

基于国产高性能计算集群实现海洋模式的加速计算

周巍¹, 吴兵成², 杨阳², 胡治平², 陈昇¹, 李毅能¹

(1.中国科学院南海海洋研究所, 广州市 510301; 2.华为技术有限公司, 广东省深圳市 518129)

摘要: 在 21 世纪, 海洋科学研究成为全球关注的焦点, 尤其是在国家安全、资源保障、防灾减灾等方面。高性能计算集群已经成为这一领域不可或缺的工具。中国科学院南海海洋研究所是我国规模最大的综合性海洋研究机构之一, 重点关注热带边缘海的多层次相互作用及其对环境和资源的影响。面对诸如气候变化预测、海洋环境保障等重大需求, 南海所采用了华为提供的国产高性能计算集群, 实现了海洋模式的加速计算。这一基于鲲鹏技术的高性能解决方案, 不仅满足了海洋预报和科学探索的应用需求, 而且极大地提升了模型的计算效率和准确性。该成果标志着国产高性能计算技术在我国海洋科学领域的成功应用, 为未来的海洋研究和应用提供了有力的技术支持, 将进一步推动我国在海洋科学和高性能计算领域的前沿研究。

关键词: 高性能计算; 海洋模式; 加速计算; 应用优化

中图分类号: P731.34

文献标识码: A

Accelerated computation of ocean models using a domestically-produced high-performance computing cluster

Zhou Wei¹, Wu Bingcheng², Yang Yang², Hu zhiping², Chen Sheng¹, Li Yineng¹

(1. South China Sea Institute of Oceanology Chinese Academy of Sciences, Guangzhou, 510301, China ;2. Huawei Technologies Co., Ltd, Shenzhen, 518129, China)

Abstract: In the 21st century, marine science research has become a focal point of global attention, particularly in the realms of national security, resource conservation, and disaster prevention and mitigation. High-performance computing clusters have emerged as indispensable tools in this field. The South China Sea Institute of Oceanology, Chinese Academy of Sciences, is one of the largest comprehensive marine research institutions in China, with a focus on the multi-level interactions in tropical marginal seas and their impact on the environment and resources. To address critical needs such as climate change forecasting and marine environmental

protection, the institute has adopted a high-performance computing cluster provided by Huawei, based on Kunpeng technology. This solution has not only met the application requirements for marine forecasting and scientific exploration but has also significantly improved the computational efficiency and accuracy of the models. This achievement marks the successful application of domestic high-performance computing technology in China's marine science sector, providing robust technical support for future marine research and applications, and further propelling China's frontier research in marine science and high-performance computing.

Keywords: High-Performance Computing, Ocean Modeling, Accelerated Computing, Application Optimization

1 背景介绍

海洋占地球表面积的 71%，为人类生产生活提供丰富的食物药物、提供更大的生存空间、提供能源和矿产，还是决定全球气候发展的主要因素。21 世纪，海洋再度成为世界的焦点，对海洋的研究对于我们关心海洋、认识海洋、经略海洋具有重要的意义。高速、大容量计算机的出现为各种复杂问题的大规模计算提供了可能性，目前高性能计算平台已经是研究海洋必不可少的工具平台，极大助力了海洋科学研究。鲲鹏为海洋领域应用提供全方位，多层次的高性能解决方案^[1]，为海洋预报、海洋科学探索等提供有力的支撑。

中国科学院南海海洋研究所（以下简称“南海所”）是我国规模最大的综合性海洋研究机构之一，重点研究热带边缘海海洋水圈—地圈—生物圈圈层结构及其相互作用特征与演变规律，探讨其对资源形成和环境变化的控制和影响，发展具有南海特色的热带海洋资源与环境过程理论体系和应用技术。人类的生存与发展离不开海洋，国家安全、资源保障、防灾减灾等全局性、战略性、长久性问题都与海洋息息相关。

面对国家海洋减灾防灾、海洋环境安全保障和气候变化预测等重大需求，南海需要一套高性能、能够满足应用程序需求的高性能计算机平台。针对海洋领域模式的主要特点：浮点计算量大、通信密集、要求系统具有高稳定性和高 IO 吞吐能力。我们选用了鲲鹏高性能解决方案，该方案基于鲲鹏全栈打造，涵盖硬件使能、基础软件、应用优化等能力。通过超算平台的建设和应用的调优，实现了 ROMS、HYCOM、CESM 三大主流海洋模式的运行和调优。

