HPC challenges for new extreme scale applications, March 07, 2023

Development of a Heterogeneous Coupling Library h3-Open-UTIL/MP

Takashi Arakawa[1,2]
Shinji Sumimoto[2]
Hisashi Yashiro[3]
Kengo Nakajima[2]
1:CliMTech Inc.
2:Information Technology Center, the University of Tokyo
3:National Institute for Environmental Studies

Content

- Introduction
- About h3-Open-UTIL/MP
  - Structure
  - How to realize heterogeneous coupling
- Performance Evaluation I
  - By Toy Models
- Case Study
  - Atmospheric model and Machine Learning library coupling
- Performance Evaluation II
  - By Real Applications

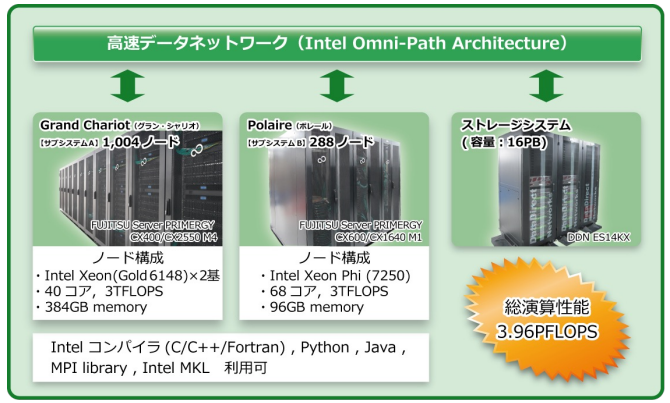
Introduction

Key word of recent HPC trend

Heterogeneity

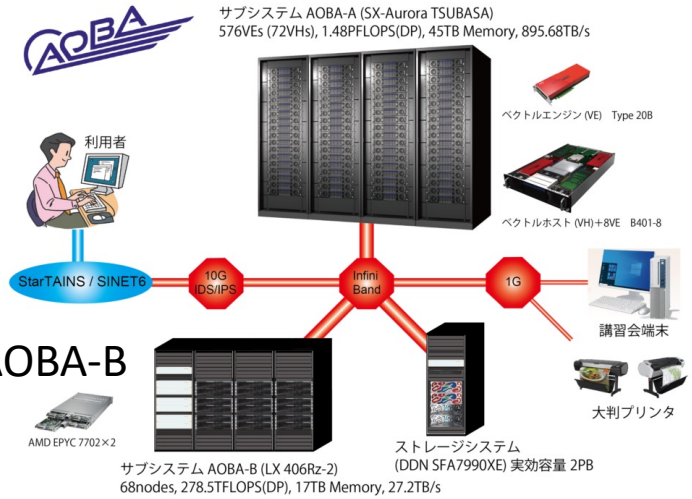
Many HPCs in Japanese universities have heterogeneous architecture

Grand Chariot　Polaire



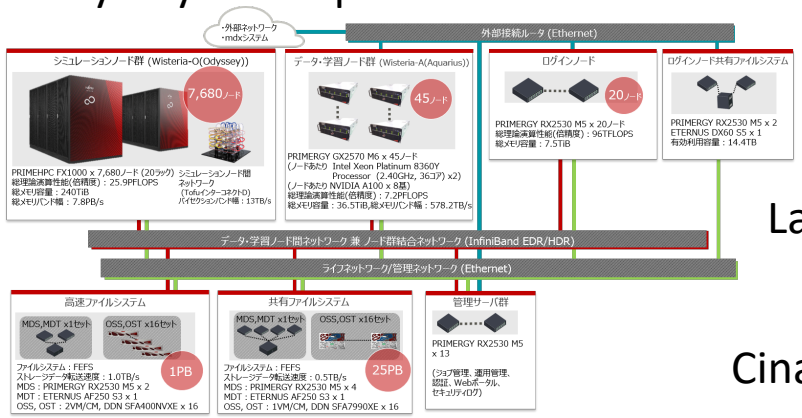
Hokkaido Univ.

AOBA-A



Tohoku Univ.

Odyssey　Aquarius

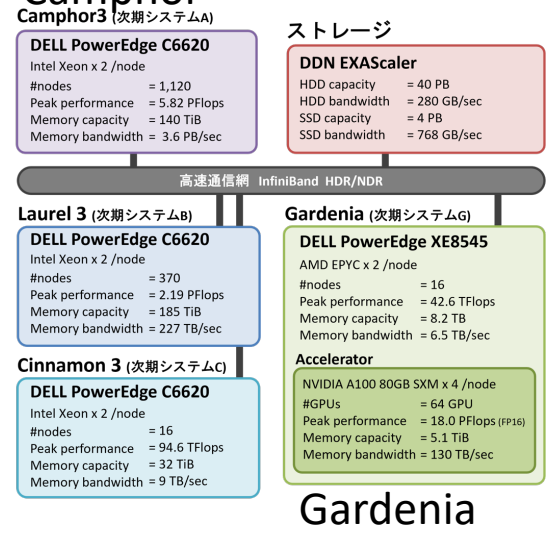


Univ. of Tokyo

Laurel

Cinamon

Camphor



Gardenia

Kyoto Univ.

System Architecture

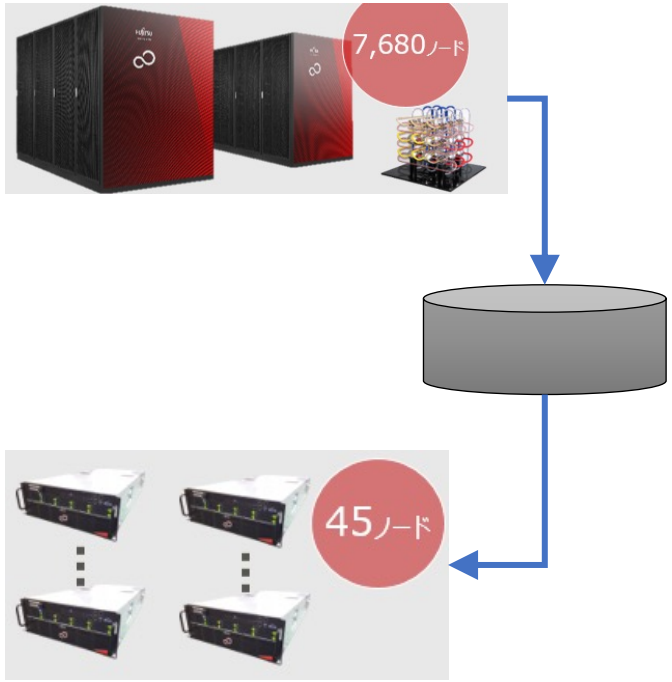
- The computer systems that comprise HPCI (High Performance Computing Infrastructure) of Japan.
- Except for Fugaku and TSUBAME, all other HPCs have multiple different architecture node group.

