，并承担南钢集团、中国移动、江苏电网、顺丰科技、中车集团等多家行业知名企业的校企合作研究项目。相关成果获得国家级教学成果二等奖以及冶金科学技术特等奖。参加了丁肇中教授领导的AMS大型物理实验，建设完成东南大学云计算中心及东南大学AMS科学数据处理中心（AMS-02 SOC）。在IEEE/ACM TON、IEEE TMC、IEEEJSAC、IEEE TSC、INFOCOM、WWW、等国际国内重要期刊及会议上发表论文100余篇，并长期担任IEEE TON、TMC、TSC、《计算机学报》等重要期刊的客座编辑及特邀审稿人，历届云计算与大数据国际会议（CBD 2013-2022）Program Co-Chair，全国高校云计算应用创新大赛组委会秘书长，申请发明专利23项、软件著作权2项。

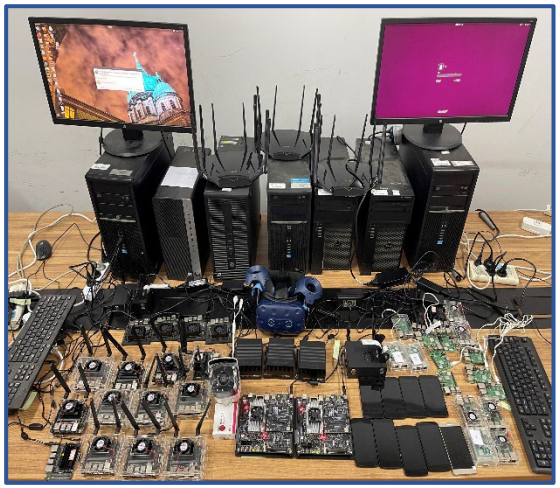
实验室情况

平台支撑

- 教育部**双一流建设重点学科**：计算机科学与技术，ESI 学科排名位列全国第2，**已进入全球前1‰**
- 计算机网络和信息集成**教育部重点实验室**
- 网络与信息安全**江苏省重点实验室**
- **东大-南钢**工业互联网联合研发中心

硬件条件

- 东方老师负责的**东南大学大数据计算中心**拥有21000+ CPU内核、260余块英伟达GPU，具有国内高校领先水平
- 实验室搭建有**异构端边云测试平台**



承担项目

项目类型	项目名称	执行时间
国家重点研发计划项目	面向工业互联网的智能云端协作关键技术及系统	2017年10月-2021年10月
国家自然科学基金重点项目	物联网智能感知与溯源方法	2023年01月-2027年12月
科技创新2030-“新一代人工智能”重大项目	元模型驱动的开放环境自适应感知	2019年12月-2022年11月
国家自然科学基金面上项目	面向深度学习应用的边缘计算执行框架与优化机制研究	2019年01月-2022年12月
江苏省前沿技术研发计划	低空多模态遥感基础模型与智能解译技术研发	2024年11月-2027年11月
深圳市科技重大专项项目	面向智慧物流的智能机器人端边云高效协同与实时自适应关键技术研发	2025年01月-2027年12月
国网江苏省电力公司重大科技项目	电力视觉云边端高效协同计算技术研究	2024年11月-2026年06月
华为2012实验室合作项目	大模型分布式、协同推理关键技术研究技术开发	2024年08月-2025年08月

小组研究方向

应用示范

- **研究内容：**基于推理和训练加速的边缘智能应用示范性研究
- **研究方向：**多目标跟踪、超分辨率视频传输、多路视频分析、VR等

推理加速

- **研究内容：**
  - ✓ 旨在从软硬件层面对智能应用提供推理加速
- **研究方向：**
  - ✓ 边缘推断、在网计算、CPU/GPU并行

训练加速

- **研究内容：**
  - ✓ 旨在从数据、模型和设备层面提供训练加速
- **研究方向：**
  - ✓ 流水线训练、联邦学习、FPGA

端边云协同

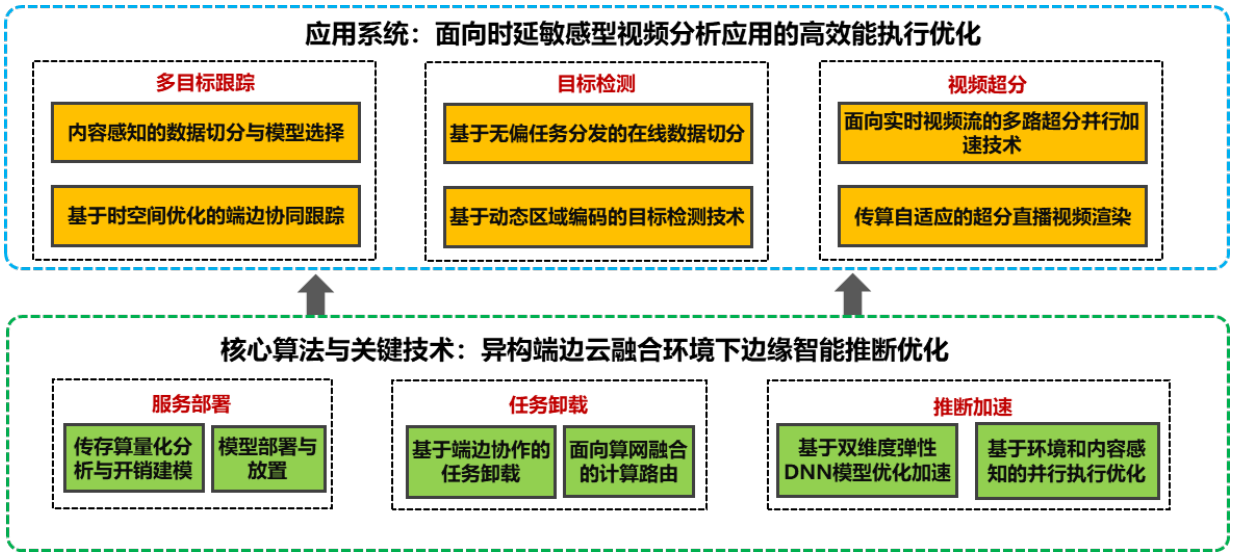
- **研究内容：**旨在高效利用端边云网络上的计算与传输能力
- **研究方向：**任务卸载、计算路由、数据缓存、数据获取、卫星计算

边缘计算与视频分析小组

边缘计算与视频分析小组

- **推理加速**：黄兆武、郭晓琳、李南翔、姜艺豪
- **无线通信**：王伟、季昀
- **视频分析**：朱浩鹏、赵文哲、陈柏均、杨昊宇
- **卫星计算**：杨琪

- 针对边缘环境下智能应用推理加速存在的问题，小组着重从**模型、计算与网络**三个方面进行优化，解决边缘智能推理性能瓶颈问题，旨在实现边缘智能应用高精度、低延时、高效能的需求。



