東京大学情報基盤センター
INFORMATION TECHNOLOGY CENTER, THE UNIVERSITY OF TOKYO

東京大学
THE UNIVERSITY OF TOKYO

Hierarchical, Hybrid, Heterogeneous
h3-Open-BDEC
Big Data & Extreme Computing

Wisteria
BDEC-01

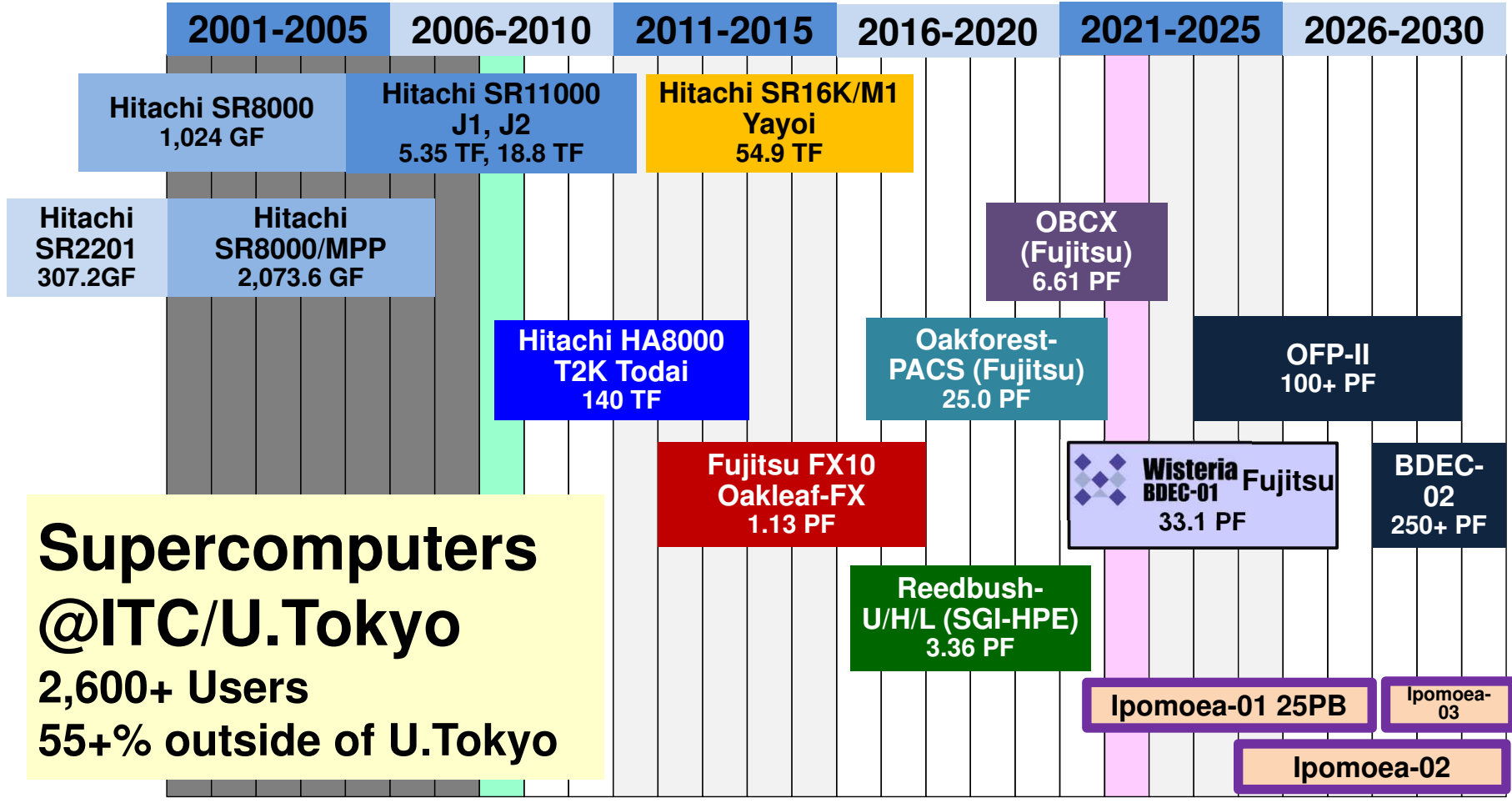
Wisteria/BDEC-01 & h3-Open-BDEC:
Innovative Scientific Computing by Integration of (Simulation + Data + Learning)

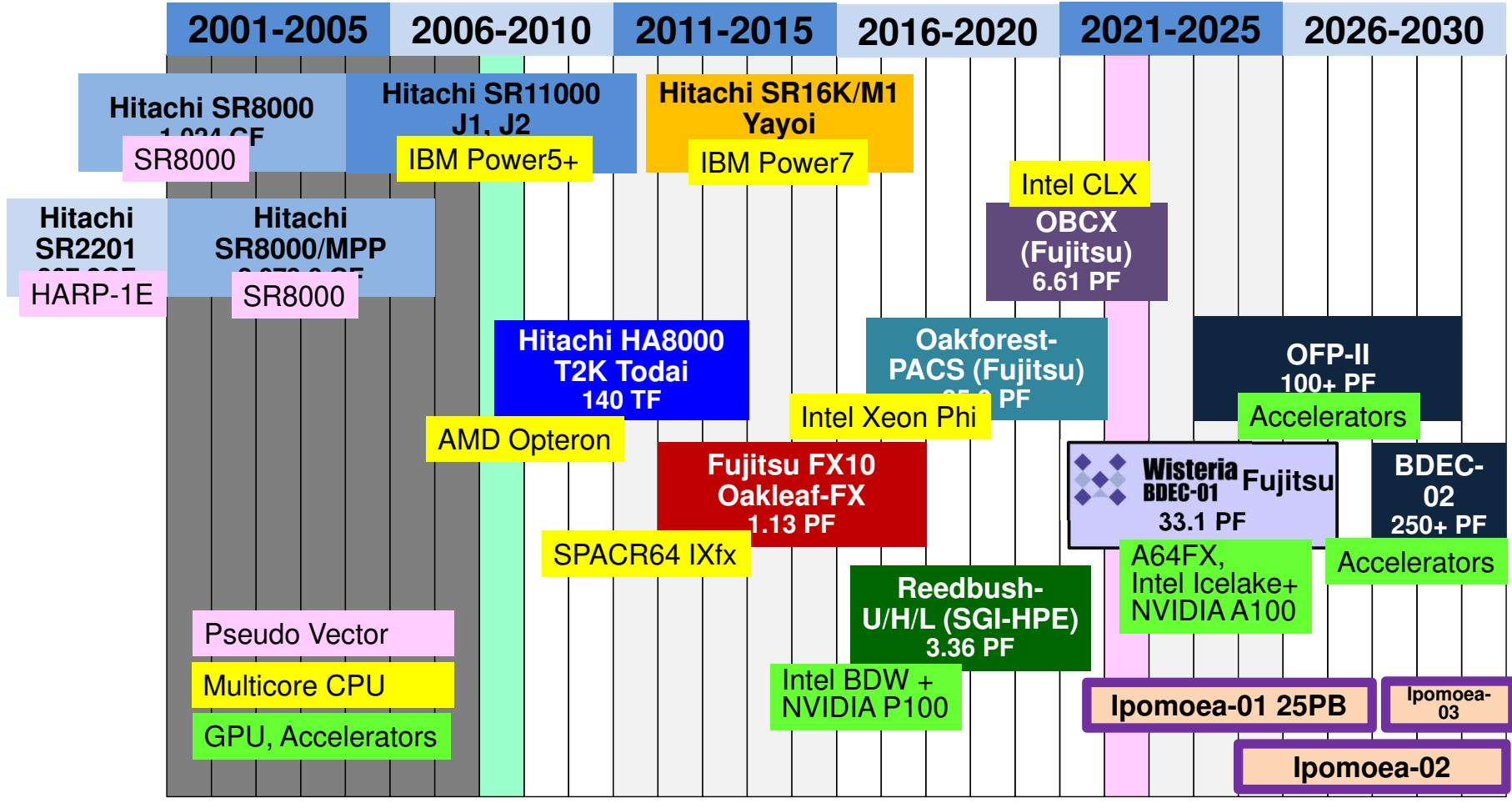
https://h3-open-bdec.cc.u-tokyo.ac.jp/

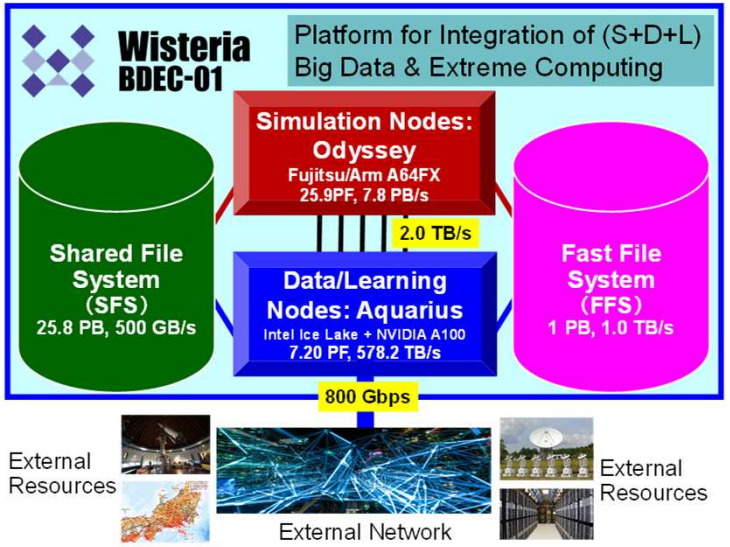
Kengo Nakajima
Information Technology Center
The University of Tokyo

MS2: Progress & Challenges in Extreme Scale Computing & Big Data
SIAM PP22, February 23-26, 2022 (online)

siam 2022 | Conference on Parallel Processing for Scientific Computing


2001-2005
2006-2010
2011-2015
2016-2020
2021-2025
2026-2030
Hitachi SR8000
1,024 GF
Hitachi SR11000
J1, J2
5.35 TF, 18.8 TF
Hitachi SR16K/M1
Yayoi
54.9 TF
Hitachi
SR2201
307.2GF
Hitachi
SR8000/MPP
2,073.6 GF
OBCX
(Fujitsu)
6.61 PF
Hitachi HA8000
T2K Todai
140 TF
Oakforest-
PACS (Fujitsu)
25.0 PF
OFP-II
100+ PF
Fujitsu FX10
Oakleaf-FX
1.13 PF
Wisteria
BDEC-01
Fujitsu
33.1 PF
BDEC-
02
250+ PF
Reedbush-
U/H/L (SGI-HPE)
3.36 PF
Ipomoea-01 25PB
Ipomoea-
03
Ipomoea-02
Supercomputers
@ITC/U.Tokyo
2,600+ Users
55+% outside of U.Tokyo


2001-2005
2006-2010
2011-2015
2016-2020
2021-2025
2026-2030
Hitachi SR8000
SR8000
Hitachi SR11000 J1, J2
IBM Power5+
Hitachi SR16K/M1 Yayoi
IBM Power7
Intel CLX
OBCX (Fujitsu) 6.61 PF
Hitachi SR2201
HARP-1E
Hitachi SR8000/MPP
SR8000
Hitachi HA8000 T2K Todai 140 TF
AMD Opteron
Oakforest-PACS (Fujitsu)
Intel Xeon Phi
OFP-II 100+ PF
Accelerators
Fujitsu FX10 Oakleaf-FX 1.13 PF
SPACR64 IXfx
Wisteria BDEC-01 Fujitsu 33.1 PF
A64FX, Intel Icelake+ NVIDIA A100
BDEC-02 250+ PF
Accelerators
Reedbush-U/H/L (SGI-HPE) 3.36 PF
Intel BDW + NVIDIA P100
Pseudo Vector
Multicore CPU
GPU, Accelerators
Ipomoea-01 25PB
Ipomoea-03
Ipomoea-02



