2022 年第三届 ACM 中国-国际并行计算挑战赛 初赛任务

大数据泛构支撑点选择性能上限的高效计算 并行优化

一、 赛题背景

面对大数据多样性（Variety）挑战所包含的数据种类多问题，一般有专用或通用两种模式。专用模式为每种数据类型量身定做专门的处理系统，因此需要为多个数据类型开发多个处理系统，性能较好但是开发维护的成本高，适用范围窄。而通用模式开发一个统一的系统来处理多种类型的数据，性能较差，但是具有较低的开发维护成本和较宽广的适用范围。商业软件为了追求最大的利润，一般都是采用通用模式的方法，通过高性价比来吸引用户。

实现通用模式的基本途径是首先把多种数据按照其共性抽象到统一数据类型和统一的匹配衡量方法，然后针对统一的数据类型和匹配衡量方法进行数据处理。

二、 赛题原理

度量空间（**Metric Space**）可以定义为一个二元组(S, d)，其中S是非空的数据集合，而d是定义在S的元素对上的具有如下性质的距离函数：

（1）非负性：对于任意$x, y \in S, d(x, y) \geq 0$，并且$d(x, y) = 0 \leftrightarrow x = y$。

（2）对称性：对于任意$x, y \in S, d(x, y) = d(y, x)$。

（3）三角不等性：对于任意$x, y, z \in S, d(x, y) + d(y, z) \geq d(x, z)$。

度量空间对数据的内部结构不作要求，仅要求定义在数据之间的满足度量空间性质的距离函数。目前相当多的常见或新型数据类型都具有或者可构建符合度量空间特性的距离函数。例如，向量数据和闵可夫斯基距离，字符串和采用满足特定性质的编辑操作代价数值的编辑距离，蛋白质序列和采用满足特定性质的替换矩阵的比对距离（Alignment）都构成了度量空间。因此，度量空间可以作为相当多数据类型及其距离函数的抽象规范。

我们把以度量空间作为统一的数据类型和匹配衡量方法，构建以度量空间作为基本数据抽象的通用大数据管理分析框架的通用数据处理模式称为大数据泛构。

大数据泛构把复杂的数据对象抽象成度量空间中的元素，而距离函数的具体实现和数据的具体表达都是透明的，用户只需要提供自定义的距离函数，同样的算法可以应用于不同的数据，实现了广泛通用的数据

管理和分析模式。

大数据泛构的高度通用性同时也是其弱点。坐标系统的缺失导致很多基于坐标系统的数学工具无法直接应用。为了给数据重建坐标，往往首先选择一些参考点作为支撑点（pivot），以数据到支撑点的距离作为坐标。支撑点选择作为大数据泛构的先导步骤，决定了后续步骤可以利用的信息量，因此**对于数据管理的性能具有决定性的影响**。支撑点选择研究的一个首要步骤是准确求出性能最好和最坏的支撑点组合，为后续研究提供标准和依据。

假设要处理的数据集是$S = \{x_i \mid i = 1, 2, \dots, n\}$，共有$n$个数据点，任两点间的距离可以由距离函数$d(.,.)$计算；要选择$k$个点作为支撑点，标记为$P = \{p_j \mid j = 1, 2, \dots, k\}$。对于$S$中任意的一个数据点$x$，其基于支撑点集合重建的坐标是其到各支撑点的距离形成的向量，

$$x^p = \big(d(x, p_1), \dots, d(x, p_k)\big)$$

本项目采用距离和目标函数，即任意两点的重建坐标间的切比雪夫距离的和，越大越好：

$$\sum L_\infty(x^p, y^p), x, y \in S$$

$$\text{其中} L_\infty\big((a_1, a_2, \dots, a_k), (b_1, b_2, \dots, b_k)\big) = \max_i(|a_i - b_i|)$$

