并行技术在神威集合数值天气预报系统中的应用*

张　贻
(北方计算中心 , 北京 100091)

提　要

文章讨论了基于神威巨型机的并行化集合数值天气预报系统中实现的各种并行算法 , 性能分析结果表明并行方案最大限度的利用了神威机的处理器资源 ,设计的并行算法效率较高 ,满足了实时业务运行的时效要求 。

关键词 : 大规模并行处理　集合数值天气预报　归约　同步

1　系统简介

神威集合数值天气预报系统以神威巨型计算机为开发平台 ,采用奇异向量法产生初始扰动 ,产生奇异向量的模式分辨率为 T21 L19 ,预报模式分辨率为 T106 L19 ,集合预报成员为 32 个 。系统的主体部分由前处理 、资料同化 、集合预报初始场生成 、模式预报 、模式后处理 、产品生成 、可视化运行监控等 7 个子系统组成 。每个样本做 10 天模式预报且按每天间隔 12 h 输出一次结果 。生成包括 :降水概率预报 ,日降水预报 ,850 hPa 逐日温度距平的概率预报 、500 hPa 高度场逐日集合平均以及 500 hPa 高度场 10 天平均等等在内的百余种预报产品 。

2　系统的并行结构与主要并行算法

集合数值天气预报在计算机软件实施上遇到的主要困难是 :样本数多 、解算量和数据传输量都很大 。为此 ,围绕多样本并行计算与控制这一核心 ,系统充分利用神威计算机并行机制的优势 ,针对系统整体和各部分的技术特点 ,尽力加大并行度 ,不断优化程序设计 ,最终在功能实现 、解算时效 、日常运行等各个方面都满足了业务运行的需求 。

2 .1　系统并行结构

解决科学计算和工程应用课题时 ,常用的并行编程模型主要包括隐式并行编程模型和显式并行编程模型[1] 。

所谓的隐式并行编程模型是指程序员不用显式地说明并行性 ,而是让编译器和运行支持系统自动的加以开发 。就目前而言 ,已推出的商品化并行编译器还很少 ,而且并行化

* 本文由国家“863”高科技研究发展计划(863-306-ZD11-03-02)项目资助 。
2001-04-28 收到 ,2001-07-16 收到修改稿 。

编译器的有效性也很不理想,所以隐式的并行编程模型很少被使用。常用的主要是显式并行编程模型。

显示并行是指在源程序中由程序员使用专用语言构造、编译器命令或库函数调用对并行性加以显式说明。目前,显式并行编程模型主要有 3 种,即数据并行模型、消息传递模型和共享变量模型。

数据并行模型主要适用于 SIMD 或 SPMD 方式,其主要思想是在多个计算结点中对不同的数据集同时执行相同的指令或程序段,开发的是数据并行性。

消息传递并行模型主要适用于 MPMD 和 SPMD 方式,其特点是,此类程序由多个进程组成,其中每个进程有自己的控制线程且可执行不同的代码,工作负载和数据均需由用户用显式方法分配给进程,工作负载的分配遵循拥有者计算法则,即由拥有数据块的进程来完成相应计算,这种模型通常用于开发大粒度并行性。

共享变量并行模型也是多线程化的,但数据是驻留在单一的共享地址空间中,不需要对数据进行显式分配,而工作负载的分配可以用显式也可以用隐式方法。因共享变量并行模型中所有进程共享单一地址空间,因此容易发生因读/写共享变量而造成的令人难以捉摸的同步错误,而且目前共享变量并行模型不存在可以广泛接受的标准,在 MPP (Massively Parallel Processor) 机上,共享变量程序可能会有更高的交互开销而比消息传递程序运行更慢,所以就目前而言,首选的并行编程模型应是数据并行和消息传递并行编