边缘智能推理加速

研究内容

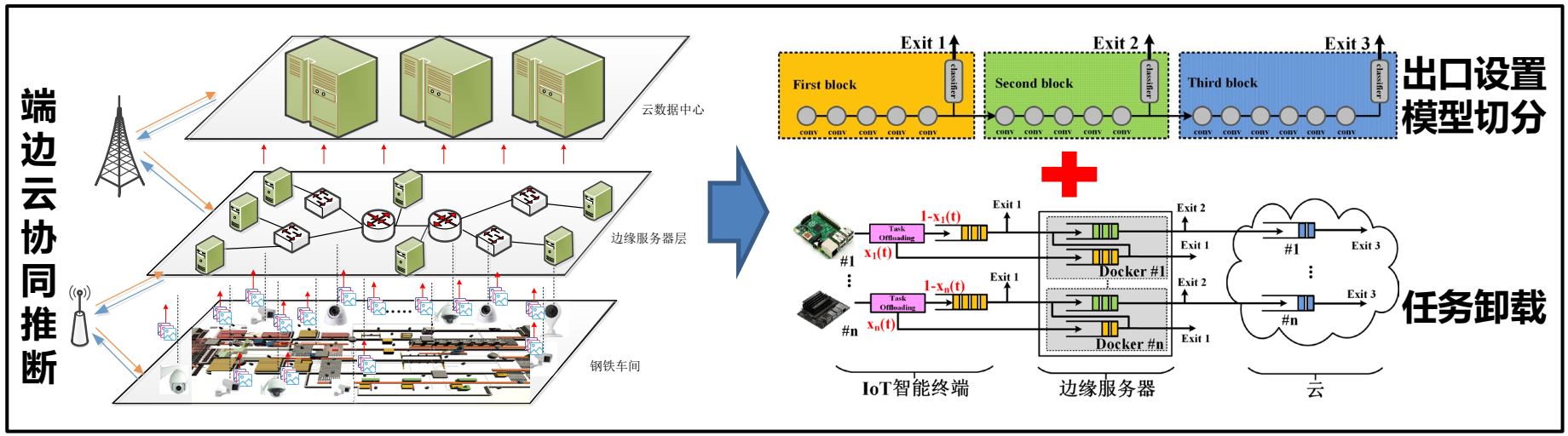
- 在动态异构的边缘计算环境中，研究高精度、低延迟的边缘智能推理加速问题。

研究范围

- 研究端边云资源的统一管理和任务调度；研究端边云间的任务卸载；研究深度学习模型的重构与切分。

待研究问题

- 边边协同、边云协同、服务部署



相关工作已被TMC、ICDCS'21、ICPP'22等会议期刊接收

算力网络计算路由

研究内容

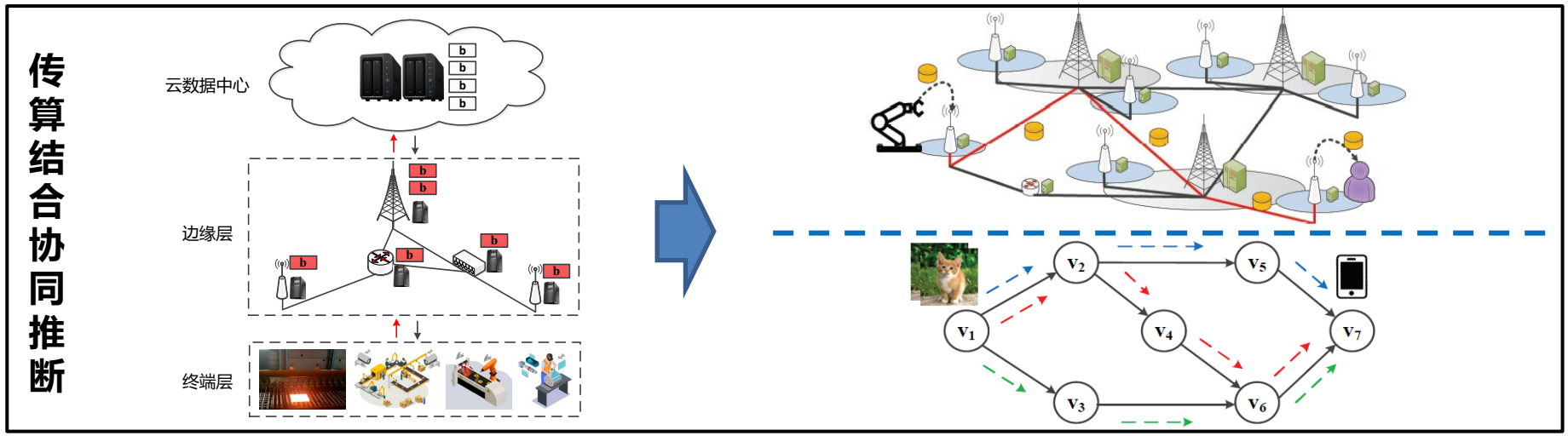
- 在算力网络环境下，研究面向智能应用的**传算结合（边传边算）**推理加速问题。

研究范围

- 研究端边云资源的感知和统一管理；研究边缘环境的任务切分和放置；研究算力网络环境下的计算路由；研究边缘环境下的服务缓存和部署；研究网络流量拼接。

待研究问题

- 边边协同、计算路由、资源分配



相关工作已被SECON'22接收

高效的边缘服务部署

研究内容

- 在边缘计算环境下，研究**低延时、资源有效**的边缘服务部署问题。

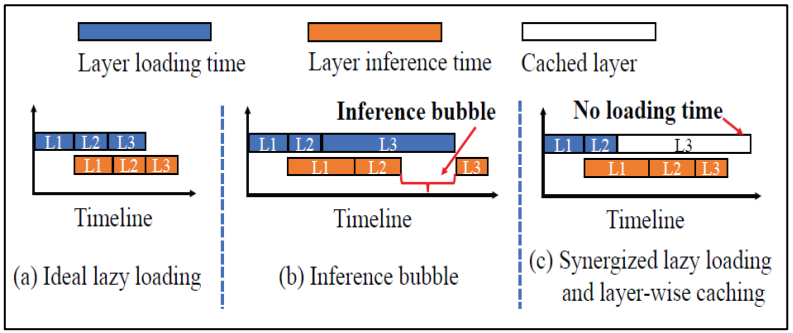
研究范围

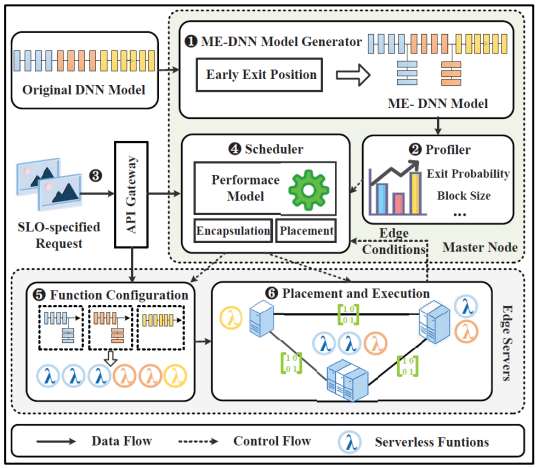
- 研究层级缓存和延迟加载的服务部署机制；研究资源有效的任务调度；研究流行度感知的边缘服务缓存与放置；

