

**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 4](#_Toc527930344)

[2研究现状 5](#_Toc527930345)

[2.1基于拓扑映射的任务布局算法研究现状 5](#_Toc527930346)

[2.1.1 节点间映射 6](#_Toc527930347)

[2.1.2 节点内映射 6](#_Toc527930348)

[2.2图划分算法以及工具 6](#_Toc527930349)

[2.2.1 Fast Approximate Quadratic（FAQ）[16] 7](#_Toc527930350)

[2.2.2 METIS[14] 7](#_Toc527930351)

[2.2.3 Chaco[11] 7](#_Toc527930352)

[2.2.4 SCOTCH[13] 7](#_Toc527930353)

[2.3尚且存在的问题 7](#_Toc527930354)

[3研究内容与解决方案 8](#_Toc527930355)

[3.1研究目标 8](#_Toc527930356)

[3.1.1调查研究现有高效的拓扑映射方法 8](#_Toc527930357)

[3.1.2实现几种拓扑映射算法 8](#_Toc527930358)

[3.1.3设计并实现原型系统 9](#_Toc527930359)

[3.2研究内容 9](#_Toc527930360)

[3.2.1调查研究现有的高效的拓扑映射方法 9](#_Toc527930361)

[3.2.2基于拓扑映射的并行计算性能优化算法实现 9](#_Toc527930362)

[3.2.3设计基于拓扑映射算法的并行计算通信性能优化系统 9](#_Toc527930363)

[3.3解决方案 10](#_Toc527930364)

[3.3.1. 调查研究现有的高效的拓扑映射方法 10](#_Toc527930365)

[3.3.2. 基于拓扑映射的并行计算性能优化算法实现与验证 10](#_Toc527930366)

[3.3.3. 设计基于拓扑映射算法的并行计算通信性能优化系统 10](#_Toc527930367)

[4关键技术及难点 11](#_Toc527930368)

[4.1关键技术 11](#_Toc527930369)

[4.1.1拓扑映射算法 11](#_Toc527930370)

[4.1.2Python程序设计与实现 12](#_Toc527930371)

[4.2技术难点 12](#_Toc527930372)

[5进度安排 12](#_Toc527930373)

[6参考文献 12](#_Toc527930374)

# 1题目背景与意义

随着科技的不断发展，高性能计算集群（High Performance Computing Cluster）得到了广泛的应用。无论是工业领域还是科研领域，如CAE（汽车、航空航天、船舶）和EDA（芯片设计单位），以及计算物理/化学/材料/气象等环境下，用计算方法模拟和分析现实应用场景等技术都取得了巨大的进展，而这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，而高性能计算系统则拥有多处理器共同完成一个任务、多用户、多作业等特点，是一种支持计算与IO密集型应用的多服务器计算机集群系统。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高性能计算的发展，从“天河二号”到“神威·太湖之光”，从云计算到GPGPU，都在使用并行计算处理巨大规模的计算问题。因此，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[1]或者Pthread[2]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，OpenMP技术根植于编译器，通过将原来串行的程序通过加入适当的编译器指令，编程并行执行从而提高程序的运行速度。而Pthread则是一套基于POSIX对线程的标准而设计的关于线程的API，提供多线程开发的库，实现多线程编程来提高程序运行速度。但另一方面，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）[3]。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开销都理想化为无开销。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[4]。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[5]。因此，要提高并行计算能力，减少通信开销起着极其重要的作用[6][7]。

而另一方面，随着网络技术的发展，分布式计算为高性能计算开辟了新的领域，人民可以使用并行计算机以便在更长和更精确的时间尺度上解决更大的问题。然而，随着应用程序并发性和输入数据大小的预期增加，未来几年要解决的最重要挑战之一是局部性，即如何改进应用程序内的数据访问和传输[8]，一个分布式系统由多台同构或异构的计算机、高速网络、接口、操作系统、通信协议和程序设计环境组成，而分布式计算则需要将一个计算问题划分成多个任务，即任务布局，并分配给各个计算节点进行并行计算。为了提高并行计算的速率，则必须将一个计算问题尽可能的划分为均匀分布在各台计算机上的单个任务进行处理，充分利用分布式系统中所有的计算资源。但如此也增加了各个任务之间的通信量，因此，如何平衡多任务平均划分和通信量两个互相矛盾的问题以达到最佳性能，便是任务到处理机的映射问题[9]。目前高性能计算机中普遍使用的任务布局算法包括EFG法以及其拓扑优化、MPIPP、METIS等，这些算法各自都在某一个方面解决了部分问题，但却存在着集成性高、可移植性弱、源码不开源等问题，因此，本课题的最终研究目的便是通过设计一个实现算法的系统来解决用户友好性问题。