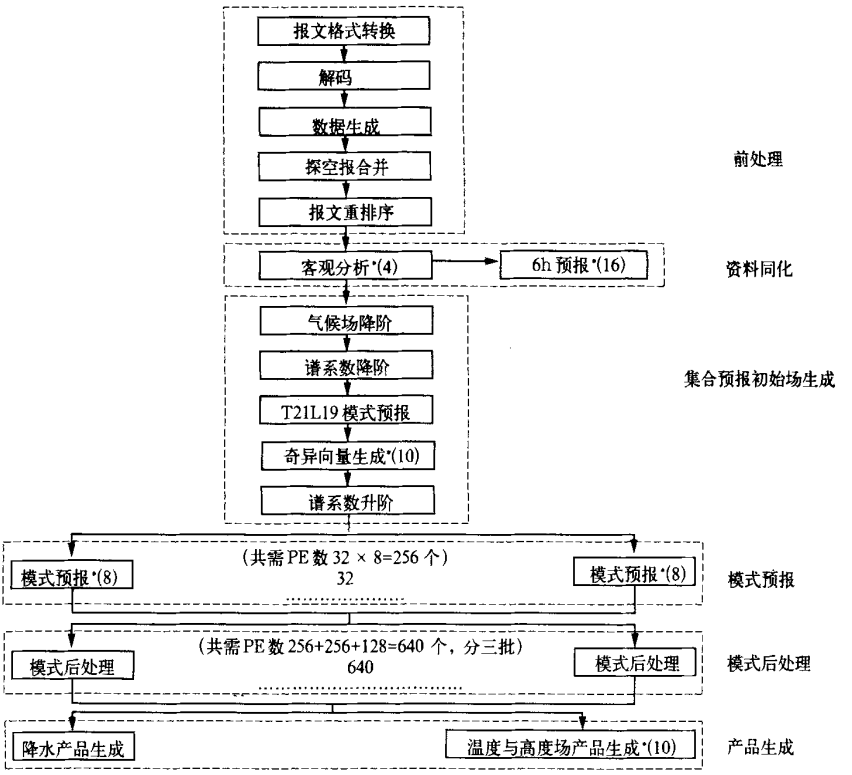

图 1　神威集合数值天气预报系统的总体并行结构框图

(其中用 * 标注的部分需要在内部做进一步并行,括号内是并行时所需的 PE 数)

程模型。

神威巨型机属可伸缩的大规模并行处理系统,通常称之为 MPP 机,它采用的是同构、分布共享主存储器、平面格栅网体系结构,共有 384 个单元处理器,每个处理器(PE)有自己的内存 128 MB,处理器间以内部互连网络的方式连接传递信息,各处理器既可独立运行也可通过相互间的网络实现并行。每个处理器还可通过平面格栅网上的 16 个 I/O节点来管理存储总量为 16 × 80 GB 的磁盘阵列(简称 IOP)。

在神威集合数值天气预报系统的研制中,结合集合数值天气预报业务需求中大规模、高分辨率的特点,充分发挥神威并行机制的众多优势,联合采用上述两种并行模型,设计了如下系统总体并行方案:①按样本、时次或产品种类划分任务,实现各样本数值预报或各时次模式后处理或产品种类间的并行;②对于集合预报中运行量大、运行时间长的部分,根据各部分的特点,在内部再分别采用不同的并行算法,提高并行解算效率,进一步缩短系统运行的墙钟时间,满足实际预报的时效要求。图 1 给出了神威集合数值天气预报系统的总体并行结构框图。

2.2 集合预报各子系统的并行算法

经程序分析和实际运行测算,发现资料同化、集合预报初始场生成子系统中的奇异向量生成部分、模式预报、模式后处理和产品生成等子系统,占用的机器资源多,计算和数据文件交换量大、运行时间长,必须做并行处理。我们结合神威并行机制特点,确定了相应子系统的并行策略。

(1) 资料同化子系统的并行算法

通过对资料同化串行程序运行性能分析发现,质量场和风场分析部分占总运行时间的 90 %左右,而其中资料检查、估算分析系数、估算格点分析值部分则几乎占尽所有时间,特别是格点分析值的计算。因此,并行应主要针对这三大部分进行。在并行化过程中发现,每个 PE 局存的大小是制约并行效果的一个主要因素,如果一些工作文件无法转换为数组形式,并行化效果将大打折扣,而神威机的局存只有 128 MB,因此,神威集合预报系统的资料同化子系统中,依据并行效果,在内存允许的范围内只对资料检查和估算格点分析值部分进行了并行。子系统中采用主从(master-slave)并行方式,它属 MPMD 并行方式的一种,由一个主进程和若干子进程共同完成任务[2]。系统运行时先创建所有的进程,然后主进程运行单任务段,子进程处于等待状态。当主进程到达资料检查的并行起点时,唤醒子进程,并向子进程分配资料盒子序号,所有子进程并行处理。并行点结束后,主进程归约收集各子进程的运算结果,继续执行后面的单任务段,子进程回到等待状态,直到下一个估算格点分析并行点再次被唤醒,这时主进程向子进程分配的是由北至南的纬圈号。这一并行点结束后,主进程再次归约收集各子进程的运算结果,执行以后的串行计算。因每一个子进程在处理完所分配的盒子或纬圈资料后,再通过主进程的计数器获取下一个需处理的盒子或纬圈,所以在每个并行点,各处理器的工作负载平衡。图 2 给出了这一子系统的并行结构框图。

