中山大學
SUN YAT-SEN UNIVERSITY

面向边缘算力网络的混合专家模型协同推理系统

Collaborative MoE Inference over Edge Computility Networks

姓名：叶盛源

导师：陈旭 教授

单位：中山大学 计算机学院

专业：计算机科学与技术

算力与算力网络

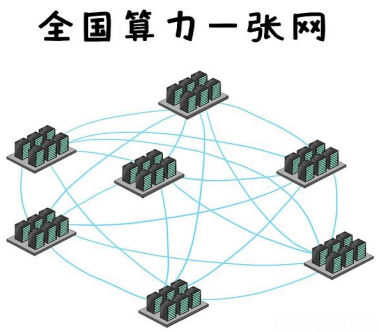
- 随着各行各业数字化、智能化转型的推进，**算力**对数字经济发展的带动作用越来越突出

智能家居

智能自动化工厂

AI赋能智慧城市

- 以**算力网络**为核心的新型基础设施建设在加快构建，实现**全国算力一张网**的愿景



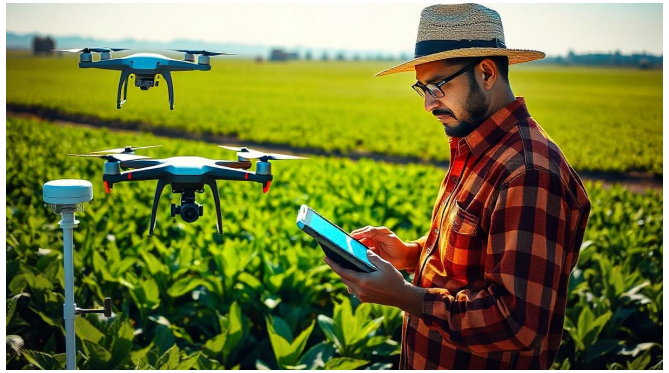
算力网络与“泛在边缘算力网络”

- **泛在边缘算力网络**是常规算力网络在终端侧的重要延展分支

智慧城市：无人车、智能路灯

智能家居：智能音箱、智能安防

智能工厂：机器人、机械臂



智慧农业：无人机、温湿监控器

端侧算力网络白皮书

（2022年）

中国移动通信集团终端有限公司、北京邮电大学

中国信息通信研究院、中国通信学会

中国通信学会
https://www.china-cic.cn › upload PDF
端侧算力网络白皮书
2022年7月28日 — 随着端侧. 算力能力的提升，边缘计算技术和

