(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 102708404 B
(45)授权公告日 2016.08.03

(21)申请号 201210042043.7

(22)申请日 2012.02.23

(73)专利权人 北京市计算中心
地址 100012 北京市朝阳区北苑路大羊坊28号北科创业大厦B座1206室

(72)发明人 曾宇

(74)专利代理机构 北京安博达知识产权代理有限公司 11271
代理人 徐国文

(51)Int.Cl.
G06N 3/08(2006.01)

(56)对比文件
CN 101520748 A,2009.09.02,
US 2002165838 A1,2002.11.07,
Jelena Pjesivac-Grbovic et al.Decision trees and MPI collective algorithm selection problem.《Euro-Par 2007 Parallel Processing》.2007,第107-117页.
王洁等.多核机群下MPI程序优化技术的研究.《计算机科学》.2011,第38卷(第10期),第281-284页.
王洁等.多核机群下基于神经网络的MPI运行时参数优化.《计算机科学》.2010,第37卷(第6期),第229-232页.

审查员 刘志敏

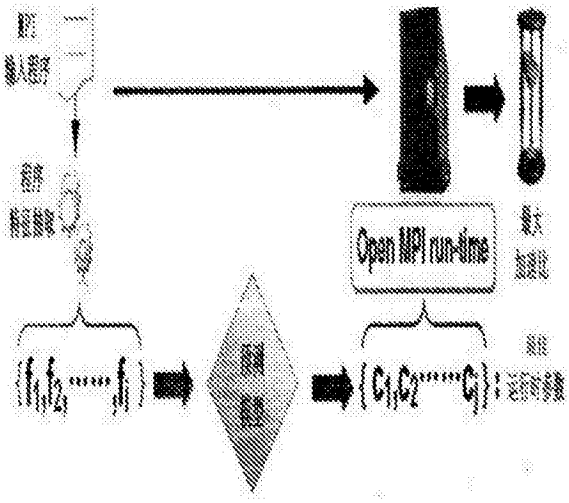
权利要求书1页 说明书6页 附图5页

(54)发明名称

一种基于机器学习的多核下MPI最优运行时的参数预测方法

(57)摘要

本发明提出了一种在多核环境下优化MPI应用的新方法:运用机器学习方法对多核机群下MPI应用的最优运行时参数进行预测。我们设计了具有不同点到点通信与集合通信数据比例的训练基准在特定的多核机群下产生训练数据,同时采用能快速输出结果的决策树REPTree和产生多个输出并具有较好抗噪性的神经网络ANN来构建运行时参数优化模型,通过训练基准产生的训练数据对优化模型进行训练,训练后的模型被用来对未知的输入MPI程序的最优运行时参数进行预测。实验证明,基于REPTree的预测模型和基于ANN的预测模型得到的优化运行时参数产生的加速比平均达到实际最大加速比的90%以上。



CN 102708404 B

1.一种基于机器学习的多核下MPI最优运行时的参数预测方法，其特征在于：

采用决策树和人工神经网络两种标准构建优化模型；

用构造的训练基准在目标多核机群上通过设置多组运行时的参数的组合生成训练数据，并对构造的模型进行离线训练；

训练后的模型用于对新的MPI程序预测最优的运行时的配置参数；

将预测所得结果与实际最优运行时参数向量做对比，评估预测模式的准确度；

所述决策树模型将训练基准的程序特征和运行时参数的配置组合作为决策树模型的输入，训练数据为：$\{F_i, C_i\}$，其中F_i为训练基准的程序特征，C_i为当前程序特征下的运行时参数组合，实际得到的加速比作为决策树的输出；

所述训练基准包括两种MPI通信方式：同步的MPI点到点通信和MPI集合操作；训练基准接收5个参数，可以分别用来控制训练基准中点到点通信的比例、集合通信的比例、两个MPI进程同步点到点通信的消息的大小、集合通信中交换的消息大小以及通信器的大小。

2.如权利要求1所述的方法，其特征在于：所述人工神经网络模型将训练基准中产生最高加速比的数据选出用来训练参数预测模型，训练数据为：$\{F_i, C_{i\_best}\}$，其中$F_i = <f_1, f_2, \ldots, f_m>$为训练基准的程序特征，$C_{i_best} = <c_1, c_2, \ldots, c_n>$为当前程序特征下的最佳运行时参数组合。

3.如权利要求1所述的方法，其特征在于：所述决策树模型在训练模型阶段，通过变换向量F与C产生不同的加速比结果；运用模型进行预测时，若F_p代表输入的MPI程序的程序特征向量，则能得到最大加速比S_{max}的运行时参数配置C_{best}将是此MPI程序的最佳运行时参数组合向量，即$S_{max} = M_{REPTree}(F_p, C_{best})$。

4.如权利要求2所述的方法，其特征在于：所述人工神经网络模型在训练模型阶段，通过变换向量F与C产生不同的加速比结果；运用模型进行预测时，若M_{ANN}是训练后的人工神经网络模型，则$C_{best} = M_{ANN}(F_p)$，其中F_p代表输入的MPI程序的程序特征向量，C_{best}是此MPI程序的最佳运行时参数组合向量。

