万修全,刘泽栋,沈飙,等. 地球系统模式 CESM 及其在高性能计算机上的配置应用实例[J]. 地球科学进展,2014,29(4):482-491,doi:10.11867/j. issn. 1001-8166. 2014. 04. 0482. [Wan Xiuquan, Liu Zedong, Shen Biao, et al. Introduction to the community Earth system model and application to high performance computing[J]. Advances in Earth Science,2014,29(4):482-491,doi:10.11867/j. issn. 1001-8166. 2014. 04. 0482.]

地球系统模式 CESM 及其在高性能计算机上的配置应用实例*

万修全,刘泽栋*,沈 飙,林霄沛,吴德星
(中国海洋大学物理海洋教育部重点实验室,中国海洋大学,山东 青岛 266100)

摘 要:通用地球系统模式(CESM)是美国国家大气研究中心最新推出的地球系统耦合模式,对解决气候(地球)系统建模中所涉及的新挑战和新问题具有很大的帮助。首先介绍 CESM 模式的结构框架以及最新版本的重要更新;然后结合具体的应用实例和使用经验,重点讨论如何在高性能计算机上对模式进行移植和合理的 CPU 配置,并比较不同配置之间的优劣性,从而确定模式最佳负载平衡和最优效率,对模式新用户的使用具有极大的帮助;最后对模式进行一系列的稳定性测试和验证,结果表明模式具有较好的稳定性,可以进行数值模拟和科学研究。同时对地球系统耦合模式的发展进行了总结,并对模式发展中存在的问题提出了一些建议。

关 键 词:CESM;高性能计算;CPU 配置;最优化;稳定性
中图分类号:P731 **文献标志码**:A **文章编号**:1001-8166(2014)04-0482-10

1 引 言

国际上对地球科学数值模式的高度重视极大促进了目前地球系统模式的快速开发应用,其中最具代表性的有:美国“共同体气候系统模式发展计划”(The Community Climate System Model,CCSM)和“地球系统模拟框架计划”(The Earth System Modeling Framework, ESMF),欧盟的“欧洲地球系统模拟网络”计划,日本的“地球模拟器” 计划。特别值得指出的是中国做为一个发展中国家,在地球系统模式领域做了大量工作,最近也启动了“全球变化研究国家重大科学研究计划”,进行我国的高分辨率气候系统模式的研制与评估,取得了一些令人瞩目的成绩。

本文介绍的是美国国家大气研究中心(NCAR)在 2010 年 6 月推出的通用地球系统模式 CESM(The Community Earth System Model)。它是在 CCSM4.0 基础上发展的地球系统模式。截至 2013 年 12 月模式更新至 CESM1.2.1 版本。CESM 模式是以海洋、大气、陆面和冰圈等为研究主体,并考虑大气化学、生物地球化学和人文过程的地球气候系统模式,在气候与环境的演变机理、自然和人类与气候变化的相互作用以及气候变化的研究和预测等诸多方面应用广泛[1,2]。

CESM 模式采用模块化框架,主体由大气、海洋、陆地、海冰、陆冰等几大模块组成,并由耦合器(CPL7)管理模块间的数据信息交换和模式运行。CESM 的各个模块都采用现阶段比较成熟的既有模式,其中大气模块采用 CAM (The Community Atmos-

* **收稿日期**:2013-11-28;**修回日期**:2014-03-14.
* **基金项目**:国家自然科学基金面上项目“天气噪声对大西洋经向翻转环流变异的作用”(编号:41276013);留学归国人员科研启动基金“太平洋经向模态对 ENSO 影响的数值研究”(教外司留 2012-1707 号)资助.
作者简介:万修全(1977-),男,山东日照人,副教授,主要从事物理海洋学和气候变化研究. **E-mail**:xqwan@ ouc. edu. cn
* **通讯作者**:刘泽栋(1987-),男,山东潍坊人,硕士研究生,主要从事物理海洋学研究. **E-mail**:zdliu@ ouc. edu. cn

phere Model),海洋模块采用 POP (The Parallel Ocean Program),陆地模块采用 CLM(The Community Land Model),海冰模块采用 CICE (The Los Alamos National Laboratory Sea-ice Model),陆冰模块采用 CISM (The Glimmer Ice Sheet Model)。模式中的各个模块都有几种不同的工作状态:active,data,dead,stub。CESM 可以根据实验目的和实验要求来选择模块组合形式(component set),不同的模块组合方式可以满足不同科学实验的要求,具有很强的灵活性和通用性。CESM 实现了模块的可插拔性,使模式操作简单,可持续发展能力较强。

我们成功将 CESM1.0.4 版本移植到中国海洋大学计算中心高性能计算机 Polaris 上,本文的所有实验都是基于这个版本(以下简称 CESM1);在实际移植及计算过程中发现不同的模式配置策略(PE layout)对其工作效率有不可忽视的影响。本文将结合这些具体的应用实例和经验,首先介绍了 CESM1 版本各个模块相对于之前的 CCSM4 版本的重要更新和改进,然后重点讨论如何在高性能计算机上对 CESM 进行合理的 CPU 配置。

虽然我国独立开发的气候耦合模式经过近 20 年的发展,在国内得到了比较广泛的应用,也取得了一系列的成果,特别是中国科学院大气物理研究所自主研发的耦合模式 FGOALS(The Flexible Global Ocean-Atmosphere-Land System Model)[3]。然而,从客观存在的总体实力和研究水平上讲,中国地球系统模式的发展与发达国家相比仍然存在一定的差距[1,4],相信本文的结果会对 CESM 的初学者和气候模式开发者起到帮助和借鉴意义。

2　CESM1 的重要更新和改进简介

CESM1 是由几个数值模块组合而成,每一个模块相对于其之前的版本 CCSM4 都有更新和改进,在其官方网站(http://www.cesm.ucar.edu/)有详细的介绍,本文在此仅做简单介绍,有兴趣的用户也可参考其他相关资料[5]。

