文章编号:1673-9469(2016)01-0086-05 doi:10.3969/j.issn.1673-9469.2016.01.019

并行I/O技术在海洋数值模式中的应用研究

黄伟建,王 鑫,韩院彬
(河北工程大学 信息与电气工程学院,河北 邯郸 056038)

摘要:在海洋数值模拟过程中,随着计算区域的扩大以及计算精度的提升,大量数据输出使得I/O效率成为系统整体性能提升的一个瓶颈。针对这一问题,使用并行I/O技术对系统的输出模块进行优化,并与传统的几种串行I/O方式在不同节点,以及不同计算规模下进行性能比较。通过实验研究数据,对不同I/O方式的不同特点和不同适用场景进行分析,证明在海洋数值模式中使用并行I/O技术切实可行,并且I/O速率得到大幅度提升。

关键词:并行I/O;海洋数值模式;并行计算;消息传递接口;I/O
中图分类号: TP391.9 **文献标识码**:A

Study and application of parallel I/O technology in numerical ocean model

HUANG Wei-jian, WANG Xin, HAN Yuan-bin
(School of Information Science and Electrical Engineering, Hebei University of Engineering, Hebei Handan 056038, China)

Abstract: In the processes of ocean numerical simulation, with the increase in the calculation domain as well as the requirement of higher accuracy, a dramatic amount of data output make the I/O efficiency become a bottleneck of the whole system performance. To solve this problem, the parallel I/O technology was used in this study to optimize the output module, and the comparison with the traditional serial I/O methods under various numbers of computing nodes and various computing scales was conducted. Experimental results demonstrate that the proposed method significantly reduces the execution time of output module.

Key words: parallel I/O; ocean numerical models; parallel computing; MPI; I/O

气候变化引起的异常大风、降水事件频发对我国近海生态系统和沿海经济、社会可持续发展带来了多重压力和严峻考验。由于海洋运动受到一定的物理化学定律支配,因此可以通过方程组予以定量表示,给定边界条件并通过数值模式可以准确客观地由当前状态模拟、预测出未来时刻的状态[1]。海洋数值模式由于其庞大的计算区域以及复杂的求解过程,因此具有浩大的计算规模[2-3]。采用并行计算方能有效提升计算效率,并在预期时间内完成计算任务。现阶段,海洋数值模式中的计算部分已实现了多节点并行计算,但是在I/O模块依然采用串行方式。因此,基于高性能并行计算技术对当前海洋数值模式中的I/O方式进行优化,提高模式数据存储效率,便成为提升系统整体性能的一个关键所在[4-6]。美国地球物理流体力学实验室采用的并行I/O方法为:N个进程并行运行,最终生成N个结果文件,在后续的程序运行过程中将N个文件进行合并;美国国家自然科学基金会以及美国国家海洋大气局共同资助的WRF数值模式系统,通过设置专用I/O进程的方法实现并行计算的数据访问。我国在海洋数值模式中应用并行I/O技术还较为罕见,多数仍采用传统的串行I/O方式或主从模式。并且,在同一个硬件和实际应用环境中针对不同I/O方式进行详细比较的例子也较为少见。本文针对海洋数值模式中的海洋水质模块,在课题组前期已完成计算模块并行化[7]的基础上,对I/O模块的算法进行优化,以提升数据的访问效率,从而进一

收稿日期:2015-11-02
基金项目:海洋公益性行业科研专项经费资助项目(201205018);河北省自然科学基金资助项目(F2015402077)
作者简介:黄伟建(1964-),男,山西交口人,博士,教授,CCF会员(E200038566M),研究方向为计算机应用。

