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

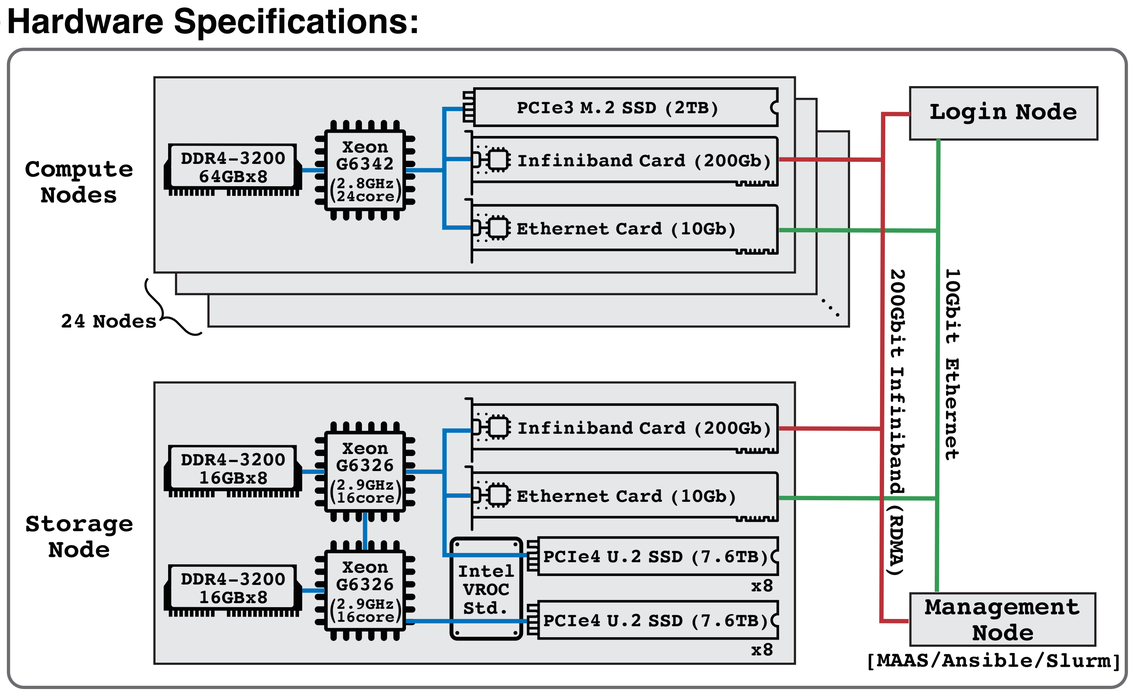
1 Introduction

- 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.
- We construct a computer cluster (Seadragon) capable of computing ocean simulations at submesoscale resolutions of up to $O(10^3)$m.
- We evaluate the parallel performance and usability of the Seadragon and supercomputer Fugaku using the ocean physics component of MITgcm [1] for varying numbers of file I/O process number, processes per node (p/n) and MPI communication methods (RDMA, TCP/IP).

2 Methods

Cluster Seadragon

- Hardware Specifications:



- Software Specifications:

OS	Ubuntu 22.04.01 LTS (Kernel v5.15.0-75-generic)
MPI library	OpenMPI 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 v5.8.3.0.4.1
Fortran compiler	gfortran v11.3.0

Benchmark Conditions

- Physics model: MITgcm Baroclinic ocean gyre [2, 3]

$$\frac{\mathrm{D}u}{\mathrm{D}t} - fv - \frac{uv}{a}\tan\varphi + \frac{1}{\rho_c a\cos\varphi}\frac{\partial p'}{\partial\lambda} + \nabla_h\cdot(-A_h\nabla_h u) + \frac{\partial}{\partial z}\left(-A_z\frac{\partial u}{\partial z}\right) = \mathcal{F}_u$$

$$\frac{\mathrm{D}v}{\mathrm{D}t} + fu + \frac{u^2}{a}\tan\varphi + \frac{1}{\rho_c a}\frac{\partial p'}{\partial\varphi} + \nabla_h\cdot(-A_h\nabla_h v) + \frac{\partial}{\partial z}\left(-A_z\frac{\partial v}{\partial z}\right) = \mathcal{F}_v$$

$$\frac{\partial\eta}{\partial t} + \frac{1}{a\cos\varphi}\left(\frac{\partial H\widehat{u}}{\partial\lambda} + \frac{\partial H\widehat{v}\cos\varphi}{\partial\varphi}\right) = 0$$