HPCI High Performance Computing Infrastructure

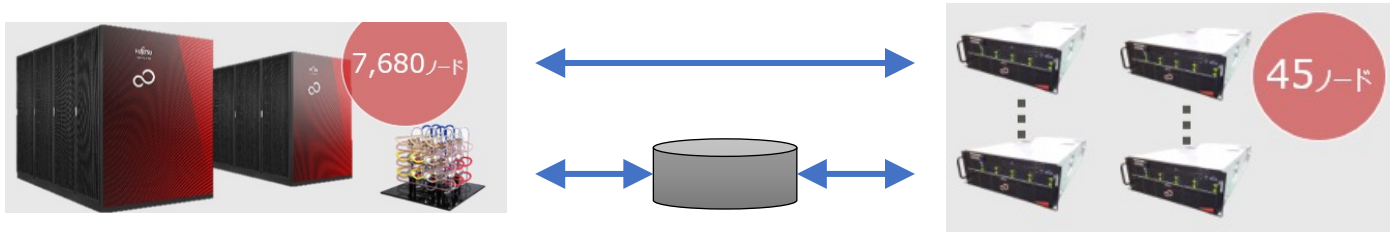
Organization	System Name	Architecture	Node Configuration
Hokkaido Univ.	Grand Chariot + Polaire	Xeon Gold + Xeon Phi	CPU + CPU
Tohoku Univ.	AOBA-A + AOBA-B	SX-Aurora Tsubasa + AMD EPYC	VE + CPU
Tsukuba Univ.	Cygnus (Deneb + Albireo)	Xeon/NVIDIA Tesla (/Nallatech 520N)	GPU + FPGA
AIST	ABCI (Node A + Node V)	Xeon Platinum/NVIDIA A100 + Xeon Gold/ NVIDIA Tesla V100	GPU + GPU
Univ. of Tokyo	Wisteria/BDEC-01 (Odyssey + Aquarius)	A64FX + Xeon Platinum/NVIDIA A100	CPU + GPU
Tokyo Tech.	TSUBAME	Xeon E5/NVIDIA Tesla P100	GPU
JAMSTEC	Earth Simulator 4	SX-Aurora Tsubasa + AMD EPYC + AMD EPYC/NVIDIA A100	VE + CPU + GPU
Nagoya Univ.	Flow [不老] (Type I + Type II + Type III)	A64FX + Xeon Gold/NVIDIA Tesla V100 + Xeon Platinum/NVIDIA Quadro	CPU + GPU + GPU
Kyoto Univ.	Camphor3 + Laurel3 + Cinnamon3 + Gardenia	Xeon + Xeon + Xeon + AMD EPYC/NVIDIA A100	CPU x3 + GPU
Osaka Univ.	SQUID	Xeon Platinum + Xeon Platinum/NGIDIA A100 + SX-Aurora Tsubasa	CPU + GPU + VE
Riken/RCCS	Fugaku[富岳]	A64FX	CPU
Kyusyu Univ.	ITO [いと](Subsystem A + Subsystem B)	Xeon Gold + Xeon Gold/NVIDIA Tesla	CPU + GPU

Heterogeneous System and Software

- The reason for the development of heterogeneous system
  - The role of HPC has expanded beyond not only simulation but also to large-scale data analysis and machine learning.
- Typical use case of Heterogeneous System
  - File sharing→One way and Sequential→Time and resource consuming