待研究问题

- 模型缓存、资源分配

层级缓存服务部署





相关工作已被IFOCOM'25、TMC、HPCC'24接收

卫星边缘计算

研究内容

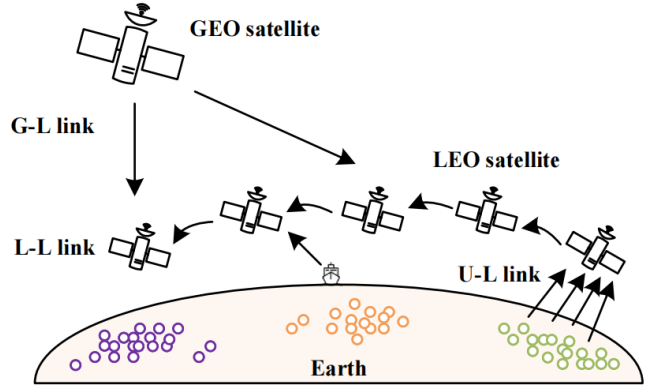
- 在卫星边缘计算网络环境中，研究低能耗、低延迟的任务调度、资源优化问题。

研究范围

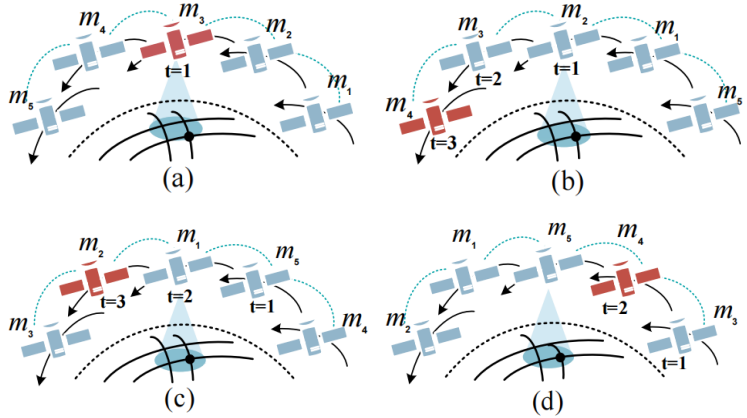
- 研究星间任务迁移与资源调度；研究用户为端，卫星位边，卫星/基站为云，端边云间的任务卸载；研究轻量化智能模型在卫星上的部署。

待研究问题

- 任务迁移、轻量化智能模型部署、大规模低轨卫星网络路由设计。



卫星边缘计算网络架构



移动场景分析

智能应用执行优化-多目标跟踪

研究内容

- 在动态异构边缘环境中，研究**自适应、高精度、低延迟**的多目标跟踪视频分析问题。

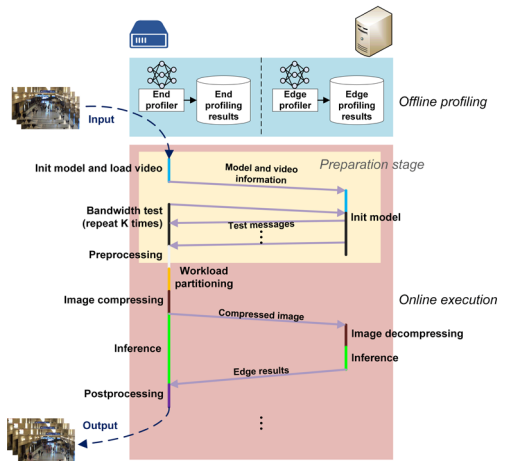
研究范围

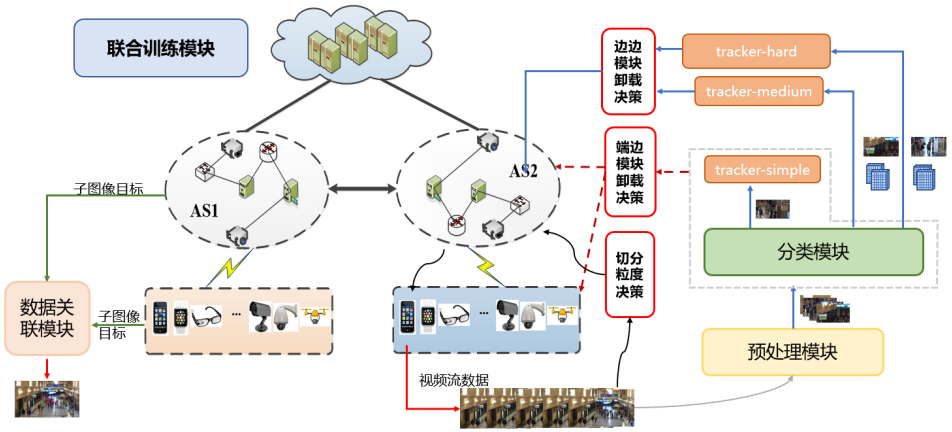
- 研究目标划分处理难度评判标准；研究图像动态切分决策策略；研究多网络特征复用联合训练；研究推理模型的切分和边端部署；研究多模型边边卸载策略。