# 2研究现状

## 2.1基于拓扑映射的任务布局算法研究现状

将并行应用程序流程映射到高性能计算机系统的物理处理器上的问题可以被视为图形映射问题[18]，且此问题的目的为查找最小化某评估标准问题，一般来说，找到最佳映射是NP问题[19]，其中图形映射指的是应用的通信模式的有向图到计算平台的物理拓扑的无向图的映射，而映射的质量可以通过若干指标来进行测量，而通常情况下，以hop-bytes作为评估指标更容易计算且能反映实际通信性能[20]。hop-bytes表示互联网的流量总负载，即处理器之间最短路径加权的通信总量。

为解决最佳映射问题，此前已经有大量的算法和方案被提出，按照映射对象主要分为节点间映射和节点内映射：

### 2.1.1 ~~节点间映射~~

节点间映射也包括两个大类，一类是物理优化技术，包括模拟退火算法[21]，遗传算法[22]，图形收缩算法[23]，粒子群优化算法[24]。一类是启发式算法，面对不同的应用场景也有各种不同的算法，其中对于静态映射问题的代表性算法之一便是递归二分算法，它已经在软件包SCOTCH中被实现，并被证明是一种推导静态映射的有效算法[25]。对于SMP集群和多集群的分层通信体系结构，图分区算法更适合来寻找最优映射，图分区算法和工具后面将会提及，而基于图分区算法的映射算法MPIPP[10]提供一种配置文件引导方法，可以自动查找优化的映射，从而最大限度的降低消息传递的通信开销。对于大规模绘图问题，chung等人提出的分层映射算法，能够使用分而治之的方式执行映射[26]。

此外，为了在实际的某些特定环境下能够方便使用，Yu等人开发出了用于BG/L超级计算机支持MPI拓扑功能的拓扑映射库[27]，Hoefler等人开发出了能够自动提取通信模式和程序集等关键模式的LibTopoMap拓扑映射库[15]，其中包含的算法包括贪心启发式算法、递归分区方法和基于RCM排序的映射策略。

### 2.1.2 ~~节点内映射~~

节点内映射，现阶段被认为是最有效的启发式算法为TreeMatch[12]，当一个高性能计算机的体系结构中的节点具有多核或者多核处理器时，开发对应的并行计算程序就需要考虑到底层的硬件特性，如存储器的层次机构。而TreeMatch便是通过将通信的模式和底层的硬件架构相匹配来优化应用程序进程之间的数据传输，从而降低通信开销的任务布局算法。

## 2.2图划分算法以及工具

图分区算法在很多拓扑映射算法中都需要用到，包括MPIPP、递归二分算法等，其中的一些图分区算法也已经集成到了工具包中，方便在设计和使用拓扑映射算法时直接调用。

### ~~2.2.1 Fast Approximate Quadratic（FAQ）[16]~~

任务布局中拓扑图的匹配问题是二次分配问题的一个个例，而在此类图匹配问题（GMP）中需要解决的是准确性和效率的平衡问题，较慢的算法可以实现更好的性能时间。而快速近似二次分配算法（FAQ）实现了准确性和效率之间的最佳可行权衡。

### 2.2.2 METIS[14]

图分区在任务布局优化过程中有着广泛的应用，是指将图的定点划分为大致相等的部分，使得连接不同部分的顶点的边的数量最小。METIS是图形划分程序，能够完成图分区算法功能。且具有速度快、划分质量高等优点。

### 2.2.3 Chaco[11]

在图分区的应用上，Chaco是Python上的交互式图表2D绘图库，其中集成了图分区算法，允许用户递归应用几种方法来在加权图中查找小边缘分隔符，包括inertial、spectral、Kernighan–Lin、multilevel methods以及其他更简单的划分策略。