Simulation Nodes (Odyssey)

Data/Learning Nodes (Aquarius)

東京大学 THE UNIVERSITY OF TOKYO

東京大学情報基盤センター INFORMATION TECHNOLOGY CENTER, THE UNIVERSITY OF TOKYO

Reedbush (HPE, Intel BDW + NVIDIA P100 (Pascal))

- Prototype of "Wisteria/BDEC-01" for Integration of (S+D+L)
- July 2016 – November 2021 (Retired)
- Our First GPU Cluster, 3.36 PF

Oakforest-PACS (OFP) (Fujitsu, Intel Xeon Phi (KNL))

- JCAHPC (U.Tsukuba, U.Tokyo), October 2016 – March 2022 (retired)
- 25 PF, #39 in 58th TOP 500 (November 2021)

Oakbridge-CX (OBCX) (Fujitsu, Intel Xeon CLX)

- July 2019 – June 2023
- 6.61 PF, #110 in 58th TOP500

Hierarchical, Hybrid, Heterogeneous h3-Open-BDEC Big Data & Extreme Computing

Wisteria/BDEC-01 (Fujitsu)

- **Simulation Nodes (Oddysey): A64FX (#17)**
- **Data/Learning Nodes (Aquarius) (#106)**
- 33.1 PF, Operation started on May 14, 2021
- Platform for Integration of "Simulation+Data+Learning (S+D+L)"
- Innovative Software Platform "h3-Open-BDEC" supported by Janapese Government (JSPS Grant-in-Aid for Scientific Res. (S) FY.2019-2023)

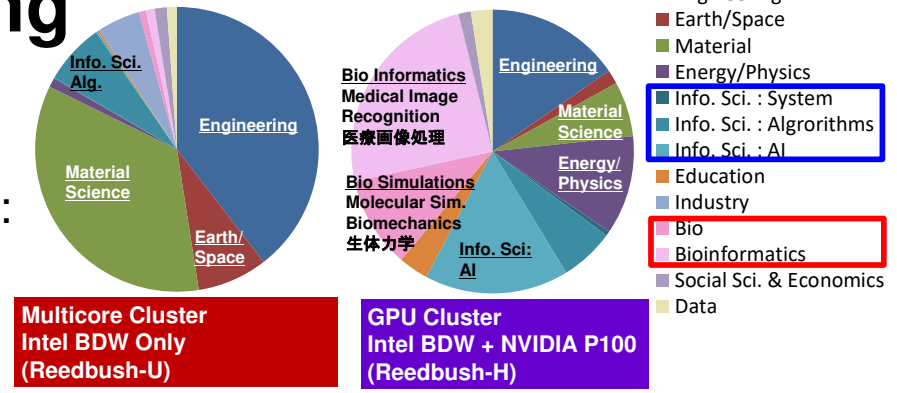
Reedbush

Oakforest-PACS

Oakbridge-CX

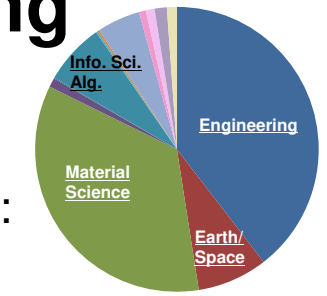
Future of Supercomputing

- Various Types of Workloads
  - Computational Science & Engineering: Simulations
  - Big Data Analytics
  - AI, Machine Learning …

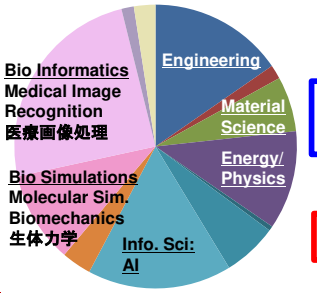


Future of Supercomputing

- Various Types of Workloads
  - Computational Science & Engineering: Simulations
  - Big Data Analytics
  - AI, Machine Learning …
- Integration/Convergence of (Simulation + Data + Learning) (S+D+L) is important towards Society 5.0
  - Super Smart & Human-centered Society by Digital Innovation (IoT, Big Data, AI etc.) and by Integration of Cyber Space & Physical Space



Multicore Cluster
Intel BDW Only
(Reedbush-U)

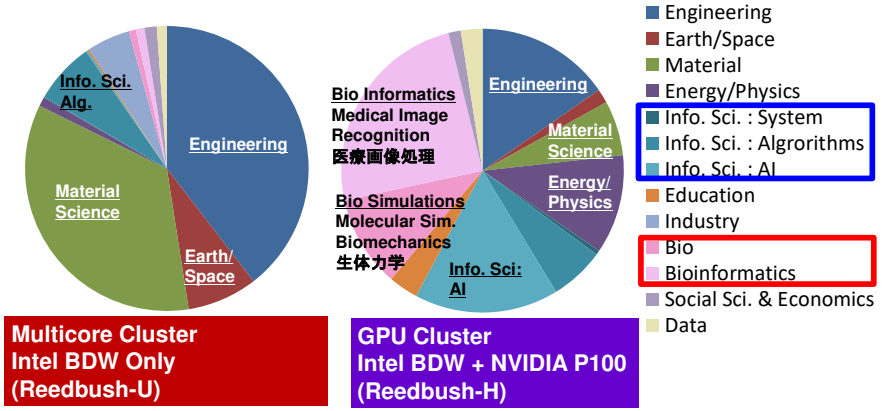


GPU Cluster
Intel BDW + NVIDIA P100
(Reedbush-H)

Engineering
Earth/Space
Material
Energy/Physics
Info. Sci. : System
Info. Sci. : Algrorithms
Info. Sci. : AI
Education
Industry
Bio
Bioinformatics
Social Sci. & Economics
Data

Future of Supercomputing

- Various Types of Workloads
  - Computational Science & Engineering: Simulations
  - Big Data Analytics
  - AI, Machine Learning …
- **Integration/Convergence of (Simulation + Data + Learning) (S+D+L) is important towards Society 5.0**
- **BDEC (Big Data & Extreme Computing)**
  - **Platform for Integration of (S+D+L)**
  - **Focusing on S (Simulation)**
    - **AI for HPC, Sophisticated Simulation**
  - **Planning started in 2015**



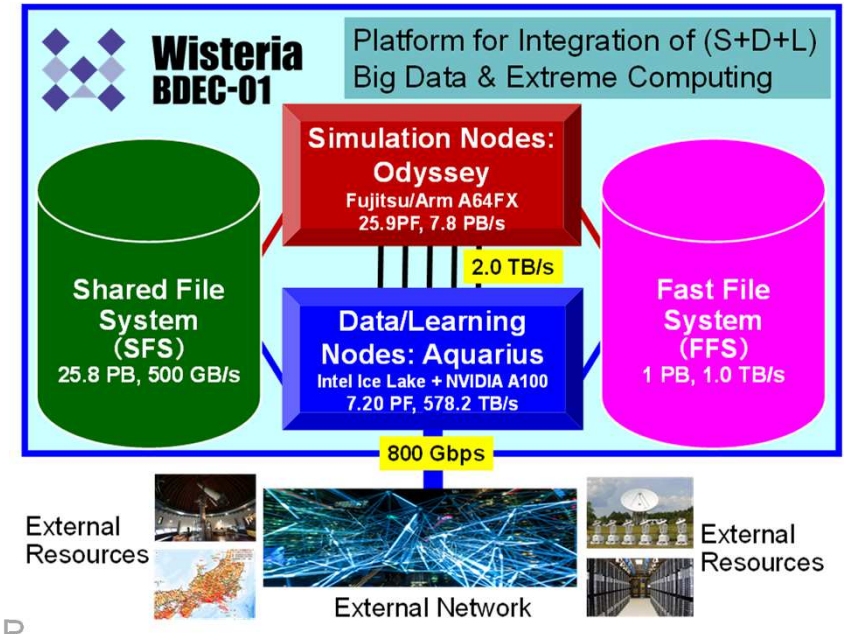
BDEC (Big Data & Extreme Computing)

S + D + L

Wisteria/BDEC-01

The 1st BDEC System (Big Data & Extreme Computing) Platform for Integration of (S+D+L)

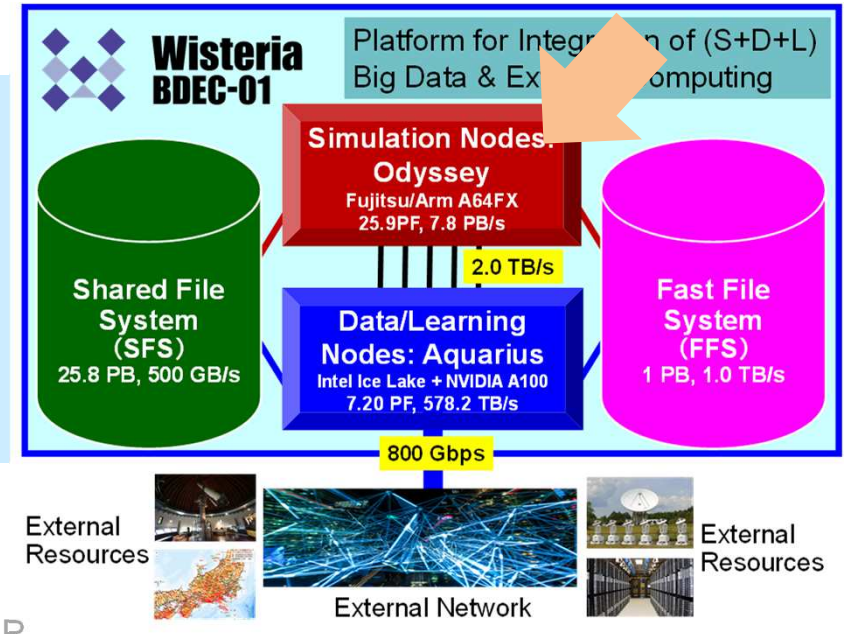
- Operation starts on May 14, 2021
- 33.1 PF, 8.38 PB/sec by **Fujitsu**
  - ~4.5 MVA with Cooling, ~360m^2
- 2 Types of Node Groups
  - Hierarchical, Hybrid, Heterogeneous (h3)
  - Simulation Nodes: Odyssey
    - Fujitsu PRIMEHPC FX1000 (A64FX), 25.9 PF
      - 7,680 nodes (368,640 cores), Tofu-D
      - General Purpose CPU + HBM
      - Commercial Version of "Fugaku"
  - Data/Learning Nodes: Aquarius
    - Data Analytics & AI/Machine Learning
    - Intel Xeon Ice Lake + NVIDIA A100, 7.2PF
      - 45 nodes (90x Ice Lake, 360x A100), IB-HDR
    - Some of the DL nodes are connected to external resources directly
- File Systems: SFS (Shared/Large) + FFS (Fast/Small)



Wisteria/BDEC-01

The 1st BDEC System (Big Data & Extreme Computing) Platform for Integration of (S+D+L)