2.1　大气模块

CESM1 中所使用的大气模块是 NCAR 的通用大气模式(The Community Atmosphere Model ,version 5,CAM5)。CAM5 相对其之前的 CAM4 版本而言,在物理过程和参数化方案等方面都有较大的修改和改善。利用改进的湿度扰动方案来模拟层云—辐射—湍流相互作用,从而有利于研究气溶胶的间接影响。利用云的宏观物理方案处理云过程,并改进层状云的微物理过程,使物理过程更加透明清晰,并且模拟结果更好。采用快速辐射通量传输方法的辐射方案,采用高效准确的 K 方法计算辐射通量和加热率,对于水蒸气宽谱的连续性和精度具有很大改善。大气模块中加入了化学过程和整层大气模块。

2.2　陆地模块

CESM1 的陆地模块(The Common Land Model,CLM)有了实质性的修改,具体包括增加了新的模型和功能、更新了模式的输入数据并修正了物理化参数方案。陆地模块中加入了碳氮循环过程、动态植被模型、城市模型和水文模型等新的物理过程和模型,首次采用动态陆地覆盖方案以保证全球能量守恒,并对陆地径流和冰山进行特殊化处理以保证全球质量守恒。特别在模式中加入了农作物的生长和灌溉等人类耕种活动,更好地反映了人类活动对地球和气候的影响[6]。从 CESM1.1 版本开始,径流模块(The River Transport Model,RTM)从 CLM 中独立出来成为一个单独的子模块,因而可以更好地模拟地球上的径流系统及其对地球系统的影响。

2.3　海冰模块

CESM1 的海冰模块(The Community Ice CodE,CICE)的主要改进是在物理过程和参数化方案以及模式运算方面。其中物理过程和参数化方案的改进主要包括:更新修正了海冰的示踪方案和短波辐射传输方案,改进了冰雪融化方案和气溶胶沉积方案。海冰模块的计算性能有很大的改进,主要包括:采用更加灵活和方便的计算方法,提高了运算速度和效率;提高了模式分辨率,使得能够模拟更小尺度的物理过程;优化了数据的输入和输出接口,使得数据传输和交换更加快捷和高效[7]。

2.4　海洋模块

CESM1 的海洋模块(The Parallel Ocean Program,POP)的主要结构功能和物理参数化方案基本没有变化,其主要改进是增加了海洋生态系统模型。海洋中植物对能量分布有不可忽视的影响,作为全球碳循环模式的一个组成部分,实现了生物地球化学过程与物理海洋过程的相互作用和反馈[8]。从 CESM1.2 版本开始,海浪模块(The Wave Model,WAV)也被加入到模式中。

2.5　陆冰模块

CESM1 增加了一个新的模块——陆冰模块,其采用通用陆冰模式(Glimmer-CISM),主要研究陆冰及其与其他地球系统的相互作用和影响,模拟大尺度的北极格陵兰岛和南极的陆地冰,也可以模拟更

小尺度的冰山、冰帽以及陆冰的变化。陆冰模式处于不断发展阶段,其物理过程和参数化方案还有待于进一步完善。现有模式只是研究冰架内部的运动,而不能研究冰架和冰的流动;并且陆冰模式与陆地模式是单向的,即陆冰模块只从陆地模块获得初始场,但是冰对地形的改变等不会进一步传递给陆地模块。

2.6 耦合器(CPL)

CESM 模式采用模块化框架,耦合器(CPL7)负责管理模块间的数据交换和模式运行。耦合器的功能主要包括:把 CESM 分割为几个独立的子模式模块,包括海洋、大气、海冰、陆冰、径流、海浪模块等,模块之间通过 MPI 交换数据;同步协调和控制各模块之间的数据流,以此来控制整个 CESM 的运行和时间积分;控制各模块之间进行界面通量的交换,并保证通量守恒。耦合器通过控制各子模式之间的数据消息交换来控制整个模式系统的运行。耦合器框架结构已经成为目前耦合气候系统模式设计的最佳方案,即将耦合器作为一个工具软件,把各子模式很方便地连接起来,构建一个完整的气候系统模式[9]。

3 模式运行环境

CESM 是一个比较复杂的地球系统模式,对运算的计算平台有较高的要求,主要包括计算平台的硬件和软件条件、并行应用的运行时环境以及机群作业管理系统。

3.1 高性能计算平台 Polaris

中国海洋大学计算中心的高性能计算机 Polaris 于 2012 年底正式启用。该计算机平台分 2 期建设,目前已有 CPU 总共 3 132 核,计算能力峰值约 33.32 万亿次/s,实测 29 万亿次/s,效率达到 89.8%;单节点采用 2 颗 Intel Xeon 5650 CPU,每颗 CPU 12 核心,主频 2.6 GHz;拥有裸容量 400 T 的高速并行文件存储系统。具体的参数指标如表 1 所示。

表 1 中国海洋大学计算中心的高性能计算机 Polaris 主要性能参数

Table 1 The main performance parameters of High Performance Computer—Polaris at Ocean University of China

节点	CPU/内存	数量/台
登陆/管理节点	2 * Intel Xeon X5670 (2.93GHz/6.4 GT/12M/6 核) 48GB Registered ECC 1333MHz DDR3	2
IO 节点	2 * Intel Xeon E5620 (2.4GHz/5.86 GT/12M/4 核) 24GB Registered ECC 1333MHz DDR3	16
计算节点	2 * Intel Xeon- X5650 (2.6GHz/6.4 GT/12M/6 核) 24GB Registered ECC 1333MHz DDR3	261