2 南海所超算平台简介

当前，南海所部署了 40 台 2 路的鲲鹏计算节点、2 台基于 NPU 的 AI 训练节点、1 套高性能分布式存储、2 台 100GE 交换机，满足了高性能计算和 AI 训练平台对国家海洋减灾防灾、海洋环境安全保障和气候变化预测的科研要求。本次建设的 HPC+AI 集群系统是为数值计算所建设的科学计算集群系统，主要包括：

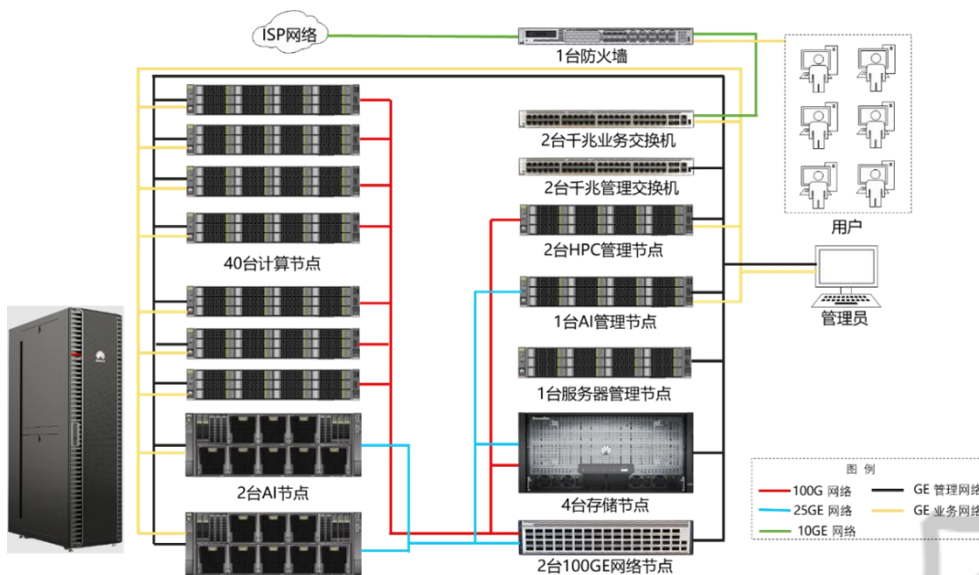
1) 科学计算子系统：计算节点是整个 HPC 平台的主体部分，其性能决定了 HPC 整体性能表现。本次 HPC 平台使用 KunLun2280 服务器，共计建设 40 台 2 路鲲鹏计算节点（单节点配置 2 个 64 核鲲鹏处理器），总物理核数达 5120 颗物理 CPU 核，高性能计算算力可达 61TFLOPS。

2) 高性能存储子系统：为保证系统性能和可扩展性，采用分布式存储。本次 HPC 平台建设使用华为 OceanStor Pacific 系列高性能分布存储，提供 1PB 可用存储容量，存储系统具备分布式并行客户端 DPC（Distributed Parallel Client），作为存储客户端运行在计算节点上，与后端存储节点进行数据交换，保证计算平台的性能。

3) AI 训练子系统：随着深度学习需求日益扩大，基于神经网络的 AI 处理器技术创新蓬勃发展。本次训练推理节点采用 NPU（Neural Processing Unit，神经网络处理单元）构建 AI 算力资源。NPU 针对矩阵运算进行专门优化设计，解决了传统芯片在神经网络运算时效率低下的问题。达芬奇架构是面向 AI 计算设计的架构，通过独创的 16*16*16 的 3D Cube 设计，每时钟周期可以进行 4096 个 16 位半精度浮点 MAC 运算，为人工智能提供强大的算力支持。本平台使用 2 台 KunLunG5680，提供 16 张 Ascend 910 B AI 芯片，可以提供 3.5PFlops AI 算力。

4) 网络子系统：为保证系统并行访问的效率与速度，采用 100Gb/s 高性能网络。本平台部署 2 台 64 端口 100Gb/s 交换机，通过 RoCE 协议组网，保障高性能计算平台和 AI 训练平台的高速访问。

