集合预报系统中客观分析方案的并行化实现*

冷亭波
(北方计算中心 ,北京 100091)

提　　要

从串行程序分析 、并行方案选择到具体算法实现 ,依次介绍了基于神威机的集合数值天气预报系统中“客观分析子系统”的并行化过程 ,并给出了并行化后的性能评测 。

关键词 : 大规模并行处理　客观分析　集合预报

引　言

“客观分析子系统”是“神威集合数值天气预报系统”的重要组成部分 ,它每天启动 4 次 ,进行资料滚动 ,为“集合预报初始场生成子系统”提供初始场 。近十年来 ,国际上的各大数值预报中心已相继采用更为先进的三维变分(3D- Var) 或四维变分同化系统 。我国从“九五”开始 ,也开展了这方面的研究开发工作 。但目前国家气象中心的业务系统采用的仍是最优插值方案 ,所以在神威集合数值天气预报系统中仍然采用此方法 。

国外对这部分并行的成熟算法一般采用多任务方式 ,国内目前在国产机“神威” 、“银河”上都实现了客观分析的分布式并行计算 ,采用的基本并行方式相似 ,但由于内存大小的差异 ,造成两台机器上的运行效率有一定的差别(这一点将在后面讨论) 。任何一个并行计算机系统都有它一定的特殊性 ,因此 ,一个高效的并行算法除了它的一般性 ,还应具备其特殊性 。在神威机上 ,客观分析采用 MPMD(Multiple Program Multiple data) 中的主从并行方式实现并行 ,并行算法具有一定的典型性 ,下面就客观分析程序的主要串行算法 、并行化设计及性能分析等几方面 ,作一些介绍 。

1　主要串行特点

原 CRAY 巨型机上运行的客观分析业务版本 ,程序大约 5 万多条 ;有 400 多个子程序 ,主要用 FORTRAN 77 语言编写 ;使用了大量 CRAY 专用程序库和多种结构的文件系统 ;为节约内存 ,还引入了内存管理系统 。

要取得良好的预报效果 ,初值质量是至关重要的 ,国家气象中心使用了最优插值技术 、时间间隔为 6 h 的非连续性的分析同化方案 ,用指定分析时间前后 3 h 内的观测资料