中国移动、中国通信学会、北京邮电大学等联合发布《**端侧算力网络白皮书**》

泛在边缘算力网络的研究意义

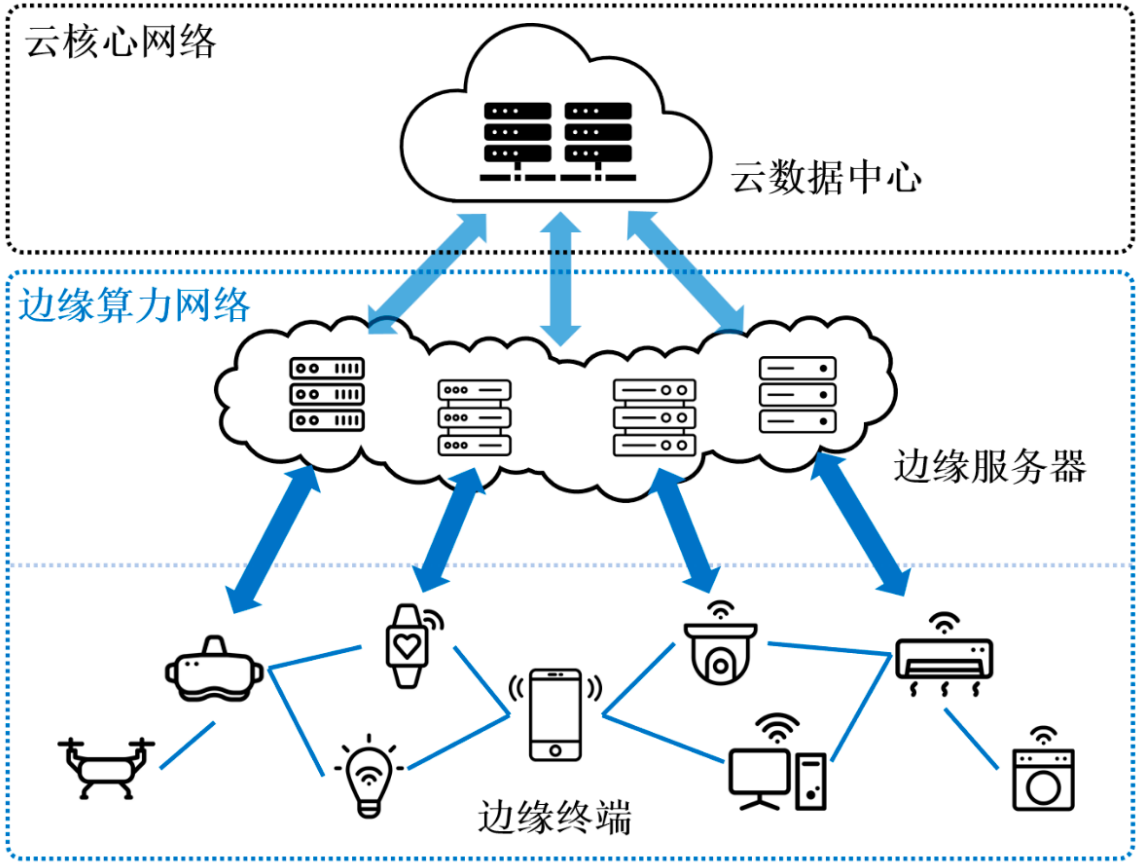
- 深入研究泛在边缘算力网络具有**重要的研究意义与实用价值**

研究意义

服务时效性
更短通信距离

资源高效性
端侧资源充分利用

隐私安全性
本地数据不出域



应用场景

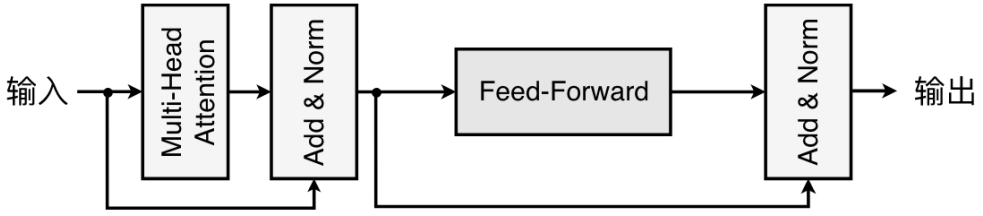

智慧城市


智能家庭

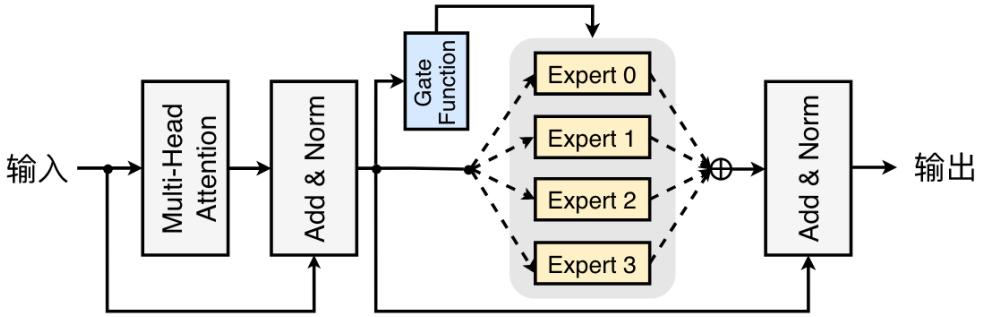

智能工厂

混合专家模型

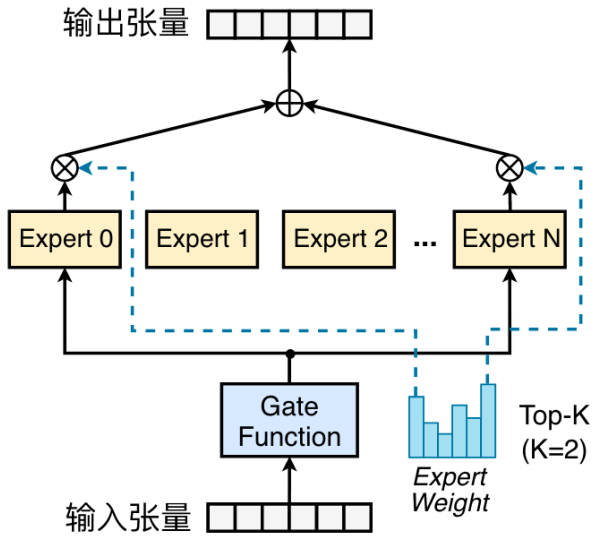
- **混合专家模型架构**
  - 当前主流的大语言模型从稠密架构向稀疏专家激活架构不断演进。
  - 凭借动态稀疏激活的特性，混合专家模型架构已成为扩展大语言模型的广泛认可范式。