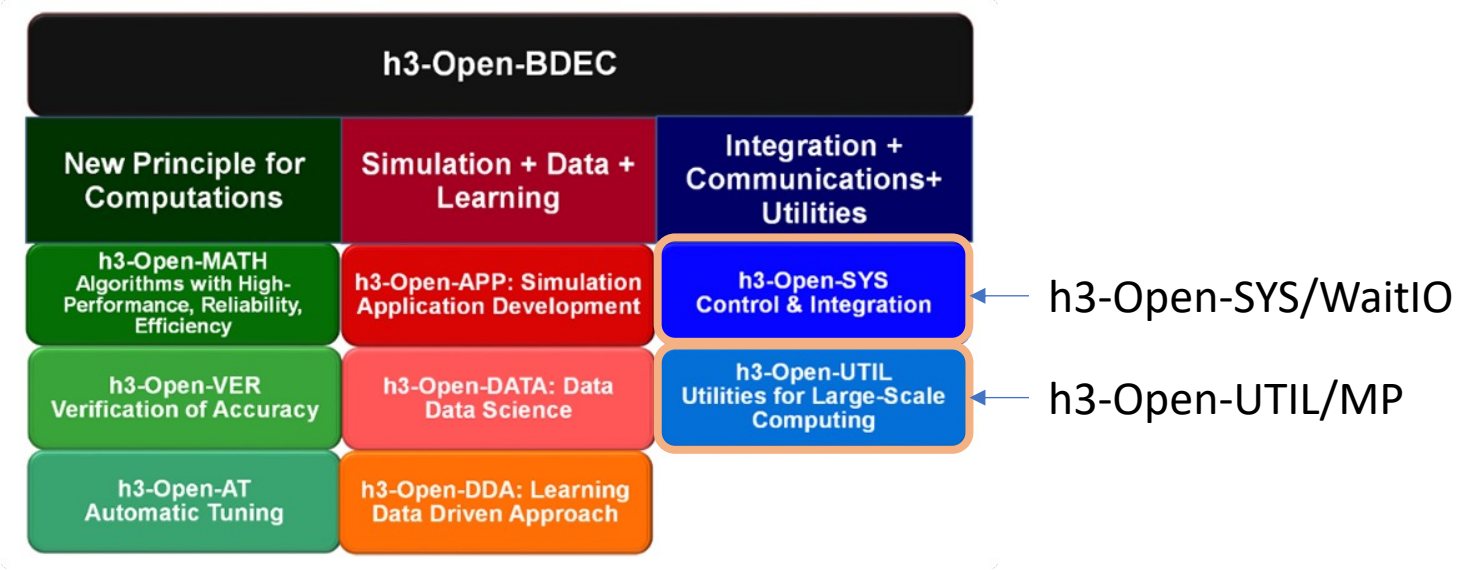


- Coupling Software which supports real time data exchange between different systems (via inter-connect or storage system)
  - Two-way Concurrent processing become possible
  - Interactive processing is essential for reproduce more realistic world in the computer system.



h3-Open-UTIL/MP

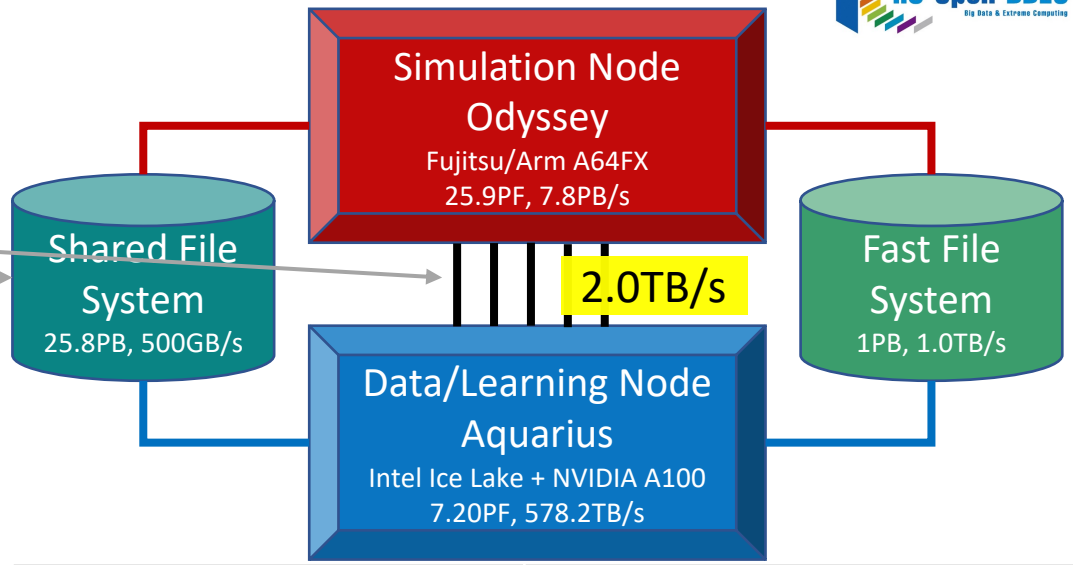
- h3-Open-UTIL/MP
  - Coupling library which supports internode data exchange and application coupling by collaborating with h3-Open-SYS/WaitIO in the h3-Open-BDEC project.
- h3-Open-BDEC project
  - h3: Hierarchical, Hybrid, and Heterogeneous
  - BDEC:Big Data & Extreme Computing
- h3-Open-BDEC software suite
  - 8 packages
  - UTIL/MP →h3-Open-UTIL, WatiIO → h3-Open-SYS



Structure of h3-Open-BDEC project/software suite
https://h3-open-bdec.cc.u-tokyo.ac.jp/index.html

h3-Open-SYS/WaitIO

- Communication Library for Heterogeneous System
  - Via Interconnect : WaitIO-Socket
  - Via File : WaitIO-File

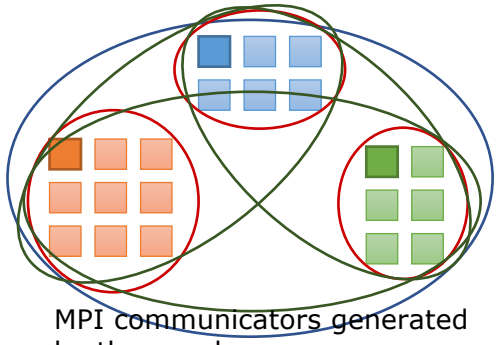


- API Overview
  - Basic data exchange API: waitio_isend, waitio_irecv, waitio_wait same as MPI
    ↑
  - easy to incorporate WaitIO into UTIL/MP

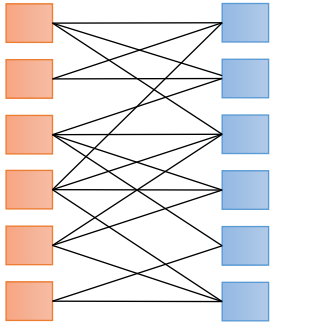
API	Description
waitio_isend	Non-Blocking Send
waitio_irecv	Non-Blocking Receive
waitio_wait	Send/Recv Wait Completion
waitio_init	WaitIO Initialization
waitio_get_nporcs	Get # of PB member
waitio_create_group	Create a PB group by function
waitio_create_group_wranks	Create a PB group by member list
waitio_group_rank	Get my group rank
waitio_group_size	Get my group size
waitio_pb_size	Get PB size
waitio_pb_rank	Get PB rank

Features of h3-Open-UTIL/MP

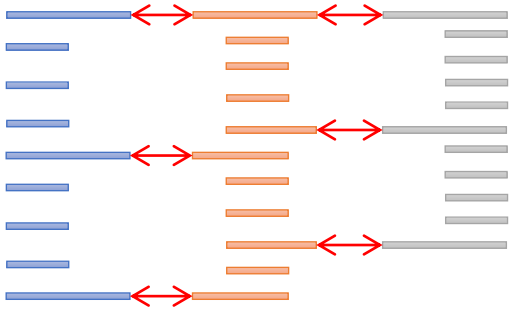
- General purpose coupling library
  - The grid system of the model dose not change in time
  - The exchange time interval of each data is constant.
- General features
  - Process group management
  - Data exchange management
    - Local to Local communication
    - Data exchange timing in the integration loop
    - Spatial remapping between the different grid systems