5) 作业调度管理软件：进行系统管理分配和资源调度，各个子系统的管理调度策略根据子系统所运行应用特点进行灵活配置。本平台采用华为多瑙管理软件，该软件具备具有鲲鹏的各种功能，包括：编译器、数学库、通信库、作业调度器、集群管理平台等，能够极大的满足全栈国产化的高性能计算台。保障研究所科研的日常应用。



总结：南海所部署了华为鲲鹏 HPC 平台+AI 训练平台，包括 40 个计算节点，2 个训练节点，5120 多个计算核心，16 张训练卡，总计算能力达到 61 万亿次/秒，AI 训练能力达 3500 万亿次/秒（FP@16）。

3 海洋模式计算优化工作

搭建好基于鲲鹏和昇腾的算力环境后,我们把海洋研究中常用的模式代码在国产化算力平台上开展优化工作。

3.1 ROMS 模式优化

Regional Ocean Modeling System(ROMS)是一个三维区域海洋模式^[2],由罗格斯大学海洋与海岸科学研究所和加利福尼亚大学洛杉矶分校共同研究开发,被广泛应用于海洋及河口地区的水动力及水环境模拟。

ROMS 是一个现代化和模块化的海洋模式，由多达 107 万行的 Fortran 代码、多个头文件以及各种输入文件组成^[3]。这一庞大的代码基础为模式提供了高度的灵活性和可扩展性。

在网格结构方面，ROMS 模式采用了三层嵌套网格^[4]，总网格数超过 6684 万。这一高度精细的网格结构使得模式能够进行更为精确的模拟。

然而，这种高度精细的网格结构也带来了计算性能的挑战。以串行计算为例，模拟一年的海洋环境需要花费 3052 天，这在实际应用中显然是不可接受的。

结合我们的应用背景和计算特性，我们在 himp 通信算法、华为毕昇编译器及 jemalloc 内存、reshape 矩阵算法、PIO 应用开展优化工作。

1) hmpi 通信算法及负载均衡优化

hmpi 通信算法是一种高效的并行通信方法，特别适用于大规模数据交换场景。这个算法通过负载均衡机制，确保每个处理器都能得到均等的计算任务。这不仅有助于提高单个处理器的效率，而且能显著提升整个计算系统的整体效率。当前优化工作实现性能提升 10%。

2) allreduce 算法优化

集合通信热点占比分析如下：

```
sh-5.0$ /workspace/home/shenchao/fangchunlin/TOOLS/hpcrunner/hpctool_1_15_1_rc2/bin/hpcreport -i ./output.v
1_15.hpcdata mpi-wait
```

Function	Wait(%)	Avg(ms)	Call Count	Data Size(bytes)
MPI_Allreduce	22.41	57.2842	4203	103444997728
MPI_Irecv	9.10	0.0523	1869848	8287459296
MPI_Bcast	2.31	1.6072	15452	2946682364
MPI_Send	0.59	0.0034	1869848	8287459296
MPI_Barrier	0.02	219.1251	1	0

图 2 集合通信热点占比分析图

目前应用的 mpi 通信热点主要在 allreduce 全局规约操作，目前 hmpi 针对该算法有实现一定的性能优化特性。

3) bind-to numa 参数负载均衡优化

ROMS 对内存带宽和内存容量都有较高的要求，因此，建议每个节点非满核 64 核运行，rank 均匀分配到节点不同 numa 上，发挥鲲鹏内存多通道的最大优势。

优化参数如下：

```
mpirun -hostfile $HOSTFILE --mca plm_rsh_agent /opt/batch/agent/tools/dstart -mca pml
ucx -x UCX_NET_DEVICES=mlx5_0:1 -mca btl ^vader,tcp,openib,uct -x UCX_TLS=self,sm,rc
-x PATH -x LD_LIBRARY_PATH --bind-to numa -x OMP_NUM_THREADS=1 -np 256 -N
64 ./romsM ocean_cold_start.in
```

4) 毕昇编译器及 jemalloc 内存优化

毕昇编译器有针对鲲鹏服务器做一些向量化、加速指令、及软件预取等优化。测试场景有大量的 malloc、memcpy 等内存操作热点，毕昇编译器内存 jemalloc 库对内存操作有一定优化。通过优化，实现性能提升 20%。

