

**学士学位论文开题报告**

**论文题目**：基于拓扑映射的并行计算通信性能优化

**专 业**：计算机科学与技术

**本 科 生**：汪圣灵

**学 号**：15061014

**指导教师**：肖利民

**北京航空航天大学计算机学院**

2018年10月16日

目 录

[1题目背景与意义 3](#_Toc527623432)

[1.1课题来源 3](#_Toc527623433)

[1.2选题的背景与意义 3](#_Toc527623434)

[2研究现状 4](#_Toc527623435)

[2.1MPI集合通信研究现状 4](#_Toc527623436)

[2.1.1 MPI Process Placement toolset（MPIPP）[6] 4](#_Toc527623437)

[2.2并行计算通信优化研究现状 4](#_Toc527623438)

[2.2.1 TreeMatch[7] 4](#_Toc527623439)

[2.2.2 SCOTCH[8] 4](#_Toc527623440)

[2.2.3 METIS[9] 4](#_Toc527623441)

[2.2.4 Topo LB[10] 4](#_Toc527623442)

[2.2.5 Fast Approximate Quadratic（FAQ）[11] 5](#_Toc527623443)

[2.3尚且存在的问题： 5](#_Toc527623444)

[3研究内容与解决方案 5](#_Toc527623445)

[3.1研究目标 5](#_Toc527623446)

[3.2研究内容 6](#_Toc527623447)

[3.2.1基于拓扑映射的并行计算性能优化算法实现 6](#_Toc527623448)

[3.2.2设计基于并行计算性能优化算法的网络拓扑图生成系统 6](#_Toc527623449)

[3.3解决方案 6](#_Toc527623450)

[3.3.1. 基于拓扑映射的并行计算性能优化算法实现与验证 6](#_Toc527623451)

[3.3.2. 设计基于并行计算性能优化算法的网络拓扑图生成系统 7](#_Toc527623452)

[4关键技术及难点 7](#_Toc527623453)

[4.1数据结构 7](#_Toc527623454)

[4.2结果测试 7](#_Toc527623455)

[5进度安排 7](#_Toc527623456)

[6参考文献 8](#_Toc527623457)

# 1题目背景与意义

## 1.1课题来源

实验室

## 1.2选题的背景与意义

随着科技的不断发展，越来越多的行业都融入到了互联网中，工业4.0、人工智能、大数据等新技术的提出对计算能力产生了巨大的需求，传统的单个主机由于其处理器计算能力以及磁盘存储能力的局限性，已经不能够适应如此大的任务量和数据量。纵观CPU体系架构的发展趋势，单个CPU的性能由于功耗和散热问题已经达到了一个瓶颈。于是，对高性能的需求促进了多计算机协同计算的发展，从现在的计算能力发展来看，仍然遵守着著名的摩尔定律在不断的提升，从“天河二号”到“神威·太湖之光”，其中发挥着重要作用的便是服务于多核计算、网格计算、云计算的并行计算。对并行计算的性能提升成为了当前研究的主要课题。

而针对不同的运算环境，如单台服务器，并行计算通常使用OpenMP技术来形成多线程程序以充分利用多核多线程的并行处理能力，但当计算的规模超过了单个节点的处理能力时，则需要多台机器同时进行计算，使用的技术则为MPI技术，即通信接口技术（Message-Passing Interface）[1]。此时所牵扯到的问题便是多台电脑并行计算时的数据通信问题，在理想状态下，并行计算机的性能应该是所有节点计算机的性能之和，但是，这种理想状态是假设任意两台计算机之间的通信距离为0，即所有计算机在进行并行计算时通信时间都为0。其中，影响应用程序计算性能发挥的主要瓶颈之一是消息传递通信[2]。但是，并行计算机是由一组处理单元组成的，这组处理单元通过相互之间的通信与协作，以更快的速度共同完成一项大规模的计算任务。因此，并行计算机的两个最主要的组成部分是计算节点和节点间的通信与协作机制[3]。因此，要提高并行计算能力，减少通信时间起着极其重要的作用[4][5]。