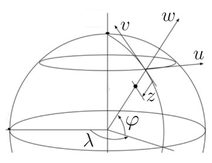
$$p' = g\rho_c\eta + \int_z^0 g\rho'\mathrm{d}z$$

$$\frac{\mathrm{D}\theta}{\mathrm{D}t} + \nabla_h\cdot(-\kappa_h\nabla_h\theta) + \frac{\partial}{\partial z}\left(-\kappa_z\frac{\partial\theta}{\partial z}\right) = \mathcal{F}_\theta$$

$$\frac{\mathrm{D}S}{\mathrm{D}t} + \nabla_h\cdot(-\kappa_h\nabla_h\theta) + \frac{\partial}{\partial z}\left(-\kappa_z\frac{\partial S}{\partial z}\right) = \mathcal{F}_S$$

$$\rho = \rho_c(1-\alpha(\theta-\theta_c)+\beta(S-S_c))$$

$$\nabla\cdot\boldsymbol{u} = 0$$



$\boldsymbol{v} = (u, v, w)$ is the flow velocity vector
A_h and A_z are the horizontal and vertical viscosity
κ_h and κ_z are the horizontal and vertical diffusivities
η is the free surface height
θ is the potential temperature
p is the pressure
$H\widehat{u}$ and $H\widehat{v}$ are the depth integrals of u and v respectively
ρ_c is the reference density
g is acceleration due to gravity
ρ' is the fluid density integrated through the water column
f is the Coriolis parameter
$\mathcal{F}_u, \mathcal{F}_v, \mathcal{F}_\theta$ and $\mathcal{F}_S$ are the forcing terms

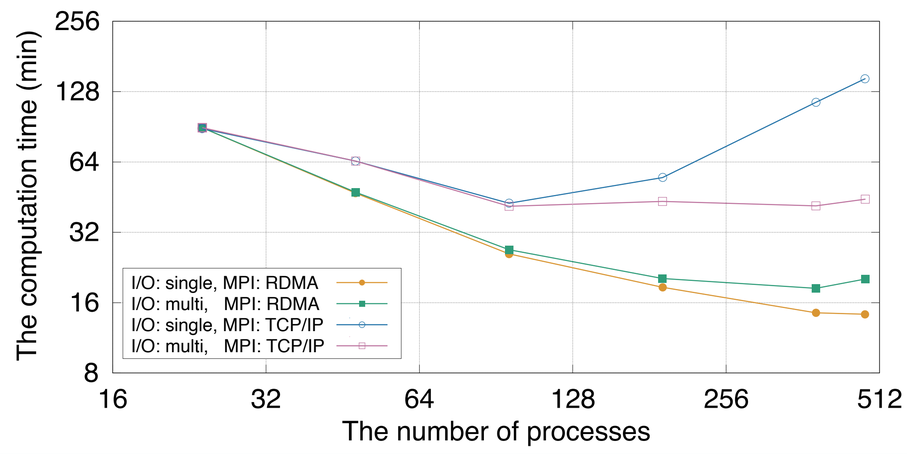
- Discretization Condition:

Domain size	$S_\lambda = 30°, S_\phi = 20°, S_z = 4\mathrm{km}$
Horizontal grid spacing	$\Delta\lambda = \Delta\phi = \{0.125°, 0.0625°\}$
Vertical grid spacing	Unequal (max. 149.4m, min. 5.9m, 49 layers)
Time step	$\Delta t = 240$ seconds
Integration Time	$N_t = 65520$ steps (~ 0.5 years)

- Parallel Computation
  - The simulation area is horizontally partitioned into grid tiles.
  - MITgcm supports running multiple processes (MPI) and multiple threads (OpenMP) in parallel.
  - In this case, only MPI is used for parallel computation (i.e. flat MPI).

