

面向任务的 TBB 多核集群混合并行编程模型

顾 慧¹, 郑晓薇², 张建强², 吴华平²

(1 内蒙古师范大学 网络技术学院, 内蒙古 呼和浩特 010022; 2 辽宁师范大学 计算机与信息技术学院, 山东 大连 116081)

摘 要: 构建了一种适用于多核集群的混合并行编程模型. 该模型融合了共享内存的面向任务的 TBB 编程和基于消息传递的 MPI 编程两种模式. 结合两者的优势, 实现进程到处理节点和进程内线程到处理器核的两级并行. 相对于单一编程方式下的程序性能, 采用这种混合并行编程模型的算法不但可以减少程序执行时间, 获得更好的加速比和执行效率, 而且明显地提高了集群性能.

关键词: 多核集群; MPI; TBB; 面向任务; 混合编程

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2011)02-0091-03

TBB Task-oriented Mixed-parallel Programming Model for Multi-Core Cluster

GU Hui¹, ZHENG Xiao-wei², ZHANG Jian-qiang², WU Hua-ping²

(1 College of Network Technology, Inner Mongolia Normal University, Hohhot 010022, China;

2 College of Computer and Information Technology, Liaoning Normal University, Dalian 116081, China)

Abstract: To take full advantage of the structural characteristics of multi-core clusters and enhance the use of CPU each core efficiently, a multi-core cluster hybrid parallel programming model is constructed. The model combines the shared memory and the TBB task-oriented programming and MPI message-passing programming mode. Using their advantages, achieve the hierarchical parallel programming of processes to nodes and threads to the processor core. Compare with the algorithm of the single parallel programming model, this programming mode not only can reduce the program execution time, get a better speedup and execution efficiency, but also improve the cluster performance greatly.

Key words: multi-core clusters; MPI; TBB; task-oriented; hybrid programming

1 引言

目前, 针对多核集群的并行编程方式尚未成熟, 若直接把传统的单一编程方式应用到这类系统中, 势必不能充分发挥多核处理器的优势. 虽然在混合编程方面国内外提出的 MPI/OpenMP 混合编程方案^[1-2]获得了很好的并行性, 但仍然存在以下不足之处: 须要有特定编译器的支持; 适合静态分析的并行程序, 不支持动态线程负载平衡等功能. 为解决以上问题, 本文引入 Intel 公司新推出的并行编程技术 TBB^[3], 利用它的面向任务编程特点与 MPI 消息传

递方式相结合, 充分利用了多核处理节点的并行特征以及多核集群任务并行分割编程的优势, 达到并行系统的高效利用. 并在 TBB/MPI 混合并行编程模型上通过 N 皇后的混合并行算法验证了本文提出的编程模型的可行性与有效性.

2 并行编程模型

2.1 消息传递编程模型 MPI

MPI 是消息传递接口的标准, 用于开发基于消息传递的并行程序, 可在分布式存储体以及共享存储体的可缩放并行机和工作站机群系统上执行. 实

现了处理机间的数据交换功能,并提供并行任务之间的同步和收/发数据的接口,具有良好的可移植性和易用性。

MPI 是面向进程级的编程. 在 MPI 中,一个独立参与通信的个体被定义为一个进程. 每个进程所运行的代码可以不同,各个进程通过调用 MPI_Comm_rank 函数来获得本进程在某指定通信器中的进程号以标识该进程,进程间的信息交互完全通过显式地调用通信函数来完成. MPI 的编程优势在于用户可控制对进程的划分与同步,其不足在于细粒度任务会导致大量通信,动态负载不均衡等问题。

