Performance evaluation of a computer cluster for the realization of submesoscale-resolved earth system models

Rin Irie, Helen Stewart, Tetsuya Fukuda, Tsuneko Kura, Masaki Hisada
Nippon Telegraph and Telephone Corporation
Musashino-shi, Tokyo, Japan
{rin.irie|helen.stewart|tetsuya.fukuda|tsuneko.kura|masaki.hisada}@ntt.com

1 INTRODUCTION

Earth system models (ESM), which model the complex interactions between physical and biological systems in the climate, are among the most computationally demanding applications of HPC today. In the 6th Coupled Model Intercomparison Project (CMIP6), the current finest horizontal resolution for ESMs computed globally is on the mesoscale ($O(10^4 \sim 10^6)$m) [1]. With the goal of resolving ESMs on the submesoscale ($O(10^3)$m), we construct a computer cluster capable of computing ocean simulations at resolutions of up to $O(10^3)$m. We evaluate the parallel performance and usability of the constructed cluster by using the ocean physics component of MITgcm [2], a commonly used ocean general circulation model.

2 METHODS

The hardware and software specifications for the cluster evaluated in this study are shown in Tables 1 and Table 2 respectively. The nodes in the cluster are internally connected by RDMA (Remote Direct Memory Access) using 200Gbit Infiniband (NVIDIA ConnectX-6). MITgcm is used for benchmarking with an eddy-permitting baroclinic ocean model building on our previous work [3]. The simulation conditions are specified in Table 3. MITgcm supports running multiple processes in parallel computation by dividing the computation area in the horizontal direction. For this evaluation, we measure the computation time for different conditions, including the process number of file I/O and MPI communication methods to validate strong scaling. Computation times are taken as the average of three repeated measurements.

Table 1: Hardware Specifications

Computing nodes (22 Nodes)	
CPU	Intel Xeon Gold 6342 (2.8GHz, 24c)
Memory	512GB (DDR4-3200 ECC 64GB x8)
Storage	PCIe 3.0 NVMe M.2 1.9TB

Storage node (1 Node)	
CPU	Intel Xeon Gold 6326 (2.9GHz, 16c) x2 sockets
Memory	256GB (DDR4-3200 ECC 16GB x16)
Storage (OS)	SATA3 SSD 1.9TB
Storage (NFS)	PCIe 4.0 NVMe U.2 SSD 7.6TB x16 (Intel VROC Std.)

Table 2: Software Specifications

OS	Ubuntu 22.04.01 (Kernel v5.15.0-75-generic)
MPI library	Open MPI (v4.1.5)
Job management	Slurm (v21.08.5-2)
Network filesystem	NFS (v2.6.1), mlnx-nfsrdma-dkms (v5.8.3.0.4.1)
HCA driver	MLNX-OFED (v5.8.3.0.7.1)
Fortran compiler	gfortran (v11.3.0)

Table 3: Simulation Conditions

Software version	checkpoint68i (Mid 2022 version)
Domain size	$S_x = 30°$, $S_y = 20°$, $S_z = 4.0$ km,
Horizontal grid spacing	$\Delta x = \Delta y = \{0.125°, 0.0625°\}$
Vertical grid spacing	Unequal (max.: 149.4 m, min.: 5.9 m, 49 layers)
Time step	$\Delta t = 240$ seconds
Integration time	$N_t = 65520$ steps (=0.5 years)

3 RESULTS AND DISCUSSION

Figure 1(a) shows the computation time depending on the process number of file I/O and MPI communication methods at a spatial resolution of 0.125°. When RDMA is used for MPI communication, the computation time decreases with the number of processes up to 384 processes. For TCP/IP, the computation time remains the same or increases as the number of processes exceeds 96. As the number of processors is increased, the effect of the bypass of OS and CPU by RDMA is increased, because the computational load on the OS and context switch increases. Above 96 processors, single I/O using RDMA gives a shorter computation time than multi I/O. But single I/O using TCP/IP is faster above 96 processors.

Figure 1(b) shows the computation time at spatial resolutions of 0.125° and 0.0625° with 16 and 24 processes per node (p/n). The computation time is shorter with 16 processes per node than with 24. This is because the number of CPU cores per node is 24, leaving the later subject to many OS and other processing interruptions. The strong scaling performance is also expected to correlate more with the number of processes used than with the number of cells per process. In this poster, results for alternative node allocations and supercomputer Fugaku will be discussed in more detail.

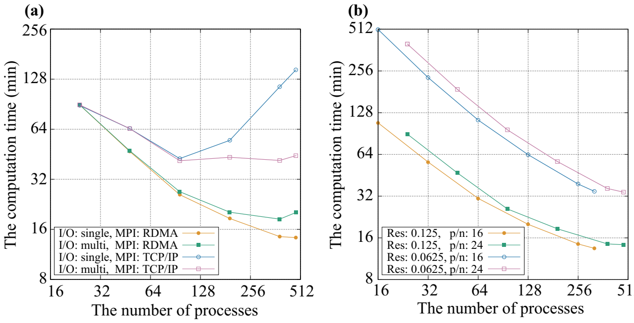

Figure 1: Benchmark of Strong Scaling.

ACKNOWLEDGEMENTS

This work used computational resources of supercomputer Fugaku provided by the RIKEN Center for Computational Science through the HPCI System Research Project (Project ID: hp230382).

REFERENCES

[1] M. C. Acosta et al. 2023. The computational and energy cost of simulation and storage for climate science: lessons from CMIP6. *Geoscientific Model Development Discussions* 2023 (2023), 1–21.

[2] J. Marotzke et al. 1999. Construction of the adjoint MIT ocean general circulation model and application to Atlantic heat transport sensitivity. *Journal of Geophysical Research: Oceans* 104, C12 (1999), 29529–29547.

[3] H. Stewart et al. 2023. Effect of submesoscale dynamics and baroclinic instabilities on phytoplankton. In *EGU General Assembly Conference Abstracts*. https://doi.org/10.5194/egusphere-egu23-11212