CCF-华为胡杨林基金
高性能计算专项
2022年申报方向与课题

CCF-华为胡杨林基金技术委员会——高性能计算领域组

二零二二年 十 月 十 日

第一条 专题设置

本专题以开放课题为主。开放课题主要资助具有前瞻性、前沿性、能为产业全面升级储备能力，实现关键基础技术底座自主、领先的相关课题

第二条 基金池

2022 年度高性能计算专项基金池共计 120 万人民币。

第三条 课题建议

单课题原则上资助不超过 30 万，为期不超过一年。鼓励开展同产业痛点、共性问题结合紧密的相关研究。鼓励进行前沿基础性的新研究方向和技术路径的探索。

第四条 课题详情

（一）支持多进程+多线程混合的 fillin-reduction 算法

课题背景：在做大规模稀疏线性方程组直接法求解时，通常需要先找到一种矩阵的对称变换，使得变换后的矩阵在直接法分解过程中有更少的非零元填充。找对称变换的算法称为 fillin-reduction 算法。经过测试发现，做和不做 fillin-reduciton，最终的非零元填充量有近 10 倍的差异，从而极大地影响了求解速度。目前 sota 的软件是 metis，包括 mt-metis 和 par-metis，但是也存在无法多进程+多线程混合，以及稳定性差的问题。希望能设计更好的 fillin-reduction 算法，一方面支持多进程+多线程混合，另一方面能比现有的软件有更好的 fillin-reduction 质量。

技术建议：基于原始的多层图分割技术（multilevel nested dissection algorithm），利用如 mpi 的单边通信技术，openmp 的 task 技术等，重新设计并行方案，以达到多进程+多线程混合的目的。期望速度上不差于 mt-metis 和 par-metis 的最快情况，质量上不差于串行的 metis（<1%fillin 误差）。

（二）支持多进程+多线程混合的 matching 算法

课题背景：在求解对称不定矩阵时，为了解决极小对角元带来的精度问题，经常需要对矩阵做 pivot 以加强对角元，从而提高求解精度。但是常规的 pivot 技术是非对称变换，这会导致 pivot 后的矩阵不再对称，从而极大地影响求解速度。为了保持对称性，可以在分解过程中将部分 pivots 从 1x1 变成 2x2，以 2x2 的矩阵逆代替 1x1 的倒数，从而在保持矩阵对称性的前提下，提高求解精度。该方案需要对原矩阵做分析，找到能够组成 2x2 pivots 的行列，目前 sota 方案是采用 Iain Duff 的 MC64 配合 symmetric weighted matching 技术。不过上述的 2 个技术均是仅支持单线程，当扩展到多机时会存在扩展性问题。希望能设计更好的 matching 算法，一方面支持多进程+多线程混合，另一方面 matching 后的求解精度不差于 sota 的方案。

技术建议：基于原始 pivot 和 matching 算法，利用如 mpi 的单边通信技术，openmp 的 task 技术等，重新设计并行方案，以达到多进程+多线程混合的目的。期望 32 核的强可扩展效率不差于 40%，128 核的强可扩展效率不差于 25%，同时 matching 后的求解精度不差于现在单线程的 mc64+symmetric weighted matching 方案

（三）在矩阵计算单元上发挥有效算力的研究

课题背景：STENCIL 计算在很多应用中是常见且重要的计算模式。近年来，随着矩阵计算单元被广泛的应用于通用处理器和加速芯片中。如何让 STENCIL 计算更好发挥矩阵计算单元的有效算力，将成为 HPC 应用软件开发者的必须思考的问题。本课题探讨如何在矩阵计算单元上发挥有效算力。

本课题限定：

1).三维 STENCIL 计算场景中 7 点，11 点格式，二维 Stencil 中 4 点（对角），5 点（十字），9 点（方块）；

2).硬件平台为支持矩阵计算单元芯片或加速器，其中矩阵单元只能限于矩阵外积操作；

3).如果没有支持矩阵计算单元的通用芯片，建议基于外积计算模式开发相关算法。

4).课题申请者可以自主选取熟悉的符合限定条件 1 的实际应用或者模板代码。

（四）稀疏矩阵基础算子加速研究

课题背景：目前多种高性能处理器已在硬件中提供矩阵加速单元，主要对稠密矩阵计算加速，如 NVIDIA GPU TensorCore、Intel AMX、昇腾的 Cube 等。但 HPC 应用的主要计算往往是稀疏矩阵运算，例如在稀疏线性方程组迭代求解中，稀疏矩阵向量乘是关键算子。这些算子无法直接使用现有矩阵计算加速单元加速。

技术建议：探索新的算法以利用稠密矩阵加速单元，例如通过矩阵变换等方法将稀疏矩阵运算拆分为多个稠密矩阵运算等；实现常见稀疏矩阵迭代求解算法如 CG 或 GMRES 在鲲鹏 920+昇腾 910 相比鲲鹏 920 实现 8 倍以上的平均加速；同时现有稠密矩阵加速单元的主要算力是 FP16 等低精度算力，进一步探索包括如何使用精度补偿、拆分计算等方法达到双精度的结果精度。