### 2.2.4 SCOTCH[13]

将并行程序的子进程分配给高性能计算机以最小化整体运行时间的优化问题称为映射，而当此类优化在程序执行之前就已经完成并且在程序运行过程中不再修改，则为静态映射。SCOTCH便是一种基于源过程图和目标体系结构图的递归二分法静态映射软件包，该软件包主要实现了将任何加权源图映射到任何加权目标图形上的功能，以及图形双分区启发式算法等。

## 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，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，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，没有开放算法源代码，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。

另一方面，作为一项实际应用于并行计算通信性能优化的任务布局技术，需要能够适用于大部分的程序集和高性能计算机集群，且能适用于各种通信模式，还要有自动选择特定算法或者最优算法的功能。前篇提到的各类算法和系统，都难以从全局上满足这些功能。

# 3研究内容与解决方案

## 3.1研究目标

### 3.1.1调查研究现有高效的拓扑映射方法

本项目将通过调查和研究相关文献资料和算法程序，进行理论分析和验证分析，理解算法的功能原理，并总结出各个算法的特性，包括算法对应适用的程序集、硬件集、通信模式等关键特性。最后通过理论验证算法的优化效率，通过对比找出现有的高效率任务布局策略或者算法。

### 3.1.2实现几种拓扑映射算法

本项目将针对调查研究的结果，对现有的高效率的拓扑映射算法进行编程设计，设计出算法的流程图、数据结构以及外部运行环境，实现出算法的功能特性，并通过编程方式实现算法程序，并设计出相应的测试程序和测试数据进行有效性验证。

### 3.1.3设计并实现原型系统

本课题针对以上已实现的拓扑映射算法程序，设计系统的流程图、输入输出的数据结构以及系统运行环境，其中流程图包括算法选择的自动化流程以及算法流程。并最终通过编程方式实现一个系统，能够通过算法，以并行程序集、通信模式、计算节点拓扑图为计算数据，生成一个并行程序集到计算节点图的最优任务布局映射方案。

## 3.2研究内容

### 3.2.1调查研究现有的高效的拓扑映射方法

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[5]。而减少通信的一个方面在于减少计算机之间的通信距离，本课题在并行计算性能优化方面的主要研究内容是通过研究和调查找到一些已经验证过高效性和可用性的算法，并分析和总结出算法的特性，包括适用的程序集、硬件集、通信模式等关键特性。

### 3.2.2基于拓扑映射的并行计算性能优化算法实现

本课题在此阶段的项目内容主要是研究如何通过编程方式，将调查研究后的具有高效性的拓扑映射算法进行设计和实现。并进行有效性验证。其中对算法进行设计时，需要设计算法的流程图、数据结构以及运行环境，在实现过程中，需要搭建相应的编程环境和测试环境，在有效性验证过程中，需要设计测试数据和测试程序，并搭建相应的测试环境。

### 3.2.3设计基于拓扑映射算法的并行计算通信性能优化系统

本课题在此阶段的主要研究内容在于设计一种能够通过算法，以并行程序集、通信模式、计算节点拓扑图为计算数据，生成一个并行程序集到计算节点图的最优任务布局映射方案的系统。系统需要至少实现一种算法，并能够自动根据配置信息调整布局方案。

## 3.3解决方案

本项目研究方案包括以下三点：

### 3.3.1. 调查研究现有的高效的拓扑映射方法

阅读相关领域论文，包括高性能计算、并行计算、拓扑映射、通信性能等，并进行理论分析和验证分析，对比理论结果和验证结果，得出现有的具有高效性的拓扑映射算法。

### 3.3.2. 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：

本项目针对调查研究的内容和不同的应用场景选择不同的算法，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。

### 3.3.3. 设计基于拓扑映射算法的并行计算通信性能优化系统

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。



图 1 系统结构图

(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关键技术

### 4.1.1拓扑映射算法

并行机器上并行程序的有效执行要求将程序的通信过程分配给机器的处理器，以便最小化其整体运行时间。当假定进程在所有程序的持续时间内同时共存时，这种优化问题称为映射。它相当于平衡机器处理器之间进程的计算权重，同时通过在附近的处理器上保持密集的通信进程来减少并行性引起的通信开销。如果映射在执行程序之前计算并且从不在运行时修改，则称为静态。在一般情况下，静态映射是NP完全的[17]。