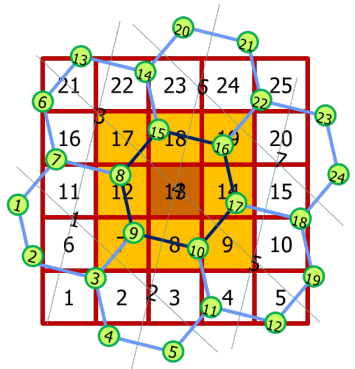
MPI communicators generated by the coupler



Data exchange pattern



Data exchange timing

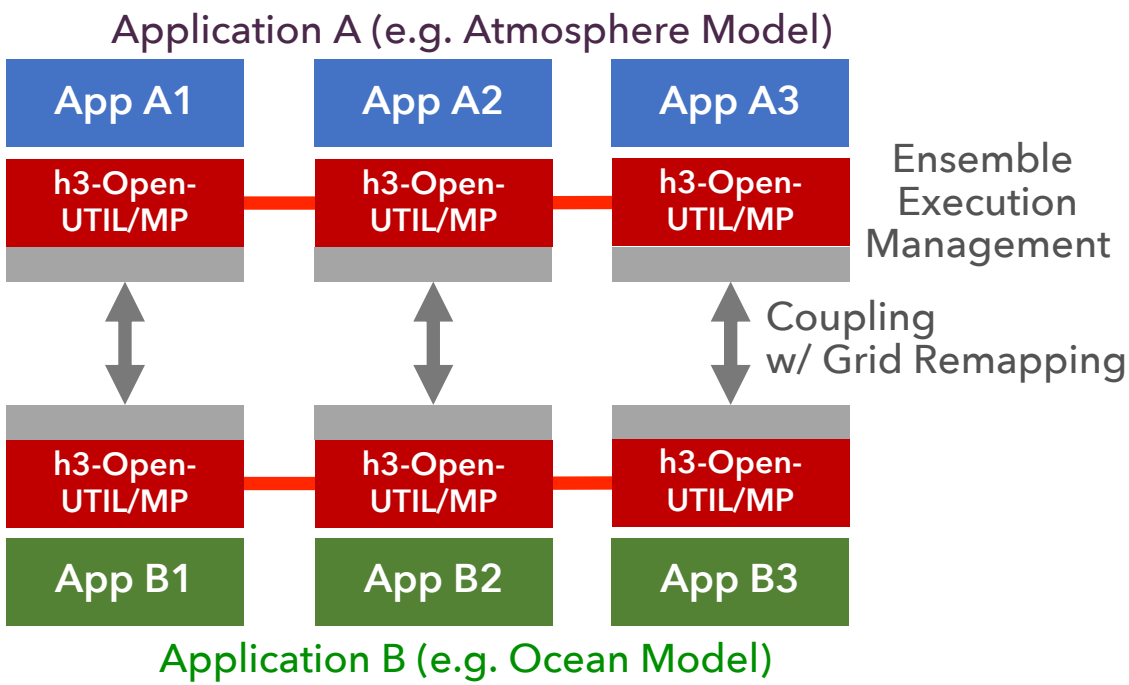


Grid remapping

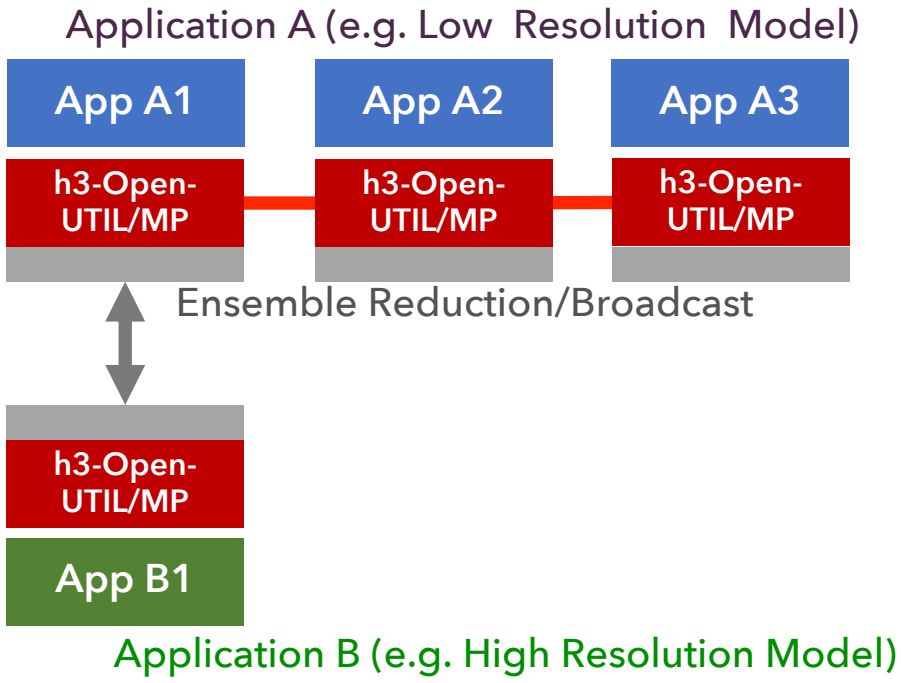
- Unique Features
  - Heterogeneous Coupling by collaborating with h3-Open-SYS/WaitIO
  - Ensemble Coupling
  - Python interface, etc.

Ensemble Coupling

- Ensemble Calculation
  - An ensemble calculation is a technique to run many identical models with slightly different conditions.