步提升系统整体运行效率。

1 I/O 技术应用分析

1.1 系统中 I/O 模块的算法分析

由于 MPI 并行方式属于分布式计算的一种,其计算结果分布式存储在各个计算节点中。传统的并行环境下的输出方式主要有以下几种:一是参与计算的通信域内的各个进程在计算完成后直接输出计算结果到文件,这样 N 个进程将会输出 N 个独立的文件,在后期需要对这些独立的文件进行进一步合并处理,显然效率将会大大降低;二是采取等待同步(Call MPI_Barrier)的方式,各个进程以追加写入的方式依次将文件串行写入到同一个文件中;三是采用收集(Gather)的方式将所有进程的数据通过消息传递收集到根进程,最后由根进程进行输出[8-10]。

上述三种 I/O 方式均使用编程环境提供的相应 I/O 函数。第一种方法的并行度是最高的,但是由于需要后期的合并处理,不但容易在文件合并过程中产生错误,而且也会消耗大量的时间;第二种方法虽然输出的为完整的文件,但由于采用了等待同步的方式来强制每一时刻只有一个进程对文件进行 I/O 操作,并行度被大大降低,在等待同步(Call MPI_Barrier)时因为频繁阻塞通信域内所有进程,也会消耗掉大量时间;第三种方法多用于采用主从模式的并行程序,但会产生较大的通信量,并且根进程会有较大的 I/O 负担。当数据量较为巨大时,通信速率会成为其系统整体性能提升的一个瓶颈,并且在本质上依然是串行 I/O 方式。

1.2 技术方法

(1)系统模式结构及 I/O 特点

近海生态环境预评估系统采用 SPMD 并行编程模型,现阶段为基于 MPI 的并行方法进行实现,采用 Fortran 语言进行编写。系统的主要结构如图 1 所示,本研究工作主要从应用层角度针对系统的数据输出模块进行优化操作。

在本系统中,一个阶段的计算结束后会产生密集的 I/O 请求将计算结果输出到文件。以胶州湾实验区域的模拟计算为例,进行 48 h 的生态环境数值模拟,每完成 3 600 次的迭代计算进行一次输出。在当前实验区的计算规模下(159 × 185 个网格)每个文件约为 5 MB 大小,共计约为 240 MB。如果计算区域扩大 200 倍,则每个文件大小约为 1 GB,48 h 的数值模拟将产生约 48 GB 的数据量。若进行两个月的数值模拟计算,总数据量将达到 1.4 TB。未来在投放使用后,随着计算精度和生态过程的复杂度的提升,以及计算区域的扩大,都将导致模式中输出数据量的进一步激增。由于模式中目前采用串行 I/O 方式,提高数据输出效率成为系统整体效率提升的一个关键。