3.2 模式运行环境配置

本文所有的 CESM 数值实验的移植、配置、测试等相关工作在中国海洋大学计算服务中心 Polaris 上完成,所用到的具体软件环境如表 2 所示。

表 2 模式移植和运行时的编译环境及运算环境

Table 2 The compiled environment and the computing environment

操作系统		Red Hat Enterprise Linux Server release 5.4 (Tikanga)
编译环境	编译器	Intel c/c++/fortran 77/90/95 编译器、调试器和性能分析工具 PGI c/c++/fortran 77/90/95 编译器、调试器和性能分析工具 GNU 编译器,支持 c/c++ fortran 77/90/95
运算环境	数学库	BLAS, LAPACK, ScaLAPACK, FFTW, NETCDF3/4, HDF, NCL_NCARG, NCO, Ncview, OCTAVE
	并行环境	OpenMPI, MVAPICH(支持 Infiniband 和以太网的 MPI 环境),MPICH(支持千兆网的 MPI), OpenMP, PVM, OCTAVE

在本文的实验运算中,我们采用 Intel c/c++/fortran 77/90/95 编译器,运用 OpenMPI 进行并行运算。

4 模式的 CPU 配置策略(PE layout)

为了追求计算高效率的同时节省计算资源,就必须进行最优化调试。模式的最优化意味着模式的输出量和消耗量达到最优化。对于一定的 CPU 核数而言,最优化意味着输出量最大。但是模式的最优化是相对的,对于不同的 CPU 核数,就需要找到一个最优化的平衡点:模式具有高输出量的同时具有低能耗。跟其他大多数模式一样,增加 CESM 的 CPU 核数会同时增加模式的输出量和消耗量。由于模式的运算不是线性的,因此核数的增加会导致

模式的消耗量增加。因此在进行较长时间的模拟实验之前,非常有必要进行 CPU 的最优化配置测试。为此,我们设计了一系列实验来测试移植以及配置。测试算例采用全耦合状态(B_1850-2000_CN),分辨率为 0.9×1.25_gx1v6,算例中各模块状态以及分辨率等详细信息如表 3 所示。

表 3　测试算例的各个模块的状态和分辨率情况

Table 3　The component sets and resolution of each model in the test case

子模块	大气模块	陆地模块	海洋模块	海冰模块	陆冰模块	耦合器
子模块模式	cam	clm	pop	cice	sglc	cpl
工作状态	active	active	active	active	stub	active
水平网格	0.9×1.25	0.9×1.25	gx1v6	gx1v6	/	/
网格类型	finite volume	finite volume	displaced pole	displaced pole	/	/

CESM 的 CPU 配置中需要设置的参数有:MPI task 的数目、线程数目和起始位置等。模块分辨率以及模块状态对每个配置参数的要求均有差别,所以参数配置对模式运行速度和效率会产生不可忽视的影响。因此在接下来的实验中我们重点探究不同的 CPU 配置策略及其对模式运算速度和效率的影响,以期找到一个最优化的配置策略。

为尽量减小偶然误差和不确定因素对模式运行的影响,我们对每个算例单独运算 5 次,本文中所用的模式数据均为多次运算模式的平均值。其中每个测试算例模拟 10 年。

4.1　配置实验一

首先,将所有模块顺序运行,这也是模式在高性能计算平台 Polaris 上默认采用的配置策略。由于是顺序运行,每个模块的 MPI task 起始位置相同,因此只需改变 MPI task 数目(即每个模块的 CPU 核数)即可。据此我们设计了实验一,模式运行数据如表 4 所示,运算速度和效率如图 1 所示。

表 4　实验一(全部模块顺序运行)的 CPU 配置情况以及模式运算速度和资源消耗情况

Table 4 Processor layout and the cost-throughput of components, with fully sequential options

算例	CPU 核数	消耗量/(核小时数/模式年)	输出量/(模式年/天)	大气模块	陆地模块	海冰模块	耦合器	海洋模块
				320	128	320	448	64
0 * *	512	1 191.21	10.32	320×1	128×1	320×1	448×1	64×1
				0	320	0	0	448
				128	128	128	128	128
1 *	128	632.82	4.85	128×1	128×1	128×1	128×1	128×1
				0	0	0	0	0
				256	256	256	256	256
2	256	1 177.18	5.22	256×1	256×1	256×1	256×1	256×1
				0	0	0	0	0
				320	320	320	320	320
3	320	1 067.04	6.50	320×1	320×1	320×1	320×1	320×1
				0	0	0	0	0
				384	384	384	384	384
4	384	1 811.24	5.09	384×1	384×1	384×1	384×1	384×1
				0	0	0	0	0
				512	512	512	512	512
5	512	2 851.06	4.31	512×1	512×1	512×1	512×1	512×1
				0	0	0	0	0
				640	640	640	640	640
6	640	4 331.25	3.55	640×1	640×1	640×1	640×1	640×1
				0	0	0	0	0

注:算例 0 * * 中的 CPU 配置策略是 NCAR 所采用的,我们按照此配置在 Polaris 上运行得到算例 0 * * 作为比对算例;算例 1 * 是模式在 Polaris 上默认的配置策略。其中每个算例中各模块的 CPU 配置中具体数字的解释:比如算例 0 * * 中大气模块的 CPU 配置为:320,320×1,0, 表示本模块占用 320 个 CPU,采用单线程(×1),从第 0 个 CPU 位置开始,占用第 0 至 319 个 CPU;陆地模块的 CPU 配置 128,128×1,320,表示本模块占用 128 个 CPU,采用单线程(×1),从第 320 个 CPU 位置开始,占用第 320 至 447 个 CPU,其余类同

在实验一中,我们均采用模式在 Polaris 上默认的配置策略,即各模块采用单线程顺序运行。据此我们设计了 6 个实验(其中算例 0 * * 仅作为比对算例),各模块核数和总核数都在增加。我们用 2 个量来表征运算速度和效率,其中将模式运算 1 天所输出的数据量作为模式运算速度的指标;将每个 CPU 所消耗的资源作为模式运算的效率。模式输出量越大,表征运算速度越快;每个 CPU 所消耗的资源越少,表征运算效率越大。我们根据实验一的表格数据计算了模式的运算速度和效率(图 1)。

