

海洋环境数值预报业务综合运维平台设计与实现

张志远, 孙立尹, 何锡玉, 王少可

(海参指挥保障大队战场环境保障室, 北京 100161)

摘 要: 分析了海洋环境数值预报运维复杂度逐渐增大的现状, 梳理出存在的6类主要问题, 给出了海洋环境数值预报业务综合运维的总体设计思路。进而研究分析了数值模式业务流程调度监控和模式运行性能特征分析等核心技术环节, 实现了海洋环境数值预报业务综合运维平台的主要功能, 并针对全球海洋预报系统开展了有效的示范应用, 为各数值预报业务中心设计开发自动化运维值班工作平台提供了一个高效可行的方案。

关键词: 海洋环境; 数值预报; 监控; 运维; 特征分析

中图分类号: TP311.52 **文献标识码:** A **文章编号:** 1003-0239(2019)03-0063-08

1 引言

随着数值预报同化、模式参数化和并行优化等技术的突破, 海洋环境数值预报业务能力得到不断提升, 新一代海洋环境数值预报业务体系中包含了数值模式业务软件、支撑环境、高性能计算机和配套设备等一系列复杂的软硬件装备。为了将该套复杂系统高效、自动化的应用起来, 并对其进行精准有效的维护, 迫切需要建设一套完善的数值业务系统运维管理平台, 实现面向数值预报业务系统性能监控、数值预报业务流程信息流运转监控及故障快速诊断、业务主机资源运维等为一体的综合运维管理。

国内外各数值预报业务中心在自身业务发展过程中, 都很注重数值模式业务综合运维监控平台的建设。欧洲中期数值预报中心早在1984年就率先利用工作流软件SMS^[1]开展业务流程监控工作, 在2011年又开始利用ecFlow^[2]软件替换SMS, 目前支持Python等新技术的ecFlow最新版为4.11.1。中国气象局数值预报中心胡江凯^[3,4]等在T213L31模式研制过程中就考虑运行监控的方案, 赵春燕等^[5]

在设计GRAPES数值预报模式集成中试支撑平台时考虑利用ecFlow软件解决作业调度问题, 王大鹏等^[6]利用消息中间件技术设计了数值预报业务系统移动监控平台, 张志远等^[7]和岳琨等^[8]也分别利用ecFlow软件对海洋环境数值预报和航空气象数值预报的业务工作流程监控进行了初步分析设计。

随着海洋环境数值预报业务的发展, 对监控运维的需求愈加强烈, 接下来本文首先对整个预报系统进行分析, 找出需要通过监控运维解决的问题; 其次构建整个综合运维架构并给出具体设计方案; 最后对其中的关键环节进行分析研究, 并初步实现综合运维平台建设内容。希望对今后数值预报业务化系统设计实现和运行服务具有一定的借鉴意义, 为各数值预报业务中心开展运维值班工作提供一个有效的方案。

2 设计目标

2.1 待解决问题

随着海洋环境数值预报业务的发展, 整个预报系统对业务稳定运行的要求越来越高, 综合高效的

收稿日期: 2019-01-25; 修回日期: 2019-03-14。

基金项目: 国家重点研发计划(2017YFC1404005)。

作者简介: 张志远(1978-), 男, 高级工程师, 博士, 从事海洋数值预报和信息化工作。E-mail: 13811119180@139.com

通讯作者: 孙立尹(1972-), 男, 高级工程师, 本科, 从事海洋环境预报保障工作。E-mail: 18910903211@189.cn

监控运维平台作为保障业务运行的主要抓手,需要解决的问题主要有以下6点:

(1)整个数值预报系统涉及的分量模式越来越多,比如某业务系统就由全球、西北太平洋和北印度洋区域嵌套组成,包括全球大气 T799 模式和全球海洋 MOM-MASNUM 模式,区域大气 WRF 和区域海洋 MASNUM-POM 耦合模式等,整个预报系统的复杂度呈几何倍数增长^[9]。整个模式系统的自动化提交、中断和断点续行,模式之间编排顺序的调整、优化,模式积分步长迭代记录,针对模式运行故障排查、诊断修复等等,这些问题急需一套高效的业务流程调度监控系统来解决;