5.如权利要求1所述的方法，其特征在于：所述离线训练通过变换训练基准的5个输入参数，控制点到点通信和集合通信的比例分别为：100％的点到点通信、100％的集合通信、50％的点到点通信及50％的集合通信，在三种不同通信比例下，分别变换点到点以及集合通信中消息大小以及MPI通信器的大小，并变换运行时参数的配置组合，共产生训练数据3000条，用来训练神经网络优化模型。

6.如权利要求1所述的方法，其特征在于：当预测模型建立并用大量的学习数据训练完毕后，就要根据实际需求来执行预测任务。

7.如权利要求1所述的方法，其特征在于：在执行预测前，需要在目标多核机群下对要预测MPI程序的进行一次instrument运行，以得到输入的MPI程序的特征向量F_p；将F_p作为模型的输入即可得到输入MPI程序的最佳运行时参数组合；当目标多核机群发生变化时，以上过程需要重复进行。

一种基于机器学习的多核下MPI最优运行时的参数预测方法

技术领域

[0001] 本发明涉及多核环境下MPI优化，具体来说，涉及一种基于机器学习的多核下MPI最优运行时的参数预测方法。

背景技术

[0002] 随着多核技术更加广泛的应用于机群，多核机群下MPI应用的性能优化成为了研究的热点。目前主流的MPI库实现(Open MPI、MPICH等)都提供了可调的运行时参数机制，允许用户根据特定的应用需求、硬件以及操作系统来调优运行时参数以提升MPI应用的性能。

[0003] 本章我们设计实现了一种基于机器学习的通用的多核环境下MPI运行时参数优化模型，能自动为给定软、硬件结构的多核机群下的MPI程序预测接近最优的运行时参数组合。我们提出的预测模型基于机器学习中的决策树和人工神经网络方法，通过对预测模型的离线训练和在线学习，能自动为未知的MPI程序预测接近最优的运行时参数。要预测的MPI程序由对源码运行一次得到的动态特征和通信器大小等静态特征来共同描述。我们提出的基于机器学习的最优MPI运行时参数预测方法在基于InfiniBand的多核SMP机群上进行验证，并运用Open MPI这一主流的MPI库作为预测MPI最优运行时参数的环境。通过NAS并行基准套件2.4中的IS和LU基准的实验证明，与Open MPI默认配置相比，基于机器学习的预测模型得到的优化运行时参数组合能为多核机群下的MPI应用带来最多约20％的性能提升。

[0004] 多核技术指将两个或多个处理内核集成到一个处理器芯片当中，并通过将负载分配到多核上来加速应用的处理性能。目前基于多核技术的机群已经成为高性能计算领域的主流平台，越来越多的机群采用多核处理器作为核心部件。消息传递接口MPI(Message Passing Interface)是机群下最常用的并行编程模型，广泛应用于分布式以及共享内存系统。

[0005] 多核处理器的新特性使多核机群的存储层次更加复杂，同时也给MPI程序带来了新的优化空间。虽然算法的数据局部性、负载均衡等是影响MPI应用性能的 因素，但其与具体的特定应用特性有关，直接将现有的MPI程序移植到多核机群平台上，应用的性能和可扩展性并没有得到多大的改进。目前对于多核下MPI的优化研究主要集中在混合MPI/OpenMP、优化MPI运行时参数、优化MPI进程拓扑、MPI集合通信的优化等方面，其中可调的运行时参数对多核环境下的MPI应用的性能有着重要的影响，但最优的运行时参数依赖于多核节点或多核机群的底层架构以及MPI程序自身的特征。

[0006] 主流的MPI库实现都提供了可调的运行时参数机制，允许用户通过调整运行时参数来获得更高性能。例如可以根据通信消息的大小来修改点到点通信采用的协议，即修改MPI库中由立即通信协议(Eager)转为集中通信协议(Rendezvous)的阀值参数。可调的运行时参数对多核机群下的MPI应用的性能有着重要的影响，但最优的运行时参数极大程度上依赖于多核机群的存储层次(包括节点内二级或三级缓存的共享方式等)、机群的网络互联方式(包括Infiniband网络、千兆以太网和Myrinet网络等)、机群的通信性能(包括内存和

网络的通信延迟与带宽)、机群内MPI应用的通信层次(包括Chip内、Chip间以及节点内通信)等因素。

[0007] 图1显示了在一个10节点,每节点8核的多核机群下五个运行时参数的不同配置组合对NAS并行基准套件中IS基准(Class B)的性能影响。在Infiniband互联的AMD双核10节点的机群下,最佳的运行时参数配置与Open MPI库默认设置相比可以带来最多约20%的性能提升,而错误的配置与默认配置相比造成约30%的性能损失。

[0008] 图2显示运行时参数对Jacobi基准的影响。实验显示在一个32核的AMD节点上且矩阵规模为4096*6096时,对于Jacobi基准,8个MPI进程时获得最大加速比的最优参数配置组合与16个MPI进程时不同(与默认配置相比)。同时实验结果也显示在8个MPI进程下,最优的MPI运行时参数可以给Jacobi基准带来约70%的性能提升。

[0009] 图1和图2说明可调的运行时参数可以对MPI应用带来可观的性能提升,但同时运行时参数的配置集合以及相应的优化空间相当庞大难以手工实现。以主流的基于模块组件结构的Open MPI为例,假设从常用的点到点通信的btl组件与集合操作的coll组件中各取一个可调的数值型参数及一个标志型参数,每个数 值型参数测试20种取值,每个标志型参数有2种取值,则使用自动迭代技术需要测试四个参数所构成1600种运行时参数的组合配置。以每种配置下MPI程序平均执行时间为5分钟计算,共需要5天时间来找到最佳的运行时参数序列。因此迫切需要一种快速自动的参数优化方法来提升多核机群下MPI应用的性能。