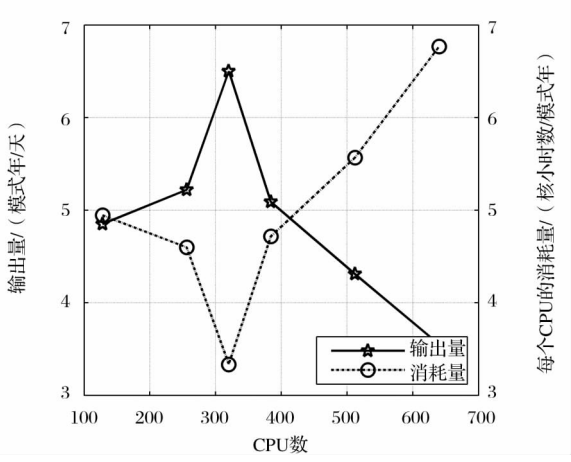

图 1 实验一的模式运算速度和效率曲线图

Fig. 1 The cost-throughput of components with fully sequential options

其中横轴是 CPU 核数;左侧纵轴是模式每天所模拟的时间(图中实线),表征模式运算速度;右侧纵轴是每个 CPU 所消耗的资源量(图中虚线),表征模式运算的效率

In the figure, the x-axis is the number of CPUs of all components; The left y-axis is the throughput, with units: Simulated_years/day, and the right is the cost, with units: Pe-hrs/simulated_year

通过分析实验一的数据(表 4)以及其运算速度和效率曲线图(图 1),可以发现,在最开始核数增加时,即从算例 1 到算例 3,模式运算速度有一定的提高,从每天能模拟 4.85 模式年提高到 7.37 模式年,其资源消耗也有一定程度的减小,即运算的效率也在提高;但是当核数继续增大时,即算例 4 ~6,模式的运行速度反而降低,并且资源消耗也在持续变大、效率降低。因此算例 3 在实验一中是最高效的,即运算速度快并且其消耗资源也少。

由于本次实验中模式分辨率约为 1°×1°,海洋模块的全球网格的格点数为 384×320。当采用核数较多时,每个 CPU 所负责计算的区域面积较小,因此可以在一定程度上增加运算速度;但是采用的核数多,使得区域划分个数增加,同时也会增加 MPI 并行运算时各个 CPU 间数据的交换量,影响总的运算速度。因此,对于并行运算要求比较高的 CESM 模式而言,并不是单纯提高 CPU 核数就会提高其运算速度和效率。并且对比发现,采用单线程、各模块顺序运行的策略(即实验一的配置策略)时,模式运行速度普遍很慢、消耗资源很大,效率很低,因此采用各模块顺序运行的 CPU 配置策略不具有实际应用价值。

4.2 配置实验二 A

通过进一步分析发现,对模式运算速度影响最大的主要是大气和海洋模块。当这 2 个模块顺序运算时,会使运算的总时间增加,从而使得模式运算的速度减小,因此考虑将海洋模块和大气模块并行运算。由于 CESM 模式中陆地模块与大气模块只能顺序运算,因此采用的策略是海洋模块与其余模块(大气、海冰、陆地和 CPL)并行运算,但大气、海冰、陆地和 CPL 之间均采用顺序运算,据此我们设计了实验二 A,详细的输出及配置如表 5 所示,运算速度和效率如图 2 所示。

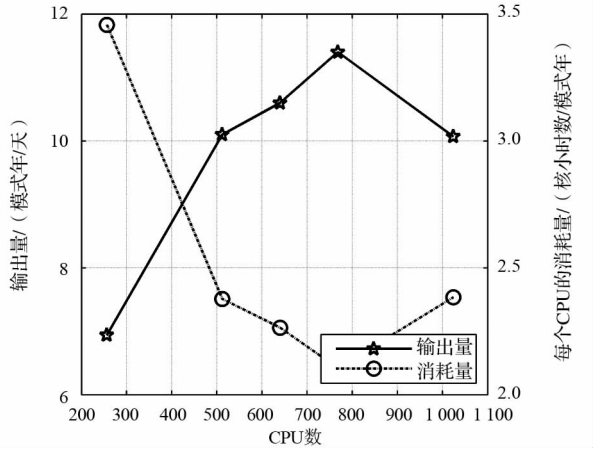

图 2 实验二 A 的模式运算速度和效率曲线图

Fig. 2 The cost-throughput of components with fully sequential options except the ocean running concurrently

其中横轴是 CPU 核数;左侧纵轴是模式每天所模拟的时间(图中实线),表征模式运算速度;右侧纵轴是每个 CPU 所消耗的资源量(图中虚线),表征模式运算的效率

In the figure, the x-axis is the number of CPUs of all components; The left y-axis is the throughput, with units: Simulated_years/day, and the right is the cost, with units: Pe-hrs/simulated_year

通过对实验一和实验二 A 的数据分析,发现实验二 A 中的模式计算速度在 10 模式年/天左右(图 2),较实验一的速度(5模式年/天左右,图1)有很

表 5　实验二 A 的 CPU 配置情况以及模式运算速度和资源消耗情况

Table 5　Processor layout and the cost-throughput of components, with fully sequential options except the ocean running concurrently

算例	CPU 核数	消耗量 /(核小时数/模式年)	输出量 /(模式年/天)	大气模块	陆地模块	海冰模块	耦合器	海洋模块
				128	128	128	128	128
1	256	885.01	6.94	128×1	128×1	128×1	128×1	128×1
				0	0	0	0	128
				256	256	256	256	256
2	512	1 216.98	10.10	256×1	256×1	256×1	256×1	256×1
				0	0	0	0	256
				320	320	320	320	320
3	640	1 449.23	10.60	320×1	320×1	320×1	320×1	320×1
				0	0	0	0	320
				384	384	384	384	384
3	768	1 616.40	11.40	384×1	384×1	384×1	384×1	384×1
				0	0	0	0	384
				512	512	512	512	512
4	1 024	2 440.65	10.07	512×1	512×1	512×1	512×1	512×1
				0	0	0	0	512