(2) 模式预报子系统的并行算法

在 T106L19 谱模式中,物理过程的计算量最大,耗时最多,占该部分总运行时间的 95 %左右,而物理过程的计算是从富氏空间→格点空间→富氏空间,在每个纬圈上依次进

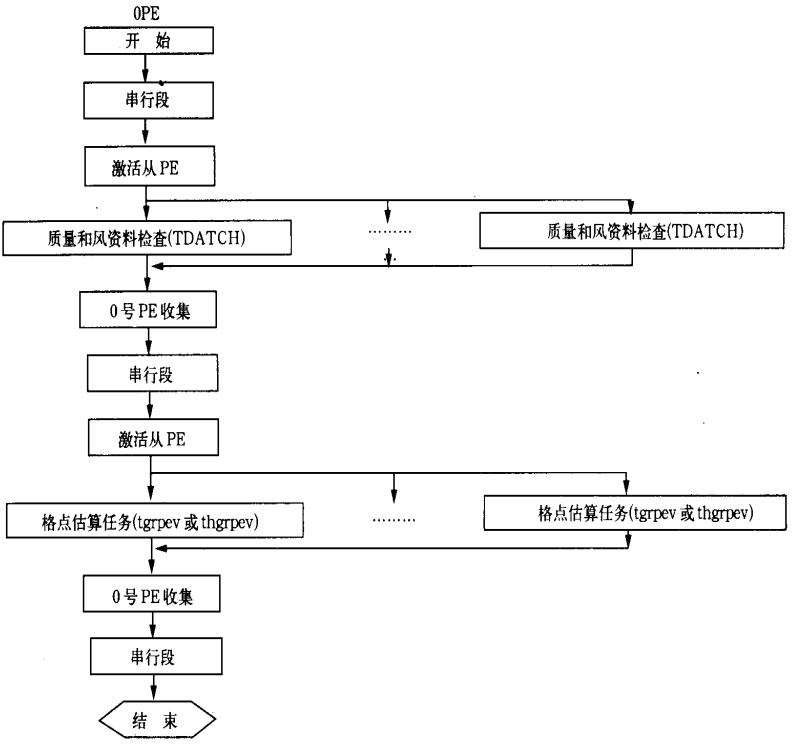

图 2　资料同化子系统并行结构框图

行的,全球的富氏空间纬圈数为 80 个,每个富氏纬圈的结构完全相同,相互之间没有顺序关系[3],非常便于实现并行处理。因此,在模式预报子系统中采用 SPMD 的并行化模式,将并行计算的目标对准两个扫描子程序中各富氏纬圈的计算部分,将相对于各富氏纬圈的格点值记录、傅立叶系数值记录及勒让德多项式值记录按块分配方式均匀分到各个处理器的局部内存中,将相应纬圈的计算对应分配到各个处理器上,各处理器在进行自己所承担的若干纬圈的计算时不需要通信,只有在完成两次扫描进入下一次积分时,才需将各处理器上需要累加的谱系数、全球物理诊断、统计、半隐式调整等物理量作归约。根据神威机所拥有的 PE 数和集合预报的样本数确定在系统中采用每 8 个 PE 一组做一个样本的模式预报,各组中每个 PE 的计算量是 10 个富氏圈,通信量是 7 兆字节。各 PE 组做各自样本的模式预报,组间不需要通信与同步。这样的数据和任务划分使得 32 个样本的模式预报最大限度的利用了神威机的处理器资源,且计算的颗粒度较大,减少了系统运行中通信开销的比重。各处理器上的工作负载基本平衡。图 3 给出了这一子系统的并行结构框图。