发明内容

[0010] 为达到上述目的,本发明提供了一种基于机器学习的多核下MPI最优运行时的参数预测方法。

[0011] 一种基于机器学习的多核下MPI最优运行时的参数预测方法,

[0012] 采用决策树和人工神经网络两种标准构建优化模型;

[0013] 用构造的训练基准在目标多核机群上通过设置多组运行时的参数的组合生成训练数据,并对构造的模型进行离线训练;

[0014] 训练后的模型用于对新的MPI程序预测最优的运行时的配置参数;

[0015] 将预测所得结果与实际最优运行时参数向量做对比,评估预测模式的准确度。

[0016] 优选的,所述决策树模型将训练基准的程序特征和运行时参数的配置组合作为决策树模型的输入,训练数据为:$\{F_i, C_i\}$,其中F_i为训练基准的程序特征,C_i为当前程序特征下的运行时参数组合,实际得到的加速比作为决策树的输出。

[0017] 优选的,所述人工神经网络模型将训练基准中产生最高加速比的数据选出用来训练参数预测模型,训练数据为:$\{F_i, C_{i\_best}\}$,其中$F_i = <f_1, f_2, \ldots, f_m>$为训练基准的程序特征,$C_{i_best} = <c_1, c_2, \ldots, c_n>$为当前程序特征下的最佳运行时参数组合。

[0018] 优选的,所述决策树模型在训练模型阶段,通过变换向量F与C产生不同的加速比结果;运用模型进行预测时,若F_p代表输入的MPI程序的程序特征向量,则能得到最大加速比S_{max}的运行时参数配置C_{best}将是此MPI程序的最佳运行时参数组合向量,即$S_{max} = M_{REPTree}(F_p, C_{best})$。

[0019] 优选的,所述人工神经网络模型在训练模型阶段,通过变换向量F与C产生不同的加速比结果;运用模型进行预测时,若M_{ANN}是训练后的人工神经网络模型,则$C_{best} = M_{ANN}(F_p)$,

其中F_p代表输入的MPI程序的程序特征向量，C_{best}是此 MPI程序的最佳运行时参数组合向量。

[0020] 优选的，所述训练基准包括两种MPI通信方式：同步的MPI点到点通信和MPI集合操作；训练基准接收5个参数，可以分别用来控制训练基准中点到点通信的比例、集合通信的比例、两个MPI进程同步点到点通信的消息的大小、集合通信中交换的消息大小以及通信器的大小。

[0021] 优选的，所述离线训练通过变换训练基准的5个输入参数，控制点到点通信和集合通信的比例分别为：100％的点到点通信、100％的集合通信、50％的点到点通信及50％的集合通信，在三种不同通信比例下，分别变换点到点以及集合通信中消息大小以及MPI通信器的大小，并变换运行时参数的配置组合，共产生训练数据3000条，用来训练神经网络优化模型。

[0022] 优选的，当预测模型建立并用大量的学习数据训练完毕后，就要根据实际需求来执行预测任务。

[0023] 优选的，在执行预测前，需要在目标多核机群下对要预测MPI程序的进行一次instrument运行，以得到输入的MPI程序的特征向量F_p；将F_p作为模型的输入即可得到输入MPI程序的最佳运行时参数组合；当目标多核机群发生变化时，以上过程需要重复进行。

[0024] 本发明提出了一种在多核环境下优化MPI应用的新方法：运用机器学习方法对多核机群下MPI应用的最优运行时参数进行预测。我们设计了具有不同点到点通信与集合通信数据比例的训练基准在特定的多核机群下产生训练数据，同时采用能快速输出结果的决策树REPTree和产生多个输出并具有较好抗噪性的神经网络ANN来构建运行时参数优化模型，通过训练基准产生的训练数据对优化模型进行训练，训练后的模型被用来对未知的输入MPI程序的最优运行时参数进行预测。实验证明，基于REPTree的预测模型和基于ANN的预测模型得到的优化运行时参数产生的加速比平均达到实际最大加速比的90％以上。

附图说明

[0025] 图1运行时参数对IS基准(Class B)的性能影响

[0026] 图2运行时参数对Jacobi基准(4096*6096)的性能影响

[0027] 图3基于机器学习的预测模型

[0028] 图4决策树预测模型

[0029] 图5神经网络预测模型

具体实施方式

[0030] 下面结合附图和具体实施例做进一步说明。

[0031] 可调的运行时参数对多核机群下的MPI应用的性能有着重要的影响，但最优的运行时参数依赖于多核机群的底层架构以及MPI程序自身的特征。本节我们介绍运用机器学习技术进行多核下MPI最优运行时参数预测的方法和步骤。

[0032] 我们的方法包括四个阶段：构造模型、模型训练、运用已训练模型进行参数预测以及模型预测准确度评估。其中第一阶段我们采用了两种标准的机器学习技术—决策树和人工神经网络被用来构建优化模型。模型训练阶段我们用构造的训练基准在目标多核机群上

通过设置多组运行时参数的组合生成训练数据，并对构造的模型进行离线训练。训练后的模型可以用来对新的未知的MPI程序预测最优的运行时参数配置。预测所得结果与实际最优运行时参数向量对比可以评估预测模式的准确度。