M x (A + B) execution

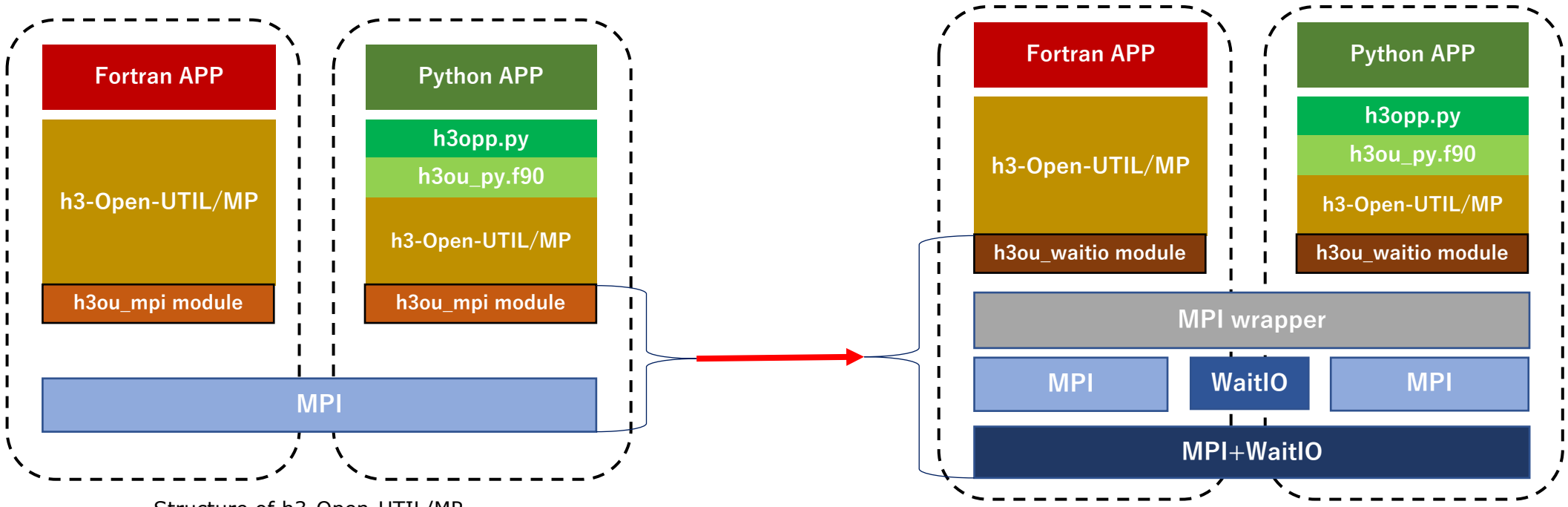


(M x A) + B execution

- Ensemble Coupling by h3-Open-UTIL/MP
  - Many to many: Ensemble of Atmosphere-Ocean Coupling
  - Many to one : Low Resolution Model Ensemble + High Resolution Model

Structure of UITL/MP with WaitIO

- h3-Open-UTIL/MP
  - A set of modules of Fortran95
  - These modules have a hierarchical structure.
  - A MPI handling module is at the bottom of this hierarchy.
  - All other modules are designed to use MPI through this MPI handling module.
- Collaboration with WaitIO was easy to achieve!
  - To make a communication library using MPI and WaitIO
  - To modify h3ou_mpi module to support this new library



Structure of h3-Open-UTIL/MP

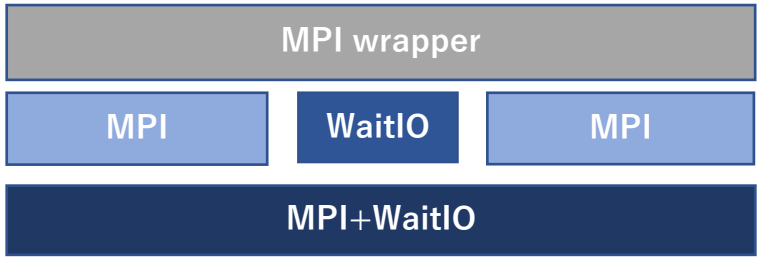
Structure of h3-Open-UTIL/MP with WaitIO

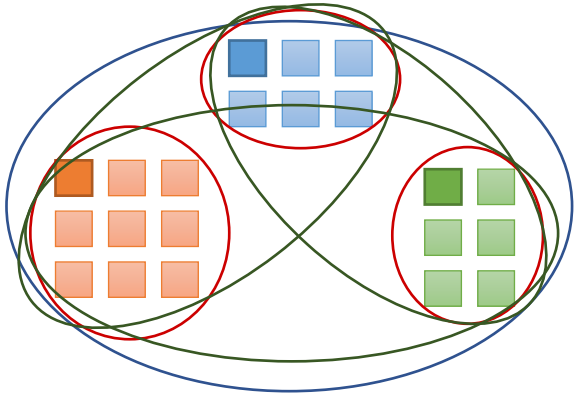
Communication Library for UITL/MP with WaitIO

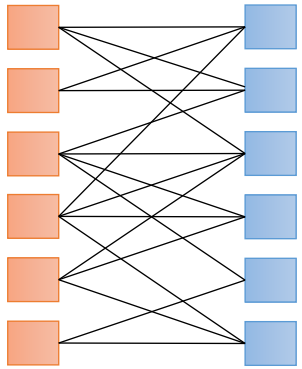
- Communication Patterns of h3-Open-UTIL/MP
  - Intra-application communication
  - One-to-one communication between applications
  - Global communication

  } Inter node communication by WaitIO

- Inter node communication
  - One-to-One
    - Pure WaitIO functions : waitio_isend, waitio_irecv, waitio_wait
  - Global
    - Combination of MPI + WaitIO
    - Global functions called in UTIL/MP
      - MPI_Bcast, MPI_Gather, MPI_Reduce, MPI_AllReduce