待研究问题

- 视频帧自适应切分、追踪网络自适应卸载、图像分类评判方法

面向MOT应用的训练和推断





智能应用执行优化-超分辨率视频增强

研究内容

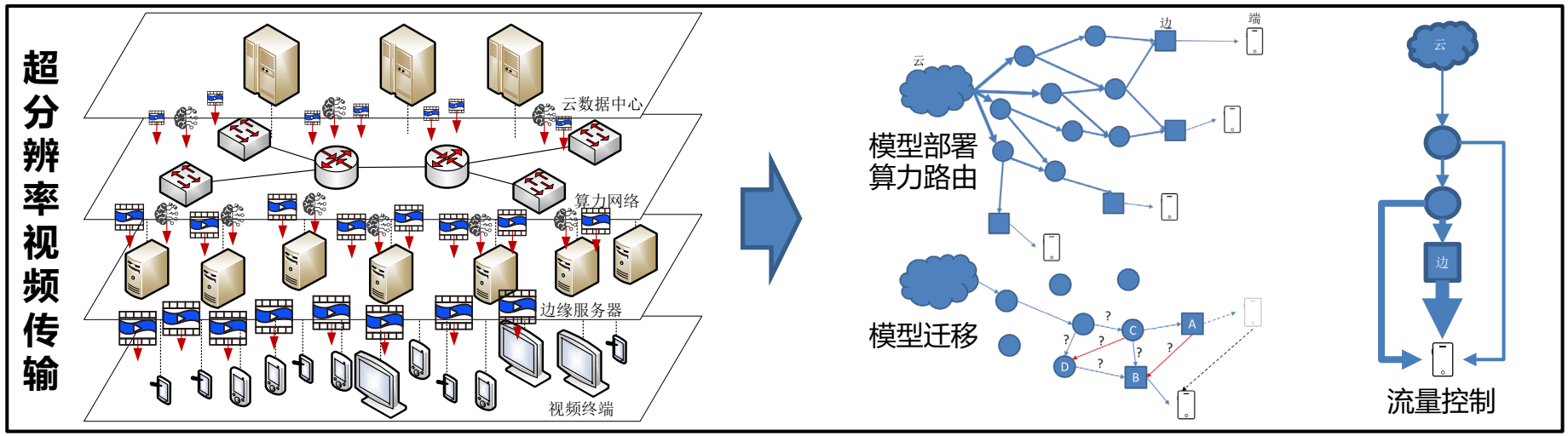
- 在算力网络环境中，研究视频在传输过程中进行**分辨率提升**的问题。

研究范围

- 研究算力网络中的带宽和算力资源分配；研究在实时推断前提下超分辨率模型的切分与部署；研究在终端移动情况下超分辨率模型的迁移。

待研究问题

- 算力路由、模型部署与迁移、超分辨率视频流量控制



相关工作已被IWQOS'23接收

智能应用执行优化-多路视频分析

研究内容

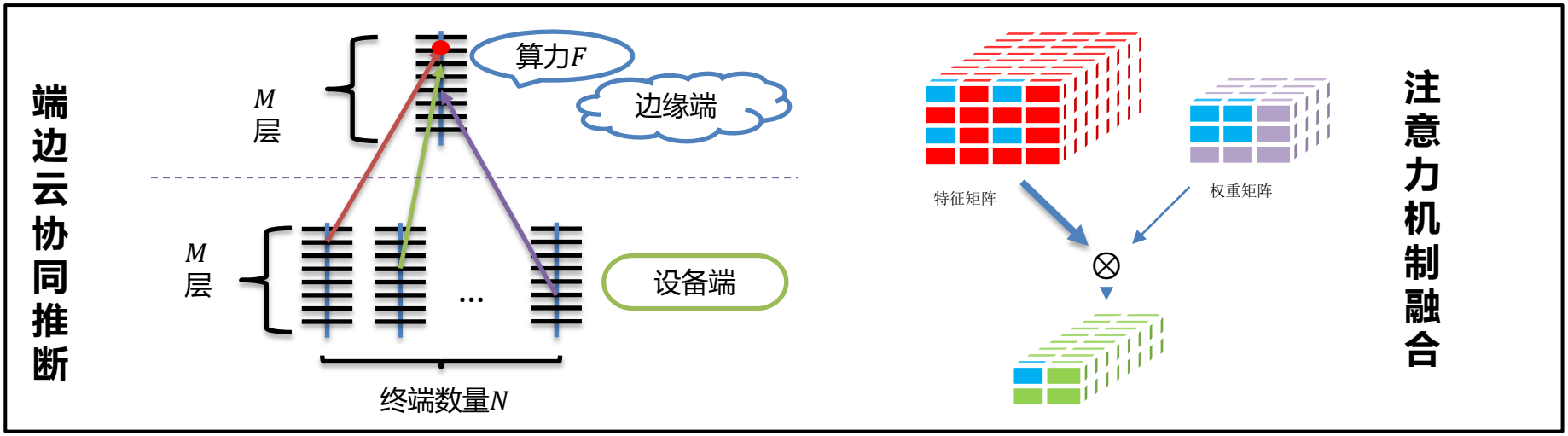
- 在动态异构的边缘环境中，研究高精度、低延迟**多路视频分析**问题。

研究范围

- 研究多视角分类技术；研究基于注意力机制的图像分类；研究多视角分类任务的卸载；

待研究问题

- 模型切分、多视角分类



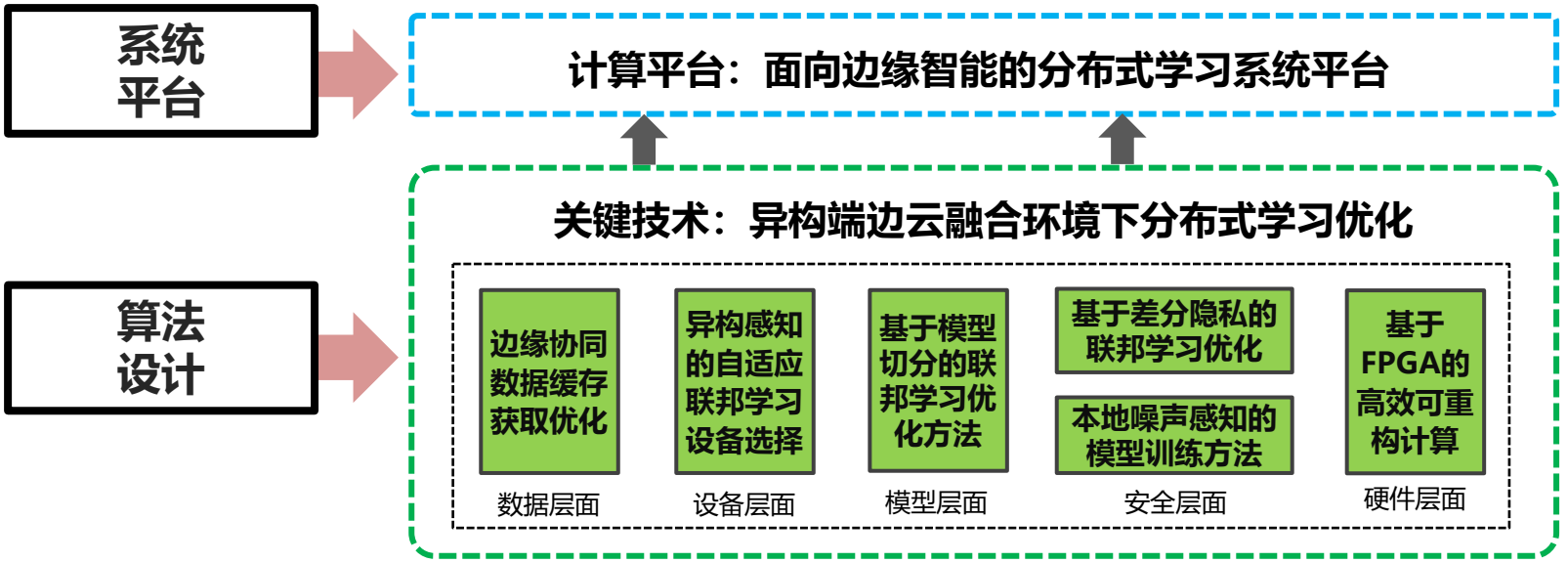
相关工作已投稿至ACM TOSN期刊

分布式学习与数据服务小组

分布式学习与数据服务小组

- **分布式训练/联邦学习：**伏舒存、郝江山、孙帆、许力丹、陈润泽、岑柄衡、檀佳玟、贺梦曦、何子衿
- **数据缓存/数据检索：**谭思雨、丘淑婷

- 针对边缘环境下部署分布式学习存在的问题，小组着重解决数据、设备、模型、安全、硬件五个层面的核心问题并构建面向边缘智能的分布式学习系统平台



边缘协同数据获取优化

研究内容

- 在数据中心和边缘环境中，研究边缘用户**高精度、低时延**的数据查询问题。

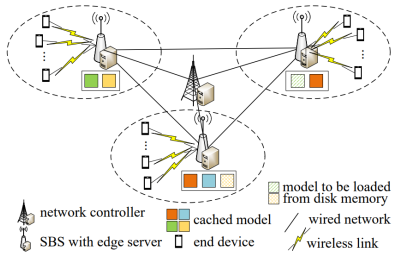
研究范围

- 研究端边云协同数据/模型缓存；研究端边云协同数据查询。

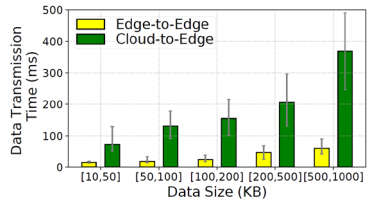
待研究问题

- 需求预测、服务部署、计算卸载、索引存储、查询机制