(2)随着海洋环境保障应用精细化需求不断提升,数值模式时空分辨率逐渐提高,一般情况下,水平分辨率每提升1倍,数值模式的计算规模要增大10倍以上,在实际数值预报业务工作中,为保证预报产品的及时性,模式运行墙钟时间受到很大限制,因此,分辨率提高所导致的巨大计算量,必须通过模式并行规模增大来解决。因此需要对数值模式业务运行过程进行性能特征分析,找到计算、访存、存储和通信等方面的性能瓶颈,以便进行模式性能优化改进进而提升模式运行效率,近几年张志远等^[10]对海浪模式 MASNUM、王延强等^[11]对海流模式 NEMO 都做了相关性能分析和优化工作;

(3)海洋数值预报系统中各模式复杂程度越来越高,仅靠人工难以区分运行中模式故障、支撑系统故障和硬件故障,实际常出现因综合故障导致整个预报系统崩溃卡死的情况,因此及时预判或发现故障并提出警示及时补救十分重要。针对上述情况就需要对数值模式运行特征进行分析监控的同时与业务流程调度监控相互联动,分析找出故障原因;

(4)数值预报系统输入输出接口保持高效稳定运行状态对系统预报准确性和运行稳定性影响很大,输入数据包括数值预报同化和检验评估用的各类观探测数据,必须保证该类数据的时效性和稳定性,才能保证数值预报初始场的效果,进而确保预报准确性;输出数据包括自身制作的各类预报、解释应用数据和图片,尤其为保障部位和信息分发服务平台推送的产品,只有这些产品长期稳定输出才能体现数值预报系统的效能。因此,需要实时准确地监控各类信息收集和分发推送服务状态,一旦出

现信息缺失或者链路堵塞情况,要直观便捷的警示故障位置,提示排除故障方法,进行继续收集数据或推送产品,以免影响业务系统的正常运行;

(5)数值预报系统业务主机的软硬件支撑环境日趋复杂,不仅种类多,且数量庞大,因此,规范化、流程化的资产管理变得越来越重要。资产管理的主要对象是IT资产、无形资产等,如服务器、交换机等属于IT资产,系统软件、应用软件等属于无形资产,传统的资产管理多依赖人工查找,效率低下,且浪费人力,各类不规范的操作难以避免;

(6)海洋环境数值预报系统的业务综合监控平台,需要有一套综合的汇聚展示平台,对系统中数值模式业务运行状况、高性能计算机运行监控信息等进行数字化展示,自动化的警示各类故障,通过数字、图表、动画等方式展现业务运行情况和支撑系统情况等。

2.2 总体设计

针对2.1节提出的问题,本文认为综合运维平台应包含业务调度监控、性能特征分析、信息流转、资产管理、故障报警以及展示集成为一体的各项功能,为数值预报业务系统可靠稳定运行提供自动化运维支持。图1展示了综合运维平台的6个组成部分及其与海洋环境数值预报系统、业务主机、数据源等的信息交互关系。

3 关键功能实现

整个综合运维平台包括数值模式业务流程调度监控、数值模式运行性能特征分析、数值预报信息流转监控、数值预报软硬件资产管理、综合故障报警、监控信息集成展示等6个部分。其中,数值预报业务流程调度监控、数值模式运行性能特征分析两个部分是实现本平台的关键,本文接下来将详细阐述这两个部分的设计思想和实现过程。