(a) 经典稠密 Transformer 架构。



(b) 基于混合专家的稀疏 Transformer 架构

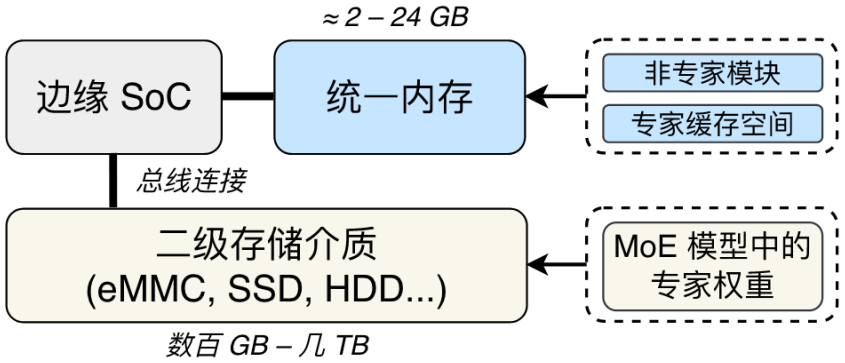


(c) MoE 模型的稀疏激活模式。

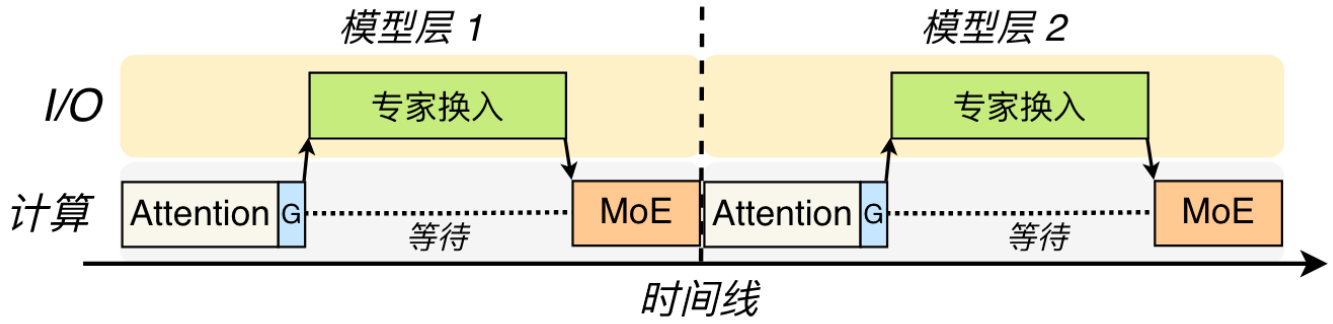
混合专家模型

混合专家模型端侧推理部署

- 现有工作通过专家交换技术来突破内存限制。
- 当输入序列很长时，依然需要全量专家换入，开销很大。
- 始终无法突破单台边缘设备有限的“资源墙”。



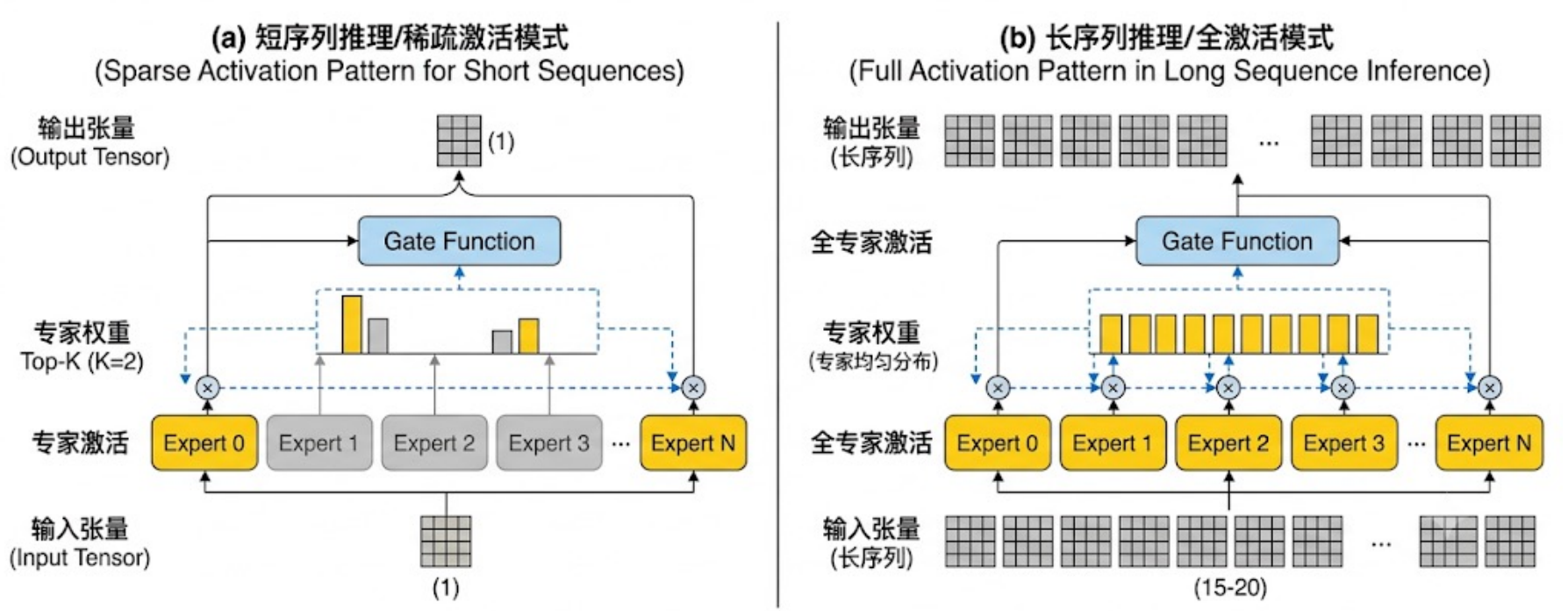
(a) 基于专家交换的 MoE 推理系统存储架构图。



(b) 基于专家交换的 MoE 推理执行时间线。

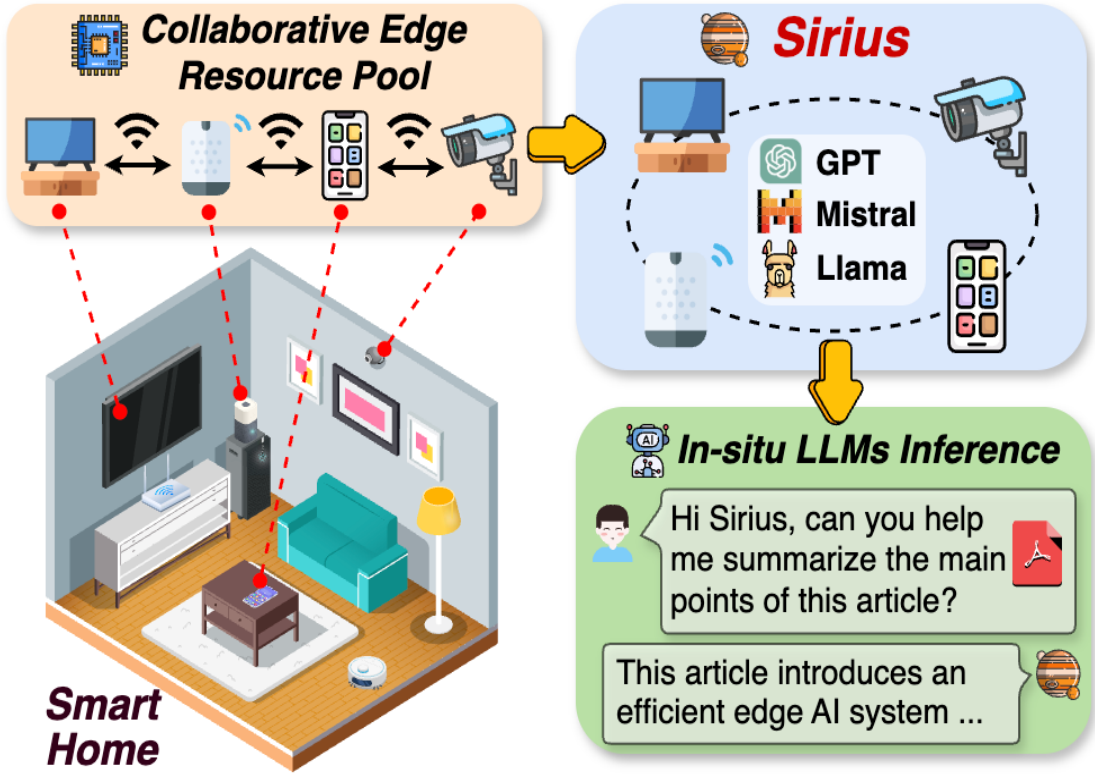
混合专家模型

- **混合专家模型端侧推理部署**