[0033] 机器学习的本质是应用计算机学习系统解决实际问题，基于机器学习的预测模型可以看作是一个映射或函数y＝F(X)，其中X是输入，而输出y是连续的或有序的值。学习的目的是得到一个映射或函数F，对X和y之间的联系建模。预测器的准确率通过对每个检验元组X，计算y的预测值与实际已知值的差来评估。

[0034] 由于手动调优多核下MPI运行时参数难以实现，因此我们采用基于机器学习的方法建立最优参数的预测模型，该模型可以对给定多核机群平台下任意未知的MPI输入程序的最佳运行时参数进行预测。

[0035] 图3描述了预测模型的工作过程，首先在目标多核机群系统上使用不同的运行时参数配置运行训练基准产生训练数据，用产生的训练数据对构造的预测模型进行离线训练，然后抽取给定的MPI程序的程序特性作为预测模型的输入，最后模型输出接近最优的运行时参数预测值，以获得接近最大加速比。基于机器学习的MPI最优运行时参数预测模型的公式形式可表示为：设M是训练后的预测模型，F＝＜f_1,f_2,...,f_i＞代表抽取输入的MPI程序的程序特征，则C＝M(F)所得向量C＝＜c_1,c_2,...,c_j＞是此程序的最佳运行时参数组合。

[0036] 决策树模型

[0037] 决策树是基于树形的预测模型，树的根节点是整个数据集合空间，每个分节点对应一个分裂问题，它是对某个单一变量的测试，该测试将数据集合空间分割成两个或更多数据块，每个叶节点是带有分类结果的数据分割。我们选择决策树来建立预测模型的原因是决策树在学习过程中不需要了解很多的背景知识，只从样本数据集提供的信息就能够产生一颗决策树，通过树节点的分叉判别可以使某一类分类问题仅与主要的树节点对应的变量属性取值相关，即不需要全部变量取值来判断对应的分类或执行预测。

[0038] 我们采用REPTree这种快速的决策树学习器来构建我们的决策树模型。REPTree使用错误约剪的树剪枝策略并且可以创建回归树，因此能有效的处理连续属性以及属性值空缺的情况。

[0039] 图4描述了我们的决策树预测模型。训练模型时我们将训练基准的程序特征和运行时参数的配置组合作为决策树模型的输入，即REPTree模型的训练数据为：{F_i,C_i}，其中F_i为训练基准的程序特征，C_i为当前程序特征下的运行时参数组合，实际得到的加速比作为决策树的输出。即我们将训练基准的程序特征、不同的运行时参数组合和实际得到的加速比作为建模REPTree的输入和输出，用来生成决策的If-then规则。通过样本数据集学习产生的决策树可以用来对抽取出程序特性的新的未知的MPI程序预测接近最优的运行时参数组合。

[0040] 对模型的公式化表述如下：设$M_{REPTree}$是决策树预测模型，则模型与训练数据及输出数据之间关系可定义为S＝$M_{REPTree}$(f_1,f_2,...,f_m,c_1,c_2,...c_n)，其中F＝＜f_1,f_2,...,f_m＞是程序的特征向量，C＝＜c_1,c_2,...,c_n＞是运行时参数组合向量，S是在输入为F与C时产生的实际加速比。训练模型阶段，我们通过变换向量F与C产生不同的加速比结果。运用模型进行预测时，若F_p代表输入的MPI程序的程序特征向量，则能得到最大加速比S_{max}的运行时参数配置C_{best}将是此MPI程序的最佳运行时参数组合向量，即S_{max}＝$M_{REPTree}$(F_p,C_{best})。

[0041] 神经网络模型

[0042] 人工神经网络ANN(Artificial Neural Network)是一类机器学习模型,可以映射一组输入参数到一组目标输出,我们采用ANN是由于它能很好应用于线 性和非线性回归问题并有很好的抗噪性。

[0043] 一个三层的前向型误差反传神经网络被用来构建预测模型,实验验证对我们预测问题性能最佳的ANN设计为:隐藏层的传输函数为正切(Sigmoid)函数:$f(n)=\frac{2}{(1+e^{-2n})}-1$,输出层的传输函数为对数正切函数(Logarithmic sigmoid):$f(n)=\frac{1}{1+e^{-n}}$,同时隐藏层有10个神经元,并且隐藏层的训练函数采用麦夸特(Levenberg-Marquardt)算法,因为它很好的结合了牛顿算法的速度与梯度下降算法的稳定性。

[0044] 图5描述了我们的神经网络预测模型。模型训练时,我们将训练基准中产生最高加速比的数据选出用来训练基于ANN的参数预测模型,即ANN模型的训练数据为:$\{F_i, C_{i\_best}\}$,其中$F_i=<f_1,f_2,...,f_m>$为训练基准的程序特征,$C_{i_best}=<c_1,c_2,...,c_n>$为当前程序特征下的最佳运行时参数组合。对应前文所描述的公式表示形式,设M_{ANN}是训练后的ANN模型,则$C_{best}=M_{ANN}(F_p)$,其中F_p代表输入的MPI程序的程序特征向量,C_{best}是此MPI程序的最佳运行时参数组合向量。