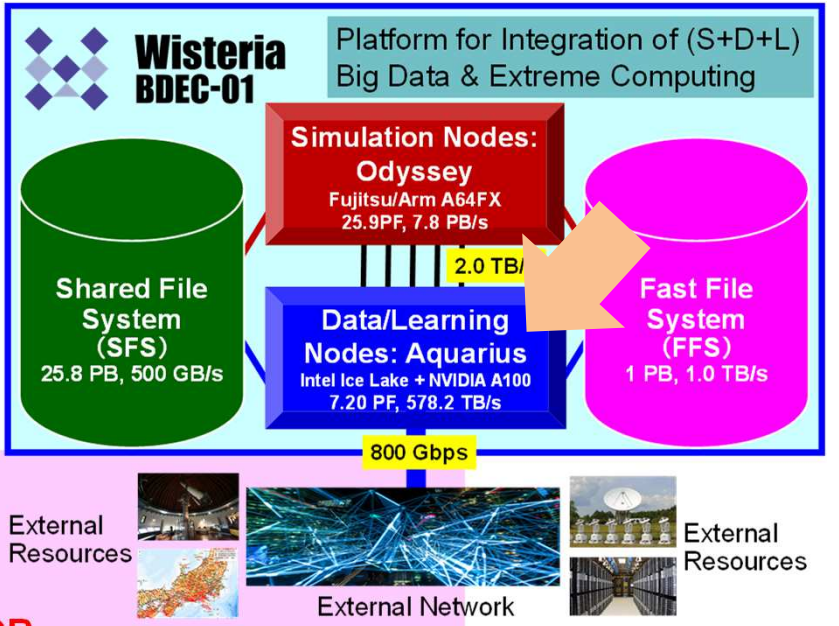
- Operation starts on May 14, 2021
- 33.1 PF, 8.38 PB/sec by **Fujitsu**
  - ~4.5 MVA with Cooling, ~360m^2
- **2 Types of Node Groups**
  - Hierarchical, Hybrid, Heterogeneous (h3)
  - **Simulation Nodes: Odyssey**
    - **Fujitsu PRIMEHPC FX1000 (A64FX), 25.9 PF**
      - **7,680 nodes (368,640 cores), Tofu-D**
      - **General Purpose CPU + HBM**
      - **Commercial Version of "Fugaku"**
  - Data/Learning Nodes: Aquarius
    - Data Analytics & AI/Machine Learning
    - Intel Xeon Ice Lake + NVIDIA A100, 7.2PF
      - 45 nodes (90x Ice Lake, 360x A100), IB-HDR
    - Some of the DL nodes are connected to external resources directly
- File Systems: SFS (Shared/Large) + FFS (Fast/Small)



Wisteria/BDEC-01

The 1st BDEC System (Big Data & Extreme Computing) Platform for Integration of (S+D+L)

- Operation starts on May 14, 2021
- 33.1 PF, 8.38 PB/sec by **Fujitsu**
  - ~4.5 MVA with Cooling, ~360m^2
- **2 Types of Node Groups**
  - Hierarchical, Hybrid, Heterogeneous (h3)
  - **Simulation Nodes: Odyssey**
    - **Fujitsu PRIMEHPC FX1000 (A64FX), 25.9 PF**
      - **7,680 nodes (368,640 cores), Tofu-D**
      - **General Purpose CPU + HBM**
      - **Commercial Version of "Fugaku"**
  - **Data/Learning Nodes: Aquarius**
    - **Data Analytics & AI/Machine Learning**
    - **Intel Xeon Ice Lake + NVIDIA A100, 7.2PF**
      - **45 nodes (90x Ice Lake, 360x A100), IB-HDR**
    - **Some of the DL nodes are connected to external resources directly**
- File Systems: SFS (Shared/Large) + FFS (Fast/Small)

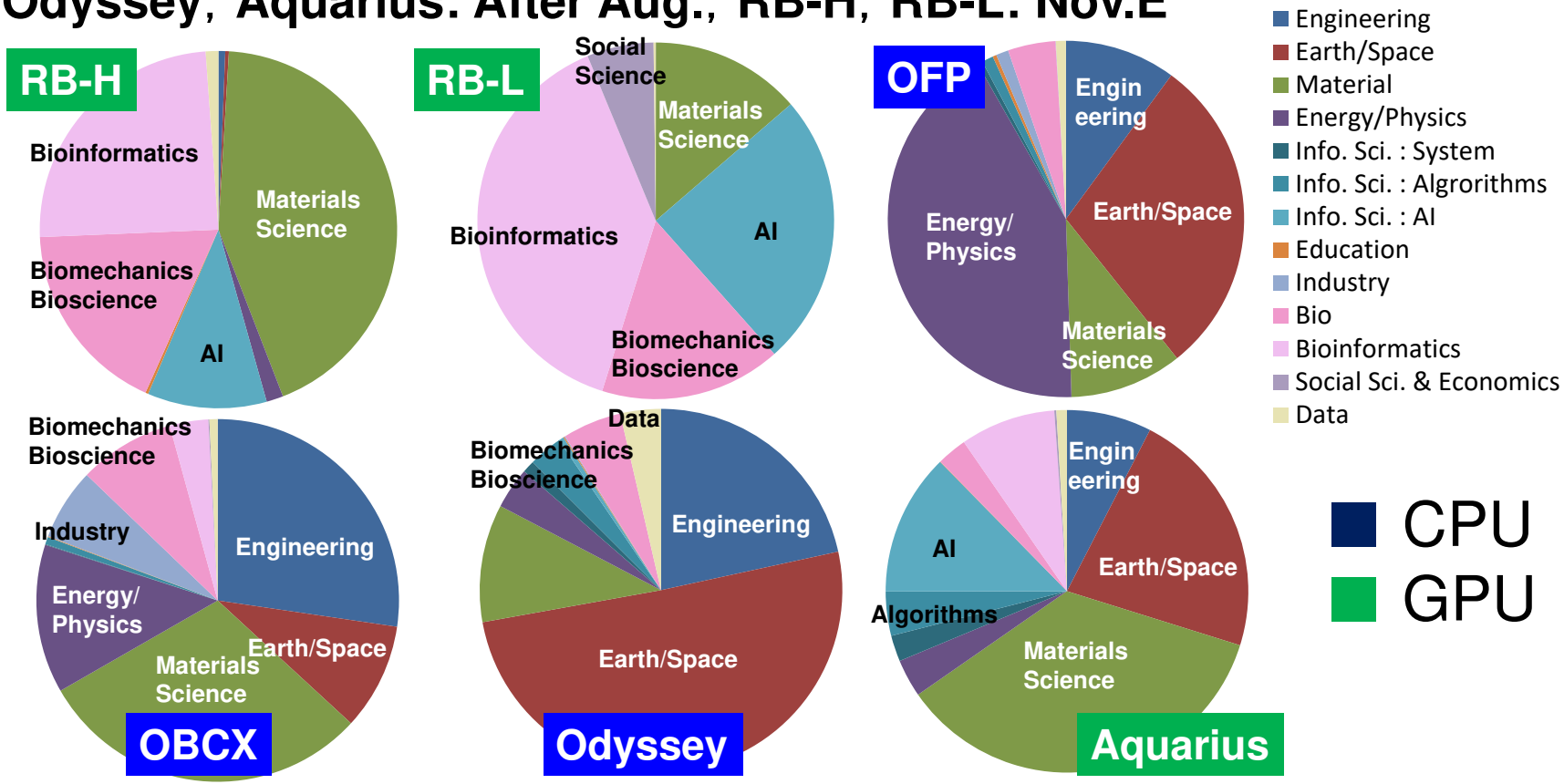


Rankings@SC21
November 2021

System	TOP500	Green500	HPCG	Graph500	HPL-AI
Oakforest-PACS	39	65	23	-	-
Oakbridge-CX	110	62	71	-	-
Wisteria/BDEC-01 (Odyssey)	**17**	**27**	**9**	**3**	**9**
Wisteria/BDEC-01 (Aquarius)	**106**	**15**	**58**	**-**	**-**

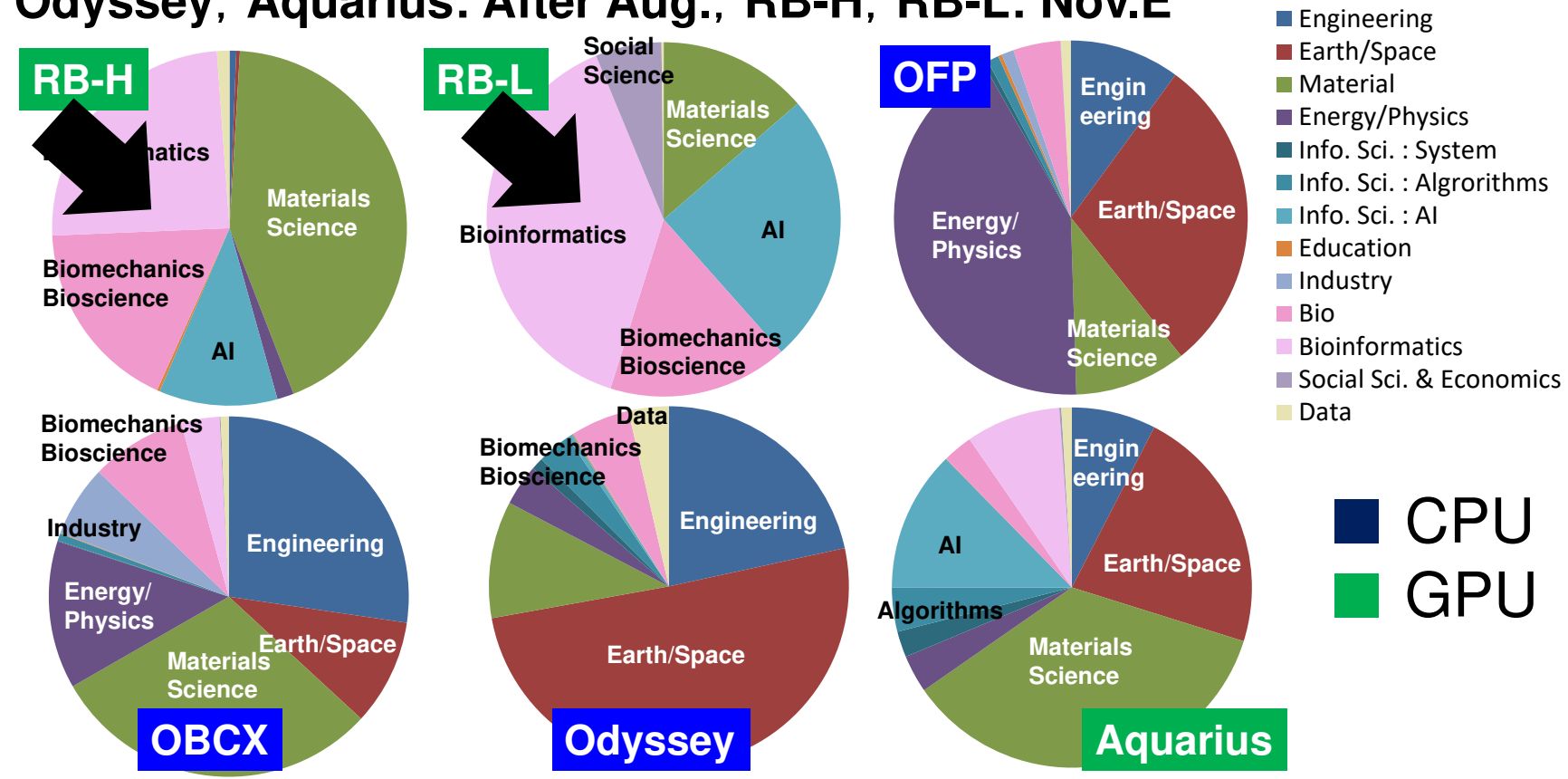
Research Area based on CPU-H's (FY.2021)

Odyssey, Aquarius: After Aug., RB-H, RB-L: Nov.E



Research Area based on CPU-H's (FY.2021)