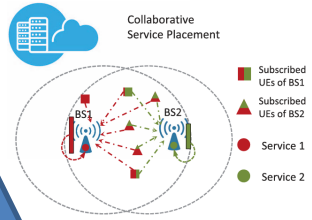
端边协同数据获取



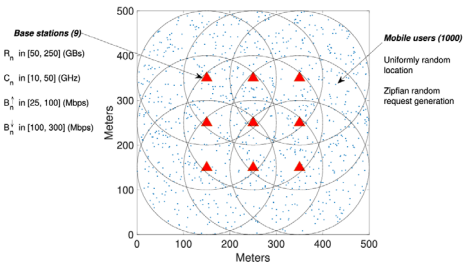
端边协同缓存示意图



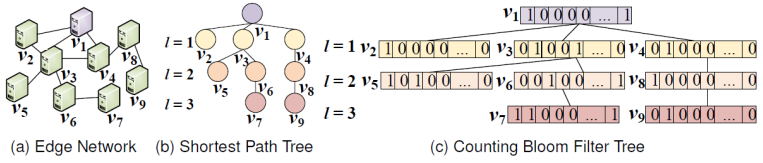
边-边和云-边数据传输时延



用户请求路由示意图



用户基站分布示意图



边缘协同索引结构

相关工作已被SIGIR'23、INFOCOM'25接收

端边协同的高效联邦学习优化

研究内容

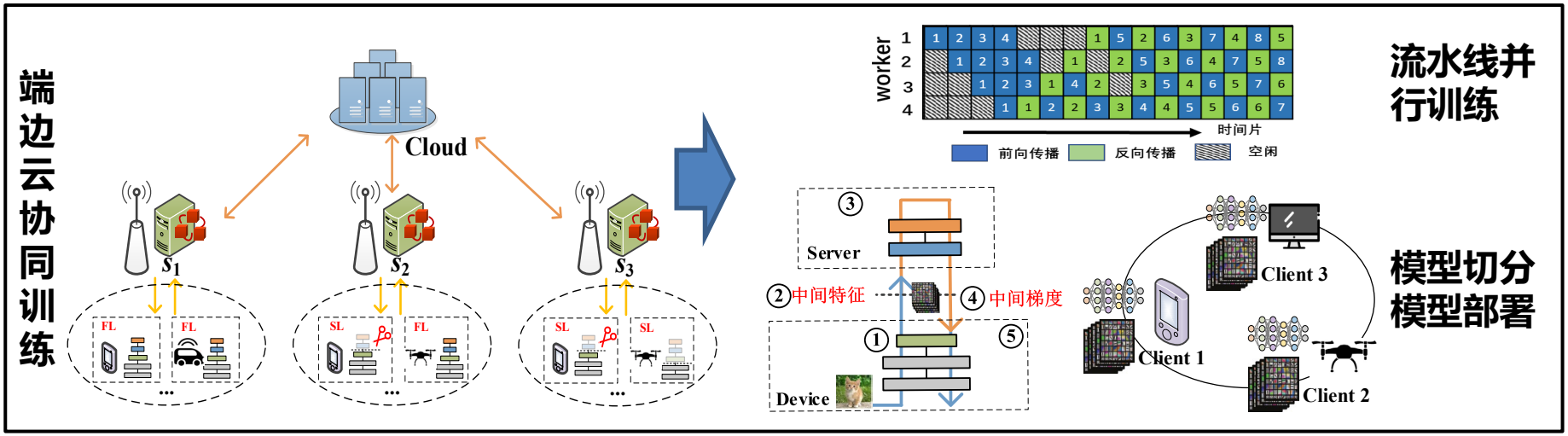
- 在端边云环境中，研究高效能、隐私保护的联邦（分割）学习训练问题。

研究范围

- 研究端边云资源的统一管理和任务调度；研究终端设备的选择和聚类；研究终端数据的隐私保护；研究训练模型的切分和部署。

待研究问题

- 大小模型协同、持续学习、服务部署



相关工作已被TMC、TSC、TIST、ICPP'24、SECON'23、WWWJ、SPE接收

基于FPGA的高性能可重构计算

研究内容

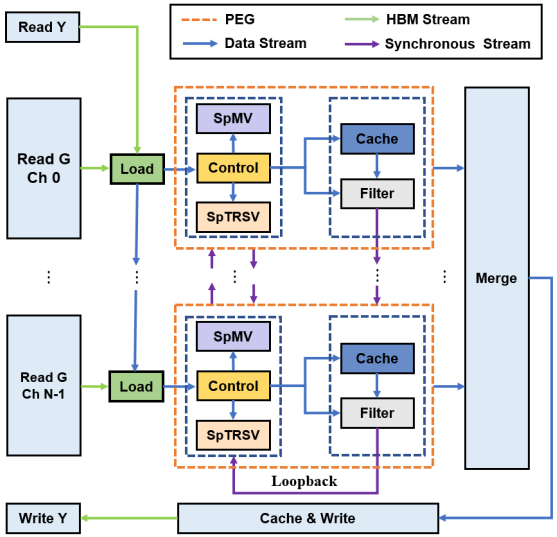
- 在端边云环境中，基于软-硬结合的视角，构建可重构、高能效的计算范式。

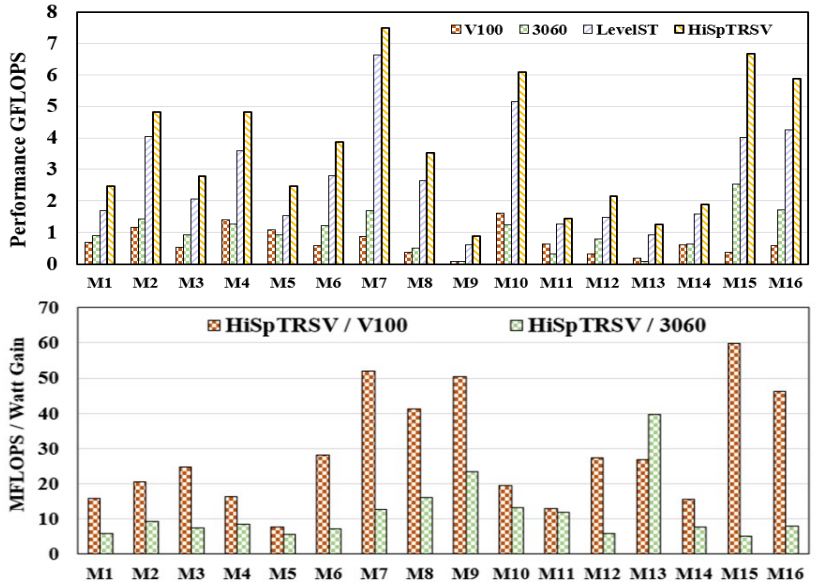
研究范围

- 研究深度学习中关键算子的优化；研究联邦学习中的通信优化；研究复杂网络环境下的高效传输。

待研究问题

- 多算子的部分可重构部署、软-硬视角的通信优化和建模、高效硬件架构设计。





相关工作已被DAC'25接收

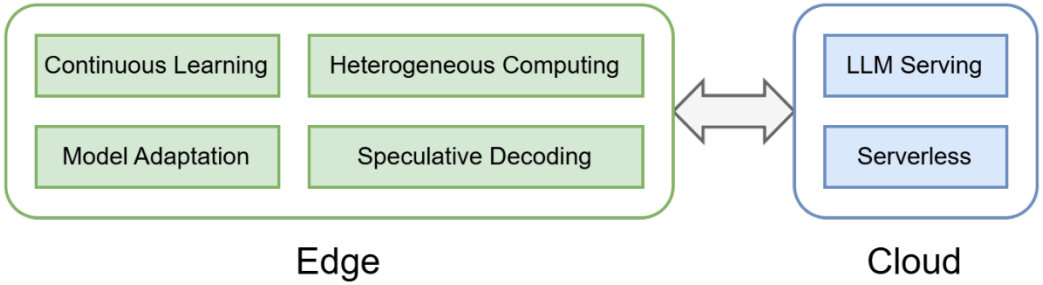
边缘智能系统小组

边缘智能系统小组

大模型分布式推理：周博文、贾金瑞、何文昊

- **持续学习**：陆辰羽
- **异构边缘设备模型推理**：李广通、万晔
- **大小模型云边协同投机推理**：刘雨宣
- **端侧大模型推理**：纪清玮、季诗尧
- **模型适应**：刘梦阳（香港理工大学）
- **无服务计算**：田昊冬（清华大学）