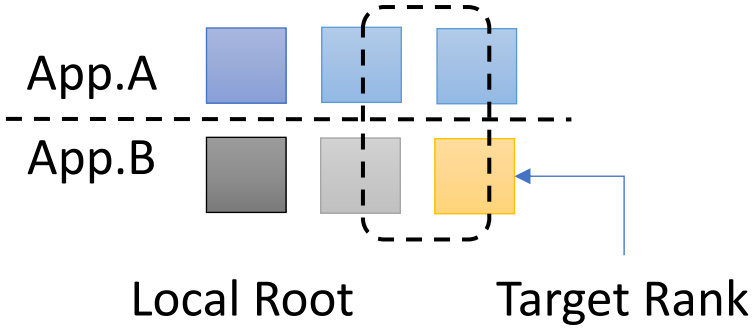


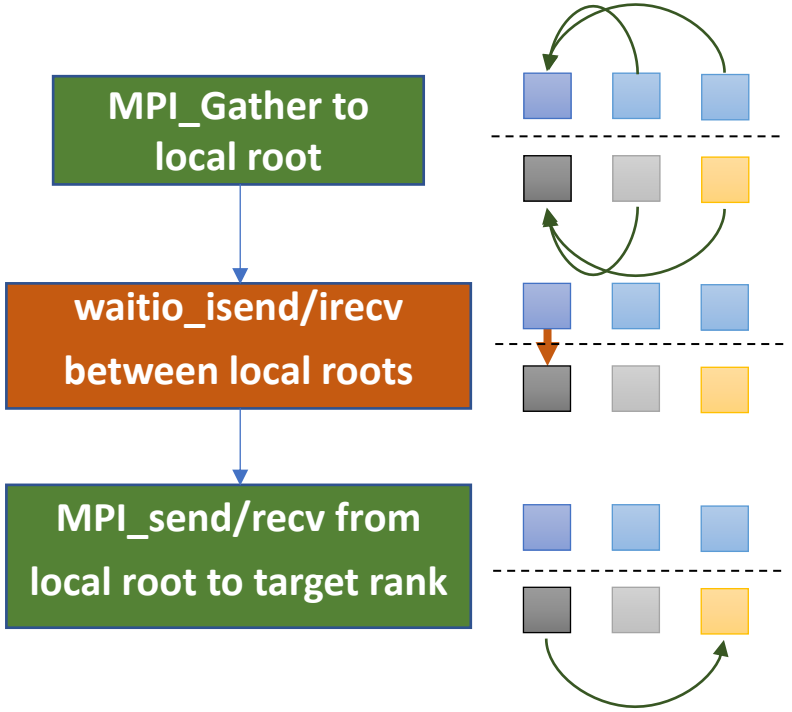




Global communication by WaitIO + MPI

- Example of Global Communication : MPI_Gahter
  - Data are gathered to local root by MPI_Gather
  - Gathered Data of App.A is sent to local root of App.B by waitio_isend/waitio_irecv
  - Merged Data is sent to target rank form local root of App.B by MPI_send/MPI_recv

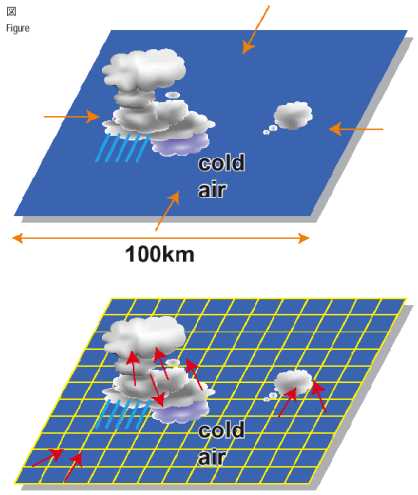




- Performance?
  - This algorithm is not efficient.
  - These global communications are used only for the initialization process.
  - ➢ The impact on total performance is small.

Case Study of Heterogeneous Coupling

- Coupling of Atmospheric Model and Machine Learning Library
- Motivation of this experiment
  - Tow types of Atmospheric models: Cloud resolving VS Cloud parameterizing
  - Could resolving model is difficult to use for long time climate simulation
  - Low resolution parameterized models have many assumptions and uncertainties.
  - Replacing low-resolution cloud processes calculation with AI trained by cloud resolving calculation.