2.2 面向任务编程模型 TBB

TBB(Threading Building Blocks)是 Intel 开发的线程构建模块,作为多核开放资源应用程序的一个解决方案,它是基于 C++ 模版库的面向任务的并行编程模式^[4]. 提供了强大的基于任务的高级并行语义功能,可充分利用多核 CPU 资源. 与 OpenMP、MPI 等传统技术相比,它具备以下特点:以多任务替代多线程,程序员只需考虑任务的设计, TBB 底层组件便可将任务以高效的方式映射到各个线程,并以任务窃取形式实现自动调度,简单易用. 任何 ISO C++ 的编译器都可直接编译 TBB 程序;程序无需过多修改就可通过线程构建模块获得并行的速度提升. 但缺点是只适用于硬件共享存储的机器,如多核芯片、多处理器节点等,而不能跨节点通信。

TBB 提供了一些基本算法、高级算法与并发性容器等功能. 基本算法中的并行循环(parallel_for),并行规约(parallel_reduce),并行扫描(parallel_scan)等泛型算法都是基于任务调度器接口的并行算法. 其中 parallel_for、parallel_reduce 是一种负载均衡的、对固定数目的独立循环迭代的并行;而 parallel_scan 是针对并行循环中存在串行依赖性的问题提出的。

3 MPI-TBB 混合编程模型

按并行程序从操作系统获得系统资源的角度,并行计算有线程并行和进程并行两种方式. 为充分利用多核集群物理硬件体系结构特点,我们将划分的任务进程映射到每个计算节点上,再将该进程所生成的每个线程映射到相应计算节点 CPU 的各个核上,如图 1 所示。

为达到计算资源高效利用以及进程与线程结合获得更好的并行性,设计了在多核集群各节点之间

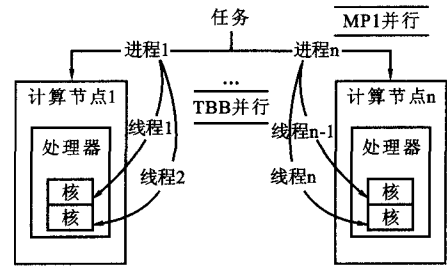


图1 多核集群下的并行任务映射

用 MPI 管理各任务物理区,节点内部用 TBB 对划分的区域进行计算. 克服了 MPI 在多核系统应用的局限性,同时避免 TBB 在节点间任务分配和通信的不足. 综合 MPI 与 TBB 各自特点将两种编程模式相结合,实现节点间和节点内两级并行. 设计的混合编程模型如图 2 所示。

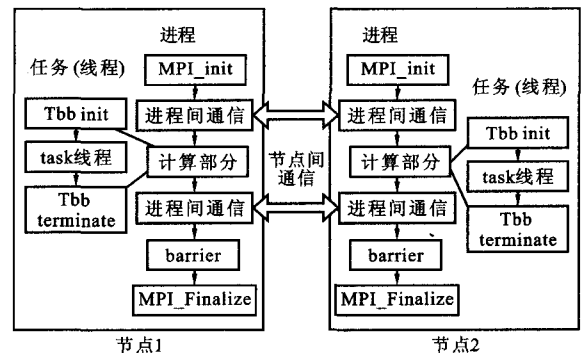


图2 MPI-TBB 混合模型

以这种混合编程模型实现的并行算法可突破单一编程模式下的计算规模、负载平衡等问题的限制,充分发挥多核并行计算的优势,实现进程级的粗粒度并行和循环级的细粒度并行,以较小的代价获取更好的性能。

4 MPI-TBB 混合编程方式

考虑用 TBB 实现并行算法可降低编程复杂度、便于程序编写,编程方式采用在 TBB 编程代码中添加 MPI 制导语句以实现多核单节点并行至多核集群的混合并行编程. 为了验证这种编程模型的可行性与有效性,以 N 皇后^[5]并行程序混合编程中 TBB 的 parallel_for 与 MPI 相结合为例描述设计过程(TBB 其他功能与 MPI 的混合编程方法与此类同)。

混合编程算法步骤:

Step1:主函数中对使用的 MPI 与 TBB 各部分进行相关声明与记录

其中: MPI_Bcast 由主进程广播数据到各从进程; task_scheduler_init 对象进行 TBB 环境初始化并启动调度器功能。

Step2:对问题执行MPI分解策略,将任务划分成通信不密集的任务进程,完成对执行任务进程到处理节点的分配功能.每个节点根据进程号获取parallel_for中的range参数指定划分block的范围;以SetQueens()指定对任务块的皇后放置操作;以auto_partitioner监测task_scheduler的工作密取行为以自动粒度划分,将子任务拆分成几个小块来并行执行,实现线程到处理器核的分配,利用共享存储进行通信,达到多线程并行.