Odyssey, Aquarius: After Aug., RB-H, RB-L: Nov.E



Simulation Nodes
Odyssey
25.9 PF, 7.8 PB/s

Fast File System (FFS)
1.0 PB, 1.0 TB/s

Shared File System (SFS)
25.8 PB, 0.50 TB/s

Data/Learning Nodes
Aquarius
7.20 PF, 578.2 TB/s

Simulation Codes

Optimized Models & Parameters

Simulation Nodes Odyssey

Results

Wisteria/BDEC-01

Machine Learning, DDA

Observation Data

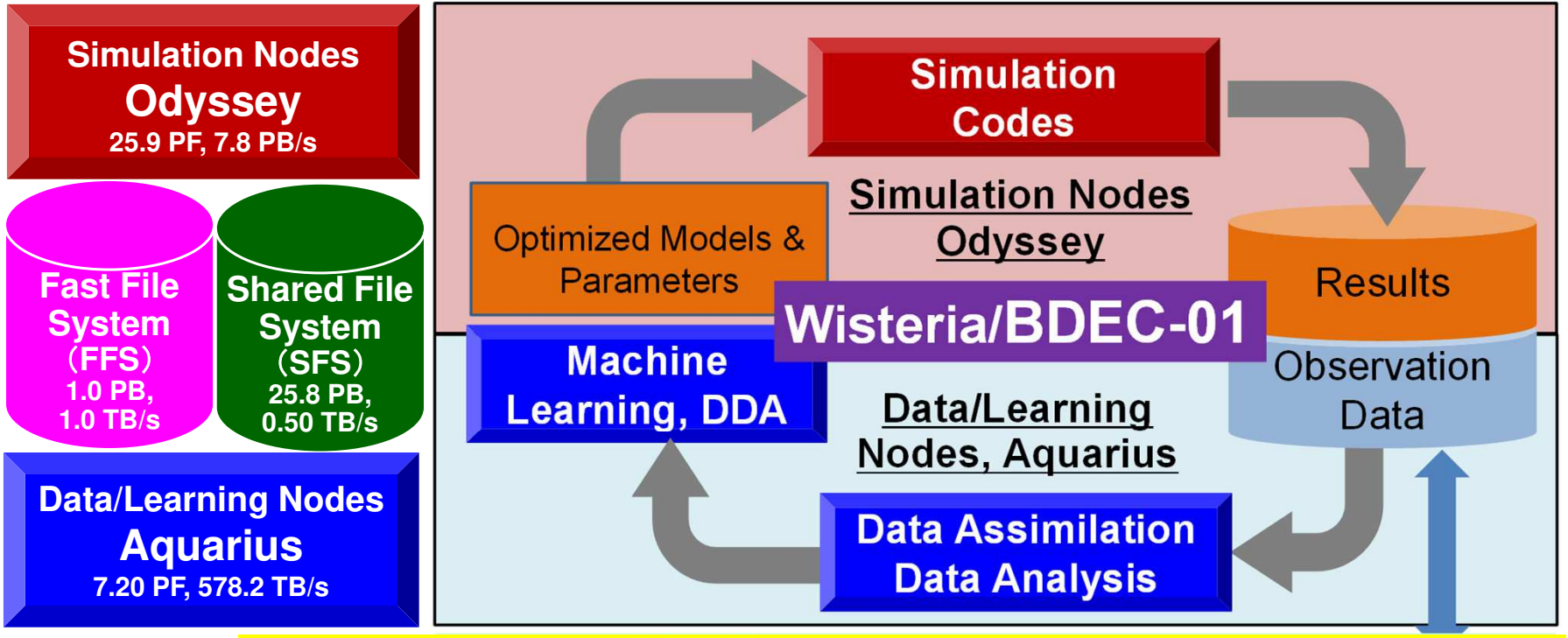
Data/Learning Nodes, Aquarius

Data Assimilation Data Analysis

Server, Storage, DB, Sensors, etc.

External Network

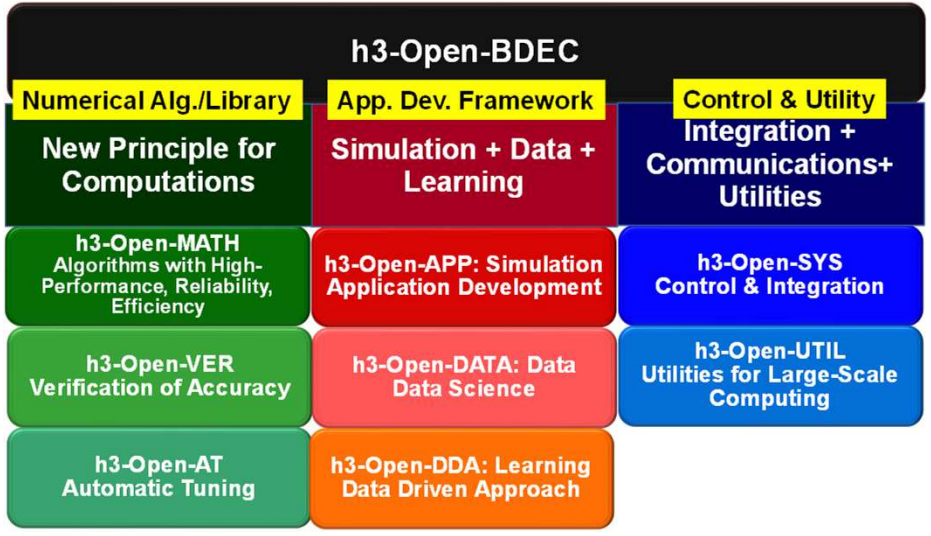
External Resources



Optimization of Models/Parameters for Simulations by Data Analytics & Machine Learning (S+D+L)

h3-Open-BDEC Innovative Software Platform for Integration of (S+D+L) on the BDEC System, such as Wisteria/BDEC-01

- 5-year project supported by Japanese Government (JSPS) since 2019
- Leading-PI: Kengo Nakajima (The University of Tokyo)
- Total Budget: 1.41M USD



Members (Co-PI's) of h3-Open-BDEC Project

Computer Science, Computational Science, Numerical Algorithms, Data Science, Machine Learning

- Kengo Nakajima (ITC/U.Tokyo, RIKEN), Leading-PI
- Takeshi Iwashita (Hokkaido U), Co-PI, Algorithms
- Hisashi Yashiro (NIES), Co-PI, Coupling, Utility
- Hiromichi Nagao (ERI/U.Tokyo), Co-PI, Data Assimilatic
- Takashi Shimokawabe (ITC/U.Tokyo), Co-PI, ML/hDDA
- Takeshi Ogita (TWCU), Co-PI, Accuracy Verification
- Takahiro Katagiri (Nagoya U), Co-PI, Appropriate Computing
- Hiroya Matsuba (ITC/U.Tokyo), Co-PI, Container

Hierarchical, Hybrid, Heterogeneous
h3-Open-BDEC
Big Data & Extreme Computing

東京大学
THE UNIVERSITY OF TOKYO

北海道大学
HOKKAIDO UNIVERSITY

国立研究開発法人
国立環境研究所
National Institute for Environmental Studies

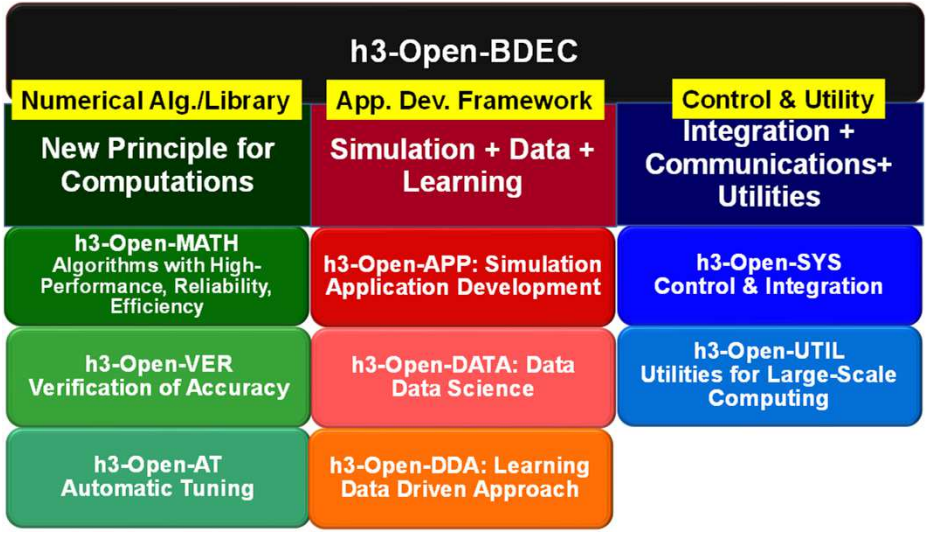
Tokyo Woman's Christian University
東京女子大学

名古屋大学
NAGOYA UNIVERSITY

RIKEN
R-CCS

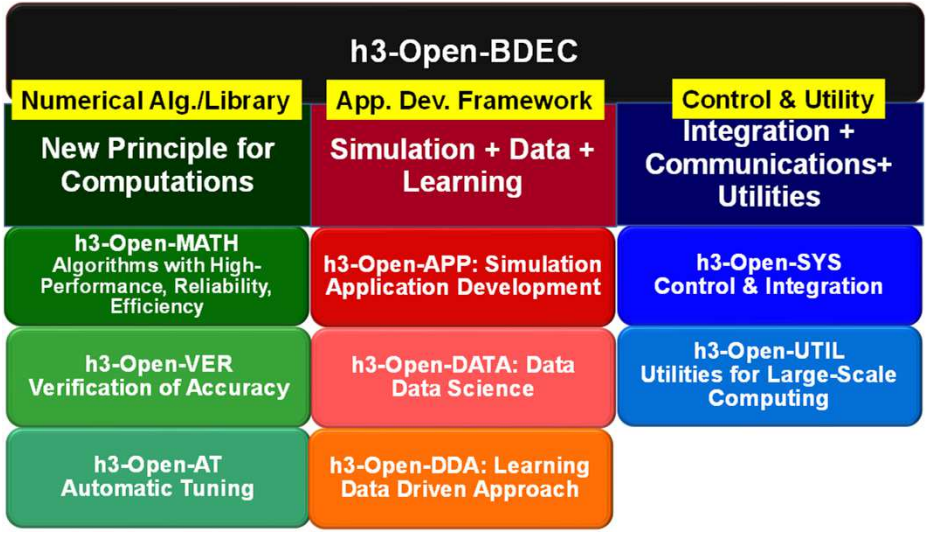
h3-Open-BDEC Innovative Software Platform for Integration of (S+D+L) on the BDEC System, such as Wisteria/BDEC-01

- "Three" Innovations
  - New Principles for Numerical Analysis by Adaptive Precision, Automatic Tuning & Accuracy Verification
  - Hierarchical Data Driven Approach (*h*DDA) based on Machine Learning
  - Software & Utilities for Heterogenous Environment, such as Wisteria/BDEC-01