# 2研究现状

## 2.1MPI集合通信研究现状

### 2.1.1 MPI Process Placement toolset（MPIPP）[6]

MPIPP是获取MPI应用通信配置的工具，也是获取目标集群的网络拓扑图工具，是一种比现有的多集群图形划分算法更有效的寻找最佳映射的算法，并且已经以工具的形式被实现。

## 2.2并行计算通信优化研究现状

### 2.2.1 TreeMatch[7]

TreeMatch是一种寻找在分布式内存托和并行计算机的计算单元上进程的映射方式的算法 第一次发表在

### 2.2.2 SCOTCH[8]

SCOTCH采用图论的方法研究静态映射，采用分而治之的方法来解决，当假设目标机器有一个完整图形形状的通信网络时，如何在合理的时间内找到次优解的算法问题。

### 2.2.3 METIS[9]

METIS是图形划分程序，能够将图的定点划分为P个大致相等的部分，使得连接不同部分顶点的边的数量最小。且具有速度快、划分质量高等优点

### 2.2.4 Topo LB[10]

Topo LB是在动态负载均衡框架下的二阶近似拓扑映射算法

### 2.2.5 Fast Approximate Quadratic（FAQ）[11]

快速近似二次分配算法：图匹配算法，能够准确高效的匹配出大数据中常见的大图形

## 2.3尚且存在的问题：

由上面的算法可以看出，部分算法仅仅只是被提出，但却没有落实被实现成可用的工具，如Topo LB，实际的开发人员或者试验人员如物理学家、生物学家，在使用过程中，无法将这些算法具体实现出来，也就无法在实验中使用和验证。还有一部分算法虽然实现出来了，如FAQ，但是却没有形成一个系统，可拓展性和可移植性缺失，在实际运用过程当中，使用人员需要花费大量的精力来移植算法或者重写算法，仍然无法满足实际需求。另一部分算法则是以程序的形式呈现，如METIS和MPIPP，没有开源代码，仅局限于部分特定领域使用，在其他领域或者环境下则无法使用其程序来进行实验操作。

因此，如果能设计出一个实现多个算法的系统，提供网络节点等配置信息，选择一个算法并输出最优的网络结点拓扑图，便能够极大的方便使用并行计算时需要构建的虚拟网络拓扑结构。

# 3研究内容与解决方案

## 3.1研究目标

本毕设针对现有的拓扑映射算法，设计并实现一个系统，能够通过算法，以网络节点和其他配置信息生成网络拓扑图，优化并行计算中各个计算机之间的通信性能，从而达到提高并行计算的性能的目的。

## 3.2研究内容

### 3.2.1基于拓扑映射的并行计算性能优化算法实现

并行计算的基本原则包含：与体系结构相结合，具有可拓展性，粗粒度，减少通信，优化性能[3]。而减少通信的一个方面在于减少计算机之间的通信距离，本毕设在并行计算性能优化方面的主要研究内容是找到一些已经验证过高效性和可用性的算法并进行实现。

### 3.2.2设计基于并行计算性能优化算法的网络拓扑图生成系统

本毕设的主要研究内容在于设计一种合理的，实现并行计算性能优化算法的网络拓扑图生成系统，使得在给出计算节点和配置信息时，能够选择一种算法，得出传输性能最优的网络拓扑图。

## 3.3解决方案

本项目研究方案包括以下两点：

### 3.3.1. 基于拓扑映射的并行计算性能优化算法实现与验证

算法实现：

本项目从MPIPP，TopoLB，FAQ，METIS中选取部分算法，根据论文描述实现算法程序，并以文件的形式产生输入规则和输出规则，对特定的算法，输入指定的计算节点数据文件，产生网络拓扑结构图。