[0045] MPI程序特征抽取

[0046] 我们所设计的方法是用离线训练的优化模型对未知的MPI应用预测最佳运行时参数,因此要从未知的MPI程序中抽取适合的程序特征作为优化模型输入,用来得到准确的预测结果。由于运行时参数主要影响MPI进程间的通信性能,因此我们在进行特征抽取时,主要考虑MPI程序的通信模式、通信交换的数据量以及通信器的大小。表1说明了MPI程序的特性,这些必要的程序特性可以通过对要预测MPI程序的一次instrument运行得到。

[0047] 表1 MPI程序特性及描述

[0048]

特性名称	特性描述
点到点通信的时间比例	点到点通信在程序所有通信操作中所占的时间比例
点到点通信的数据量	所有进程间点到点通信所交换的平均数据量
集合通信的时间比例	集合通信在程序所有通信操作中所占的时间比例
集合通信的数据量	所有进程间集合通信所交换的平均数据量
通信器的大小	MPI 应用的通信器大小

[0049] 训练基准构造与训练数据生成

[0050] 为了产生训练预测模型的数据,我们设计了训练基准程序。在目标体系结构的多核机群上对训练基准使用可调运行时参数的多种不同组合可以产生训练数据。同时,训练基准可接受多个输入参数来控制训练基准中点到点、集合操作传输的数据量以及通信器大小。

[0051] 我们根据表1所定义的MPI程序特征来设计训练基准。基准主要包括以下两种MPI

通信方式：同步的MPI点到点通信和MPI集合操作。训练基准接收5个参数，可以分别用来控制训练基准中点到点通信的比例、集合通信的比例、两个MPI进程同步点到点通信的消息的大小、集合通信中交换的消息大小以及通信器的大小。

[0052] 通过变换训练基准的5个输入参数，控制点到点通信和集合通信的比例分别为：100％的点到点通信、100％的集合通信、50％的点到点通信及50％的集合通信。在三种不同通信比例下，分别变换点到点以及集合通信中消息大小以及MPI通信器的大小，并变换运行时参数的配置组合，共产生训练数据3000条，用来训练神经网络优化模型。

[0053] 执行预测

[0054] 当预测模型建立并用大量的学习数据训练完毕后，就要根据实际需求来执行预测任务。由于我们的决策树预测模式$S_{max}=M_{REPTree}(F_p,C_{best})$和神经网络预测模型$C_{best}=M_{ANN}(F_p)$中，都需要用待预测程序的程序特征向量$F_p$来作为输入，因此在执行预测前，我们需要在目标多核机群下对要预测MPI程序的进行一次instrument运行，以得到输入的MPI程序的特征向量F_p。将F_p作为模型的输入即可得到输入MPI程序的最佳运行时参数组合。但当目标多核机群发生变化时，以上过程需要重复进行。