3.1 数值预报业务流程调度监控

3.1.1 基于ecFlow的业务流程调度监控

针对2.1节中提出的1)问题的分析,数值预报业务系统除了包含多个模式外,整个流程包含资料预处理、资料同化、模式积分、数据后处理、数据检

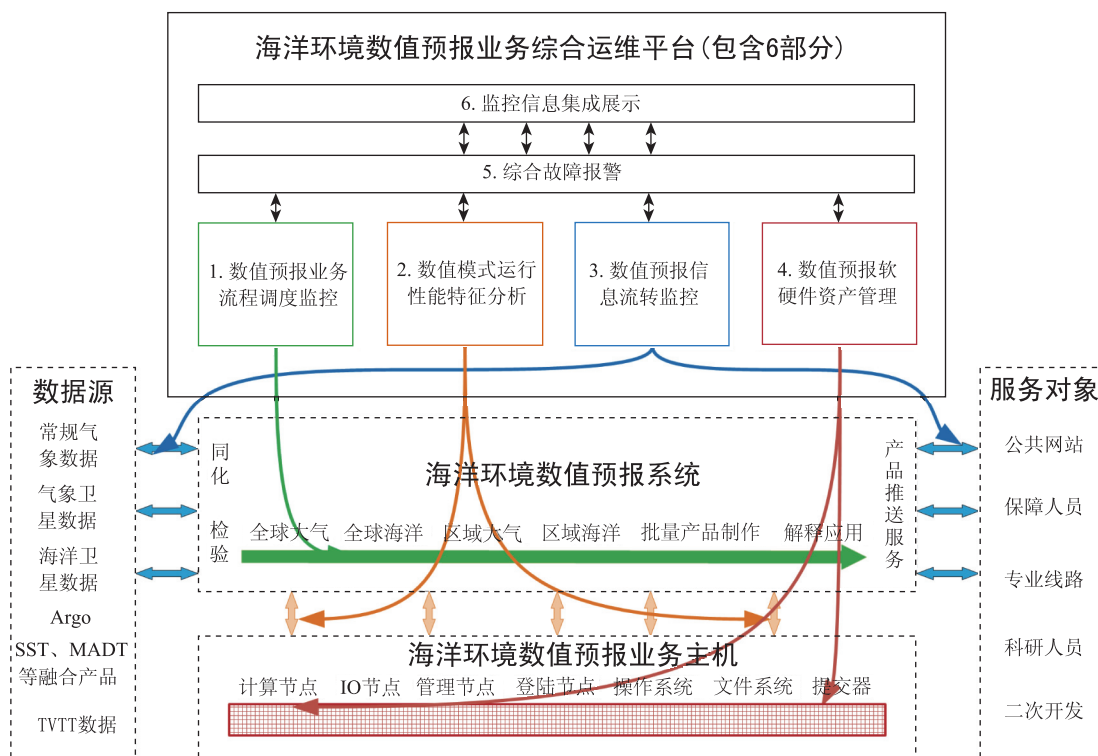


图1 平台组成及与外部关系图

验与产品制作分发等环节,涉及了大量科学计算和复杂工作流。ecFlow软件包是专门针对气象海洋模式应用开发的工作流软件,是一种解释型功能化编程语言,提供面向工作流的开发组件,用于刻画数值预报业务系统的各种作业运行过程、执行调度运行和监控管理。ecFlow兼容64位操作系统,可在Windows/Linux等各类操作系统上运行。本文基于ecFlow开发数值预报业务流程调度监控,解决数值预报业务系统中业务流程编排、作业自动调度与管理、故障诊断与处理等问题。

业务流程调度监控由服务层、调用层和展示层3部分组成。服务层负责调度作业和响应调用层的请求;调用层是系统用于和服务层通信的命令程序,通过命令级接口(CLI)实现;展示层作为人机交互界面,实现系统的监视与可视化功能。服务层根据用户对作业流程的定制情况和任务所处状态调度作业,并对作业运行情况和执行状态进行实时跟踪与记录。调用层和展示层作为客户端与服务层交互,获取整个作业流程的工作状态,并通过命令信息或图形显示方式,将监控信息提供给操作人

员。同时,操作人员可通过命令行或图形界面的鼠标命令,向服务层发送命令请求。

本文根据数值预报业务系统流程的设计需求,按照ecFlow软件结构设计了作业包(suite)、作业族(family)、作业(task)3层目录结构,实现对业务系统中多个预报模式的管理。作业包对应整个要监控管理的数值预报系统,每一个业务流程只能定义唯一一个作业包。作业族包含若干个为实现某特定功能的作业,作业是业务流程定义中的最小单位。其中,作业包由多个作业族组成,作业族又可以包含多个子作业族,最终由一系列作业构成每一个作业族,只有作业族允许嵌套定义。为实现多个预报模式的管理,本文将整个数值预报业务系统作为作业包,每个模式为一级作业族,各模式的资料同化、模式积分、后处理等环节为二级作业族。根据各模式特点,设计出多级作业族以及最小管理单元作业。图2为预报系统流程监控结构示意图。