  - 只交换被激活专家的思想在解码阶段效果显著，因为每轮推理只有少量专家被激活，交换开销低。
  - 长序列预填充阶段，每轮推理激活所有专家都必须要被交换到内存中，交换成本高。



边缘算力网络与混合专家模型

- **优化机会：**
  - 边缘算力网络中存在分布式的可用计算设备。
  - 汇集这些设备的闲置资源，将其计算能力、内存容量、通信资源与 I/O 带宽进行协同。
  - 缓解单个边缘节点推理的资源约束，从而降低预填充阶段的 TTFT。

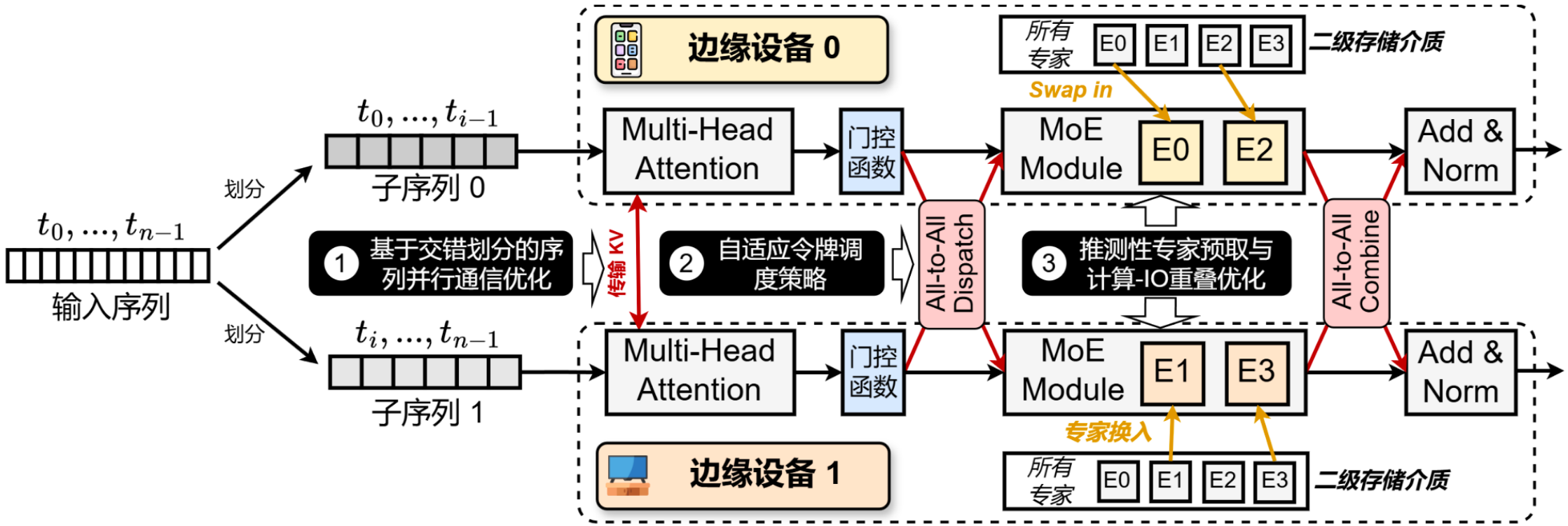


边缘算力网络与混合专家模型

- **系统设计目标：**
  1. 在多个边缘节点之间并行化单条序列序列推理请求，以降低预填充阶段的首词时延。
  2. 支持高效的专家交换，充分利用协作边缘设 备的集体内存容量和 I/O 带宽。
  3. 实现计算与 I/O 重叠的细粒度优化，以最小化 I/O 开销，降低延迟。