关键代码如下:

```
if (myid==k)
for(k=0;k<numprocs;k++)
if(my_rank==k)
{parallel_for(blocked_range<size_t>(my_rank* size/
numprocs,(my_rank+1)*
size/numprocs),SetQueens(),auto_partitioner());}
```

(应注意:划分的任务块必须是相互独立的才可以采用parallel_for来实现并行.)

Step3:皇后放置计算功能函数有关部分.首先各独立任务通过体对象中的operator()方法对setQueen()函数进行调用.其次,采用预置与回溯机制查找 $N \times N$ 上的所有可放位置,若按规则皇后被全部放置,则解决方案nrOfSolutions自增1以记录方法总数.

关键代码如下:

```
void operator()( const blocked_range< size_t > & r
) const
{for( size_t i=r.begin(); i!= r.end(); ++i )
{setQueen(new int[size], 0, (int)i);}}
void setQueen(int queens[], int row, int col)
{ .....queens[row]=col;
if(row==size-1)
{ {NrOfSolutionsMutexType::scoped_lock mylock
(NrOfSolutionsMutex);
nrOfSolutions=nrOfSolutions++;}}
else {for(int i=0; i<size; i++)
{setQueen(queens, row+1, i);}}}
```

5 实验测试

为简便起见,本文仅用三台多核节点构建小型集群系统.节点机为主频2.8 GHz的双核Intel Xeon CPU、2 GB内存,每个节点机安装Linux Red-Hat9.0操作系统,并配置MPICH-2.10.8和TBB22的混合并行编程环境.

实验以12至14皇后放置方法并行程序为例进行对比.表1为多核环境下MPI单一编程与混合编

程方式的执行时间比较.

表1 各种编程模式下N皇后问题执行时间 s

编程模式	节点数	各种皇后数下的程序执行时间		
		12 皇后	13 皇后	14 皇后
TBB	1	0.320 000	2.000 000	11.920 000
	1	0.791 440	4.924 603	33.360 946
MPI	2	0.925 463	5.341 632	35.371 346
	3	0.480 727	2.713 292	17.687 258
MPI-TBB	1	0.315 832	2.047 307	13.476 229
	2	0.177 567	1.020 702	5.790 100
	3	0.132 755	0.636 083	3.400 035

混合编程方式采用在TBB程序上加入MPI消息传递机制,一方面编程代码简洁、节省了程序运行时间;另一方面CPU处理能力的充分发挥更有效的实现了并行分配.从表1及图3中可观察到混合编程模型在时间上要优于MPI单一编程模式,加速比也显著增加,且随着问题规模的不断扩大,混合模型与单一MPI编程下的加速比差距不断加大,MPI-TBB混合编程拥有更优的并行执行效率.

12皇后各编程模式加速比 14皇后各编程模式加速比 13皇后各编程模式加速比

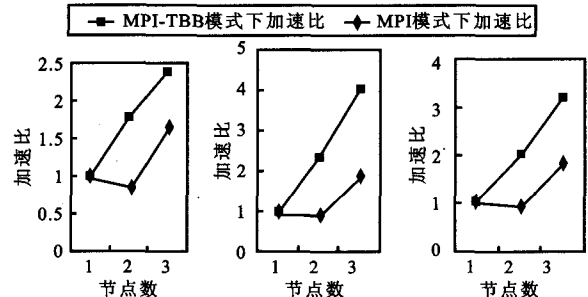


图3 MPI与MPI-TBB混合编程并行算法加速比

6 结束语

混合模型的设计与实验测试表明,本文提出的MPI消息传递与面向任务并行的TBB混合编程模型是可行的.该方案充分利用了多核集群计算节点的运算能力及两种单一模型编程的优势,大幅度提高了并行程程序的执行效率,具有良好的可行性、有效性与可移植性,为多核集群并行编程提供了一种有效途径.