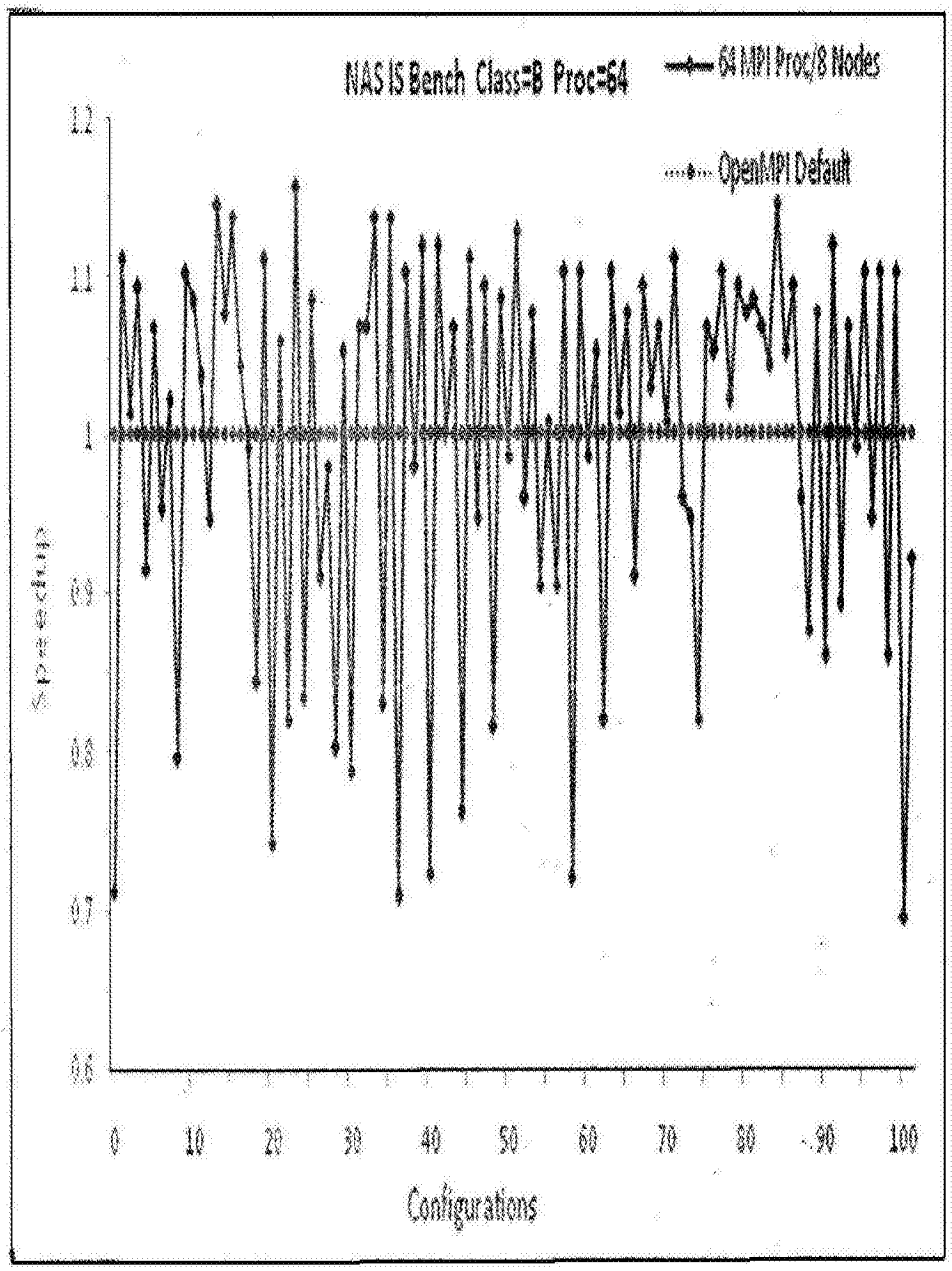

NAS IS Bench Class=B Proc=64
64 MPI Proc/8 Nodes
OpenMPI Default
1.2
1.1
1
0.9
0.8
0.7
0.6
0
10
20
30
40
50
60
70
80
90
100
Configurations

图1

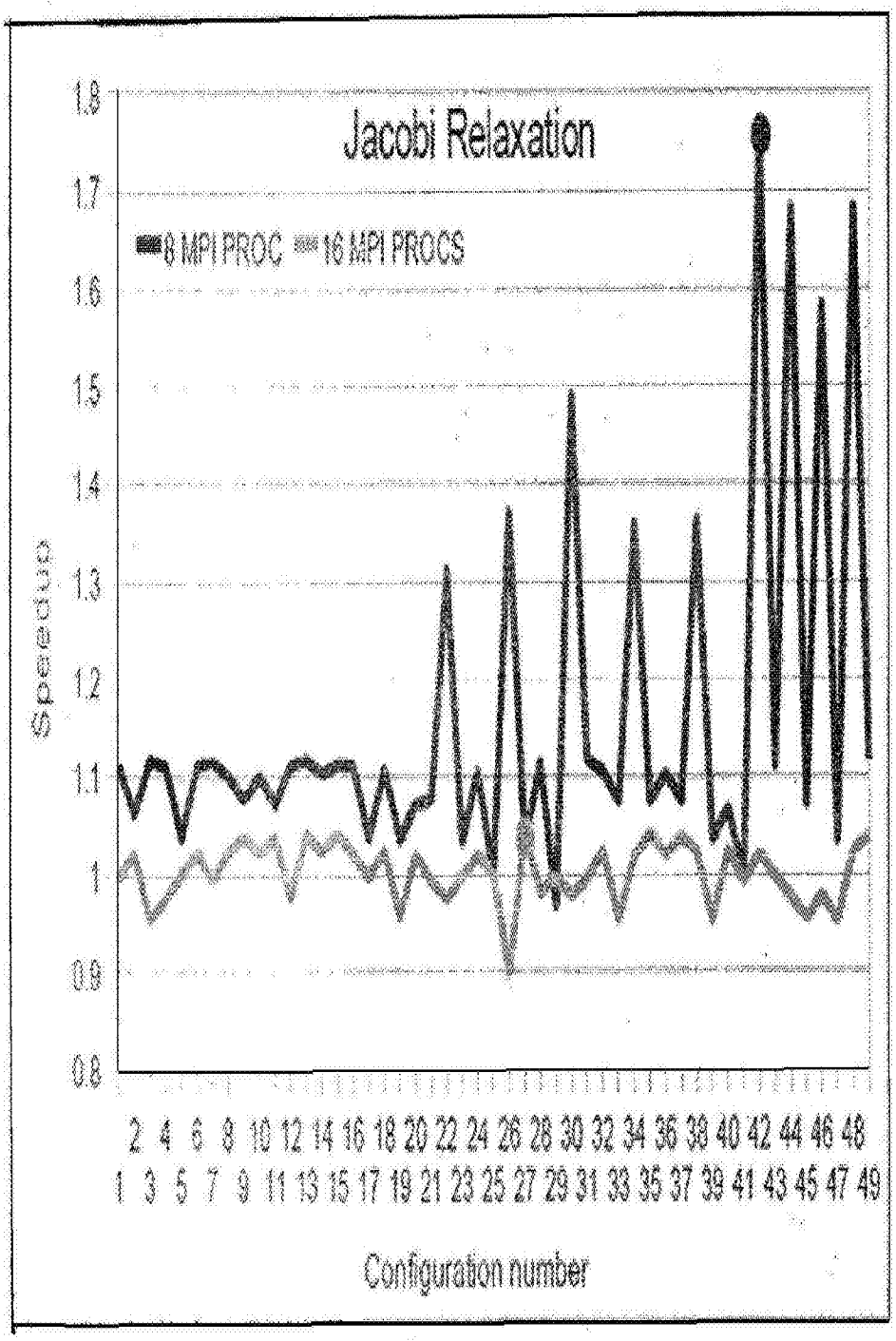

Jacobi Relaxation
8 MPI PROC
16 MPI PROCS
Speedup
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
Configuration number