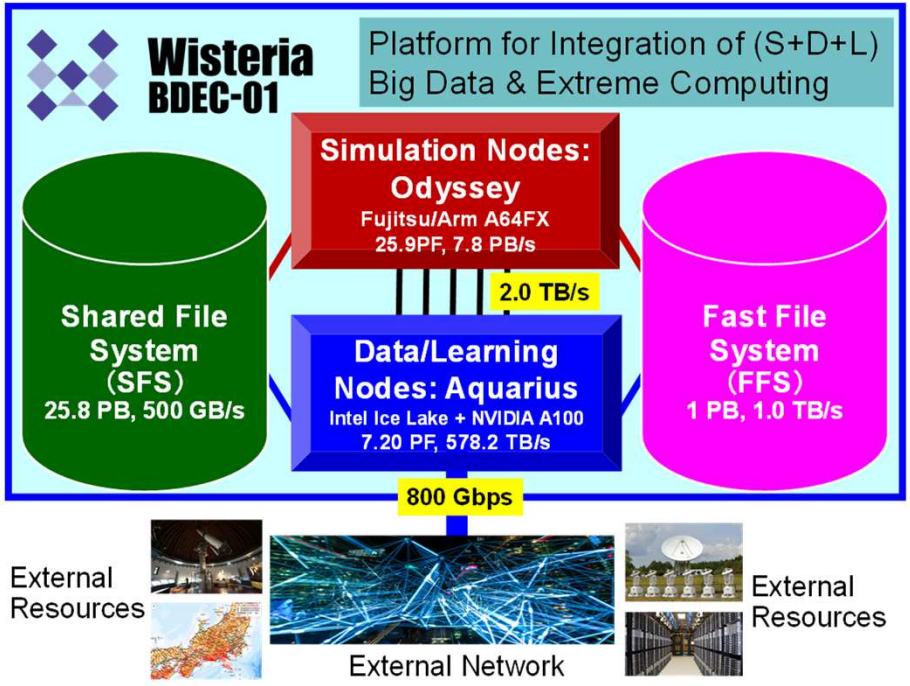


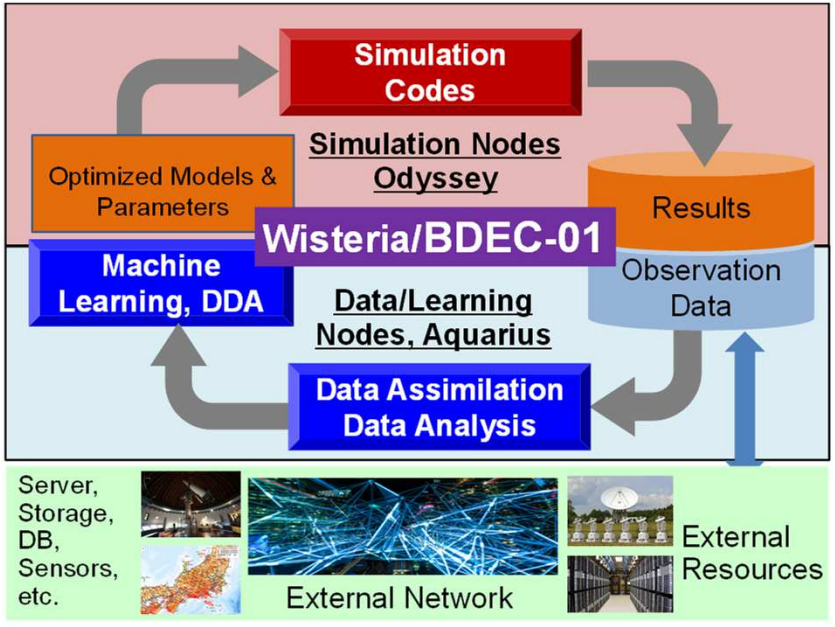
h3-Open-BDEC Innovative Software Platform for Integration of (S+D+L) on the BDEC System, such as Wisteria/BDEC-01

- "Three" Innovations
  - New Principles for Numerical Analysis by Adaptive Precision, Automatic Tuning & Accuracy Verification
  - Hierarchical Data Driven Approach (*h*DDA) based on Machine Learning
  - Software & Utilities for Heterogenous Environment, such as Wisteria/BDEC-01



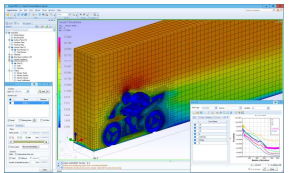
Wisteria/BDEC-01: The First "Really Heterogenous" System in the World

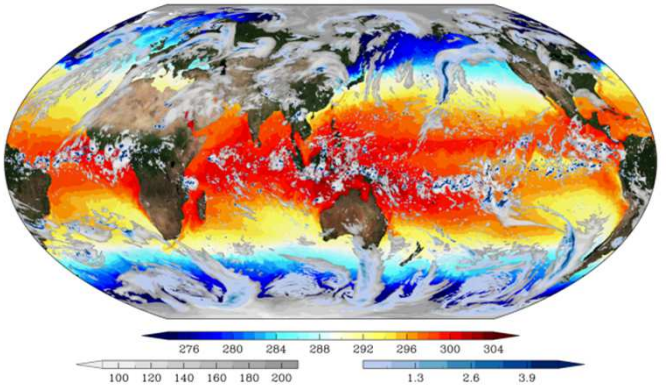


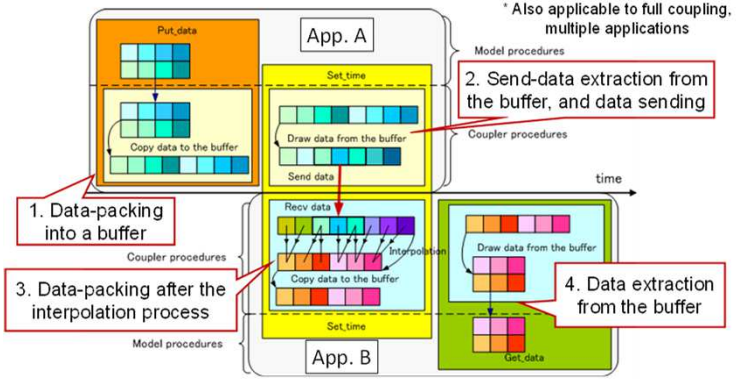


Possible Applications (S+D+L) on Wisteria/BDEC-01 with h3-Open-BDEC

- Simulations with Data Assimilation
  - Very Typical Example of (S+D+L)
- Atmosphere-Ocean Coupling for Weather and Climate Simulations
  - AORI/U.Tokyo, RIKEN R-CCS, NIES
- **Earthquake Simulations with Real-Time Data Assimilation**
  - **ERI/U. Tokyo**
- Real-Time Disaster Simulations
  - Flood, Tsunami
- (S+D+L) for Existing Simulation Codes (Open Source Software)
  - OpenFOAM

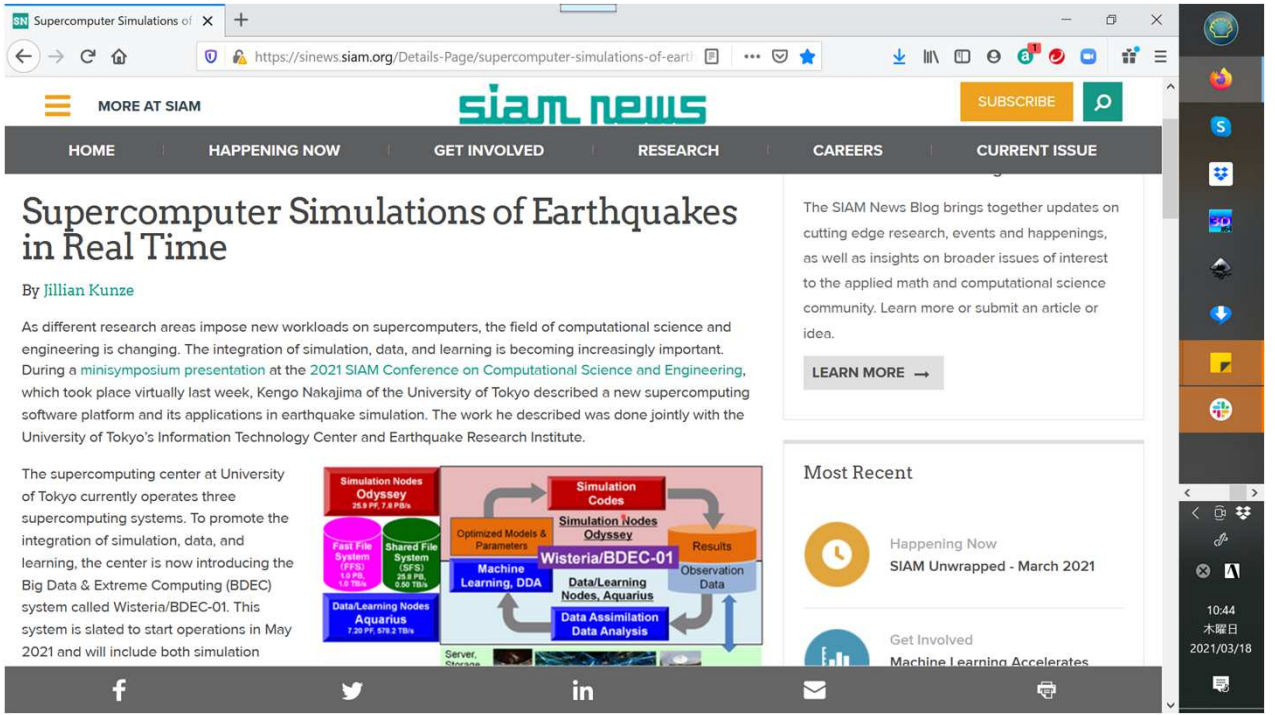






The article on my related presentation@SIAM CSE21 appeared in *SIAM News*

https://sinews.siam.org/Details-Page/supercomputer-simulations-of-earthquakes-in-real-time

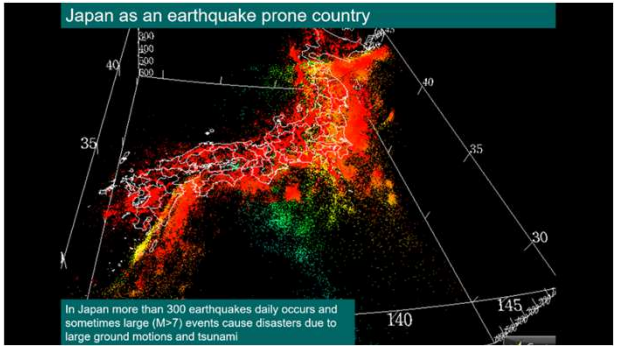


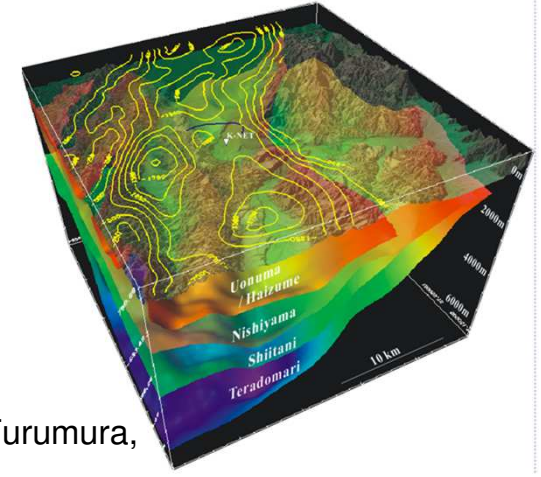
Early Forecast of Long-Period Ground Motions via Data Assimilation of Observation and Simulations [Furumura et al. 2019]

- New method for the early forecast of long-period (> 3–10 s) ground motions generated by large earthquakes based on the data assimilation of observed ground motions and FDM simulations of seismic wave propagation in a 3-D heterogeneous structure (**Seism3D/OpenSWPC-DAF**(**Data-Assimilation-Based Forecast**)).
- **This approach uses the dense nationwide network in Japan and supercomputers to perform forecasts using the assimilated wavefields at speeds much faster than the actual wave propagation speed.**
- **An early alert can be issued prior to the occurrence of strong motions due to large, distant earthquakes.**
- Validation of the effectiveness of this data-assimilation-based forecast approach via numerical tests for the early forecast of long-period ground motions in central Tokyo using the observed waveform data from the Mw6.6 2007 Off Niigata and Mw9.0 2011 Off Tohoku earthquakes.

