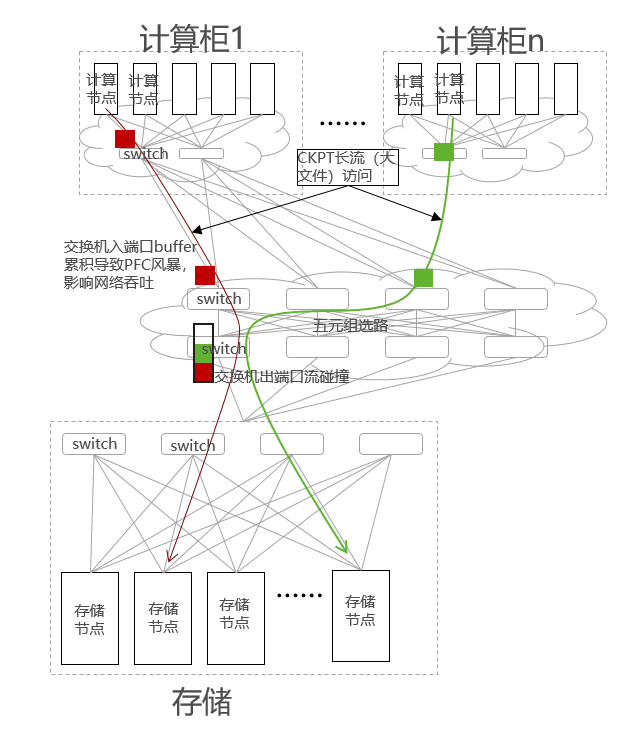
# 赛题名称

大规模组网下存储系统网络均衡算法设计与实现

# 赛题描述

数据中心大规模AI训练场景下，由于计算、存储节点数量多，需要多层Fat tree组网构成。AI训练过程中，AI集群会进行Checkpoint操作，以应对故障场景下，对集群计算进行快速恢复，该操作涉及到大量计算节点并行访问存储系统。在这种大规模组网下，针对存储访问的流量以大文件流量为主。本课题的主要挑战是，如何降低存储访问特征的流量在网络中碰撞造成的带宽损耗，提升系统有效吞吐能力。



* + - * 1. 节点规模及组网：假设计算节点规模为160节点（每个节点2x25G网络接口），存储节点规模为8节点（每个节点12x25G网络接口），采用4层Fat tree组网拓扑。交换机采用32x25G规格。拓扑：参见附件1。
        2. 业务访问模型：Checkpoint全局打散的大文件（共6400文件，每个文件5MB），使用的网络协议为RoCE v2.。
        3. 文件打散规则：文件以file\_idx（例如file\_0, file\_1……）形式命名，计算节点以server\_idx（例如server\_0……）形式命名，存储节点以stor\_idx（例如stor\_0……）形式命名。文件和计算/存储节点对应关系：参见附件2。
        4. 资源约束：交换机采用五元组ECMP选路机制，可以配置选用不同Hash算法。网卡MTU限制为4KB。

# 相关要求

* + - * 1. 基于以上要求，设计和实现一种网络多路径均衡算法结合拥塞控制方案， 使得在流完成时间内，写场景（流的方向为从计算到存储）达到较高的网络带宽利用率和较低的波动率。
        2. 选手基于开源NS3仿真平台进行验证，路径：https://github.com/bobzhuyb/ns3-rdma。参赛人员最终提交算法代码库、方案设计文档（含实验效果说明），方案设计文档需完整描述方案设计思路，达成效果及原理。
        3. 提交的代码以专家组验证效果为最终效果，方案设计文档需经过专家组评审，以确保结果符合参赛的实验效果。

# 评分标准

基于仿真平台，测试算法达成的网络带宽利用率，波动率指标，根据评分排序最佳方案。

最终得分=（网络平均带宽利用率\*1– 0.5\*波动率1 –1\*波动率2 –0.5\*波动率3 ）\* 100。 （网络平均带宽利用率=（总数据量/实际整体流完成时间）/（存储理论总带宽，即8\*12\*25G）；波动率= (最大带宽-最小带宽）/最大带宽) 。波动率采样时间段：0.1ms-5ms，40ms-60ms, 90ms -100ms