* 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 。

对用前一次分析所作的 6 h 预报进行订正,分析观测场与预报场之间的偏差得到一个增量场,再把增量场加到预报场上[1]。

图 1 是分析程序的简要流程图,给出了程序的总体结构。

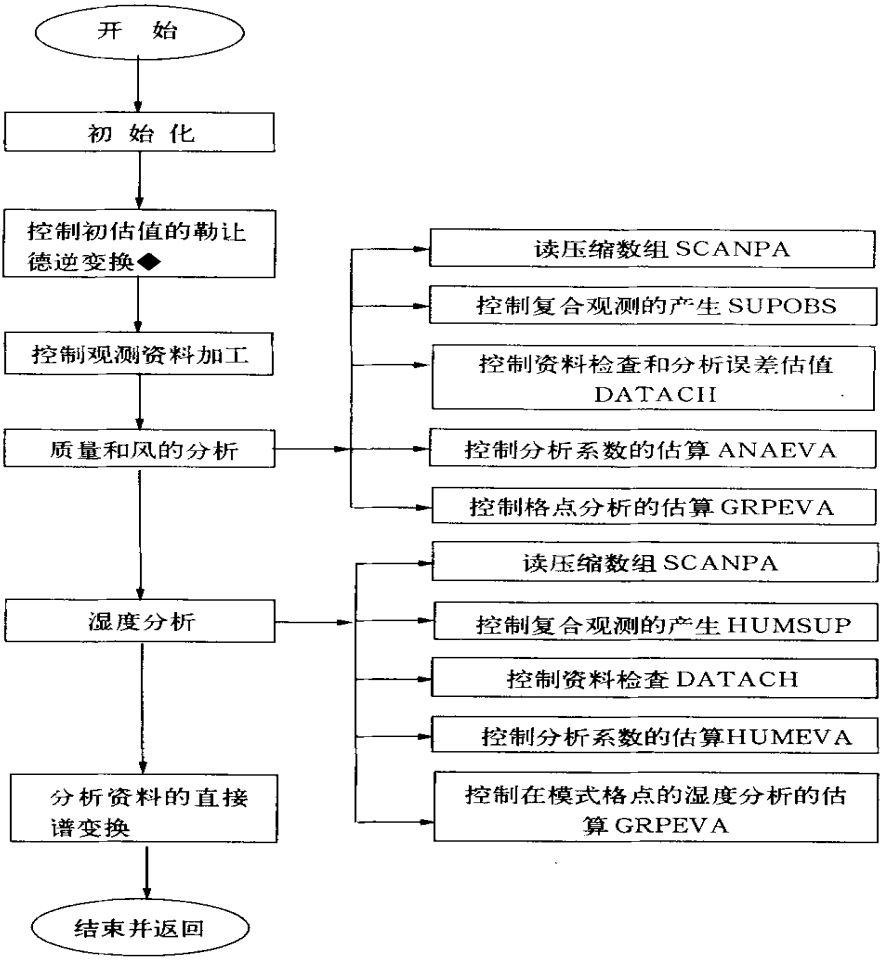

图 1 客观分析简要流程图

客观分析方案采用了三维、多变量、与预报场归一化偏差的统计内插技术。若用 A 表示任一个标量;用 E 表示它的估值均方根误差;用上标 i、p、o、t 分别表示插值、预报值、观测值和真值。基本插值方程为:

$$\frac{A_k^i - A_k^p}{E_k^p} = \sum_{n=1}^{N} W_{kn} \frac{A_n^o - A_n^p}{E_n^p}$$

设:$a_n^o = \frac{(A_n^o - A_n^t)}{E_n^o}$,$a_n^p = \frac{(A_n^p - A_n^t)}{E_n^p}$,$a_k^i = \frac{(A_k^i - A_k^t)}{E_k^i}$,$\varepsilon_n^o = \frac{E_n^o}{E_n^p}$,$\varepsilon_n^i = \frac{E_k^i}{E_k^p}$,假定预报误差与观测资料误差的相关为零,该假定对现有的观测类型是合理的。采用使归一化内插误差方差的期望值达到最小的方法,推导得到最优内插方程为:

$$\frac{A_k^i - A_k^p}{E_k^p} = B^T W_K = B^T M^{-1} P_k = C^T P_k$$

其中 W_k 是权重 W_{kn}的列向量; $M = P + O$, P 是预报误差的相关矩阵[〈α_m^p, α_n^p〉], O 是观测误差的相关矩阵[ε_m^o〈α_m^o, α_n^o〉ε_n^o]; B 是归一化增量值$\frac{(A_n^o - A_n^p)}{E_n^p}$的向量; P_k 是预报误差相关〈α_k^p, α_n^p〉向量。

2　并行化设计及方法概述

"神威集合数值天气预报系统"的开发平台是神威高性能计算机,它属于 MPP 型巨型机,采用了对称同构、分布共享主存储器、平面格珊网体系结构,每个处理单元都拥有自己的内存,通过网络联接传递消息。每个处理单元可独立运行,不受其它处理单元的干扰,相互间可采用 MPI、HPF 和并行 C 等提供的并行化方法和消息通信类函数,进行通信和实现并行。

我们首先解决 CRAY 机与神威机系统的不兼容之处,给客观分析程序增补了几十个数学函数,完成 CRAY 随机 I/O 库子程序的替换等等,移植成功。

2.1　并行化前的分析

神威机的体系结构是适合多任务的,多任务并行计算是能使多个准依赖性任务共享处理机资源,它一般(但不一定)是与多处理环境相联系的。题目是否适合于多任务并行计算,有某些重要的基本要求:首先,该题目可以分成若干个任务,这些任务能平衡地装填在各个处理机上;其次,任务要足够长,值得进行多任务并行计算。

通过分析,客观分析子系统从算法上很适合使用多任务并行。首先从结构上,程序的几个主要循环运算部分,可以按分析盒子或纬圈行划分,将计算分配给若干个任务完成;其次从运行时间上,十个可以采用多任务的程序段运行时间占总时间的 84.3 %,值得使用多任务模式实现并行。

2.2　并行化方案

一般进行数据分配的方法有两种:静态分配和动态分配,动态分配可以根据各个处理机运行过程中的计算情况分配任务,能达到较好的负载平衡。本子系统采用动态多任务的并行化思想较好,通过使用 HPF 语言的多任务模式(MP MD),在神威机上实现本部分的并行。总体并行方案是:客观分析程序单任务执行的部分由主 PE(处理单元)执行,每遇到多任务状态,主 PE 激活其它从 PE,然后主 PE 和从 PE 各自执行分配到的任务,多任务段工作完成后,从 PE 睡眠等待下一个多任务段,主 PE 继续执行当前单任务段。