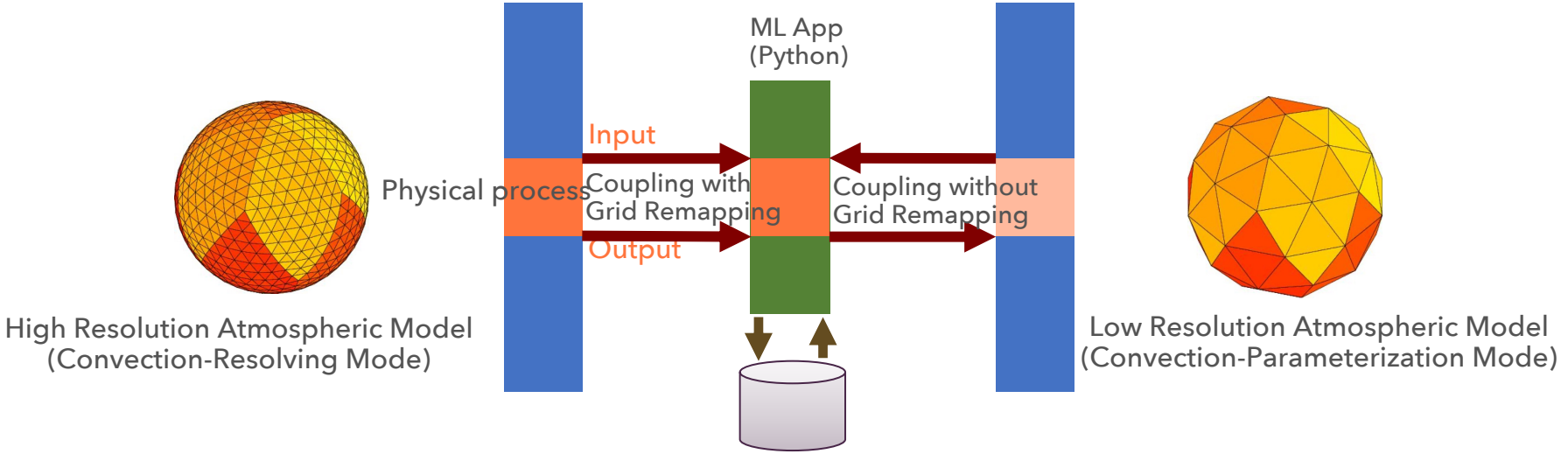
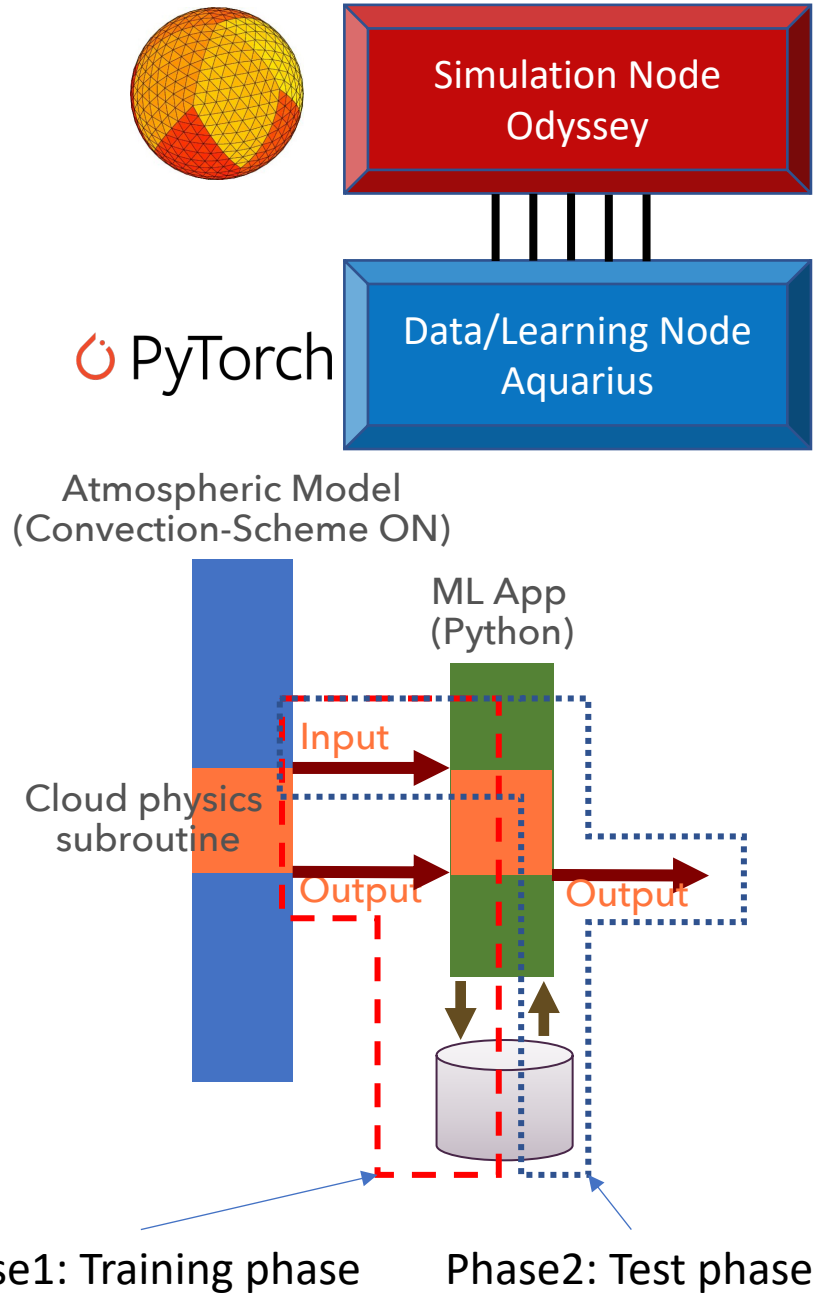


Diagram of applying ML to an atmospheric model

Experimental Design

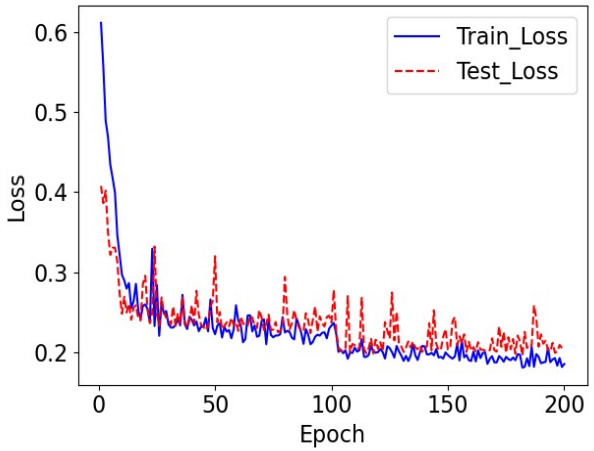
- Atmospheric model on Odyssey
  - NICAM : global non-hydrostatic model which has an icosahedral grid
  - Resolution : horizontal : 10240, vertical : 78
- ML on Aquarius
  - Framework : PyTorch
  - Method : Three-Layer MLP
  - Resolution : horizontal : 10240, vertical : 78
- Experimental design
  - Phase1: PyTorch is trained to reproduce output variables from input variables of cloud physics subroutine.
  - Phase2:Reproduce the output variables from Input variables and training results
- Training data
  - Input : total air density (rho), internal energy (ein), density of water vapor (rho_q)
  - Output : tendencies of input variables computed within the cloud physics subroutine

$$\frac{\Delta rho}{\Delta T} \quad \frac{\Delta ein}{\Delta T} \quad \frac{\Delta rho\_q}{\Delta T}$$