3.1.2 针对FIOCOM海洋模式的业务流程调度监控实现

本文将以FIOCOM模式为例,简述基于ecFlow

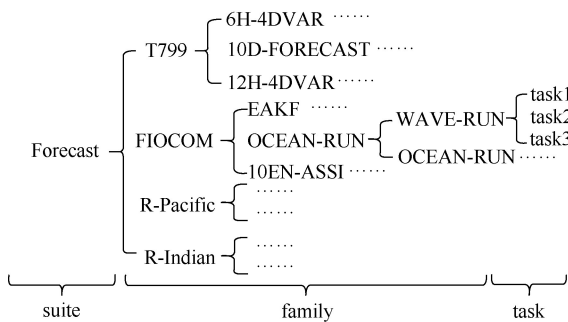


图2 预报系统流程监控结构示意图

的数值预报业务流程调度监控的开发过程。FIOCOM模式作为整个数值预报系统作业包下的一个作业族,来实现对FIOCOM的调度监控。主要包括编排模式作业流程、编写作业脚本文件、优化调度监控功能3个步骤。

(1) 编排模式作业流程

通过变量定义、结构控制和作业定义语句,实现作业组成、作业控制以及复杂的逻辑结构设计,本文主要通过定义描述性脚本def文件实现。在分

析各作业的执行次序、作业之间的依赖关系、以及各变量的定义内容后,编写def定义文件,主要使用family、task、event、meter、label等结构定义语句,trigger、date、time、cron等依赖定义语句,edit、complete、late、repeat等属性定义语句来实现作业编排。根据FIOCOM的运行特点,在00时与12时次的基础上,将作业流程分为EAKF同化、OCEAN_RUN模式积分和10EN_ASSI集合同化3部分,细化作业流程,并对每一部分添加输入输出的检查(check)流程,最后完成作业依赖设计。在编排作业流程过程中,不仅要考虑作业的编排效果,更要考虑相应作业执行脚本的编写难度和可维护性。综合考虑后,FIOCOM作业编排的效果如图3所示。

(2) 编写作业脚本文件

将模式移植到业务流程监控平台时,需要在原模式的基础上拆分代码、添加头文件和基于ecFlow命名规则修改相关代码与变量。本文将对对应def文件中定义的每个作业task的执行脚本放到ecf脚本文件中。ecf的头文件head.h与tail.h用来进行初始

