Galaxy:一个于边缘设备上高效执行 Transformer 推理的协作 AI 系统。

介绍部分介绍了背景问题：

1. **Transform 模型的挑战：** Transformer 模型在 NLP 领域表现卓越，但其推理过程计算密集，传统的云端推理方法会带来网络压力和隐私问题。
2. **边缘推理的优势和挑战：** 边缘设备上的原位推理能够保持数据本地化，减少网络传输和隐私风险，但边缘设备资源有限，难以应对 Transformer 推理的高计算需求。

贡献:

Galaxy: 是一个协作边缘 AI 系统，利用异构边缘设备实现高效的 transformer 推理

关于用户隐私问题：在没有远程协助的边缘设备上进行原位推理，通过将数据保存在本地并避免网络传输，以实现边缘智能。但是边缘设备资源有限，对于 transformer 这种计算密集型和资源消耗型的推理有着很大挑战。

解决办法：

协作推理：利用物理邻近的多台边缘设备的闲置资源进行协作推理，从而打破单个设备的资源限制，实现资源共享和负载均衡。(比如智能家居环境：通常有多台闲置的可信设备（智能音箱、平板电脑、电视等），利用这些设备进行协同工作，分担推理任务。将这些设备的计算资源和内存资源整合，形成一个更强大的计算资源，共同处理 transformer 推理任务)。

第二部分：A 讲了基础 Transformer 模型，包括其应用在自然语言处理领域的卓越表现，然后就是 Transformer 层的结构、B 边缘设备上的 Transformer 模型的推理，实验结果从两个方面，一个是推理延迟：推理延迟明显高于数据中心的 GPU 推理；另一个是内存占用情况，实验中内存需求超出设备预算，导致无法进行推理. C 分析了协作推理的潜力,利用边缘环境中闲置资源的设备协同工作,分担推理任务,然后就是考虑并行的策略:包括数据并行,将数据分割到不同设备独立推理,局限:数据无法利用多设备资源； 流水线并行:将模型按层级分割,每层分配到不同设备,局限:推理中,流水线并行需要等待前一层完成,不能充分利用设备并行计算能力； 模型并行:在模型内部按操作进行水平分割,多个设备并行执行， 比较适合,但是需要解决同步点的通信延迟问题. 这三种对比分析在图三

第三部分:

A:Galaxy 系统设计的工作流

预处理阶段：在物理边缘设备上进行推理过程的运行时分析，收集必要的运行时数据，如计算延迟和内存占用情况。

并行规划阶段：结合张量并行（TP）和序列并行（SP）的混合模型并行架构，协调分布式推理工作流程。

执行阶段：应用并行规划配置，进行分布式推理。

B:其中的 HMP（混合模型并行）架构通过结合张量并行（TP）和序列并行（SP），将 Transformer 模型的计算任务在多个边缘设备之间分割和并行处理，TP 用于多头注意力和多层感知器块的并行计算，SP 用于连接块的并行处理，同时通过插入同步点（ReduceScatter 和 AllGather）确保计算结果一致性，从而高效利用计算资源和内存资源，实现低延迟、高效率的推理

C:这部分是 其中的异构 和内存感知的工作负载规划(任务分配的过程)：通过综合考虑每个边缘设备的计算能力和内存预算，利用两步启发式算法，先按计算能力均衡分配工作负载，然后根据内存限制进行调整，确保负载均衡和资源高效利用，避免内存溢出问题，从而优化

多设备协作的 Transformer 推理性能。
D：做了一种通信优化，通过将矩阵计算分块（瓦片化），使计算和通信操作可以并行进行。具体来说，将矩阵按瓦片分段处理，在每个瓦片上独立进行计算，同时进行 AllGather 和 ReduceScatter 通信操作，这样实现了计算与通信的重叠，大大减少了同步延迟，从而提高了整体推理效率并降低通信开销。

第四部分:
这部分介绍了关于这个 Galaxy 系统的实现和评估,选择了五种典型的 Transfirmer 模型，通过在同构和异构边缘设备环境中的实验验证，显示了 Galaxy 在利用多设备协作、优化计算资源和内存使用方面的优越性能，显著减少了推理延迟，并且在 GPU 环境中也表现出色，证明其在计算密集型 Transformer 推理中的有效性和可扩展性。

总的来说,这个系统通过在边缘设备上进行原位推理来实现隐私保护，避免将用户数据传输到远程云服务器，从而减少隐私泄露的风险。认为本地是可行的,也没有利用可信执行环境这种技术.