在分布式内存结构中,各个 PE 使用各自的内存空间来贮存数据,本 PE 不可能改变同时传送给别的 PE 的数据,因此计算的独立性和储存的独立性都是必须小心谨慎考虑的问题。CRAY 机是共享内存,而神威机却是分布式内存,这一不同造成它们在多任务的具体实现上有明显的差异,需要解决临界区内的多任务 I/O 和其它共享程序段的处理;各种共享变量的处理;多任务控制三个问题。

(1) 临界区　几段多任务段中,有三类 I/O 临界区:

① 子任务先输出到不同的子文件,在多任务状态结束前,利用锁,将多个子文件合并到一个主输出文件。

② 子任务将报警或出错信息直接输出到同一个输出文件。

③ 多个任务读/写同一个文件的不同纪录。

在合理使用 HPF 的加锁操作,对这几类 I/O 操作的临界区加以保护后,达到了预想目的。

对其它共享程序段,则针对各自的具体情况,做了适当的代码的添加和修改,并通过调用 HPF 的计数器子程序,使原来的共享程序段能去掉锁操作,成为各个 PE 的私有程序段,其中的操作成为各个 PE 的私有行为,从而去掉彼此的相关性。

(2) 共享变量的分析 多任务段中使用的变量有两种:一种是在多任务状态前就已经存在的共享变量;另一种是在进入多任务状态后,申请、使用并释放的局部变量,即各任务的私有变量。针对这两种变量的不同特征,需要改写与内存管理有关的所有过程。另外,共享变量在多任务状态下还有两种使用形式:一部分共享变量在多任务状态下仅有只读属性,另一部分共享变量在多任务状态下却既有读访问,又有写访问。有写访问的共享变量,在进入多任务状态时,各个子任务都有一份自己的拷贝,需要在多任务状态结束前,由主 PE 根据具体变量的不同特点,采用不同方式收集处理。

(3) 多任务控制 HPF 语言不支持与事件相关的系统调用,好在需要调用与事件相

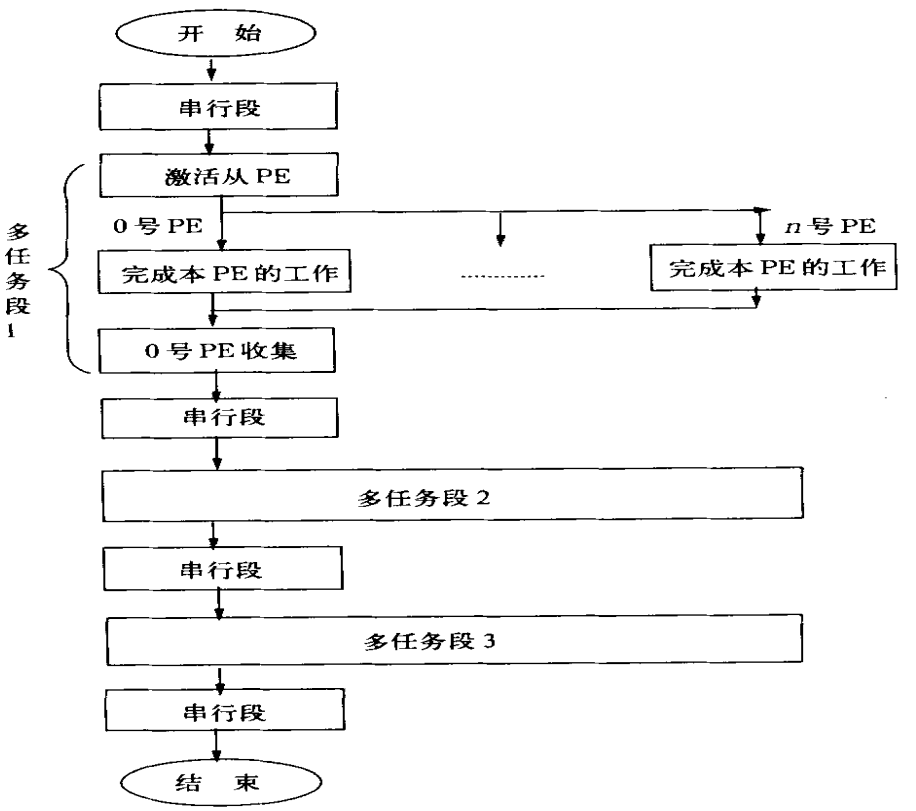

图 2 并行方案示意图

关的系统调用的程序段,只有一处,其运行时间占的比例又很小,这段程序,我们使用单任务就可以了。HPF 语言支持使用锁以及与任务相关的各种系统调用,合理、适当地使用它们,就能实现多任务的控制。