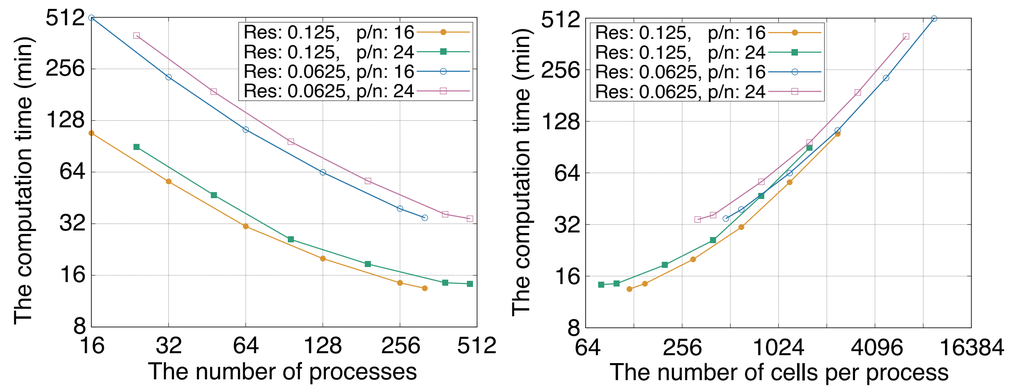
3 Results & Discussions

Result 1 The computation times depending on the process number of file I/O and MPI communication methods at a horizontal resolution of 0.125° with 24 processes per node (p/n).

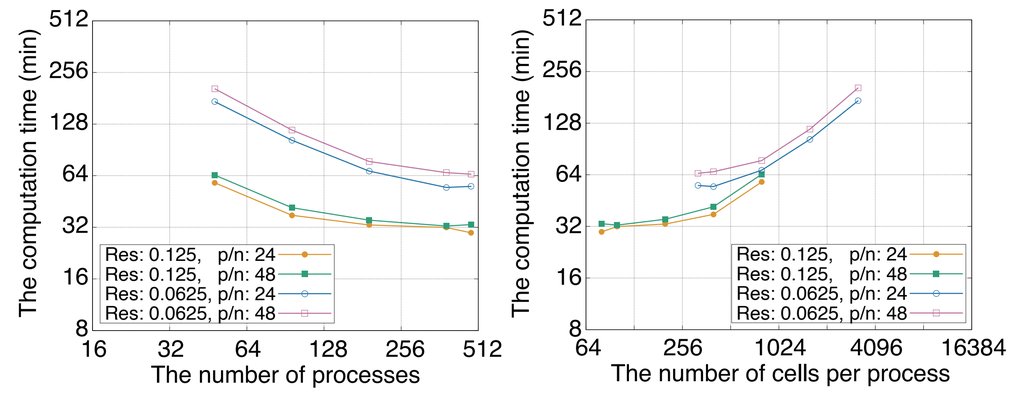


- For TCP/IP, the computation time remains the same or increases as the number of processes exceeds 96.
- For RDMA, the proportional effect of the OS and CPU bypass increases as the number of processors is increased.
- Single I/O with RDMA is faster than multi I/O exceeding 96 processors.
- Single I/O with TCP/IP is slower than multi I/O exceeding 96 processors.

Result 2 The computation time at resolutions of 0.125° and 0.0625° (a) on the constructed cluster with 16 and 24 p/n,



(b) on Fugaku with 24 and 48 p/n.



- The computation time is longer when the process number is the maximum number of CPU cores per node (Seadragon = 24, Fugaku = 48) on the respective machines.
  - For the Seadragon Cluster, this is because of OS and other processing interruptions.
  - As Fugaku is not affected by processing interruptions, the is thought to be limited by p/n.
- Strong scaling performance is seen be more determined by the number of processes used than the p/n.

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. J. Marotzke *et al.*, Construction of the adjoint MIT ocean general circulation model and application to Atlantic heat transport sensitivity, *Journal of Geophysical Research: Oceans* 104, C12, 29529–29547 (1999).
2. H. Stewart *et al.*, Effect of submesoscale dynamics and baroclinic instabilities on phytoplankton, In *EGU General Assembly Conference Abstracts*, https://doi.org/10.5194/egusphere-egu23-11212.
3. R. Irie *et al.*, Ultra-High-Resolution Simulations for Resolving Submesoscale Ocean Eddies, In *34th IUPAP Conference on Computational Pyhsics*, C07-4-04 (2023).