对应的优化编译选项如下：

```
FFLAGS := -fpack-arrays -fvectorize -funroll-loops -mllvm -unroll-indirect-loads=true
-mcpu=tsv110 -mllvm -prefetch-loop-depth=3 -mllvm -min-prefetch-stride=4 -mllvm
-prefetch-distance=1600 -fstack-arrays -ljemalloc -lstringlib -Wl,--wrap=memcpy
-Wl,--wrap=memset
```

5) reshape 矩阵算法优化

在 ROMS 的主函数 mp_assemblef_3d 中，参加大量矩阵转维函数 reshape 的调用，将 3D data 数组转为为 1D 数组，内置 reshape 函数针对简单的多维转一维数组性能较差，向量化率低，可考虑算法优化，更好发挥编译器的自动向量化，如下：

```
!-----
! Collect data from all nodes.
!-----
! Reshape input 3D data into 1D array to facilitate communications.
!
! Asend=RESHAPE(A, (/Npts/))
do k=1,Asize(3)
do j=1,Asize(2)
do i=1,Asize(1)
Asend((k-1)*Asize(2)*Asize(1) + (j-1)*Asize(1) + i) = A(i,j,k)
end do
end do
end do
```

图 3 reshape 矩阵算法优化结果图

当前实现性能提升 10%

6) PIO 优化

ROMS 模式的输入和输出文件均采用 netcdf 文件格式，这些文件的大小通常超过 800GB。在如此大规模的数据处理场景下，传统的串行读写方法明显不适用，因此我们引入了并行输入输出技术 PIO。通过应用 PIO 技术，我们实现了文件读写时间的大幅度减少。这不仅加速了数据的流动，而且也优化了数据处理的整体效率。

通过测试，PIO 在鲲鹏平台上实现完整运行。

7) 优化总结

经过性能优化，使用 4 节点 256 核，可 3403s（56 分钟）内，完成 2 天的计算量，性能提升 1.4 倍。

ROMS 性能优化测试数据如下：

节点数	进程数	运行时间	迭代时间	提升比例	优化措施
4	256	2042s	12h	100%	基础性能
4	256	1851s	12h	110%	Hmpi负载均衡优化
4	256	1570s	12h	130%	编译器jemalloc内存优化
4	256	851s	12h	240%	reshape矩阵算法优化

图 4 ROMS 性能优化测试数据

3.2 HYCOM 模式性能调优

HYBRID COORDINATE OCEAN MODEL（HYCOM）作为美国全球海洋数据同化实验

(GODAE)的一部分，旨在开发和评估数据同化混合坐标海洋模式。模式结合了不同类型坐标的优势，以优化模拟沿海和公海环流特征。

1) 编译选项优化

在鲲鹏平台上，增加如下编译选项，可以充分发挥鲲鹏芯片的性能，对 hycom 性能提升效果大约有 10%。

```
FCFLAGS = -Ofast -march=armv8.2-a -fdefault-real-8 -Mflushz
```