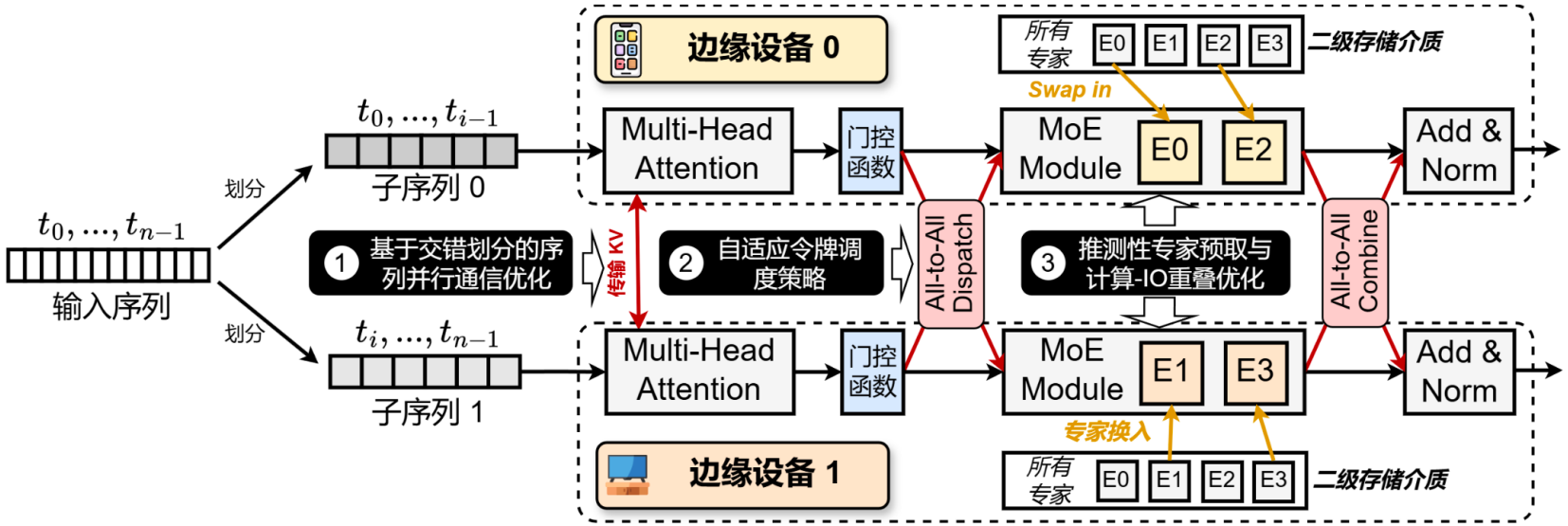
面向混合专家模型的边缘协同推理系统：Sirius

- Sirius 系统并行架构概览：



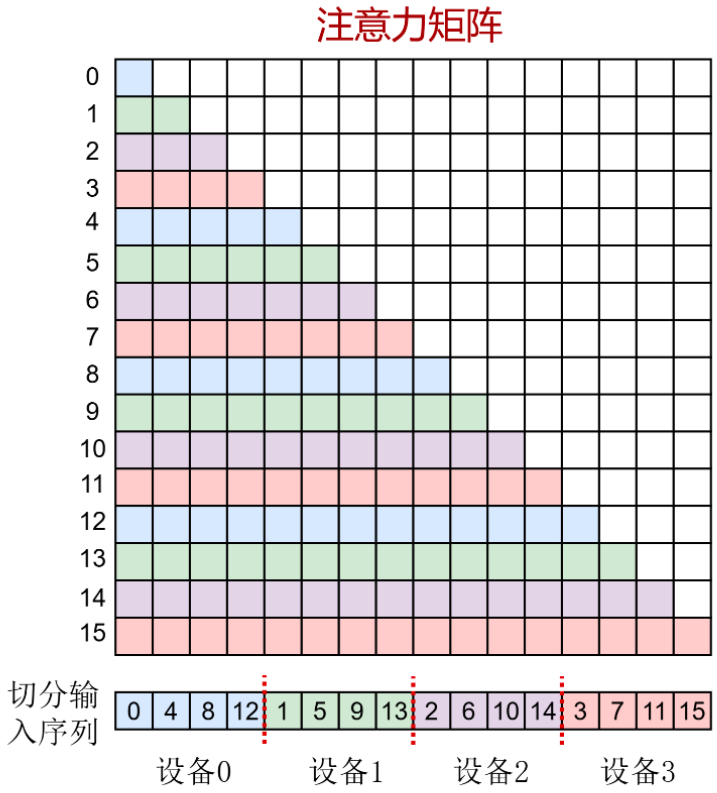
面向混合专家模型的边缘协同推理系统：Sirius

- **优化难点和挑战：**
  1. 多头注意力模块中的单向掩码机制使得序列并行会产生负载不均衡问题。
  2. 多头注意力模块到 MoE 模块之间需要寻求最优的动态 Token 分发规则。
  3. 专家交换引入的 I/O 开销仍可能成为系统瓶颈，需要优化专家换入延迟。