Earthquake simulation is always with uncertainty

- Subsurface/Underground Structure
  - Heterogenous, Random, Stochastic
  - Fluctuations
- **Integration of Simulation/Observation is essential**
- Traditional Simulations
  - Forward Simulations
- **New Types of Methods for Simulations combined with Data Assimilation/Real-Time Observation is under development**
  - **Forecast by Simulations, Correction by Data Assimilation**



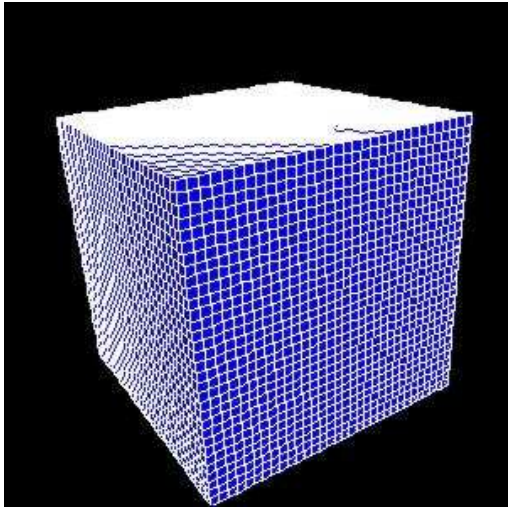


[c/o Prof. T. Furumura, ERI/U.Tokyo]

Simulations of Long-Period Ground Motion [Furumura et al.]

- 3D Equation of Motions solved by FDM (Finite-Difference Method)

$$v_p^n = v_p^{n-1} + \frac{1}{\rho}\left(\frac{\partial \sigma_{xp}^{n-1/2}}{\partial x} + \frac{\partial \sigma_{yp}^{n-1/2}}{\partial y} + \frac{\partial \sigma_{zp}^{n-1/2}}{\partial z} \right) \Delta t \quad (p = x, y, z)$$

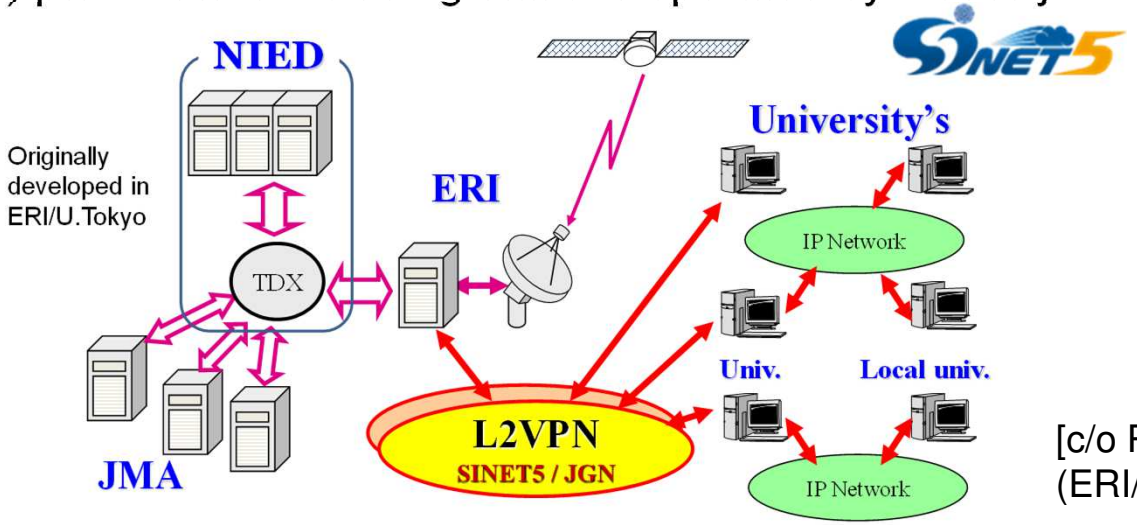


- Seism3D
  - Staggered Discretization in Space/Time
  - 4th order in Space
  - 2nd order in Time (Explicit Time Marching)
  - OpenMP + MPI, Fortran

Real-Time Sharing of Seismic Observation is possible in Japan by JDXnet with SINET

Japan Data eXchange network

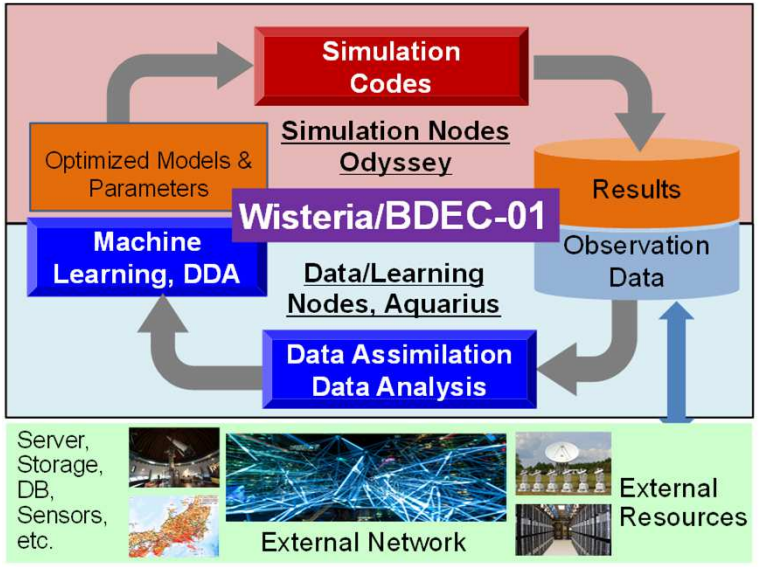
- Seismic Observation Data (100Hz/3-dir's/O(10^3) observation points) by JDXnet is available through SINET in Real Time
  - O(10^2) GB/day: available at Website of NIED
  - O(10^5) pts in future including stations operated by industry

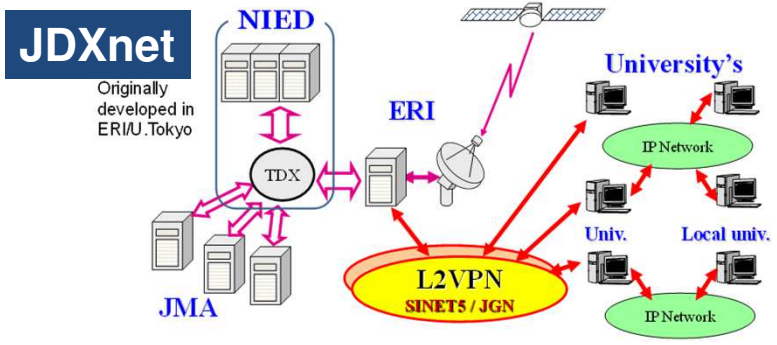


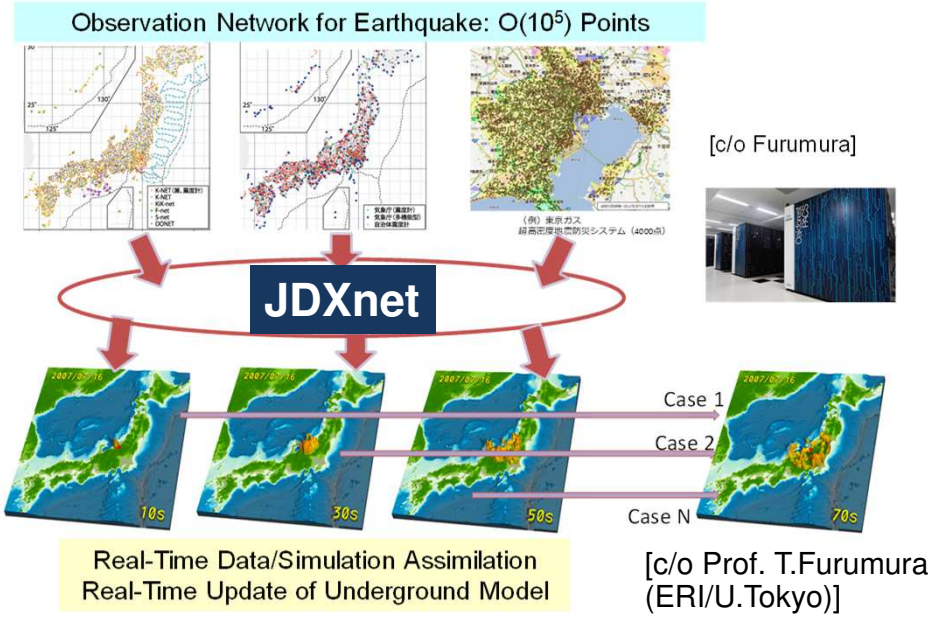
[c/o Prof. H.Tsuruoka (ERI/U.Tokyo)]

3D Earthquake Simulation with Real-Time Data Observation/Assimilation

Simulation of Strong Motion (Wave Propagation) by 3D FDM







[c/o Prof. T.Furumura (ERI/U.Tokyo)]

Real-Time Assimilation of "Observation+Computation" in Seismic Wave Propagation [c/o Oba & Furumura]

- **Data Assimilation of Wave Propagation by "Optimal Interpolation Technique"**

Assim. Comp. Residual Obs. Comp.

$$x_n^a = x_n^f + W(y_n - Hx_n^f)$$

Comp. Assim.

$$x_{n+1}^f = Fx_n^a$$

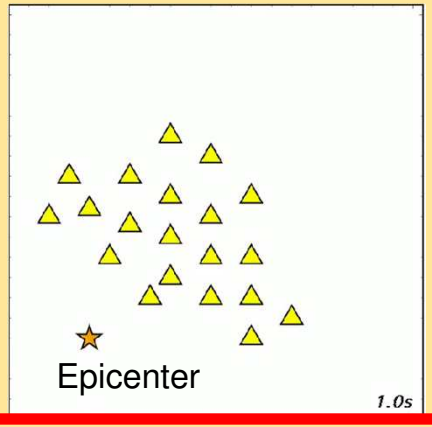
n: Time Step

W: **Weighting Matrix**

F: Wave Propagation simulation

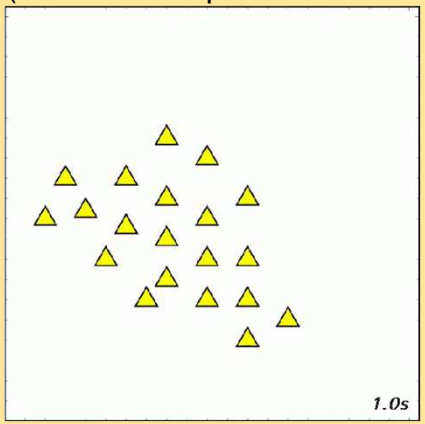
(A) Pure Simulation

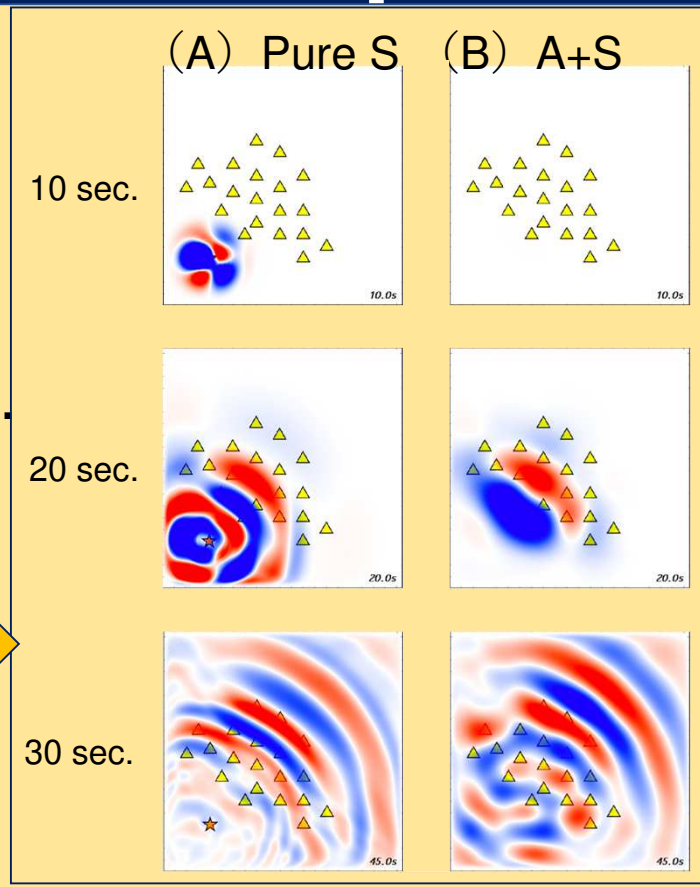
△: Obs. Pts.