- 从**System**角度解决现有端边云环境下的机器学习任务挑战；相比算法研究，更侧重系统设计与实现。



边缘设备模型部署

研究内容

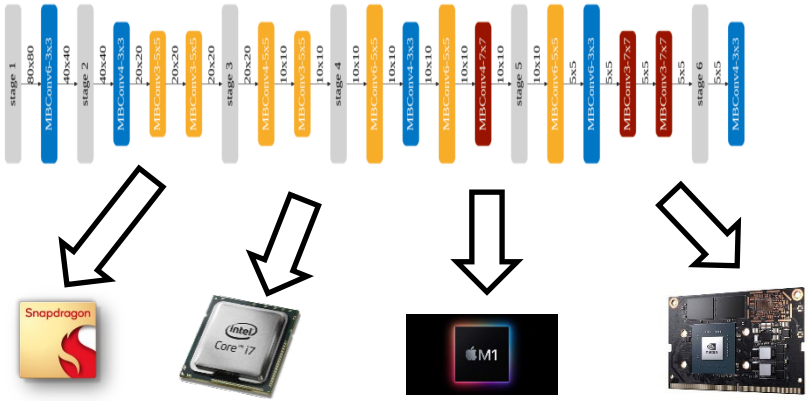
- 面向资源受限场景的自适应轻量化模型部署

研究范围

- 高效自适应轻量化模型部署旨在面对应用、需求、软件框架和硬件设备纷繁多样的边缘环境，提高模型适应效率。

待研究问题

- 模型结构搜索算法、自适应适配机制。



已投稿 ASPLOS 2025

无服务计算

研究内容

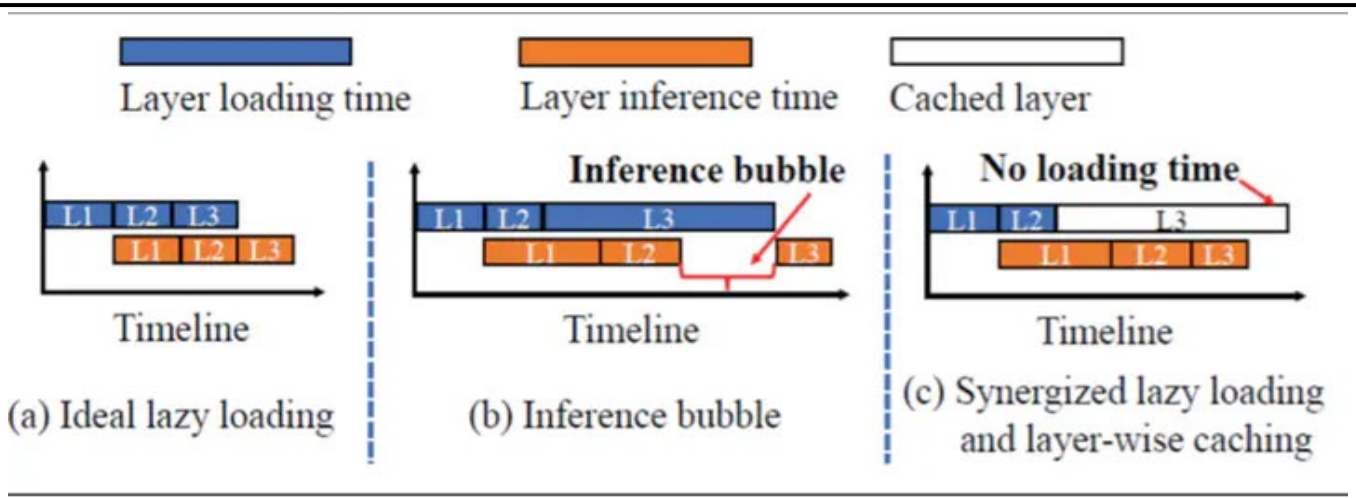
- 基于 Serverless 架构的边缘服务部署

研究范围

- 针对冷启动时间长、资源利用率低的问题，设计高效的边缘服务部署机制，以降低冷启动延迟并提升资源利用率。

待研究问题

- 服务部署机制、冷启动算法。



已录用IWQoS'24

持续学习

研究内容

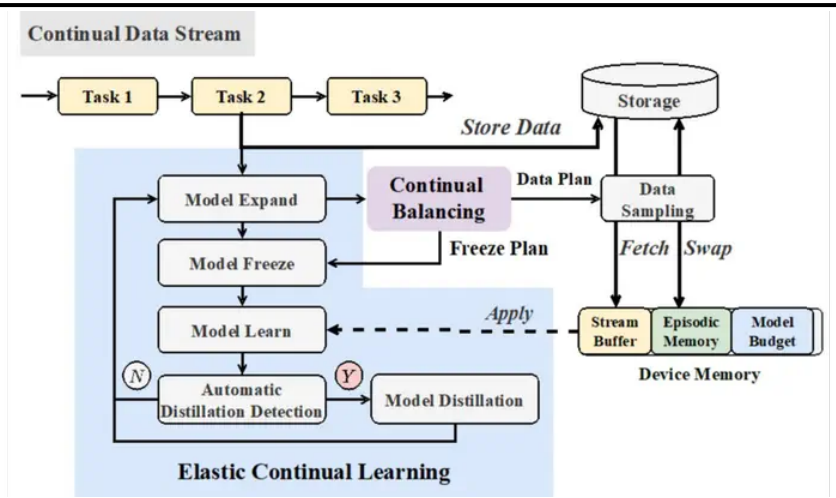
- 基于边缘设备的可持续学习

研究范围

- 针对资源受限的边缘场景中持续学习面临内存开销激增和数据、模型争用问题，在有限内存下实现动态模型扩展和合理数据分配，以支撑高效长期的持续学习。

待研究问题

- 动态模型扩展、数据分配。



已投稿IWQoS'25

异构边缘设备模型推理

研究内容

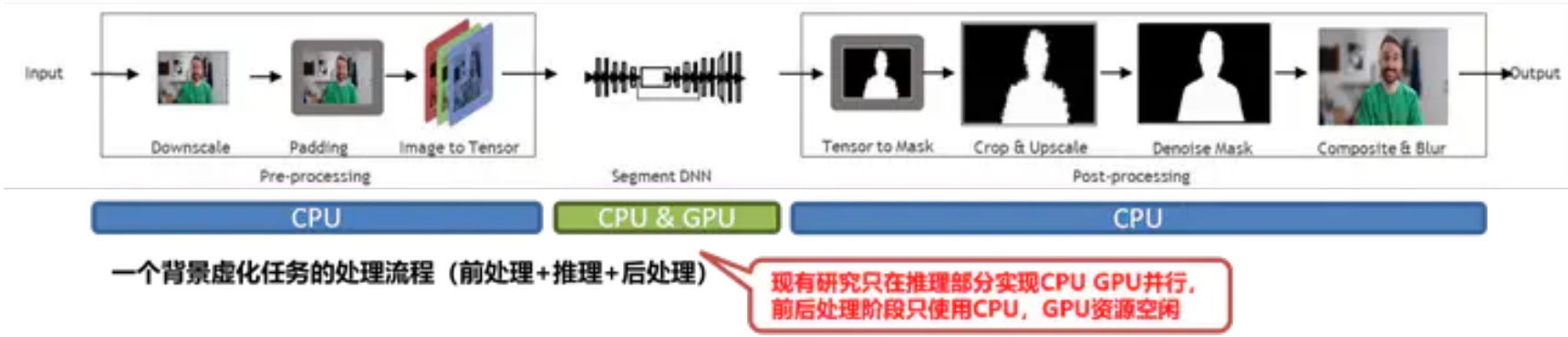
- 面向端设备的异构资源并行推理加速与能效优化

研究范围

- 研究面向异构处理器的并行推理和基于DVFS的高效节能推理，分别从推理延迟和推理能耗研究，提高移动端端侧模型推理的性能。

待研究问题

- 延迟预测、能耗优化。



大模型分布式推理

研究内容

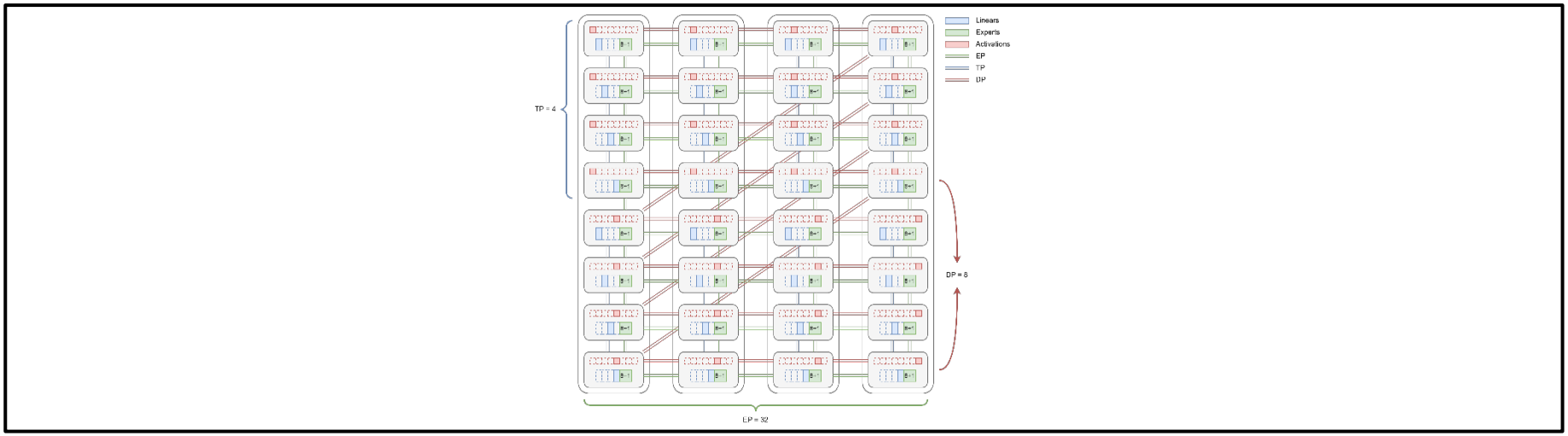
- 面向数据中心的**多节点**大模型分布式推理系统

研究范围

- 针对LLM和MoE模型的特点，设计高效的分布式并行策略和推理机制，以降低TTFT和TPOT，提升系统吞吐率。

待研究问题

- 并行策略、P/D分离机制、batch调度策略。



拟投稿ASPLOS'26

大小模型云边协同投机推理

研究内容

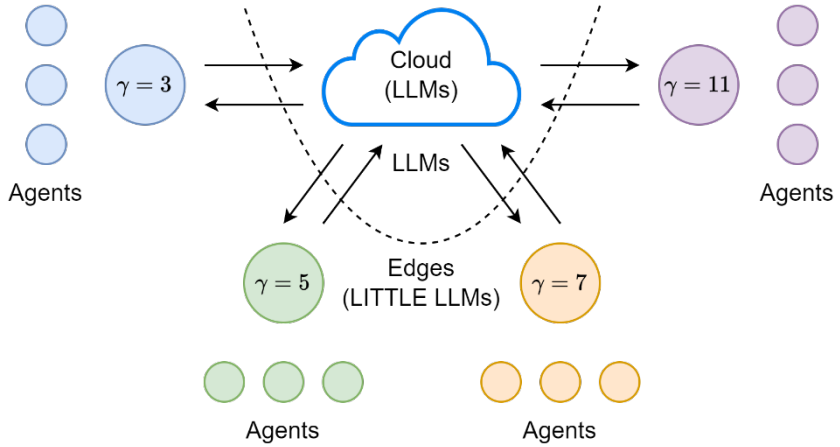
- 基于云边协同的大小模型投机推理

研究范围

- 基于LLM解码的投机推理机制，充分利用边缘设备计算资源，构建大小模型云边协同投机推理系统。

待研究问题

- 推理任务卸载、资源分配。



端侧大模型推理

研究内容

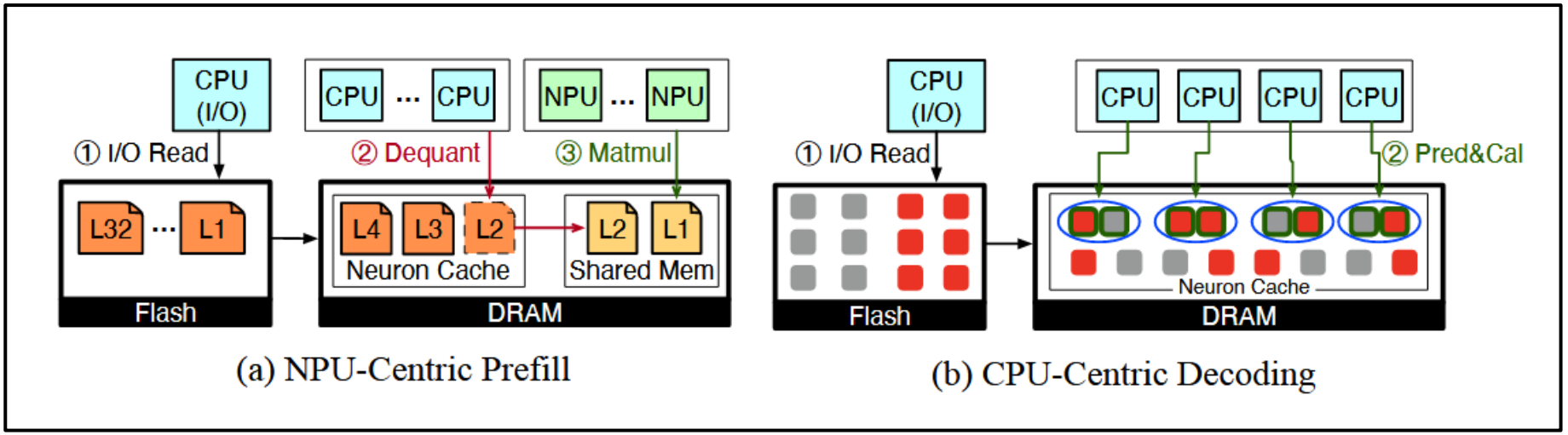
- 基于终端设备的CPU/GPU混合大模型推理

研究范围

- 为了充分降低大模型部署成本和保障用户数据隐私安全，利用终端设备现有的CPU/GPU和存储资源部署大模型推理服务。

待研究问题

- 推理任务在CPU/GPU间的卸载、成本最优存算资源配比。



发表论文

■ 团队在CCF A/B类、ACM/IEEE Transaction等国内外重要期刊和会议上发表一系列代表性工作

1. Fan Sun, Fang Dong, Dian Shen. HiSpTRSV: Exploring Tile-Level Parallelism for SpTRSV Acceleration on FPGAs. **Design Automation Conference(DAC), 2025**. **(CCF A)**