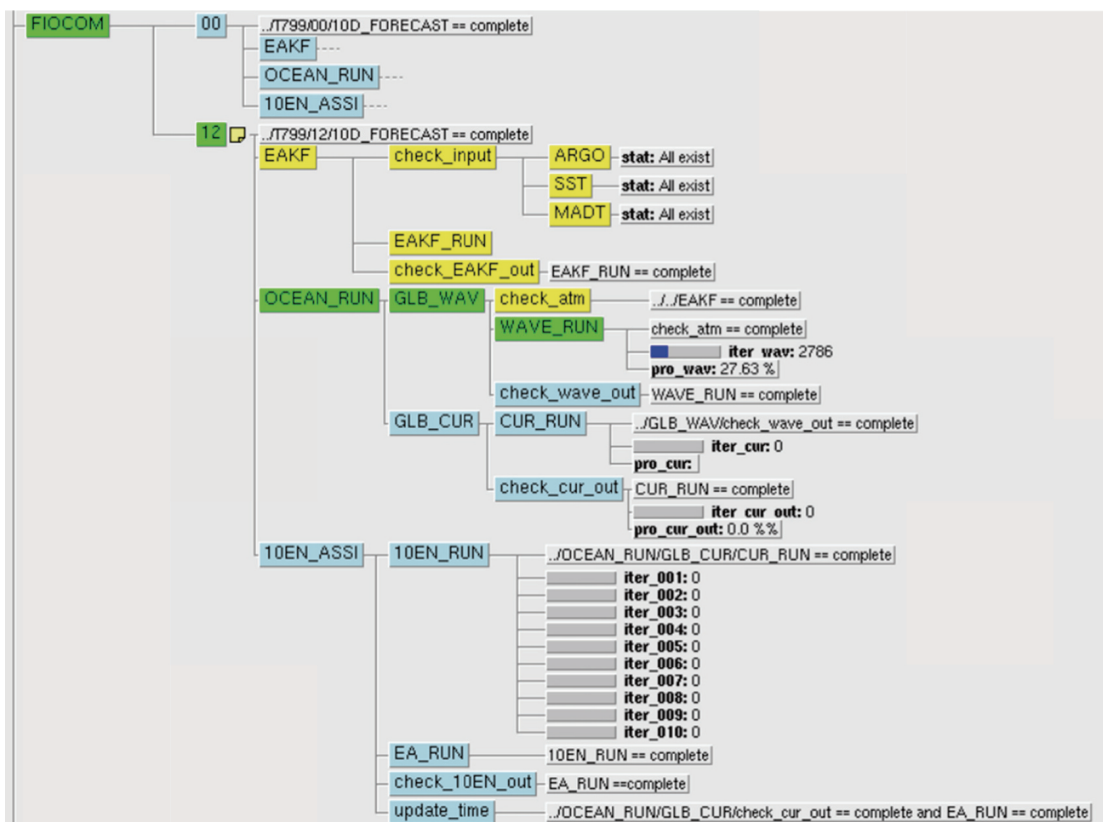


图3 FIOCOM模式流程监控效果图

化操作、定义异常信号和扫尾操作等,如无特殊要求,在ecf文件中通过include命令直接调用即可。编写的ecf文件与def文件中定义的作业要相对应,在业务流程监控平台中提交作业后,系统将生成一个用于提交的job文件。

编写针对性的ecf文件是将模式移植到监控平台的核心步骤,ecf脚本文件与模式代码融合,要求移植人员清楚模式代码结构,并进行多次运行调试测试。

(3) 优化调度监控功能

在完成前面两个步骤后,监控平台已初步实现对FIOCOM模式的调度监控功能。为便于值班业务人员直观掌握系统的运行情况,可通过进一步添加标签、进度条等形式来丰富图形界面的显示效果,如图3所示。业务系统中还有发作业至后台或业务主机的操作,为保证每一个作业报错后的重新运行功能,需要确认已发送的作业是否在相应平台上结束运行。在必要情况下,可通过添加相应的终止作业命令来解决此类问题。

3.2 数值模式运行性能特征分析

3.2.1 运行性能特征分析

针对2.1节中提出的2)和3)问题的分析,海洋环境数值模式是典型的复杂科学计算问题,具有计算密集型、IO密集型、访存密集型和通信密集型等特点,必须以高性能计算机为载体,对机器综合性能要求较高,用于海洋环境预报保障的数值模式具有极高的时效性要求,必须在指定的时间内完成特定运算。因此,在模式中试或业务化运行过程中,需要不断运行测试分析,找到模式运行中时间延迟的瓶颈,并有针对性的进行性能优化,进一步缩短模式运行的墙钟时间,更高效的满足时效性要求。在整个往复循环测试过程中,最耗时的步骤就是收集性能数据、分析数据定位问题和提出解决方案,这就是运行性能特征分析,目前并行系统性能评测的解决方案有很多^[12],如实际测量法、Benchmark法、性能模拟器、性能模型。针对数值模式的性能测试,主要采用采样测量法来实现^[13]。目前商业化的高性能分析软件包括Intel的分析工具Intel VTune Amplifier XE^[14]、二进制插装工具Pin、开源软件gprof (GNU profiler)等。

针对海洋环境数值预报模式的性能分析测试不需要商用软件那么复杂,海洋模式一般都是多任务多流程任务,是由Fortran70/90编写的程序,流程主要包括模式初始化、数值计算和结果输出三部分。1)初始化部分主要负责配置并行环境、设置垂向层数、读取地形数据、进行网格剖分和读取重启文件等;2)数值计算部分通过模式时间步长控制程序循环次数,是核心计算部分,此处会涉及负载均衡和并行通信等模块;3)结果输出部分主要包括预报数据文件输出和写重启文件等,I/O算法负责数据输出功能。