(B) Assimilation+Sim.

(No info for Epicenter needed)





Real-Time Assimilation of "Observation+Computation" in Seismic Wave Propagation [c/o Oba & Furumura]

- **Data Assimilation of Wave Propagation by "Optimal Interpolation Technique"**

Residual

Assim. Comp. Obs. Comp.

$$x_n^a = x_n^f + W(y_n - Hx_n^f)$$

$$x_{n+1}^f = Fx_n^a$$

Comp. Assim.

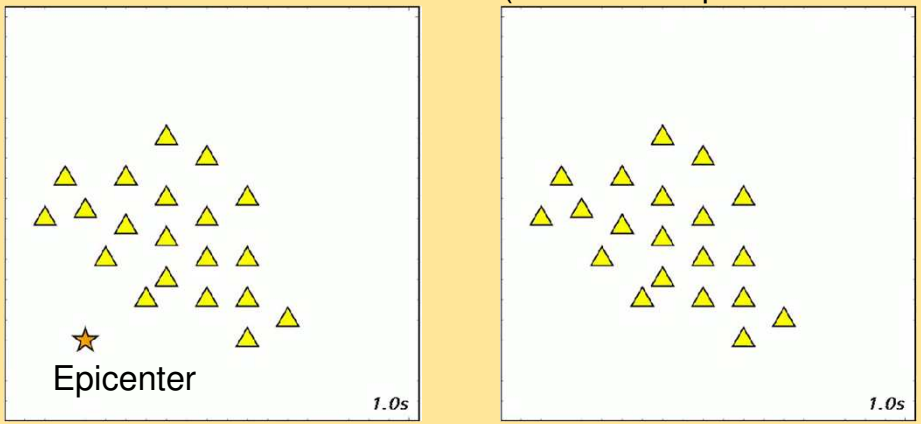
n: Time Step

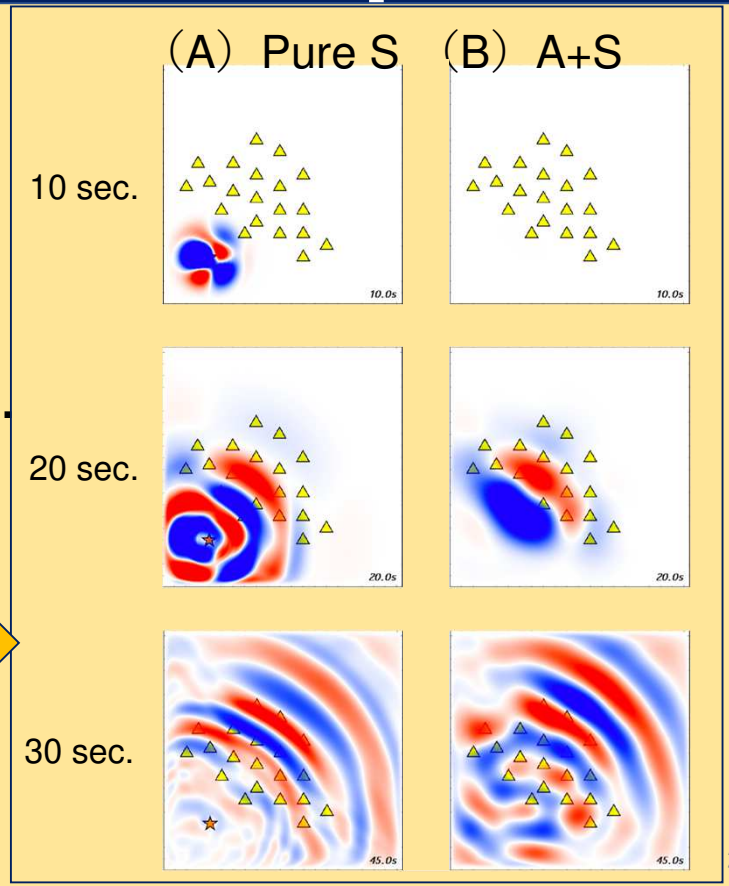
W: **Weighting Matrix**

F: Wave Propagation simulation

(A) Pure Simulation **(B) Assimilation+Sim.**

△: Obs. Pts. (No info for Epicenter needed)





Starting from (A+S: Assim+Sim.) to (Pure S: Pure Simulation)

Residual

Assim. Comp. Obs. Comp.

$$x_n^a = x_n^f + W(y_n - Hx_n^f)$$

$$x_{n+1}^f = Fx_n^a$$

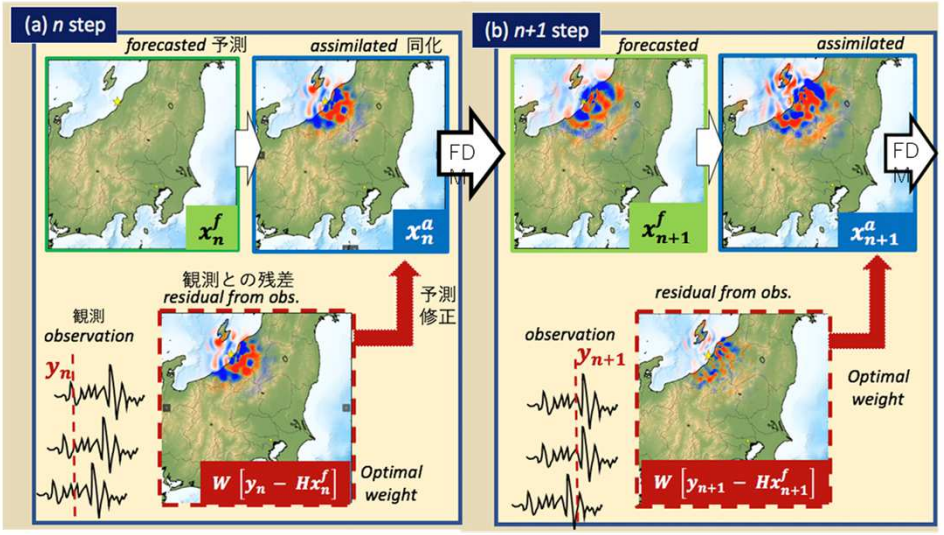
Comp. Assim.

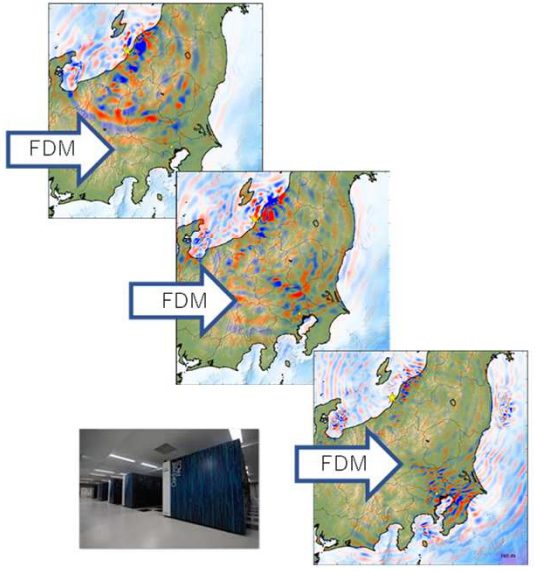
n: Time Step

W: **Weighting Matrix**

F: Wave Propagation simulation

(A+S) Assimilation+Simulation → **(Pure S) Pure Simulation/Forecast**

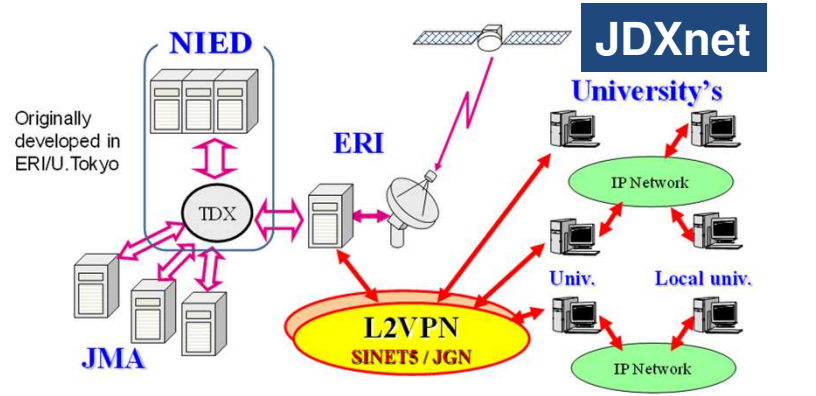




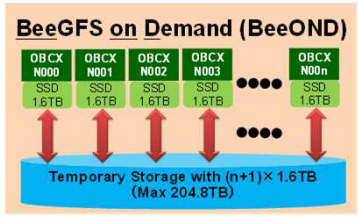
[c/o Prof. T. Furumura, ERI/U.Tokyo]

Preliminary Works on Oakbridge-CX (OBCX)

- Intel Xeon Platinum 8280 (Cascade Lake, CLX), Fujitsu
  - 1,368 nodes, 6.61 PF peak, 385.1 TB/sec, 4.2+ PF for HPL **#110 in 58th Top500 (Nov.2021)**
  - Fast Cache: SSD's for 128 nodes: Intel SSD, BeeGFS: 200+TB Fast FS
    - 1.6 TB/node, 3.20/1.32 GB/s/node for R/W
    - 16 of these nodes can directly access external resources (server, storage, sensor network etc.) through SINET
- Switching to Wisteria/BDEC-01 after May 2021



Oakbridge-CX (OBCX)
Total: 1,368 nodes
Intel Xeon Platinum 8280 (Cascade Lake, CLX)
128 nodes with SSD 16

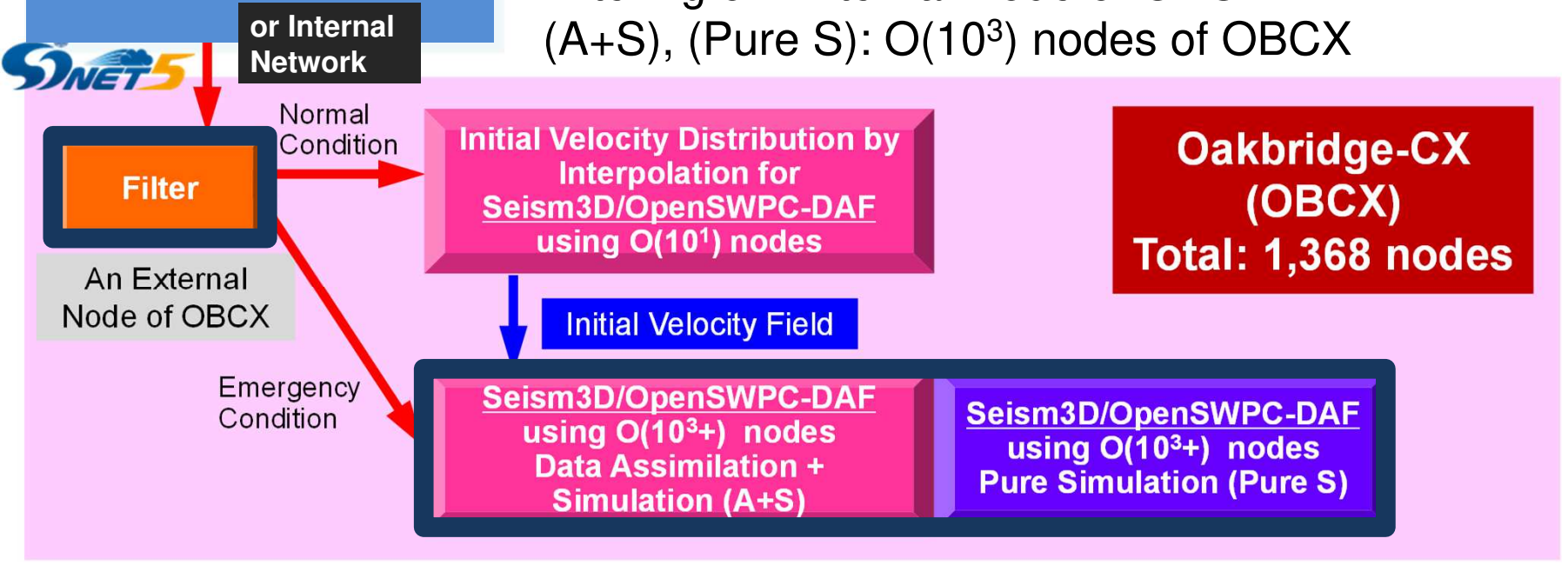


16 of 128 nodes with SSD can access external resources directly through SINET (**External Nodes**)