图2

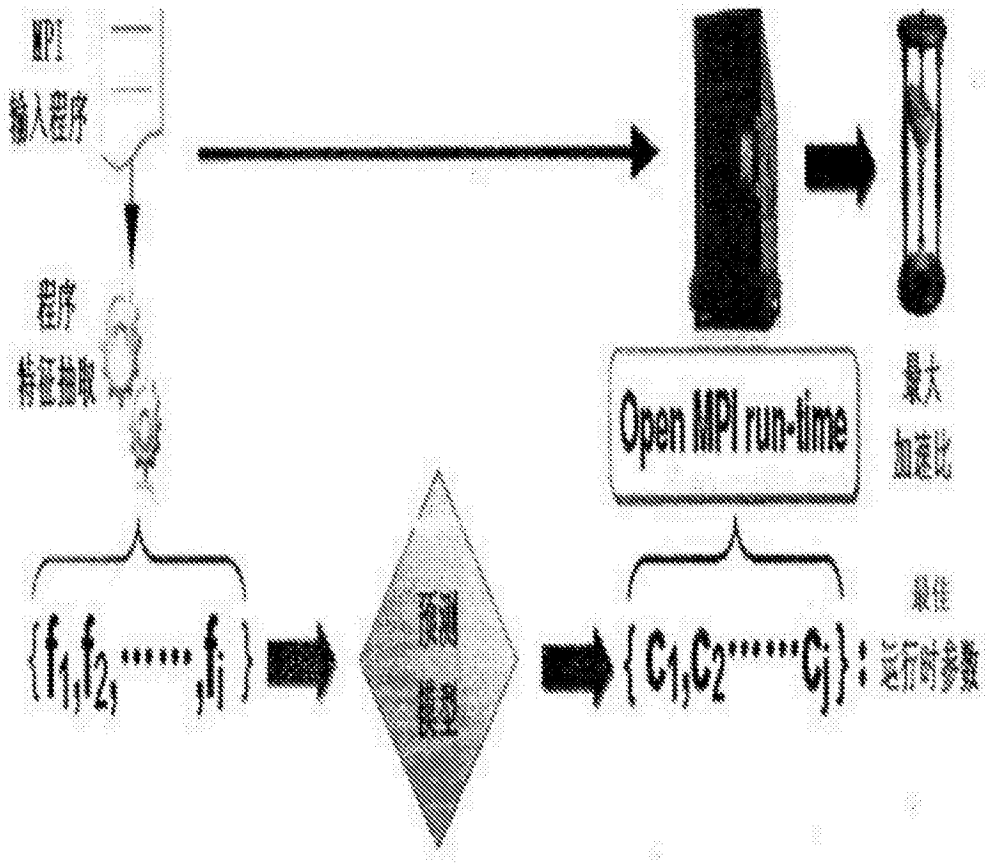


图3

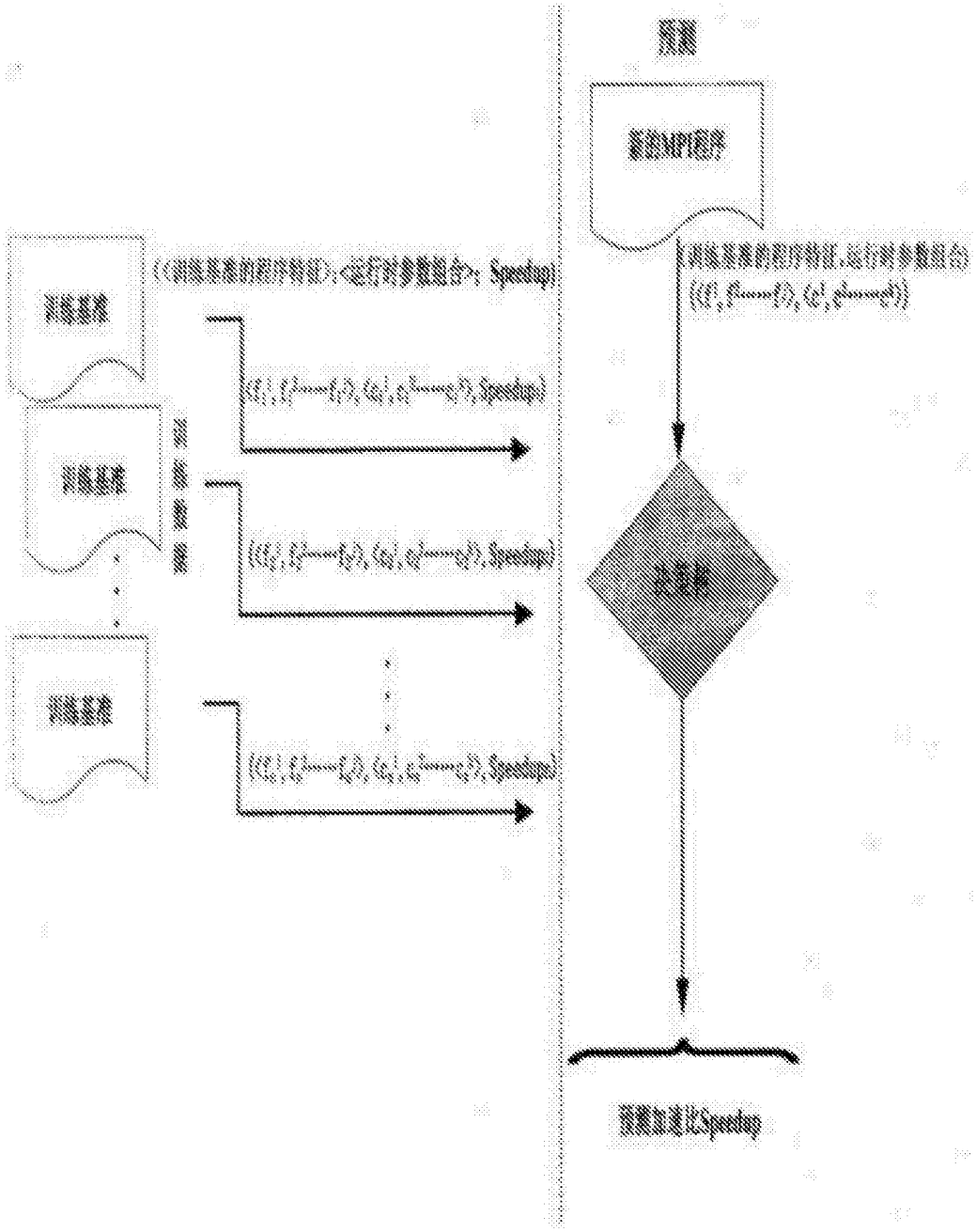