注:其中 CPU 配置策略采用海洋模块与其余模块(大气、海冰、陆地和 CPL)并行运算,但大气、海冰、陆地和 CPL 之间均采用顺序运算

大提高,并且模式运算的资源消耗也有所降低。在本实验中随着核数的增大,从算例 1 至算例 4,运行速度和效率均有了较大的提高,到算例 5 时速度和效率都减小,这一点在实验一中这也得到验证(核数增大到一定程度时,运算速度和效率均降低)。进一步分析发现算例 3,4 的核数远比算例 2 多,但其运算速度和效率变化却不是特别明显,因此算例 3,4 的实际应用价值不如算例 2 好。

因此对于分辨率和状态确定的算例,其运算速度和效率不仅仅与分配核数(总核数、各个模块的核数)有关,还与模块间的串并行方案有关。增加核数仅在一定范围内使得模式的运算速度和效率增加,当超过这一范围后,反而会使模式运算速度和效率降低。

4.3　配置实验二 B

进一步分析实验二 A 的详细输出数据,发现算例 2 中大气模块的核数不是模式运行的限制因素,因此进一步增加大气模块的核数对模式的影响已经不大,因此实验二 A 中的算例 3,4 虽然增加了使用的核数,但是相对于算例 2 而言,其运算速度和效率增加不明显,原因与海洋模块所分配的核数有关。

因此以实验二 A 中的算例 2 为基准设计了实验二 B 来测试海洋模块分配核数对模式运行速度和效率的影响。在实验二 B 中海洋模块和其他模块同时运行,但是通过改变海洋模块分配的核数进行了一系列实验,得到算例 1 至算例 5,其中本次实验中的算例 5 就是实验二 A 中的算例 2,详细配置见表 6,运算速度和效率如图 3 所示。

在实验二 B 中(表 6,图 3),我们发现减少海洋模块的 CPU 之后,大部分算例都比算例 5(即实验二 A 中的算例 2)的运算速度快并且效率也要高,其中运算最高效的是算例 2,总用核数为 336CPU,海洋模块为 80 核,其余模块均为 256 核,运算输出量为 13.56 模式年/天,模式消耗为 594.74 核小时数/模式年;而其中只有算例 1 运算速度比算例 5 慢,原因是海洋分配的 CPU 过少,导致其运算速度和效率都变慢。因此,海洋模块分配的 CPU 不能过多,也不能过少。在保证每个模块分配的 CPU 核数不是计算瓶颈的前提下,适当减少海洋模块的核数反而在一定程度上使得模式运行速度提高、资源消耗减少。

对于全球 1°分辨率的全耦合算例而言,海洋模块分配 80 ~ 120 个 CPU 就可以足够保证其运算速度和效率维持在一个高水平上。但是由于 CESM 是计算密集型应用,对资源的消耗相当大,计算平台的内存可能是模式运算的瓶颈。因此在不同的计算平台上运行模式时,也不容忽视内存对模式运算的速度和效率的影响。所以要综合考虑模块分配的核数以及内存的影响,不能单纯增加或者减少 CPU 数量。

结合实验一和实验二,我们发现模式的运行速度与模式分配的总核数、各个模块分配的核数以及模块之间的并行策略等均有关。CPU核数在一定

表 6 实验二 B 的 CPU 配置情况以及模式运算速度和资源消耗情况

Table 6 Processor layout and the cost-throughput of components, with fully sequential options except the ocean running concurrently

算例	CPU 核数	消耗量 /(核小时数/模式年)	输出量 /(模式年/天)	大气模块	陆地模块	海冰模块	耦合器	海洋模块
				256	256	256	256	64
1	320	783.19	9.81	256×1	256×1	256×1	256×1	64×1
				0	0	0	0	256
				256	256	256	256	80
2	336	594.74	13.56	256×1	256×1	256×1	256×1	80×1
				0	0	0	0	256
				256	256	256	256	96
3	352	710.86	11.88	256×1	256×1	256×1	256×1	96×1
				0	0	0	0	256
				256	256	256	256	128
4	384	782.77	11.77	256×1	256×1	256×1	256×1	128x1
				0	0	0	0	256
				256	256	256	256	256
5	512	1 216.98	10.10	256×1	256×1	256×1	256×1	256×1
				0	0	0	0	256

注:其中 CPU 配置策略采用海洋模块与其余模块(大气、海冰、陆地和 CPL)并行运算,但大气、海冰、陆地和 CPL 之间均采用顺序运算

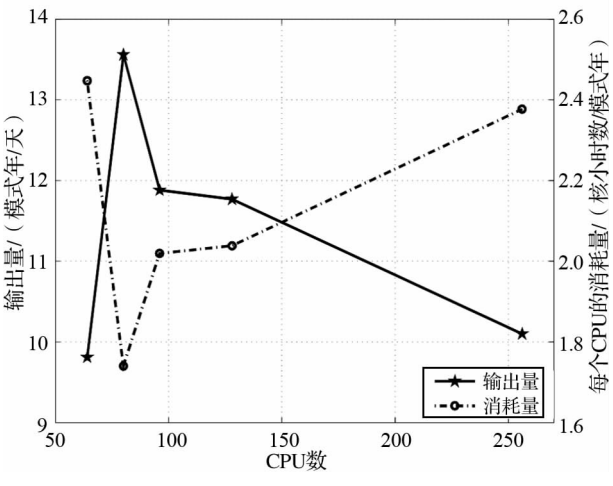

图 3 实验二 B 的模式运算速度和效率曲线图

Fig. 3 The cost-throughput of components with fully sequential options except the ocean running concurrently

其中横轴是海洋模块的 CPU 核数;左侧纵轴是模式每天所模拟的时间(图中实线),表征模式运算速度;右侧纵轴是每个 CPU 所消耗的资源量(图中虚线),表征模式运算的效率