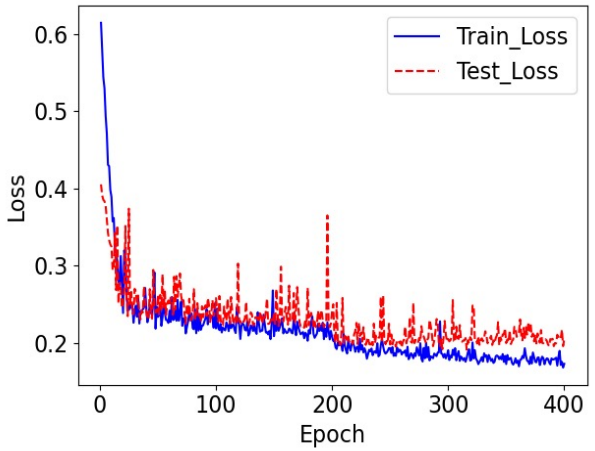


Phase1, Training process

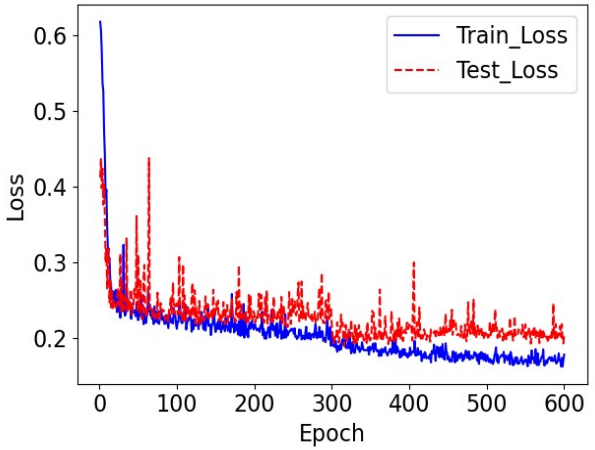
- Training
  - 3 cases : 100, 200, 300 epochs per time step
  - Convergence is very good in any case



Epoch = 100/step



Epoch = 200/step



Epoch = 300/step

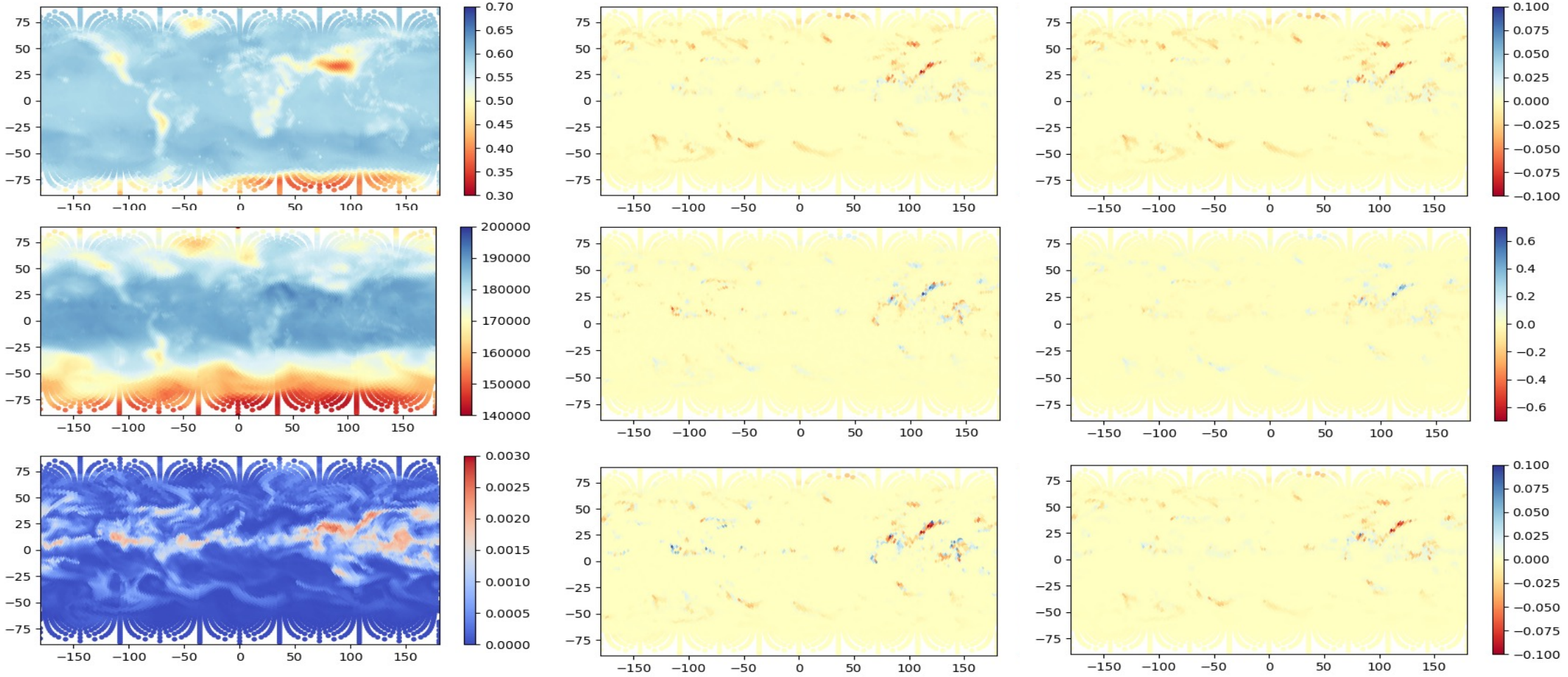
Test calculation

- Compute output variables from input variables and PyTorch
  - The rough distribution of all variables is well reproduced
  - The reproductivity of extreme values is no good→next table

Total air density

Internal energy

Density of water vapor



Summary of ML test

Correlation between reference and AI output

	slope	intercept	correlation coef.
air density	0.53847887	0.00017136	0.6194595
internal energy	0.58866578	-0.00019053	0.6769882
water vapor	0.56042855	0.000008773	0.6494912

- for more accurate reproductivity
  - Cloud physics is a complex system
    - NICAM subroutine mp_driver has INPUT:23, OUTPUT: 27, INOUT: 11
    - Variable selection is important!
  - Atmosphere has spatial (and temporal) structure
    - MLP: using only point-to-point relationship
    - Using an algorithm reflects spatio-temporal structure such as CNN

Heterogeneous coupling is successfully completed

Conclusion

- Summary
  - h3-Ope-UTIL/MP is general purpose coupling library
  - Enables heterogeneous coupling by collaborating with WaitIO
  - Atmospheric model NICAM and ML APP was coupled on Odyssey + Aquarius
  - Inter-node connectivity needs to be faster
- Future Work
  - Heterogeneous coupling
    - larger scale
    - better reproductivity
  - Ensemble coupling
    - Application case study is ongoing