针对海洋模式,应该主要通过串行和并行两个方面开展性能特点分析^[15]。

导致串行浮点效率低的主要因素主要有2类:1)指令级并行性,是否保证每个周期都发射浮点指令;2)浮点指令是否能充分利用浮点计算部件,如向量和乘加计算部件。这就需要提供系统级和微架构级数据分析指标,便于分析访存性能、分支性能、长延迟指令、SIMD性能、乘加性能等。

导致并行性能效率低的主要因素主要包括串行开销、通信带宽、写I/O带宽、读I/O带宽和负载平衡等随着并行度变化的情况,因此需要分析CPU、内存、磁盘、网络通信等数据。

3.2.2 针对海流模式CUR_RUN的运行性能特征分析实现

本文以在国家超级计算广州中心“天河二号”高性能计算机上的FIOCOM数值预报系统中海流模式运行CUR_RUN部分为例,展示分析运行性能特征分析的过程,具体软硬件环境如表1所示:

本测试采用独占计算节点的方式进行,使用50个节点共12 00核,墙钟时间为42 min。监控软件可以提供CPU、浮点计算、内存占用、IB Send\Recv、

表1 运行环境参数表

类别	内容
CPU	Intel(R) Xeon(R) CPU E5-2692 v2 @ 2.20GHz,12core/socket 2sockets
内存	64 GB
操作系统	Red Hat Enterprise Linux Server release 6.5 (Santiago)
编译器	Intel composer 14.0.2.144 Build 20140120
MPI	MPI3

VEC等指标和雷达图等展示效果。资源使用率分析主要从CPU资源、访存资源和网络资源使用等方面入手,对比平均值和峰值信息,如图4所示。本文主要在阐述该模式运行性能特征分析实现效果,后续另有其他文章根据具体的性能分析结果进行优化分析。

海流模式运行过程中,使用CPU(a11)达到

97.72%,表明在程序运行时CPU绝大部分时间都在处理执行程序,CPU使用率较高。VEC%平均值为80.73%,表明程序大部分代码能够充分利用CPU向量化浮点计算部件。内存运行方面,模式内存使用率平均值为43.9%,没有达到饱和状态,说明程序的访存使用不存在瓶颈,能够满足业务需求。网络通信方面IB Recv值为89.01 MB/s,IB Send值为