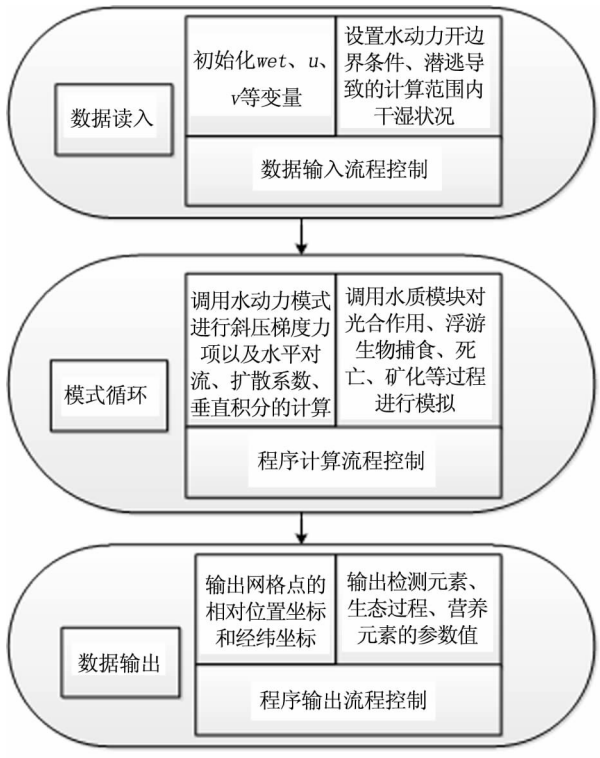

图1 水质模块结构示意

Fig.1 Structure schematic of water quality module

(2)模式中数据格式分析

模式的最终输出文件以 DAT 格式保存,在文件内部每行有 33 列,每个文件的首行为标题行,第一至四列为该数据点在整体计算网格内的坐标以及其经纬度坐标值。后面 29 列则为不同检测元素的参数值。在数据输出时,根据潮汐变换模拟出当前时刻该数据点的干湿情况(是否因海水涨落露出漫滩)来决定是否输出该数据点。在当前实验区域每模拟 1 h 的输出量包含约 13 000 左右个数据点。由于采取并行计算技术对计算区域进行了网格划分,这些数据点的数据分布式存储在各个进程中。

(3)系统开发环境

本系统运行环境配置如表 1 所示。

表 1 集群配置概要

Tab. 1 Overview of cluster configuration

名称	配置
节点数量	20 台 IBM Blade Center HS21
处理器(每节点)	四核 Intel Xeon E5405 2GHz ×2
内存(每节点)	8 GB
操作系统	Red Hat Enterprise Linux Server release 5.2
编译器	MPIF90 ver. 10.1 ; gcc ver. 4.1.2

(4)应用并行 I/O 需要解决的主要问题

由于输出数据场中数据的分布式存储,导致数据在内存中存放顺序和最终文件需要的顺序存在一些差异。在输出数据时需要确保各个进程中数据在最终文件中的顺序与计算网格之间正确对应,以保证最终输出数据场的正确性。

使用并行 I/O 方式输出会将内存中的数值类型数据以二进制码格式直接输出到文件,最终输出文件为二进制文件,在不进行格式转化或者使用专用的数据查看工具的情况下,并不能直接读取或查看数据的值。这将给之后的部分工作带来一些麻烦,这与我们所追求的高效运行便有些背道而驰了。因此需要解决系统输出数据格式与并行 I/O 之间的兼容问题,保证输出文件与原文件格式上的一致性。

2 实验设计与算法实现

实验区域胶州湾位于黄海中部、山东半岛南岸,介于东经 120°04′ ~ 120°23′、北纬 35°58′ ~ 36°18′之间。其整体形状类似椭圆,东西之间宽约 27.78 km,南北之间长约 33.336 km(低潮位),总的面积达 446 km^2,划分为 159 ×185 个计算网格,垂直 5 层,积分步长 1 s,积分 48 h(172 800 步)。当前输出频率为每完成 3 600 步积分输出一次。

2.1 三种串行 I/O 模式的实现

阻塞(Barrier)模式在每个进程的输出模块后加入 Call MPI_Barrier 语句进行强制同步等待,阻止调用直到通信域内所有进程完成调用,控制各个进程按照计算网格划分的顺序依次输出数据,并保证每个时刻只有一个进程对文件进行操作,从而避免写入的混乱。收集(Gather)模式则在输出之前通过调用 MPI 的 MPI_Gather 函数将所有进程需要输出的数据按顺序通过进程之间的通信收集到 root 进程,最后由 root 进程进行输出。分开输出模式则不进行任何控制操作,由各个进程直接将结果输出到彼此独立的文件中。

2.2 并行 I/O 模式的实现

并行 I/O 模式使用设置视口(Set_View)的方式,通过定义每一个进程在文件中的视口,使得各个进程可以通过视口对文件并行操作。首先通过分析每个文件的内部数据排列,计算出了每输出一行数据需要的文件指针的偏移量(410 byte);然后通过判定数据点干湿状况确定每个进程需要输出的数据点个数,并通过 MPI 的组收集(Gatherv)方式将每一个进程内的统计数据传递到其余进程的专用数组中,以控制每个进程在文件中视口的偏移量,既要保证不与其它进程产生文件地址上的冲突,同时也不能占用多余的空间,在各视口中,每个进程输出采用指定偏移方式控制数据在文件中的位置,保证数据正确的同时,也提高了寻址效率;接着将所有数值数据进行数据类型转换,并保存到字符数组中,以最小的时间代价实现数据格式的兼容性;最后通过通信域内的所有进程并发操作,将数据输出到各自的文件视口中,完成数据的输出,并且保证输出结果的正确性。