In the figure, the x-axis is the number of CPUs of ocean component; The left y-axis is the throughput, with units: Simulated_years/day, and the right is the cost, with units: Pe-hrs/simulated_year

的范围内对模式运算有一定的影响,但是模块间的并行策略也起到很重要的作用;模式的运算速度和效率主要受大气模块和海洋模块的影响,其中大气的影响可能占的比重更大一些。模式的 PE 配置时,在满足大气模块的 CPU 核数分配的前提下,将海洋模块跟其余模块并行,并适当分配海洋模块的 CPU 数就可以使满足模式运行的速度和效率[10]。

5 模式移植稳定性验证

对于一个成熟的模式,其运算结果应该认为与所使用的运算工具无关,即模式的结果不受计算平台硬件、软件等计算环境的影响。但是由于 CESM 模式的复杂性,不能排除硬件以及计算配置等对结果的影响,因此需要验证移植的准确性和精确度。为此我们进行了一系列的敏感性试验,并将结果与 NCAR 的 CESM 模式组已经测试成功的实验结果进行对比,以此验证模式移植的准确性和精确度。

在稳定性试验中,这里主要测试海洋模块(POP)的准确性和精确度,因此我们采用 C-compset,即只有海洋模块为 active 状态、其余模块为 data 或者 stub 状态;分辨率为 T62_gx1v6[11]。数据输出是以每模式计算步(step)作为输出间隔。本次试验共有 5 个子实验算例,分别验证模式对模式收敛度、模式迭代次数、机器 PE layout 配置等参数的敏感度。如表 7 所示是各个实验的配置,其中算例 1 是模式默认的配置,算例 2 中改变了海洋模块分配的 CPU 数,算例 3 中将模式的收敛度由 1.0e-13 改为 1.0e-14,算例 4 改变了收敛度、迭代次数,算例 5 在算例 4 的基础上又改变了海洋模块分配的 CPU 数。

模式运行 1 年后,取最后 5 天计算各个实验中海表面高度的均方根进行对比(图 4)。

表 7　稳定性测试实验中各个算例的情况

Table 7　Port-validation information of POP2 on Polaris

算例	模式收敛度	模式迭代次数	CPU 核数配置
1	1.0e-13	1 000	NTASKS_OCN = 64
2	1.0e-13	1 000	NTASKS_OCN = 48
3	1.0e-14	1 000	NTASKS_OCN = 64
4	0	500	NTASKS_OCN = 64
5	0	500	NTASKS_OCN = 48

图 4a 中验证了改变海洋模块 CPU 数对模式结果的影响,发现SSH的均方差在1.0e-10 ~ 1.0e-7 之间,即当将模式移植到计算平台上运行时,CPU 的分配数量对模式运算的结果影响很小,甚至可以忽略;图 4b 中验证了模式收敛度变化对海洋模式结果的影响,发现 SSH 的均方差在 1.0e-7 ~ 1.0e-6 之间,即由于计算平台计算能力的不同使得模式的计算迭代次数以及收敛度均不同,但是其对模式计算的影响也很小;图 4c 中验证了同时改变收敛度和迭代次数后,不同的 CPU 分配对模式的影响,发现 SSH 的均方差在 1.0e-13 ~ 1.0e-10 之间,即使同时改变计算平台的计算能力和 PE layout,模式运算的准确性仍然可以得到保障。

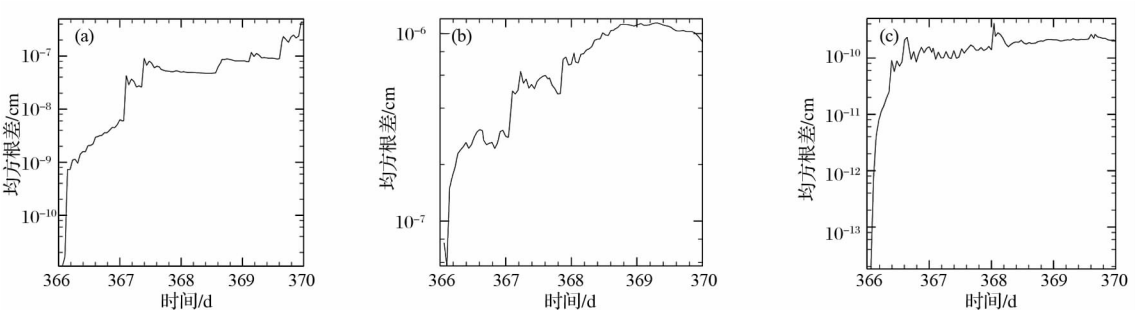

图 4　稳定性测试实验中各个算例间的海表面高度的均方根的差值

Fig. 4　Timeseries of RMS differences of the SSH field in different port case on the new system Polaris

(a)是算例 1 与算例 2 的差,(b)是算例 1 与算例 3 的差,(c)是算例 4 与算例 5 的差

In the figure, it respectively shows the differences of case1 with case2(a), case1 with case3(b), case4 with case5(c)

为了检验模式在不同的计算平台上的模拟效果,我们将 Polaris 上的敏感性试验与 NCAR 官方的 Bluefire 机器的运算相同算例的结果进行对比,结果如图 5 所示。

从图 5 中可以发现,不管是改变收敛度、迭代次数还是 PE 配置,2 个计算平台模拟的结果很相近,差别比较小,可以看出计算环境的变化对模式的运算的影响基本可以忽略。

从以上的一系列敏感性试验以及其他的验证实验中发现,计算平台的计算能力、运算环境以及 PE 配置等对 CESM 模式的结果不会产生严重的影响。因此 CESM 模式是一个移植性较好、比较成熟的模式,可以应用在不同的计算平台上进行大规模的科学实验和研究。