2. Shucun Fu, Fang Dong, Runze Chen, Dian Shen, Jinghui Zhang, Qiang He. Multi-Dimensional Training Optimization for Efficient Federated Synergy Learning. **IEEE Transactions on Mobile Computing(TMC),** 2025. **(CCF A)**
3. Zhaowu Huang, Fang Dong, Xiaolin Guo, and Daheng Yin. FaSei: Fast Serverless Edge Inference with Synergistic Lazy Loading and Layer-wise Caching. **IEEE INFOCOM**, **2025**. **(CCF A)**
4. Shuting Qiu, Fang Dong, Siyu Tan, Dian Shen, Ruiting Zhou and Qilin Fan. CoCaR: Enabling Efficient Dynamic DNN-based Model Caching and Request Routing in MEC. **IEEE INFOCOM**, **2025**. **(CCF A)**
5. Xiaolin Guo, Fang Dong, Dian Shen, Zhaowu Huang, and Jinghui Zhang. Resource-Efficient DNN Inference with Early Exiting in Serverless Edge Computing. **IEEE Transactions on Mobile Computing(TMC),** 2024. **(CCF A)**
6. Shucun Fu, Fang Dong, Dian Shen, Jinghui Zhang, Zhaowu Huang, Qiang He. Joint Optimization of Device Selection and Resource Allocation for Multiple Federations in Federated Edge Learning. **IEEE Transactions on Services Computing(TSC),** 17(1): 251-262, 2024. **(CCF A)**
7. Fang Dong, Huitian Wang, Dian Shen, Zhaowu Huang, Qiang He, Jinghui Zhang. Multi-exit DNN Inference Acceleration based on Multi-Dimensional Optimization for Edge Intelligence. **IEEE Transactions on Mobile Computing(TMC),** 2022. **(CCF A)**
8. Feng Shan, Jianping Huang, Runqun Xiong, Fang Dong, Luo Junzhou, Suyang Wang. Energy-Efficient General PoI-Visiting by UAV with a Practical Flight Energy Model. **IEEE Transactions on Mobile Computing(TMC),** 2022. **(CCF A)**
9. Dian Shen, Junzhou Luo, Fang Dong. Enabling Distributed and Optimal RDMA Resource Sharing in Large-scale Data Center Networks: Modeling, Analysis, and Implementation. **IEEE/ACM Transactions on Networking(TON)**, 2023. **(CCF A)**