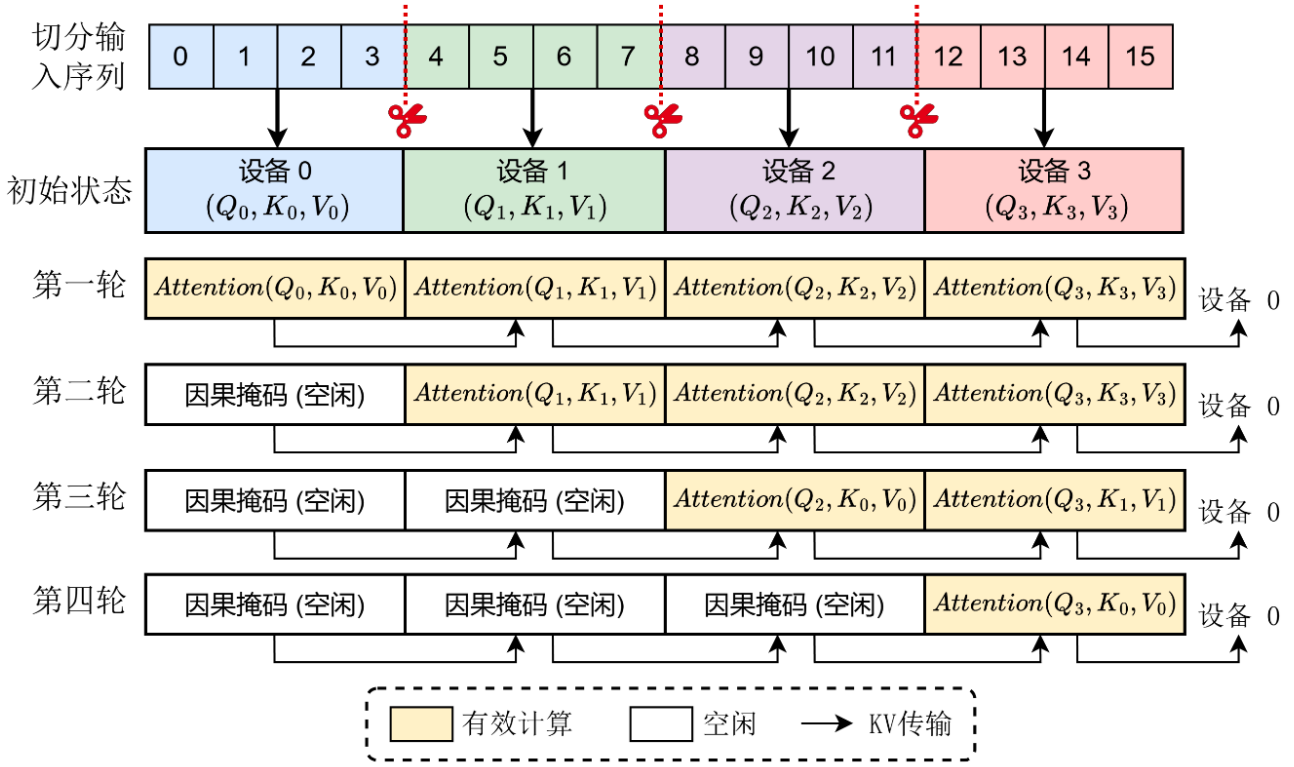


面向混合专家模型的边缘协同推理系统：Sirius

- **设计一：基于交错划分的序列并行通信优化**



(b) 基于交错划分的序列并行注意力分布情况.



(a) 基于环状通信的注意力机制计算流程。

面向混合专家模型的边缘协同推理系统：Sirius

- **设计二：动态 Token 分派机制**
  - **优化目标：**为每一组输入序列找到最优的 Token-边缘节点分派策略，使得通信+计算+专家交换的总开销最小。