需要准确地求出（可采用但不限于穷举法，但要保证结果的正确性）目标函数值最大和最小的各 1000 个支撑点集合。

三、 赛题说明

1. 源码包包括以下文件：
   a) pivot.c 源代码文件
   b) uniformvector-2dim-5h.txt 输入数据文件
   c) refer-2dim-5h.txt 基准输出文件

2. 程序使用方法：
   a) 源码包位置：/public1/soft/IPCC/2022/first/pivot.tar
   b) 参考编译命令：gcc pivot.c -lm -o pivot
   c) 集群参考运行命令：srun -p IPCC -N 1 ./pivot

3. 比赛考核程序求出目标函数值最大和最小的各 1000 个支撑点组合所用总时间，以程序输出“Using

time”时间为准，不包括读写文件的时间，不得修改计时函数的位置。

4. 输入数据文件 uniformvector-2dim-5h.txt 不可修改。
5. 可以改变源代码的数据结构和数据类型，优化方法需要对满足三角不等式的距离函数适用。
6. 以 result.txt 结果文件作为评判标准，所选出的目标函数值最大和最小的各 1000 个支撑点集合及其顺序须与 refer 基准文件完全相同。即先按目标函数值降序排序，输出目标函数值前 1000 大的点集；再按目标函数值升序排序，输出目标函数值前 1000 小的点集。
7. 参赛队员可自行更改编译方式，但需要留存脚本文件或 Makefile 文件。
8. 后续将发布多组参数和数据用于结果验证，各组数据对最终成绩权重占比相同。

四、 作品内容及要求

1. **优化版源代码**
   - **1) 包含编译、运行方式。可进行重新编译，并且能够正确生成可执行文件。**
   - **2) 不涉及版权问题，大赛组不负责保障源代码安全。**
2. **性能优化过程记录表（模板请见附件：1）**
3. **技术报告 PPT（含讲解录音，模板请见附件：2）**
   - 1) 应用程序运行的硬件环境和软件环境，其中软件环境至少包括操作系统、并行环境、相关依赖软件、所运行的应用负载等。
   - 2) 提供参赛应用程序的代码结构，从设计思路到主要流程设计及主要功能模块。
   - 3) 详细介绍参赛应用程序中采用的优化方法，基于优化方法达到的优化结果和性能指标。
   - 4) 详细描述程序运行结果。
   - 5) 参赛作品讲解录音（不多于 5 分钟），注意录音环境安静，确保作品质量。
4. **请于 2022 年 8 月 15 日前压缩以上文件上传至百度云盘（注意文件分享选择“永久有效”），登录官网个人主页，在“我的队伍”界面选择对应赛事队伍后提交。**

四、 竞赛平台

1. 北京超级云计算中心（https://cloud.blsc.cn/）

五、 竞赛形式及规则

1. **所有赛区初赛组织专家评审会，针对所有参赛方案进行评分，参赛队无需出席。**
2. 参赛队需在作品提交截止前（8 月 15 日）于组委会指定平台（https://cloud.blsc.cn/）运行初赛程序。（注意：请提前注册平台账号并申请试算核时）
3. 提交方式：上传百度网盘，登录官网个人主页，在我的队伍界面提交链接及提取码。
4. 组委会收到参赛队程序后，将以程序运行 5 次时间的均值作为上机成绩最终评分依据。
5. 初赛成绩中，上机成绩占比 80%，技术报告 PPT 讲解占比 20%。
6. 如参赛队发生任何学术不端、违反组委会规定的行为，组委会有权取消其参赛资格，并视情况向所在单位通报。

六、 联系我们

1. 官网：www.paraedu.org.cn
2. 微信：北京超级云计算中心（ID：BJBLSC）
3. 组委会：18310726311 余老师（QQ916034114）
4. IPCC-QQ 群：1046805935（学生/参赛选手）；1095416620（指导老师）
5. 邮箱：ACM_IPCC@163.com

注意：

1. 本赛题涉及的技术内容已获得作者授权，如需商业用途请同作者联系。因此产生的问题，IPCC 组委会不承担法律责任。
2. 以上内容最终解释权归 IPCC 组委会所有。