2.3 数据正确性验证

将最后输出的文件用验证程序读入内存,并与原始程序的输出结果进行逐个单元数据的数值对比,保证输出结果的正确性以及数据格式的一致性。

3 效率测试及分析

上述四种 I/O 方式分别在 2、4、8、16 个计算节点的规模下运行 3 次,统计时间为整个输出模块的运行时间,取中间值结果为最终结果(见图 2 ~ 图 5)。为了测试较大规模数据下不同 I/O 方式的效率,将计算区域扩大 200 倍并使用 8 个节点运行,并记录各模式的时间消耗(见图 6)。其中,并行 I/O 模式的时间取参与运算进程中时间最长者为有效时间,收集模式和分开输出模式 I/O 方式则将各进程时间相加作为有效时间。阻塞模式由于阻塞同步,最终有效时间综合了串行输出与阻塞等待两部分时间。

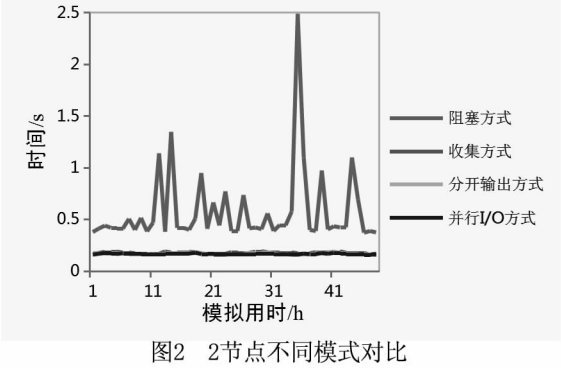

图2　2节点不同模式对比

Fig. 2 Comparison of different model runs with 2 nodes

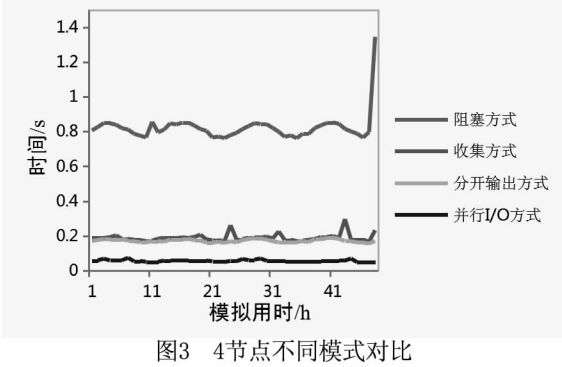

图3　4节点不同模式对比

Fig. 3 Comparison of different model runs with 4nodes

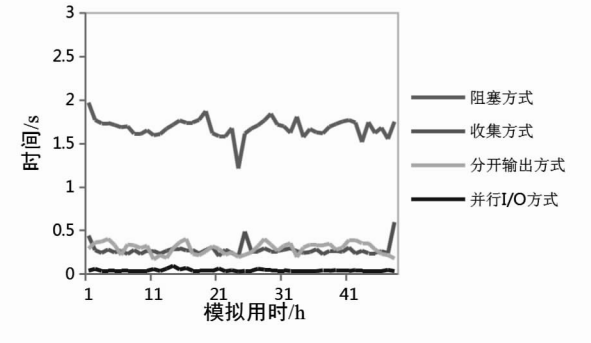

图4　8节点不同模式对比

Fig. 4 Comparison of different model runs with 8 nodes

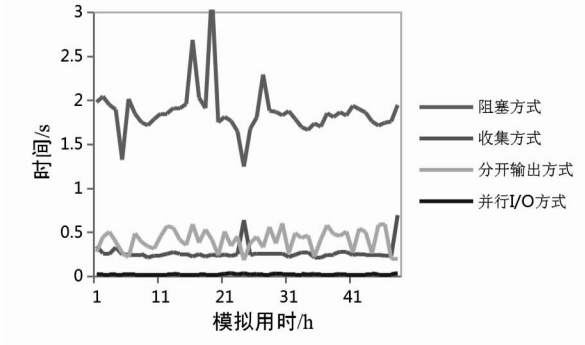

图5　16节点不同模式对比

Fig. 5 Comparison of different models run with 16 nodes

对比图 2 ~ 图 5 可以发现,尽管参与计算的节点数不同,并行 I/O 模式的输出时间非常稳定,并且进程数越多越具有效率上的优势。阻塞模式和收集模式相对于并行 I/O 模式的总输出时间倍数如表 2 所示:

通过数据可以发现,阻塞模式正如所预计一样,由于阻塞等待以及串行 I/O,消耗时间是最多的,并且随着进程数的增加效率呈递减趋势。收集模式则在进程数较少时具有一定的优势,随着进程数的增加,通信开销逐步增大,导致效率逐渐降低。分开输出模式在此只起到一定的参照作用(可以作为串行 I/O 最理想的状况),因为其输出结果仍需要后期的合并处理,所以一般不采用这种方式进行 I/O 操作。

从图 6 的数据可以看出在大量数据输出的情况下,阻塞模式已经完全不能满足高性能的需求,阻塞模式完成每段积分输出时间约 450 s 左右;收集模式约 55 s 左右,I/O 消耗已经成为降低系统的整体性能的一个严重瓶颈。收集模式则由于需要大量的数据传递,通信开销巨大,且由于只由一个进程进行 I/O 操作,又进一步加剧了时间消耗;而并行 I/O 模式则由于由多个进程同时进行 I/O 操作,速率得到大幅提升,并且在本系统中并行 I/O 采取了指定地址偏移写入,而不用每次写入前都进行寻址操作,也减少了时间开销。而在其余三种 I/O 中,由于使用编程语言中的语句输出函数进行逐行循环输出,每次调用输出函数时都需要进行寻址操作。

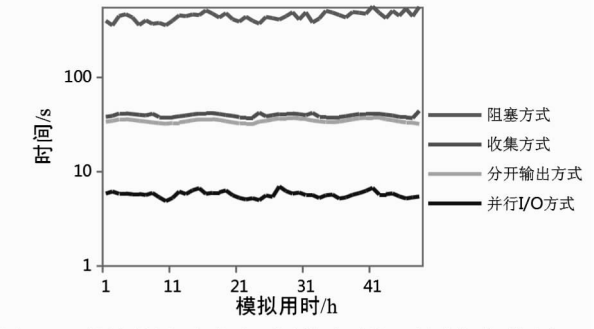

图6　200倍计算量下8节点不同模式对比(纵坐标指数坐标)

Fig. 6 Comparison of different model runs with 8 nodes under 200X computation (index coordinates)

表 2 不同 I/O 模式的总输出时间比

Tab. 2 Time of total output

节点数量	阻塞模式	收集模式
2 节点	3.5 倍	1.06 倍
4 节点	14.43 倍	3.33 倍
8 节点	39.6 倍	6.45 倍
16 节点	80.34 倍	11.51 倍
8 节点(200 倍计算量)	77.14 倍	6.93 倍

4 结论

文章提出的并行 I/O 模式,时间消耗最少,且效率稳定,随着节点的递增可以保持良好的加速比。但在编写并行 I/O 程序时比较复杂,需要对文件系统有详细的了解方能在保证数据正确的前提下成功应用。

参考文献:

[1]穆望舒. 海洋数值预报产品综合处理平台开发研究[D]. 上海:华东师范大学,2014.
[2]GUO CHENG,LU LIU,NING JING,et al. General purpose optimization method for parallelization of digital terrain analysis based on cellular automata [J]. Computers & Sciences,2012,45(8):57 - 67.
[3]李 亮,聂瑞华. 高性能计算平台的 I/O 性能测试与分析[J]. 计算机与现代化,2011,29(5):160 - 164 .
[4]董文睿,刘光明,刘 欣. 高强度 I/O 的应用对并行存储系统的挑战和解决方法研究[J]. 计算机研究与发展,2012,49(Z1):47 - 52.
[5]晏益慧,张 辉. 高性能计算机性能评测基准 HPCC 应用研究[J]. 计算机工程与科学,2009,31(1):279 - 282.
[6]于忠亮. 并行计算中的 I/O 问题研究[D]. 呼和浩特:内蒙古大学,2010.
[7]HUANG WEIJIAN,NIU PEI,DU WEI. Technology to water quality forecasting model of Jiao Zhou Bay [J]. World Journal of Engineering,2011,12(4):395 - 399.
[8]DENNIS J M,EDWARDS J,LOY R,et al. An application - level parallel I/O Library for Earth system models [J]. International Journal of High Performance Computing Applications,2012,26(1):43 - 53.
[9]张武生,薛 巍,李建江,等. MPI 并行程序设计实例教程[M]. 北京:清华大学出版社,2009.
[10]周建鑫,陈 荦,熊 伟,等. 地理栅格数据并行 I/O 的研究与实现[J]. 地理信息世界,2013,20(6):62 - 65.

(责任编辑 王利君)

(上接第 80 页)

地利用类型,对生物及周围环境产生一定影响。在未来建设中可继续鼓励正确的植树造林、植草,加强该地区蓄水能力,减缓地下水开采;减缓城市化进程,保护道路及水渠、水库附近的植被。

参考文献:

[1]王秀兰,包玉海. 土地利用动态变化研究方法探讨[J]. 地理科学进展,1999,18(1):81 - 87.
[2]傅伯杰,陈利顶,马克明. 黄土丘陵区小流域土地利用变化对生态环境的影响[J]. 地理学报,1999,54(3):241 - 246.
[3]朱运海,张百平,曹银璇,等. 土地利用/覆被变化遥感检测方法与应用分析[J]. 地球信息科学,2007(3):116 - 122.
[4]HAO HUI MEI. Land use/land cover change(LUCC) and eco - environment response to LUCC in Farming - Pastoral Zone,China[J]. Agricultural Sciences in China,2009(1):91 - 97.
[5]黄宝荣,张慧智,王学志. 城市扩张对北京市城乡结合部自然和农业景观的影响——以昌平区三镇为例[J]. 生态学报,2014(22):6756 - 6766.
[6]陈松林. 基于 GIS 的土壤侵蚀与土地利用关系研究[J]. 福建师范大学学报:自然科版,2000(1):106 - 109.
[7]邹亚荣,张增祥,周全斌,等. 基于 GIS 的土壤侵蚀与土地利用关系分析[J]. 水土保持研究,2002(1):67 - 69.
[8]陈龙泉,郑海金. 基于 Markov - CA 的土地利用/土地覆盖变化动态模型研究[J]. 测绘信息与工程,2004(1):36 - 38.
[9]李忠锋,王一谋,冯毓荪,等. 基于 RS 与 GIS 的榆林地区土地利用变化分析[J]. 水土保持学报,2003(2):97 - 99.
[10]GU XIAOHE. Dynamic monitoring and driving power analysis of LUCC based on remote sensing in Beijing in recent thirty years[J]. SPIE,2013,89(21):1 - 9
[11]史培军. 土地利用顺盖变化与生态安全响应机制度[M]. 北京:科学出版社,2004.
[12]赵东波,梁 伟,杨勤科,等. 陕北黄土丘陵区近 30 年来土地利用动态变化分析[J]. 水土保持通报,2008,28(2):22 - 28.

(责任编辑 王利君)