(3) 奇异向量生成部分的并行算法

奇异向量生成部分是集合预报的核心。它主要采用正、反时间积分的模式,正、反积分运算时结构相似,顺序相反。积分中需进行两次扫描。根据串行程序运行时间统计,正反积分的运行时间占总运行时间的 95.25 %,两次扫描时间则占积分时间的 92.63 %。因此,并行化的工作主要放在正反积分的两次扫描上。

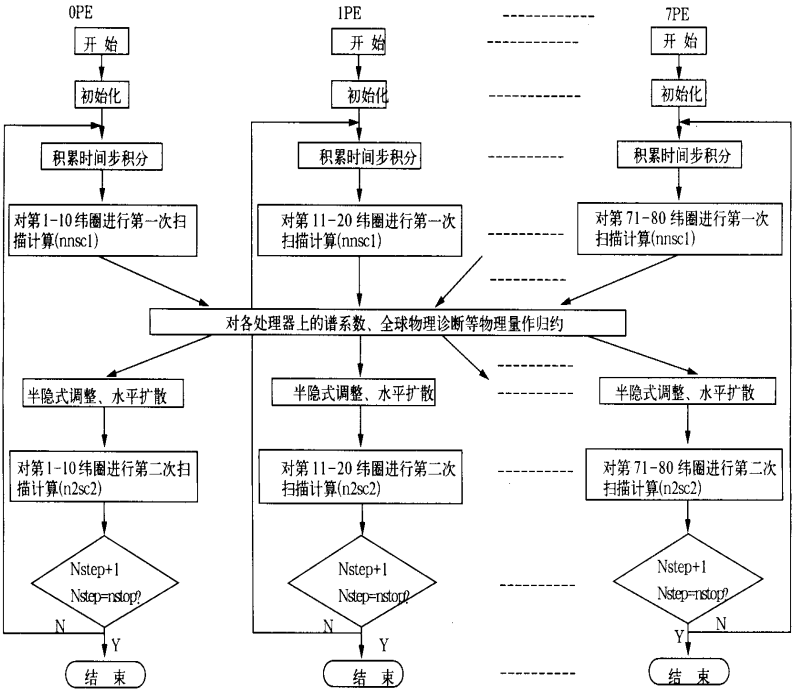

图 3 模式预报子系统框图

因这一子系统中,正反积分过程与模式预报子系统中的类似,所以其中我们采用了与模式预报子系统中相同的并行方法与策略,只是在反向积分过程中注意归约的位置和时间。另外在深入分析解算需求、数据结构的基础上,对串行程序作了结构上的修改,收集每次正、反积分时都要进行的资料读入过程,将它们从相应的段落中提取出来,一次读入,存放于内存,解决了并行循环中频繁读写,严重影响并行效率的问题;分析正反积分前的初始化过程中各种变量的初始化情况,在反向积分时巧妙借用正向积分中已初始化且积分过程中未做修改的变量,大大减少了系统中串行部分占有量;利用系统所提供的底层消息库,自行编写出所需要的高效通信函数,最大限度的降低通信时间。

(4) 模式后处理和产品生成子系统的并行算法

32 个样本 10 天模式预报总共产生 640 个时次的资料输出,单个资料后处理的时间不长,但总体工作量很大,与此相关,产品生成所要读入的资料量也很多,这两部分处理不好将极大的影响集合预报的整体时效。我们采用神威机所提供的 SPMD 并行模式,每个处理器运行相同的程序处理各自所拥有局部存储器内的不同时次的资料,从而充分的利用了神威机处理器多的优势,使不同时次的资料得以并行处理。

产品生成子系统中,在认真分析温度和 500hPa 高度场产品生成中的数据结构和流向的基础上,将原本分开处理的两部分程序进行了优化合并,一次读入资料,进行两个产品的处理,去除了占用时间很长的重复数据读入,较大幅度的提高了系统的运行效率。所采用的并行策略分两个方面:一方面将降水产品生成和温度与高度场产品生成两部分并行