6　结　语

CESM(通用地球系统模式)是在 CCSM(通用气候系统模式)的基础上发展起来的,是研究海洋和大气等地球系统的一个很有力的工具,对解决地球系统的新挑战和新问题具有很大的帮助。模式的程序框架和物理方案比较先进,并且具有较强的可移植性和高并行运算能力。科学家们能够获得对地球系统更广泛更清晰的研究和认知,能够更好地描述客观世界[1,12]。

在本文的一系列的移植测试中,我们使用 f09_g16 分辨率的 CESM1.0 模式的 B_1850-2000_CN 算例,针对其各模块的运行配置情况统计了各模块的计算量,对各模块所分配的核数和并行方案进行了测试。在 Polaris 的计算平台上,该算例最优化运算配置为使用 336 个 CPU 核,其中大气模块 CAM 分配 256 核,海洋模块 POP 采用 80 核,陆地模块 CLM 使用 256 核,海冰模块 CICE 使用 256 核,耦合器 CPL 使用 256 核;运算速度约为 13.56 模式年/天。

在对 CESM 的配置测试试验中,我们发现整个地球系统模式中大气模块是运算最大的瓶颈,其计算量最大、耗时最长,并且随着大气模块分辨率的提高,模式的计算量和消耗会明显增加,因此必须首先保证大气模块的运算,才能使整个模式运算速度和效率提高。CESM 对计算平台的并行运算能力要求比较高,各个模块的并行方案和 CPU 核数的选取对

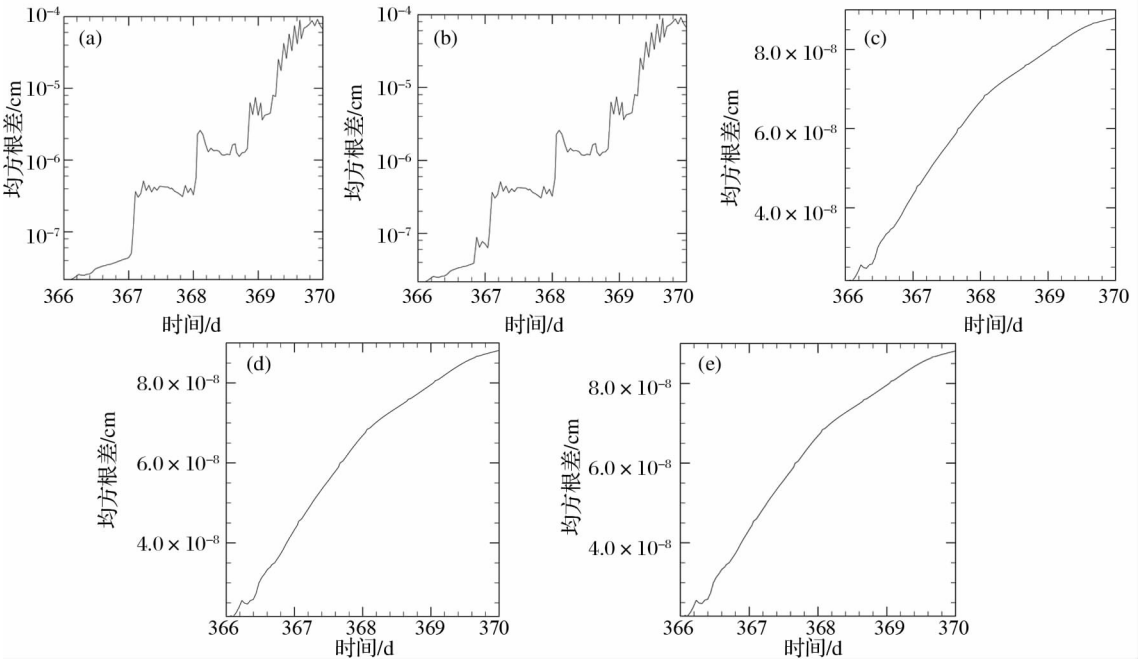

图 5 Polaris 与 NCAR 机器 Bluefire 相同算例(C-compset)的海表面高度的均方根的差值

Fig. 5 Timeseries of RMS differences of the SSH field in each port case between the Polaris solutions and those generated on a "trusted machine," the NCAR IBM bluefire

其中(a)~(e)分别对应算例 1~5

In the figure, it respectively shows case 1 to case 5; figure (a) is case 1, and so on

整个模式的运算速度和效率都会有很大的影响。通过测试发现,在 Polaris 计算平台上,采用海洋模块与其余模块并行的方案是最优选择,既能保证所用计算资源最少,同时还能确保计算速度和效率能够达到相对最优化。在确保大气模块不是运算瓶颈的前提下,海洋模块的 CPU 核数对模式的运算也有很大的影响,过多的海洋模块 CPU 核数,会使运算速度和效率降低。

由于地球系统模式的复杂性和高计算量,属于典型的计算密集型程序,其对计算平台的内存也提出了很高的要求。计算平台的多线程运算,也会使得运算速度和效率有进一步的提高。而随着模式分辨率的提高,地球系统模式在高性能计算机上要使用上千 CPU 核、数千 CPU 核甚至数万 CPU 核进行计算[13,14],对计算平台的计算性能和技术又提出了新的需求和挑战。

地球本身就是一个很庞大和复杂的系统,对其的研究就要涉及到更多系统和因素,其复杂性和艰难性不言而喻[12]。现在包括 CESM 在内的模式都还存在很多的问题,模式的研究还不是很完善,也需要结合观测资料对模式的物理过程和参数化进行进一步改进[15]。这需要科学家和相关技术人员进行研究和改进,使之能够更好的模拟现实并预测未来。

参考文献(References):

[1] Wang Bin, Zhou Tianjun, Yu Yongqiang. A perspective on Earth system model development[J]. *Acta Meteorologica Sinica*, 2008, 66(6): 857-869. [王斌, 周天军, 俞永强. 地球系统模式发展展望[J]. 气象学报, 2008, 66(6): 857-869.]