从并行化效果、内存需求量等方面综合考虑,我们只对运行时间花费较大的三部分进行并行化工作:估算格点分析值"质量和风的分析",资料检查和分析误差估值"质量和风的分析",估算格点分析值"湿度分析",这三部分的运行时间占去十个多任务段的 90 %。图 2 是并行化方案的总体示意图,它给出"分析子系统"的整体并行化思路,能清晰地看出几次单任务状态、多任务状态之间的转换过程。

2.3　合理使用局存的优化措施

神威机每个 PE 的本地局存为 128 M,为了更好地利用这些空间,提高运行效率,采取了将工作文件以数组方式存放于局存的方法。但若将所有的工作文件都转成数组,需要几百兆的空间。我们选择了一个在多任务段中有很多次读写的文件,采用压缩大数组形式,将其存于内存。该数组的大小取决于观测资料的多少,可能会超出局存的许可,针对这种情况,我们采取数组和文件共存的方法,倘若超出,便将超出的部分存在工作文件中。事实上超出的可能性很小。这种改动取得了很明显的效果。

3　结果分析与结论

本程序在神威机上是采用 4 个 PE(处理单元)实现并行的。因内存不够大,无法将多任务段中的工作文件都用数组实现,这样对此类文件的读、写依旧需要锁操作。若有几个 PE 都运行到这种位置时,只能有一个 PE 进行读或写,其它 PE 需排队等待,这种可能性又很高,浪费了不少 CPU 时间,所以虽然并行程序使用的 PE 数可缩放,但 PE 数增加到一定数量后,运行时间反倒不理想,经综合考虑,我们采用了 4 PE 方案。

本文简单分析了程序整体加速比不够理想的原因:(1) 随并行段运行时间的减少,串行段运行时间占总运行时间的比例加大了;(2) 通信、同步和访存冲突花费了不少时间;(3) 与银河机相比,神威机的内存较小,只有 128 MB,造成需在多任务段中频繁读写的工作文件无法转换成数组,较大地影响了并行效果。

下面是选取 2000 年 7 月 6 日的 4 个时次数据,做的单 PE 与 4 PE 的运行时间比较,表 1、表 2 分别记录了程序整体和某段多任务的并行效果(由于条件限制,表中测试是程序对机器资源(处理器、磁盘阵列等)不完全独占下测得的)。

表 1　"客观分析程序"运行时间和加速比(时间单位:s)

	00:00	06:00	12:00	18:00
1PE	883.4	500.9	1047.6	623.97
4PE	509.7	363.7	603	470.3
加速比	1.73	1.38	1.74	1.33

表 2　TDATCH 子程序使用单任务、多任务的运算效果(时间单位:s)

	00:00	06:00	12:00	18:00
1PE	136.0	33.5	114.1	30.8
4PE	35.9	11.0	31.8	10.13
加速比	3.78	3.05	3.59	3.04

从上表可以看出,采用上述并行化方法是可行的。尤其对于局部程序,例如 TDATCH 段,局部加速比,在资料少的时次为 3.0,资料多的时次过了 3.5,取得了很好的

效果。总的来说,几段多任务段,通过使用计数器,实现自动负载平衡,都达到了一定的并行效果。

参 考 文 献

1 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.

PARALLEL IMPLEMENTATION OF THE OBJECTIVE ANALYSIS SCHEME IN ENSEMBLE PREDICTION SYSTEM

Leng Tingbo
(*North Computation Center, Beijing* 100091)

Abstract

The parallel processing of the Objective Analysis subsystem in "Shenwei" massively parallel processing system in respect of serial program analysis, parallel scheme selection and algorithms implement is introduced. The application performance results are given.

Key words: Massively parallel processing Objective analysis Ensemble weather forecast