处理,另一方面在温度与高度场产品生成部分中再采用 10 个处理器并行生成 10 天产品的并行处理方法。

3　性能测试与结果分析

通常,对并行计算的检测主要分两个方面:一是正确性;二是效率。

神威集合预报系统研制完成后,由国家气象中心组织专家对系统进行了全面的测试和试用,证明并行系统的解算是正确的,开发是成功的,该系统于 1999 年底在中国气象局投入准业务运行,于 2000 年 3 月正式投入业务运行。

在神威巨型机上,对并行化的各部分的效率所进行的测试结果如表 1 所示。需要说

表 1　并行化子系统测试结果

	PE 数	串行运行时间(s)	并行运行时间(s)	加速比
资料同化	4	501 ~ 1048	364 ~ 603	1.37 ~ 1.73
奇异向量生成	16	16348	1932	8.46
32 个样本的模式预报	256	1681920	10410	161.6
640 个时次的模式后处理	640(分 3 次)	74048	2344	31.5
降水	1	383	458	8.84
温度与高度场	10	3667		

明的是:1) 由于资料同化子系统每天四个时次资料量不同,因此处理时间也不尽相同,表中给出的只是对某一天 00:00、06:00、12:00、18:00(UTC) 资料处理时的运行时间范围;2) 因模式预报子系统在神威机上单处理器运行时所需的内存量超过了 128 MB,所以表中给出的是 32 个样本的模式预报在前端机上独占运行时的总时间。

从表中可以看出,除产品生成子系统外,其它各部分子系统的效率都不是很高,原因主要有两方面:1) 受神威机内存(128 M) 和高速缓存大小的限制,各子系统纯计算部分的速度无法取得令人满意的效果,我们曾经在主频相同,内存为 512 M 的机器上做过实验,每个子系统纯计算部分的速度都有 3 倍以上的提高;2) 由于具体算法的限制使得各子系统中存在着无法并行化的串行段,影响了总体加速比。就各子系统中的并行段而言,加速比还是比较令人满意的。

从表 1 中还可以看出,模式后处理的效果尤其不理想。这是磁盘阵列拥堵造成的。32 个样本输出 640 个时次的资料,分布在 16 个磁盘上,模式后处理时,264 个 PE 同时去处理这 16 个磁盘上的资料,拥堵不可避免。

解决这个问题有两种方法:一是将模式预报和模式后处理两个子系统合二为一,模式预报生成每个时次的资料后,无需输出而直接传给后处理,由专门的 PE 计算后处理过程,这种模式预报与后处理的捆绑处理,不仅节省了磁盘空间,去除了用于后处理的时间,而且因没有了磁盘读写,运行时间得以进一步降低。我们作了这方面的尝试,但因合并后系统所需的内存量为 190 MB,所以在目前的神威机上无法运行。为此我们尝试了第二种解决方法,即模式预报后仍然输出每个时次的资料,当第一个时次的资料输出后,就开始由专门的 PE 进行资料的后处理,这种方法也使后处理的时间与模式预报时间基本重叠。

目前,在神威机上已实现了这种解决方案,经测试发现效果与预想相同,模式预报与模式后处理时间基本叠加,整个系统的总体运行时间减少了 45 min。

4 结束语

综上所述,针对如集合数值天气预报这样子系统众多,算法各异的大的应用系统,在并行实现时应根据各子系统中不同算法的特点以及开发平台的类型选择适当的并行模型,采用相适应的并行算法,并综合考虑负载平衡策略和计算颗粒度大小的选择,才能最大限度的发挥并行效率,取得令人满意的并行效果。

参 考 文 献

1 陆鑫达. 并行编程中的方法学. 见:高性能计算机及应用学术研讨会论文集. 2001. 374~382.

2 国家并行计算机工程技术研究中心. 高性能 FORTRAN 语言参考手册, 神威计算机系统技术资料(13). 1999. 29.

3 国家气象中心编译. 资料同化和中期数值预报. 北京: 气象出版社,1991. 80~83.

PARALLEL ALGORITHM IN SW ENSEMBLE NUMERICAL WEATHER PREDICTION SYSTEM

Zhang Yi

(*North Computation Center, Beijing* 100091)

Abstract

The parallel algorithm and the implementation of the SW ensemble numerical weather prediction system are described. Performance results indicate that the parallel algorithm makes the best use of the CPU resources of SW and acquires higher efficiency.

Key words: Massively parallel processing Ensemble numerical weather prediction Reduce Synchronization