2) 使用优化的数学库

本次测试过程中，发现应用存在 libm 相关数学函数的热点，例如 exp 函数。使用华为自研 kml 数学库，测试有 10%的性能优化，如下：

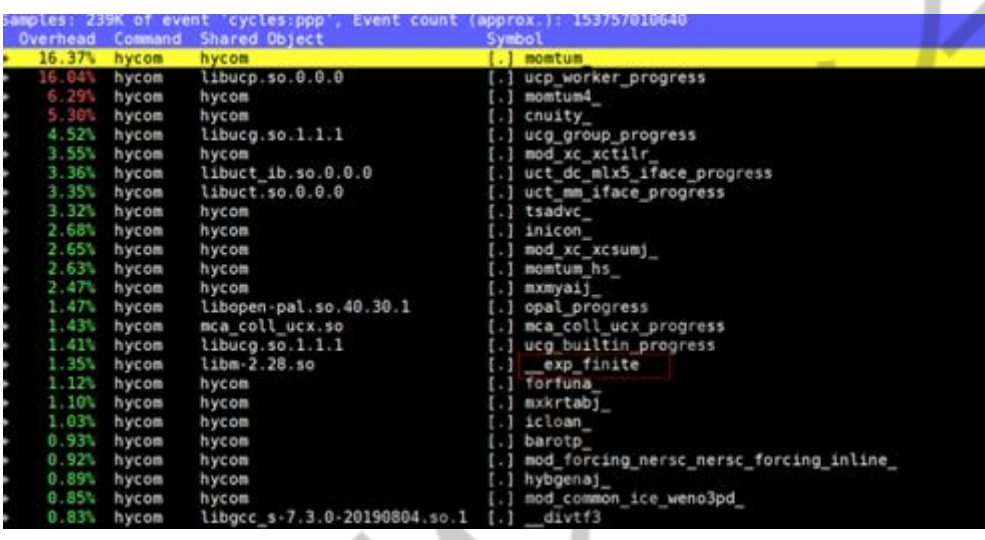


图 5 HYGOM 模式热带函数

3) mpirun 参数优化

测试发现，hycom 存在大量的内存访问，mpirun 的使用使用 -bind-to socket 绑核参数，让节点内的进程分配更加均匀，充分发挥鲲鹏平台的大内存通道的优势，性能提升约 20%

```
mpirun --allow-run-as-root --mca plm_rsh_agent /opt/batch/agent/tools/dstart -hostfile $HOSTFILE -bind-to socket -x PATH -x LD_LIBRARY_PATH -x OMP_NUM_THREADS=1 -np 1032 -N 104 ./hycom
```

4) 优化总结

经过应用移植和性能调优，软件在鲲鹏集群+ROCE 网络+OceanStor 分布式存储环境下运行稳定，性能测试结果良好。

性能结果：

同等核数下，鲲鹏 7260 性能是 Intel E5-2680 V4 的 125%。

同等算力下，鲲鹏 7260 性能是 Intel E5-2680 V4 的 230%。

同节点数下，鲲鹏 7260 性能是 Intel E5-2680 V4 的 333%。

表 1 鲲鹏集群+ROCE 网络+OceanStor 分布式存储环境下性能对比表

CPU	节点数	CPU 数	运行时间 (s)	性能对比
Intel E5-2680 V4	10	255	13200	100%
	10	1032	3960	333%
鲲鹏 7260	5	492	5898	223%
	3	255	10512	125%

3.3 CESM 模式性能调优

Community Earth System Model (CESM) 是由美国 NCAR 于 2010 年 07 月推出的新一代地球系统模式，是一个完全耦合的气候模式，在 CESM 中包括大气、陆地、海洋、海冰、陆冰等几大模块。其主要用于研究地球过去、现在、和未来的气候状况。

1) 编译选项优化

在鲲鹏平台上，修改默认编译选项如下，可以充分发挥鲲鹏芯片的性能，对 CESM 性能提升效果大约有 10%。

```
FFLAGS := -mno-fcmla -fconvert=big-endian -O3 -mcpu=tsv110 -Hx,54,0x08 -flto=full
-fuse-ld=lld -Hx,54,0x800 -Mx,54,0x1000 -mllvm -alias-set-saturation-threshold=10000 -mllvm
-licm-disable-gep-const-hoist=true -mllvm -lsr-enable-gep-const-erase-limit=1 -lstringlib
-Wl,--wrap=memcpy -Wl,--wrap=memset

CFLAGS := -std=gnu99 -O3 -mcpu=tsv110 -fuse-ld=lld -mllvm
-alias-set-saturation-threshold=10000 -mllvm -licm-disable-gep-const-hoist=true -mllvm
-lsr-enable-gep-const-erase-limit=1

LDFLAGS := $(LDFLAGS) -fuse-ld=lld
-L/bisheng-HPC/lizekun/CESM/CESM_LLVM/dependence/OPENBLAS/lib -lopenblas
-mcpu=tsv110 -Wl,--mllvm="-licm-disable-gep-const-hoist=true"
-Wl,--mllvm="-lsr-enable-gep-const-erase-limit=1"
-Wl,--mllvm="-alias-set-saturation-threshold=10000"
```