10. Fang Dong, Junzhou Luo, Jiahui Jin, Jiyuan Shi, Ye Yang, Jun Shen. Accelerating skycube computation with partial and parallel processing for service selection. **IEEE Transactions on Services Computing(TSC),** 2020. **(CCF A)**
11. Shen Dian, Luo Junzhou, Dong Fang, Jin Jiahui, Zhang Junxue and Shen Jun. Facilitating Application-aware Bandwidth Allocation in the Cloud with One-step-ahead Traffic Information. **IEEE Transactions on Services Computing (TSC)**, 2020. **(CCF A)**
12. **……**

在读学生

博士13人（其中2人为非全日制工程博士）、硕士17人

姓名	类别	年级	姓名	类别	年级
郭晓琳	学博	2018	黄兆武	学博	2019
王苏扬	专博（非全）	2019	伏舒存	学博	2020
王芳	专博（非全）	2021	周博文	学博	2022
王伟	专博（全）	2022	郝江山	学博	2022
谭思雨	学博	2023	孙帆	学博	2023
杨琪	学博	2024	纪清玮	专博（全）	2024
许力丹	专博（全）	2024			
陆辰羽	学硕	2022	陈柏均	学硕	2022
朱浩鹏	学硕	2022	赵文哲	专硕	2022
季昫	专硕	2022	李广通	学硕	2023
丘淑婷	学硕	2023	陈润泽	学硕	2023
万晔	学硕	2023	岑炳衡	专硕	2023
刘雨宣	专硕	2024	李南翔	专硕	2024
姜艺豪	专硕	2024	贾金瑞	专硕	2024
季诗尧	专硕	2024	杨昊宇	专硕	2024
贺孟曦	专硕	2024			

学术活动及毕业去向

毕业时间	姓名	毕业去向	毕业时间	姓名	毕业去向
2019届	张欢欢	华为南研所（SP）	2024届	徐振轩	中国电信
2019届	黄兆武	本校硕博连读	2024届	孙星	南京某部军队文职
2022届	王慧田	华为（省优秀毕业生）	2024届	刘巳琪	本校大数据中心
2022届	蔡光兴	华为南研所	2025届	朱浩鹏	中国移动
2022届	朱立群	涉密单位	2025届	季昀	上海选调生
2023届	尹达恒	西蒙菲莎大学读博	2025届	陈柏均	字节跳动
2023届	唐安然	阿里巴巴			
2024届	周萌	华为南研所（SP）			
2024届	刘梦阳	香港理工大学读博			

招生说明

我们的优势

- 紧跟学术前沿，聚焦产业需求
  - 团队长期聚焦系统与网络领域的核心问题，面向当前学术界与工业界高度关注的研究热点，致力于探索如何基于端-边-云协同架构支撑智能应用的高效执行与落地。
- 依托重点平台，资源保障充分
  - 团队依托教育部与江苏省重点实验室，具备强大的软硬件支撑能力，提供稳定可靠的资源环境。指导教师现任东南大学大数据中心主任，具备丰富的科研积累与项目经验，能够为学生提供系统的科研指导与多元化的成长路径。

我们的期望

- 热爱科研，基础扎实
  - 对学术研究充满热情，有计算机网络、分布式计算、优化理论基础者优先。
- 责任心强、主动性高
  - 对自己的学习和科研负责、做科研要沉得住气，耐得住寂寞。具备良好的团队意识，积极参与团队科研任务，与团队同进步、共发展。

東南大學
SOUTHEAST UNIVERSITY

欢迎报考！

东方，fdong@seu.edu.cn