[2] Zheng Peinan, Song Jun, Zhang Fangran, *et al*. Common instruction of some OGCM[J]. *Marine Forecasts*, 2008, 25(4): 108-120. [郑沛楠, 宋军, 张芳苒, 等. 常用海洋数值模式简介[J]. 海洋预报, 2008, 25(4): 108-120.]

[3] Zhou Tianjun, Yu Yongqiang, Liu Hailong, *et al*. Progress in the development and application of climate ocean models and ocean-atmosphere coupled models in China[J]. *Advances in Atmospheric Sciences*, 2007, 24(6):729-738.

[4] Zou Liwei, Zhou Tianjun. A review of development and application of regional ocean-atmosphere coupled model[J]. *Advances in Earth Science*, 2012, 27(8):857-865. [邹立维, 周天军. 区域海气耦合模式研究进展[J]. 地球科学进展, 2012, 27(8): 857-865.]

[5] Vertenstein M, Craig T, Middleton A, *et al*. CESM-1. 0. 4 User's guide[R/OL]. Boulder: National Center for Atmospheric Research, 2012. [2013-12-22]. http://www.cesm.ucar.edu/models/cesm1.0/cesm/cesm_doc_1_0_4/book1.Html.

[6] Lawrence D M, Oleson K W, Flanner M G, *et al*. Parameterization improvements and functional and structural advances in version 4 of the community land model[J]. *Journal of Advances in Modeling Earth Systems*, 2011, 3(1):1-27, doi:10.1029/2011MS00045.

[7] Lipscomb W H, Hunke E C, Maslowski W, *et al*. Ridging, strength, and stability in high-resolution sea ice models[J]. *Journal of Geophysical Research: Oceans* (1978-2012), 2007, 112(C3), doi:10.1029/2005JC003355.

[8] Smith R, Gent P, Briegleb B, *et al*. The Parallel Ocean Program (POP) reference manual[R]// Technical Report LAUR-10-01853. Los Alamos: Los Alamos National Laboratory, 2010.

[9] Zhou Tianjun, Yu Yongqiang, Yu Rucong, *et al*. Coupled climate system model coupler review[J]. *Chinese Journal of Atmospheric Sciences*, 2004, 28(6):993-1 008, doi:10.3878/j.issn.1006-9895.2004.06.16.[周天军, 俞永强, 宇如聪, 等. 气候系统模式发展中的耦合器研制问题[J]. 大气科学, 2004, 28(6): 993-1 008, doi:10.3878/j.issn.1006-9895.2004.06.16.]

[10] Dowd K, Severance C R, Loukides M K. High Performance Computing[M]. California: O'Reilly, 1998.

[11] Danabasoglu G, Bates S C, Briegleb B P, *et al*. The CCSM4 ocean component[J]. *Journal of Climate*, 2012, 25(5): 1 361-1 389.

[12] Zeng Qingcun, Lin Zhaohui. Recent progress on the Earth system dynamical model and its numerical simulations[J]. *Advances in Earth Science*, 2010, 25(1): 1-6.[曾庆存, 林朝晖. 地球系统动力学模式和模拟研究的进展[J]. 地球科学进展, 2010, 25(1): 1-6.]

[13] Pu Ye, Li Lijuan. The application of thousands of CPU cores in high resolution Earth system model[J]. *e-Science Technology & Application*, 2010,1(4): 69-75.[普业, 李立娟. 高分辨地球系统模式的千核应用[J]. 科研信息化技术与应用, 2010,1(4): 69-75.]

[14] Wang Bin. A typical type of high-performance computation: Earth system modeling[J]. *Physics*, 2009, 38(8):569-574.[王斌. 一种典型的高性能计算: 地球系统模拟[J]. 物理, 2009, 38(8):569-574.]

[15] Wu Lixin, Chen Zhaohui. Progresses and challenges in observational studies of physical oceanography[J]. *Advances in Earth Science*, 2013, 28(5):542-551.[吴立新, 陈朝晖. 物理海洋观测研究的进展与挑战[J]. 地球科学进展, 2013, 28(5): 542-551.]

Introduction to the Community Earth System Model and Application to High Performance Computing

Wan Xiuquan, Liu Zedong, Shen Biao, Lin Xiaopei, Wu Dexing

(*Physical Oceanography Laboratory of the Ministry of Education, Ocean University of China, Qingdao* 266100, *China*)

Abstract: The Community Earth System Model (CESM) is a fully-coupled global climate model, and is maintained by the National Center for Atmospheric Research (NCAR). Composed of several separate models simultaneously simulating the earth's atmosphere, ocean, land surface, sea-ice, land-ice, river transport and wave, and one central coupler component, the CESM allows researchers to conduct fundamental research into the earth's past, present and future climate states. CESM1 contains totally new infrastructure capabilities, the implementation of a coupling architecture, and model parameterization development. These permit new flexibility and extensibility to address the challenges involved in earth system modeling with ultra high resolution simulations on High Performance Computing (HPC) platforms using tens-of-thousands of cores. Firstly, the infrastructure of the model is introduced, and also the notable improvements. The CESM1 coupling architecture provides "plug and play" capability of data and active components and includes a user-friendly scripting system and informative timing utilities. Then, the processor (PE) layout is customized for the load balancing on high-performance computers to optimize the throughput or efficiency of a CESM experiment. At the end of the paper, the port validation and model verification are made for the ocean model—the Parallel Ocean Program version 2 (POP2) which has properly ported to the machine—Polaris at Ocean University of China. The POP2 model output is subsequently verified to be a successful port, and CESM1 POP2 ocean-model solutions are the same as solutions generated on a trusted machine—bluefire at NCAR. Together, it enables a user to create a wide variety of "out-of-the-box" experiments for different model configurations and resolutions and also to determine the optimal load balance for those experiments to ensure maximal throughput and efficiency. The results and experiments will provide useful experience and method to the new CESM users to make simulations and load balancing of the model.

Key words: The community Earth system model; High performance computing; CPU Processor layout; Optimization; Stability.