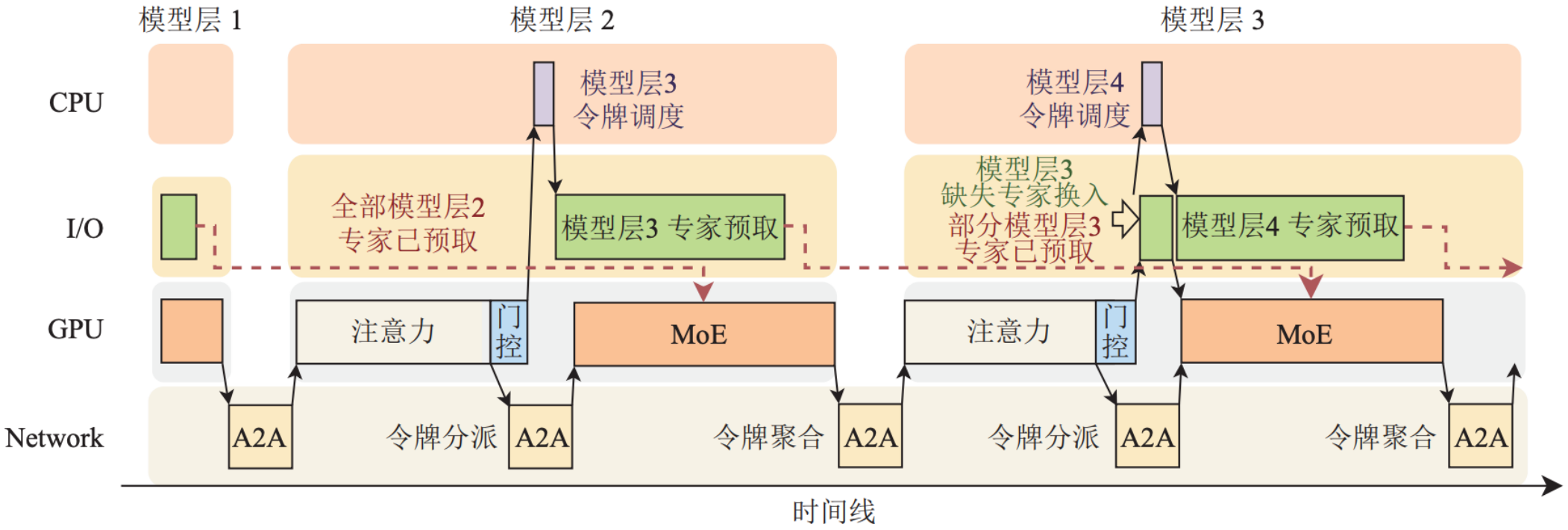
$$T_j = T_j^{\text{swap}} + T_j^{\text{comp}} = \frac{|\mathcal{E}_j^{\text{swap}}| \cdot W_e}{B_j} + |\mathcal{T}_j| \cdot K \cdot \tau_e ,$$

$$\min_{\phi} \max_{j \in \{0,1,\cdots,N-1\}} T_j \circ$$

  - **启发式贪心算法：**逐个遍历待分配的 Token，对于每个 Token 评估将其分配至 各候选设备后产生的累积时延增量，选择使当前最 大设备时延最小的分配方案。

面向混合专家模型的边缘协同推理系统：Sirius

- **设计三：推测性专家预取策略**
  - 观察发现：相邻两层的隐藏状态具有极高的相似性，可以提前对专家进行预取。
  - 通过细粒度计算-调度-I/O 并行优化，隐藏 Token 调度和专家换入的系统开销。