### 4.1.2Python程序设计与实现

Python是当前环境下主流的编程语言之一，具有高可读性、高一致性、高质量等特性，能够显著提升程序开发效率。Python开发的程序具有可移植性，Python还提供了大量的标准库，其中包含众多应用级编程库，大大降低了开发难度。在进行算法实现过程中，Python的标准库能够极大的减少系统开发时间，能够使研究的主要重心放在算法实现上而不是代码编写上。

## 4.2技术难点

拓扑映射算法有很多种，无论是传统的还是最新的，不同的算法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，在进行理论分析和验证分析时，需要综合参考各项参数，而如何寻找出具有高效性的拓扑映射算法将会是一个难点。

拓扑映射算法部分仍然停留在理论层次，没有具体的代码实现，而本课题需要通过分析和验证算法，并使用编程方法将算法设计和实现出来，将会使一个难点。

# 5进度安排

表格 1进度安排

|  |  |
| --- | --- |
| 2018年10月20日—2018年10月31日 | 开题报告撰写 |
| 2018年11月01日—2018年11月30日 | 算法调查和研究 |
| 2018年12月01日—2018年12月31日 | 算法程序实现 |
| 2019年01月01日—2019年02月31日 | 算法系统实现 |
| 2019年03月01日—2019年04月01日 | 论文撰写 |

# 6参考文献

|  |  |
| --- | --- |
|  | Chandra R, Dagum L, Kohr D, et al. Parallel programming in OpenMP[M]. Morgan Kaufmann Publishers, 2001. |
|  | Buttlar D, Farrell J, Nichols B. PThreads Programming[J]. Oreilly Media, 1996. |
|  | 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. |
|  | Jeannot E, Mercier G, Tessier F. Process Placement in Multicore Clusters:Algorithmic Issues and Practical Techniques[J]. IEEE Transactions on Parallel & Distributed Systems, 2014, 25(4):993-1002. |
|  | 胡金初. 并行计算中的任务分配算法[C]// 全国理论计算机科学学术年会. 2005. |
|  | 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. |
|  | B. Hendrickson and R. Leland, “The Chaco User’s Guide: Version 2.0,” Technical Report SAND94–2692, Sandia Nat’l Laboratory, 1994 |
|  | 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. |
|  | Garey M R, Johnson D S. Computers and Intractability: A Guide to the Theory of NP-Completeness[M]. W.H. Freeman and company, 1979. |
|  | Chan, M.Y, Chin, F.Y.L. On embedding rectangular grids in hypercubes[J]. Computers, IEEE Transactions on, 1988, 37(10):1285-1288. |
|  | Bokhari S H. On the Mapping Problem[J]. IEEE Transactions on Computers, 2006, C-30(3):207-214. |
|  | Bhatele A. Automating topology aware mapping for supercomputers[M]. University of Illinois at Urbana-Champaign, 2010. |
|  | Lee C, Bic L. On the mapping problem using simulated annealing[M]. 1989. |
|  | Chockalingam T, Arunkumar S. A randomized heuristics for the mapping problem: The genetic approach[J]. Parallel Computing, 1992, 18(10):1157-1165. |
|  | Berman F, Snyder L. On mapping parallel algorithms into parallel architectures ☆[J]. Journal of Parallel & Distributed Computing, 1987, 4(5):439-458. |
|  | Salman A, Ahmad I, Al-Madani S. Particle swarm optimization for task assignment problem[J]. Microprocessors & Microsystems, 2002, 26(8):363-371. |
|  | Ercal F, Ramanujam J, Sadayappan P. Task allocation onto a hypercube by recursive mincut bipartitioning[J]. Journal of Parallel & Distributed Computing, 1990, 10(1):35-44. |
|  | I-HSIN CHUNG, CHE-RUNG LEE, JIAZHENG ZHOU, et al. HIERARCHICAL MAPPING FOR HPC APPLICATIONS[J]. Parallel Processing Letters, 2011, 21(03):279-299. |
|  | Yu H, Chung I H, Moreira J. Topology mapping for Blue Gene/L supercomputer[C]// ACM/IEEE Conference on Supercomputing. ACM, 2006:116. |