算法验证：

(1). 有效性验证：

按照生成的拓扑结构图布置超级计算机的计算节点，并运行并行计算程序，检测程序的运行有效性。

(2).对比试验：

分别按照本项目算法和实验室已有算法生成的拓扑结构图布置超级计算机的计算节点，并多次运行同一个并行计算程序，运算时间取平均值，比较每个算法生成的拓扑图的运行时间。

### 3.3.2. 设计基于并行计算性能优化算法的网络拓扑图生成系统

系统采用python来实现，根据选择的算法，将读取的数据文件分析并按照算法计算出最优的拓扑结构图，并输出为文件。

(1). 系统有效性检测

输入指定的数据文件，选择每种算法一次，检测程序是否正常运行。检测输出文件是否符合输出格式和是否正确

(2). 系统界面优化

将系统的交互式界面进行优化，体现较好的用户友好性

# 4关键技术及难点

## 4.1数据结构

拓扑映射算法有很多种，不同的模型适用于不同的场景，也会有不同的输入数据和参数选择，要集成到一个系统中并通过适配调整数据结构会是一个难点。

## 4.2结果测试

算法的运行结果需要借助其他程序和硬件进行测试，如何构建一套合理而准确的测试过程会是一个难点

# 5进度安排

|  |  |
| --- | --- |
| xxxx年xx月xx日—xxxx年xx月xx日 | 内容 |
| xxxx年xx月xx日—xxxx年xx月xx日 | 内容 |
| xxxx年xx月xx日—xxxx年xx月xx日 | 内容 |
| xxxx年xx月xx日—xxxx年xx月xx日 | 内容 |
| xxxx年xx月xx日—xxxx年xx月xx日 | 内容 |

# 6参考文献

|  |  |
| --- | --- |
|  | Li Q, Huo Z, Sun N. Optimizing MPI Alltoall Communication of Large Messages in Multicore Clusters[C]// International Conference on Parallel and Distributed Computing, Applications and Technologies. IEEE, 2011:257-262. |
|  | 罗红兵, 张晓霞. MPI集合通信性能可扩展性研究与分析[J]. 计算机科学与探索, 2017, 11(2):252-261. |
|  | 迟学斌, 赵毅. 高性能计算技术及其应用[J]. 中国科学院院刊, 2007, 22(4):306-313. |
|  | 刘青昆, 王佳, 韩颖,等. 一种并行计算通信优化策略[J]. 辽宁工程技术大学学报(自然科学版), 2011, 30(2):268-271. |
|  | Guo Q, Paker Y. Concurrent communication and granularity assessment for a transputer-based multiprocessor system[M]. CRL Publishing Ltd. 1990. |
|  | Chen H, Chen W, Huang J, et al. MPIPP:an automatic profile-guided parallel process placement toolset for SMP clusters and multiclusters[C]// International Conference on Supercomputing. ACM, 2006:353-360. |
|  | Jeannot E, Mercier G. Near-Optimal Placement of MPI Processes on Hierarchical NUMA Architectures[M]// Euro-Par 2010 - Parallel Processing. Springer Berlin Heidelberg, 2010:199-210. |
|  | Pellegrini F, Roman J. Scotch: A software package for static mapping by dual recursive bipartitioning of process and architecture graphs[J]. 1996, 1067(1067):493-498. |
|  | G. Karypis and V. Kumar. MiETiS - Unstructured Graph Partitioning and SparseMatrix Ordering System - Version 2.0. University of Minnesota, jun 1995. |
|  | Agarwal T, Sharma A, Laxmikant A, et al. Topology-aware task mapping for reducing communication contention on large parallel machines[C]// International Conference on Parallel and Distributed Processing. IEEE Computer Society, 2006:145-145. |
|  | Vogelstein J T, Conroy J M, Lyzinski V, et al. Fast approximate quadratic programming for graph matching.[J]. Plos One, 2015, 10(4):e0121002. |