参考文献:

- [1] Smith L Bull M. Development of Mixed Mode MPI/Mode[J]. Scientific Programming, 2001,9:83-98.

(下转第97页)

采用 DSDV 协议传输通信流时,自相似程度最高,突发性最强,由计算可得的自相似参数 H 为 0.841 1. 主要原因在于 DSDV 需要周期性更新路由,消耗了较多的带宽。

DSR 的按需特性克服了这一问题. 但是随着网络拓扑结构的复杂,DSR 路由中断频率增大,网络重传次数变多,突发性增强,其自相似参数为 0.802 6.

而采用 CBRP 协议传输导致的通信流的自相似突发性程度最小,自相似参数为 0.771 9. 主要是因为 CBRP 对节点进行分治管理,能够较好地适应网络状况的变化,维持较好的通信性能。

4 结束语

本文研究了在无线网络中,采用不同的路由算法传输多媒体数据时,对通信流自相似特性的影响. 结果是平面路由协议中的基于表驱动的路由的 DSDV 的自相似参数最高,不适合实时流媒体的传输,DSR 协议和 CBRP 都优于 DSDV,其中分簇路由协议 CBRP 的 H 参数相对较小,用于实时流媒体传输时,能够提供较好的通信性能. 通过比较,为我们选择路由协议又提供了参考,同时这对我们根据其突发参数,选择更好的模型对其建模,进一步提高网络性能有重要意义。

参考文献:

- [1] Tang D, Baker M. Analysis of a local-area wireless network[C]// Proceedings of the 6th Annual International Conference on Mobile Computing and Networking. Boston:IEEE, 2000: 1-10.
- [2] Balachandran A, Voelker G M, Bahl P, et al. Characterizing user behavior and network performance in a public wireless LAN[C]// Proceedings of ACM SIGMETRICS. Marina Del Rey:ACM, 2002: 195-205.
- [3] 胡英,谭献海,肖中卿,等. 自相似业务下资源分配研究[J],微电子学与计算机,2008,25(4):22-25.
- [4] Crovella M, Bestavros A. Self-similarity in World Wide Web traffic: evidence and possible causes[J]. IEEE/ACM Transactions on Networking, 1997, 5(6): 835-846.
- [5] Park K, Kim G, Crovella M. On the relationship between file sizes, transport protocols, and self-similar network traffic[C]// Proceedings of ICNP'96. Columbus:IEEE, 1996: 171-180.
- [6] 刘向东,张俊星,张金海,等. TCP 批量业务的自相似特性[J]. 微电子学与计算机,2008,25(5):30-34.
- [7] Video traces research group [EB/OL]. [2009-12-30]. <http://trace.eas.asu.edu/index.html>.

作者简介:

赵露女,(1986-),硕士研究生. 研究方向为网络通信流的特性分析与建模。

(上接第 93 页)

- [2] 单莹,吴建平,王正华. 基于 SMP 集群的多层次并行编程模型与并行优化技术[J]. 计算机应用研究. 2006, 10:254-260.
- [3] Renders. Intel Threading Building Blocks 编程指南[M]. 北京:机械工业出版社,2009.
- [4] 胡斌,袁道华. TBB 多核编程及其混合编程模型的研究[J]. 计算机技术与发展,2009, 19 (2):98-104.
- [5] 田翠华. 算法设计与分析[M]. 北京:冶金工业出版社 2007.

作者简介:

顾慧女,(1985-),硕士研究生. 研究方向为并行计算、网络系统。

郑晓薇女,(1957-),教授. 研究方向为网络系统、并行计算。

张建强男,(1981-),硕士研究生. 研究方向为并行计算。

吴华平男,(1984-),硕士研究生. 研究方向为并行计算。