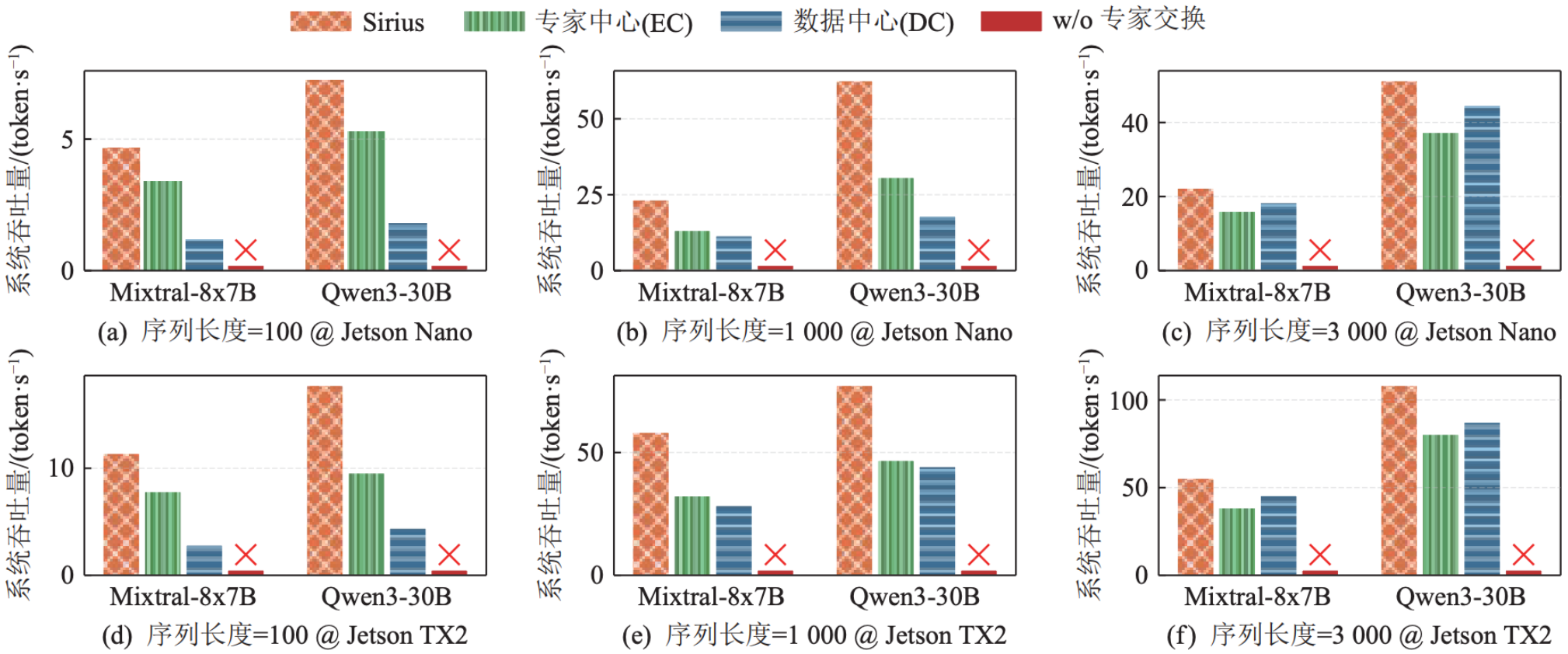
实验与结果

- **实验设置**
  - **模型**：本文选用 2 种主流混合专家架构 LLM Mixtral8x7B 与 Qwen3-30B。
  - **数据集**：使用 GLUE 数据集，构造了 3 组不同长度的子集，用来评估在不同序列长度下的系统性能表现。
  - **边缘环境**：采用 2 种现成的边缘计算设备 Jetson Nano与 Jetson TX2 构建仿真边缘环境。一个由 4 台 Jetson Nano 组成，另一个集群由 4 台 Jetson TX2 组成。

硬件参数	Jetson Nano	Jetson TX2
CPU	Cortex-A57	Denver 2 + Cortex-A57
GPU	128-core NVIDIA Maxwell	256-core NVIDIA Pascal
单精度浮点算力/GFLOPS	472	1 330
显存/GB	8	8
功耗/W	10	20

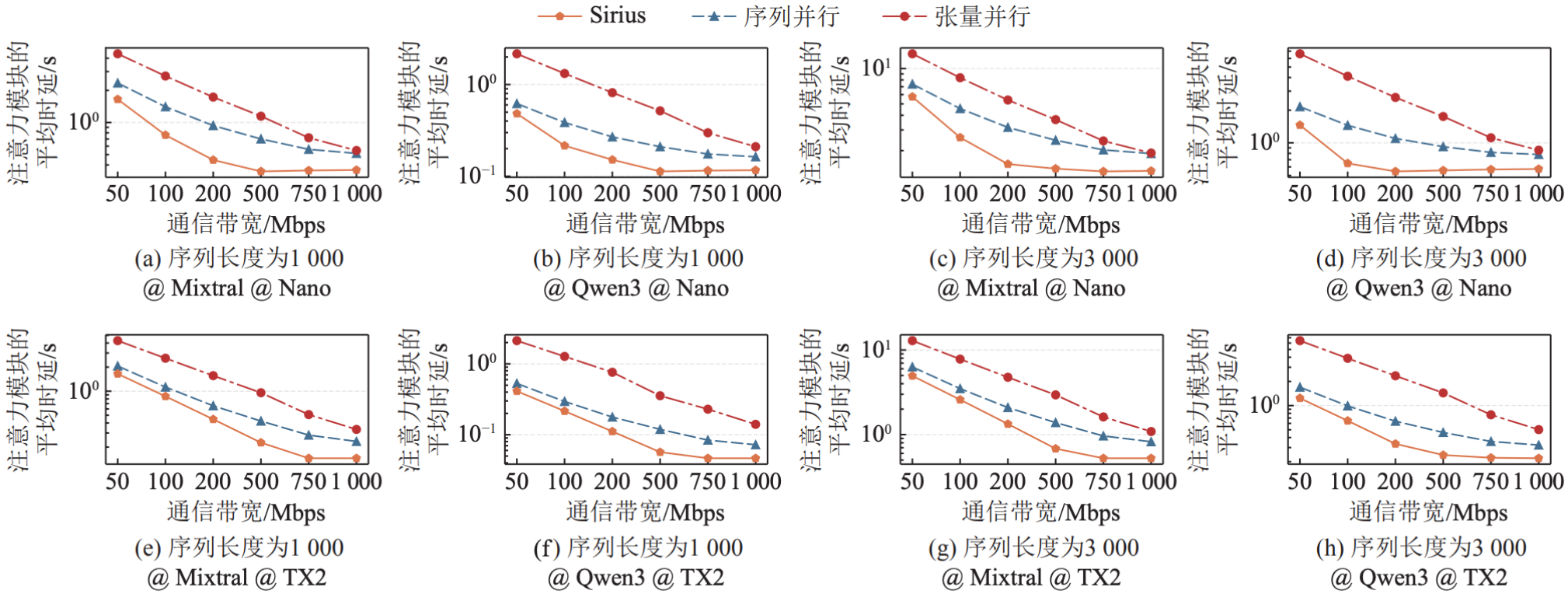
实验与结果

- **端到端性能对比**
  - ➢ 在全部模型与硬件配置下均取得最优性能，推理吞吐量提升幅度为 1.2~4.0 倍。



实验与结果

- **注意力模块序列并行优化效果分析**
  - 系统在全部实验配置下均取得最优性能。与张量并行相比，系统实现了 1.42~7.03 倍的时延降低；与序列并行相比，系统实现了 1.25~2.24 倍的时延降低。



注：Mixtral 是 Mixtral-8x7B 简写，Qwen3 是 Qwen3-30B 简写，Nano 是 Jetson Nano 简写，TX2 是 Jetson TX2 简写。

中山大學

SUN YAT-SEN UNIVERSITY

感谢

姓名：叶盛源

导师：陈旭 教授

单位：中山大学 计算机学院