2) 优化结论

经过应用移植和性能调优，使用 8 点 1024 核，4 分钟内，完成 f09_f09 的 5 天模拟的计

算量。整体性能提升 44%。下图是鲲鹏平台 CESM 测试结果：

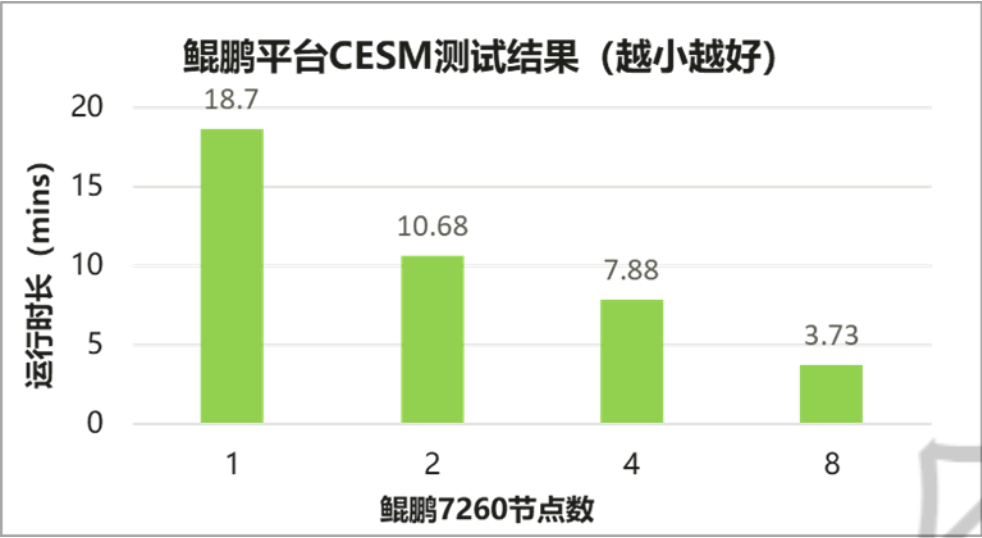


图 6 鲲鹏平台 CESM 测试结果

4 总结

在基于鲲鹏昇腾的算力平台上，我们通过对自研毕昇编译器、鲲鹏数学库、华为通信库 HMPI、应用热点函数的调优，实现 ROMS、HYCOM 和 CESM 模式在鲲鹏昇腾算力平台上的适配和优化，分别实现了：

ROMS 模式：4 节点 256 核，3403s 完成 2 天的计算量；鲲鹏软件栈及热点函数优化性能提升 140%。

HYCOM 模式：10 节点 1032 核，3960s 完成计算；鲲鹏软件栈性能提升 20%，是同节点 Intel E5-2680 V4 性能的 330%。

CESM 模式：8 节点 1024 核，224s 完成计算；鲲鹏软件栈性能提升 44%。

气候变化加剧,各种极端天气和海洋灾害频发,这些对海洋和大气过程的精准和精细化预报提出了更高要求。因此,追求海洋模式更快的计算速度是永无止境的。

在这个背景下，通过优化海洋模式代码并充分利用鲲鹏处理器的性能优势，大幅提升了海洋模拟的计算效率。这不仅缩短了模拟时间,还可以支持更高精度的海洋预报,为应对极端天气和海洋灾害提供关键支撑。

我们相信,通过计算性能的持续提升,海洋模式可以更好地服务于气候预测、灾害预警以及海洋生态保护等领域。我们会继续与业界伙伴积极开展前沿的高性能计算优化研究,以支持海洋科学事业的发展。

参考文献:

- [1] Praharenka W , Pankratz D , Carvalho J P L D ,et al.Vectorizing divergent control flow with active-lane consolidation on long-vector architectures[J].Journal of supercomputing, 2022.DOI:10.1007/s11227-022-04359-w.
- [2] Shchepetkin A F, McWilliams J C. The regional oceanic modeling system (ROMS): a split-explicit, free-surface, topography-following-coordinate oceanic model[J]. Ocean Model, 2005, 9(4): 347–404.
- [3] Lupo C , Pantoja M , Choboter P .Enhancing regional ocean modeling simulation performance with the Xeon Phi architecture[C]//OCEANS 2017 - Aberdeen.California Polytechnic State University, San Luis Obispo, California, USA 93407;California Polytechnic State University, San Luis Obispo, California, USA 93407;California Polytechnic State University, San Luis Obispo, California, USA 93407;, 2017.
- [4] Warner J C , Geyer R R , Arango R G .Using a composite grid approach in a complex coastal domain to estimate estuarine residence time[J].Computers & Geosciences, 2010, 36(7):921-935.