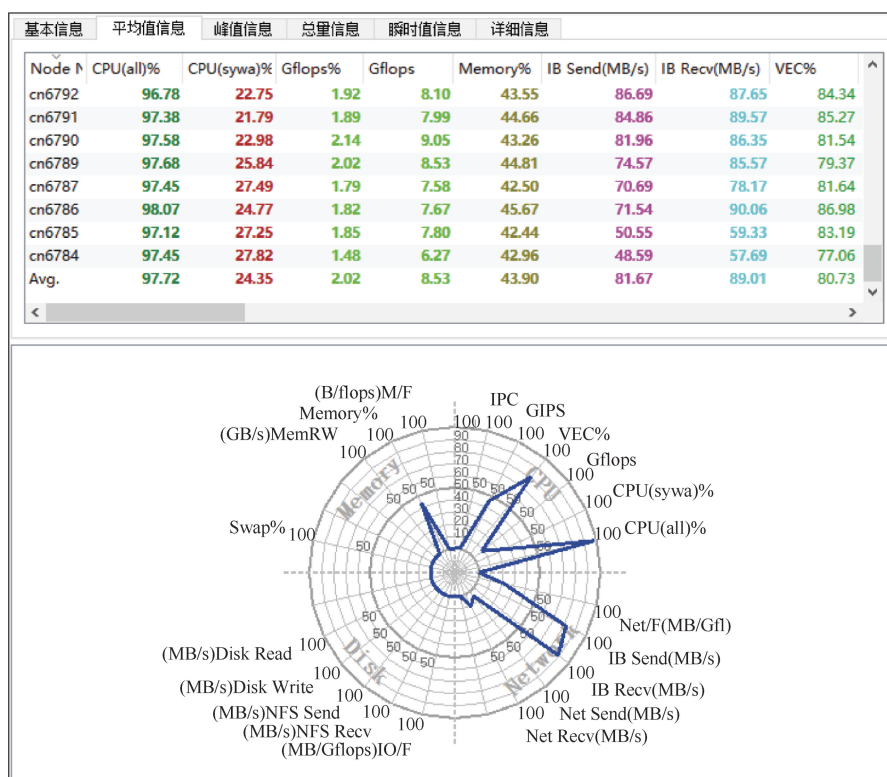


图4 海流模式性能分析数据平均值分析实现效果图

基本信息	平均值信息	峰值信息	总量信息	瞬时值信息	详细信息				
Node Name	CPU(all)%	CPU(sywa)%	Gflops%	Gflops	Memory%	IB Send(MB/s)	IB Recv(MB/s)	VEC%	
cn6894	100.00	98.10	6.22	26.28	46.69	140.62	132.04	100.00	
cn6893	100.00	98.29	6.55	27.66	46.88	151.41	122.90	100.00	
cn6892	100.00	98.44	6.77	28.60	51.44	135.77	127.70	100.00	
cn6891	100.00	98.29	5.85	24.69	47.03	146.30	133.94	100.00	
cn6890	100.00	99.22	5.78	24.40	50.37	127.76	124.81	100.00	
cn6889	100.00	98.03	6.64	28.06	47.15	126.19	122.10	100.00	
cn6888	100.00	98.57	6.67	28.19	49.82	121.93	119.07	100.00	
cn6887	100.00	99.56	6.43	27.16	47.23	117.96	117.91	100.00	
cn6886	100.00	98.87	5.63	23.79	49.72	122.01	123.18	100.00	
cn6885	100.00	97.72	5.96	25.16	47.49	118.19	122.89	100.00	
cn6883	100.00	98.18	6.74	28.46	47.61	123.03	123.58	100.00	
cn6882	100.00	99.90	6.67	28.16	49.81	128.53	132.60	100.00	
cn6862	100.00	94.45	5.43	22.94	47.64	124.18	131.70	100.00	
cn6861	100.00	96.45	6.71	28.32	49.27	126.51	125.68	100.00	
cn6860	100.00	98.40	6.80	28.72	47.60	122.91	127.18	100.00	
cn6859	100.00	96.94	5.26	22.21	49.13	136.14	128.91	100.00	
cn6857	100.00	98.83	6.76	28.55	48.91	129.00	127.62	100.00	
cn6838	100.00	99.70	6.55	27.67	48.66	125.68	119.82	100.00	
cn6830	100.00	98.44	5.34	22.55	50.70	125.09	123.91	100.00	

图5 海流模式性能分析数据峰值分析实现效果图

81.67 MB/s,说明程序运行过程中网络通信速度较快,数据流运转流畅。

海流模式运行特征分析峰值部分信息如图5所示。节点内存使用峰值在46%~52%之间,平均值在43.9%,可见海流模式对内存使用率平稳,且资源使用未达到饱和。IB Recv最大峰值为131.04 MB/s,IB Send峰值为151.41 MB/s,资源使用未达到饱和。VEC%向量化程度为100%,可以看出海流模式在密集计算部分能够有效利用系统向量化处理能力。

4 小结

随着海洋环境数值预报应用需求的不断增强,业务综合运维平台成为保证业务系统准确、高效、稳定运行的有力抓手,本文设计并初步实现的海洋环境数值预报业务综合运维平台,能够较好地提高海洋环境数值预报业务监控能力,同时最大限度地减轻业务值班人员在各类故障预警防范方面的工作强度,有效提升数值预报系统产出的效益。

随着多模式耦合、嵌套等技术的发展,海洋环境数值预报系统的模式种类和复杂程度将不断增大,作为一个完整的系统,每个模式、每个环节之间有着强依赖关系,无论哪个环节出了问题都将影响整个系统正常运行。不同的模式又可能因为开发团队、实现方案的不同导致接口关系、命名规则等出现不同,使整个业务系统变得复杂,不利于管理和维护。因此,业务综合运维平台的设计也不是一蹴而就的,未来将进一步扩展和完善功能,通过接入不同模式或者预报系统,提炼运维过程中出现的共性问题,力争对复杂的技术实现细节进行屏蔽,使所有的数值业务系统能够在运行层面上保持相对一致,为其他数值预报系统运行维护提供经验,进而简化海洋环境数值预报系统维护和运行的复