图4

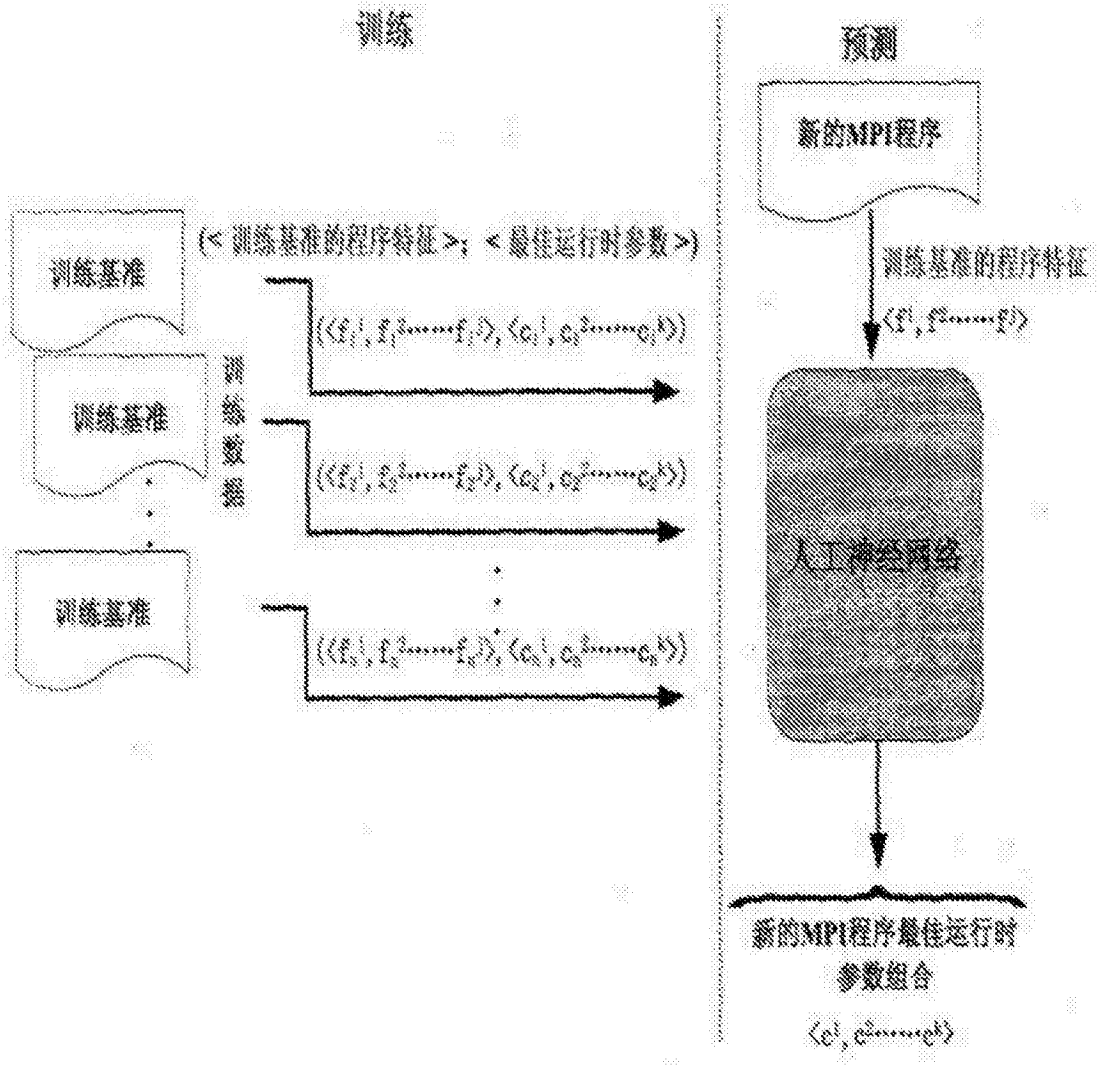


图5