杂度,从而提高数值预报的时效性和稳定性,有效提升数值预报业务系统的支撑服务能力。

参考文献:

- [1] Hollmann F. SMS-supervisor monitor scheduler[R]. Reading: ECMWF, 1984.
- [2] Bahra A. Managing work flows with ecFlow[R]. Reading: ECMWF Newsl, 2011: 30-32.
- [3] 胡江凯, 李娟, 宗翔. 监控管理调度系统与数值预报运行的应用[C]//数值天气预报新理论新方法及应用学术研讨会论文摘要汇编. 乌鲁木齐: 中国气象学会, 2004: 176-183.
- [4] 胡江凯, 王雨, 王毅涛. 国家气象中心 T213L31 数值预报运行监控方案及预报效果评估[J]. 应用气象学报, 2005, 16(2): 249-259.
- [5] 赵春燕, 郭锋, 孙婧, 等. GRAPES 数值预报模式集成中试支撑平台的设计与应用[J]. 气象科技进展, 2018, 8(1): 292-293.
- [6] 王大鹏, 崔应杰, 胡江凯. 数值预报业务系统移动监控平台的设计与实现[J]. 气象科技, 2018, 46(5): 892-898.
- [7] 张志远, 江玮. 基于 ecFlow 的海洋环境数值预报业务工作流程管理[J]. 电脑编程技巧与维护, 2017, (10): 74-76, 88.
- [8] 岳琨, 王廷芳, 王伟峰. 基于 ECFLOW 的数值预报业务流程监控系统开发[C]//第32届中国气象学会年会论文集. 天津: 中国气象学会, 2015.
- [9] 王辉, 万莉颖, 秦英豪, 等. 中国全球业务化海洋学预报系统的发展和应用[J]. 地球科学进展, 2016, 31(10): 1090-1104.
- [10] 张志远, 周宇峰, 刘利, 等. MASNUM 海浪模式的性能特点分析与并行优化[J]. 计算机研究与发展, 2015, 52(4): 851-860.
- [11] 王延强, 张宇, 林波, 等. 基于 NEMO 的全球海洋环境预报模式在超算集群的计算性能优化[J]. 海洋预报, 2018, 35(3): 41-47.
- [12] 陈永然. 面向高性能计算的性能评价模型技术研究[D]. 长沙: 国防科学技术大学, 2007.
- [13] 魏敏, 孙婧, 沈瑜, 等. 高性能计算系统性能评测方法及其应用[J]. 应用气象学报, 2013, 24(6): 753-760.
- [14] Marowka A. On performance analysis of a multithreaded application parallelized by different programming models using Intel VTune[C]//11th International Conference on Parallel Computing Technologies. Berlin, Heidelberg: Springer-Verlag, 2011.
- [15] 张志远. 浪流耦合模式 MASNUM-POM 的性能分析与优化[D]. 北京: 清华大学, 2015.

Design and implementation of marine environment numerical forecast integrated operation and maintenance platform

ZHANG Zhi-yuan, SUN Li-yin, HE Xi-yu, WANG Shao-ke

(Battlefield Environment Support Office of Naval Staff Command and Support Brigade, Beijing 100161 China)

Abstract: This paper analyses the increasing complexity for the operation and maintenance of marine environment numerical forecast and summarizes six areas of key issues it faces, based on which the overall design of an integrated operation and maintenance platform for marine environment numerical forecast is developed. We also solve the key technical issues including the dispatching and monitoring of operation flow and the performance analysis of the numerical model. The integrated operation and maintenance platform for marine environment numerical forecast is implemented in the global ocean forecasting system as a demonstration, which provides a feasible and efficient solution for numerical operational organizations to develop their own operation and maintenance platforms.

Key words: marine environment; numerical forecast; monitoring; operation and maintenance; characteristic analysis