Emergency Operations + Data Assimilation & Forecast

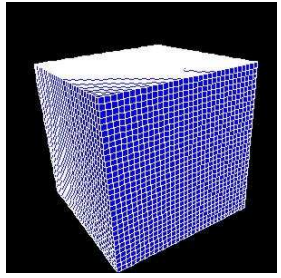
Experimental Environment

Filtering on External Node of OBCX
(A+S), (Pure S): O(10^3) nodes of OBCX



Example: Off Niigata 2007 Mw6.6 Earthquake

- Observed Data: Stored in External Server (Mini-mdx)
- An external node of OBCX receives observed data, and apply filtering
- “Data Assimilation + Simulation (A+S)”, and “Forecast by Simulation (Pure S)” are separated codes, while same number of computing nodes were used
- Movies were created after simulations (O(10) sec.)



- Seism3D/OpenSWPC-DAF
  - 3D FDM + Optimal Interpolation Technique for Data Assimilation
  - Each Mesh: 240m × 240m × 240m
  - 1,920 × 1,920 × 240 meshes (8.85×10^8)
  - 460.8 km × 460.8 km × 57.6 km

$$v_p^n = v_p^{n-1} + \frac{1}{\rho}\left(\frac{\partial \sigma_{xp}^{n-1/2}}{\partial x} + \frac{\partial \sigma_{yp}^{n-1/2}}{\partial y} + \frac{\partial \sigma_{zp}^{n-1/2}}{\partial z}\right)\Delta t \quad (p = x, y, z)$$

n: Time Step

Residual

Assim. Comp. Obs. Comp.

$$\boldsymbol{x}_n^a = \boldsymbol{x}_n^f + \boldsymbol{W}(\boldsymbol{y}_n - \boldsymbol{H}\boldsymbol{x}_n^f)$$

$\boldsymbol{W}$: Weighting Matrix

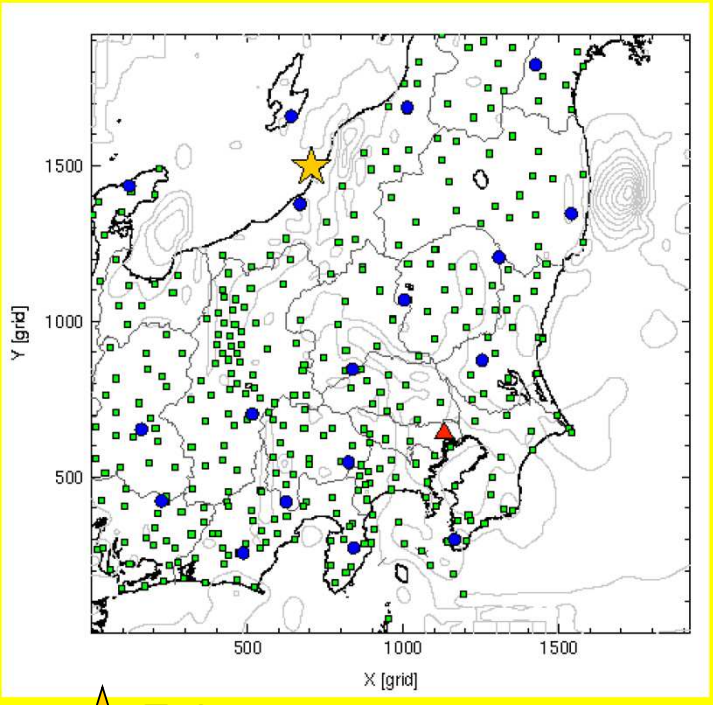
Comp. Assim.

$$\boldsymbol{x}_{n+1}^f = \boldsymbol{F}\boldsymbol{x}_n^a$$

$\boldsymbol{F}$: Wave Propagation simulation

Off Niigata 2007 Mw6.6 Earthquake

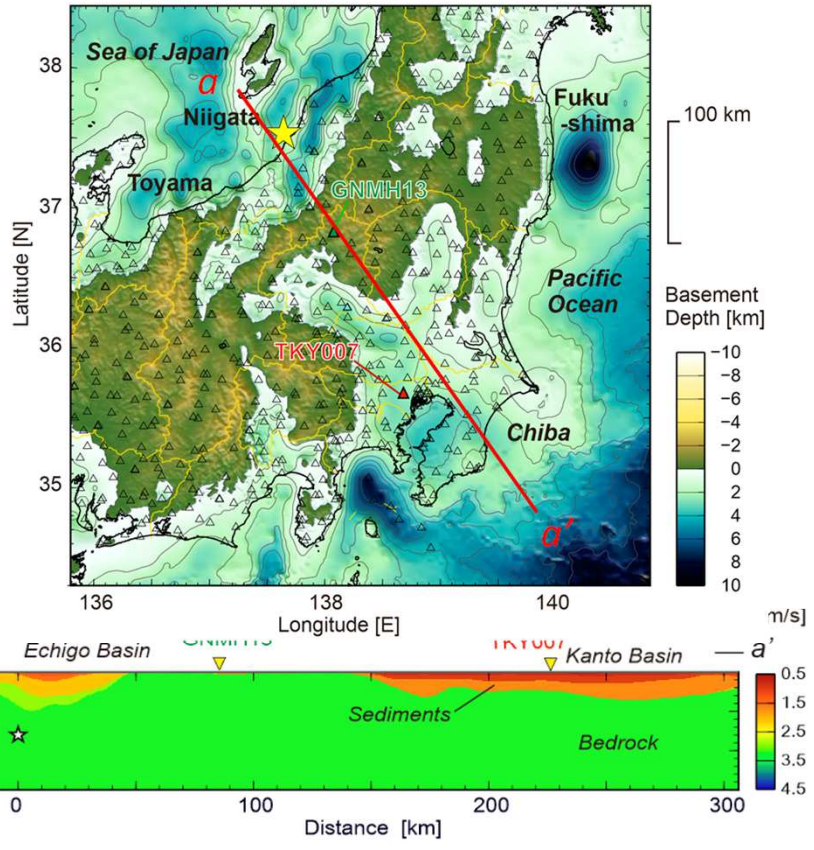
[c/o Prof. T. Furumura, ERI/U.Tokyo]



Epicenter

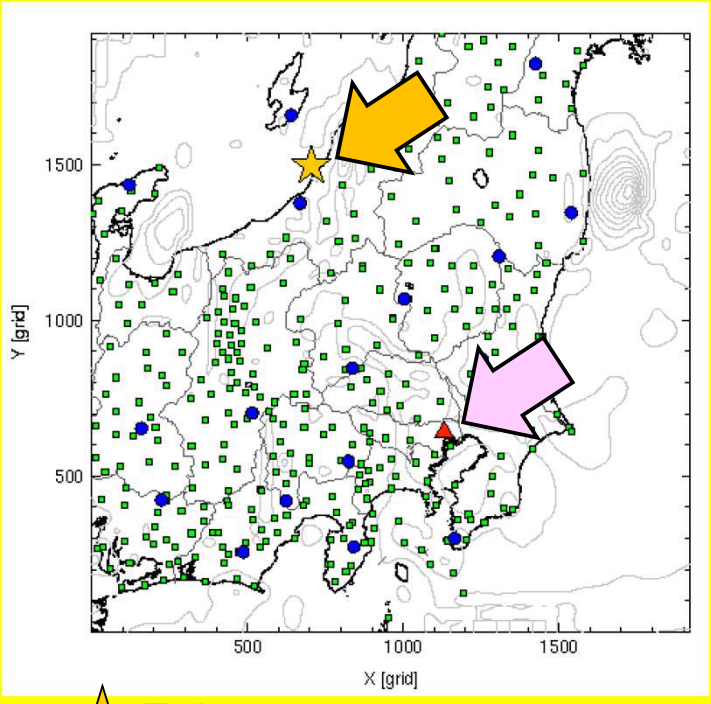
Hi-net (Short Period) 349 pts

F-net (Broadband) 18 pts



Off Niigata 2007 Mw6.6 Earthquake

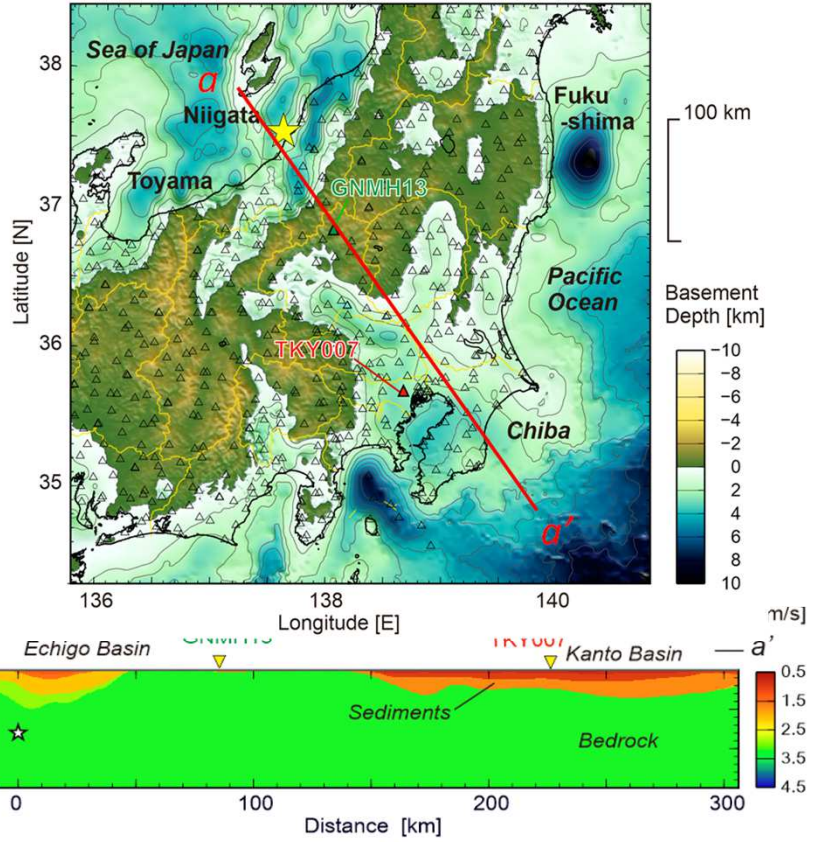
[c/o Prof. T. Furumura, ERI/U.Tokyo]



☆ Epicenter

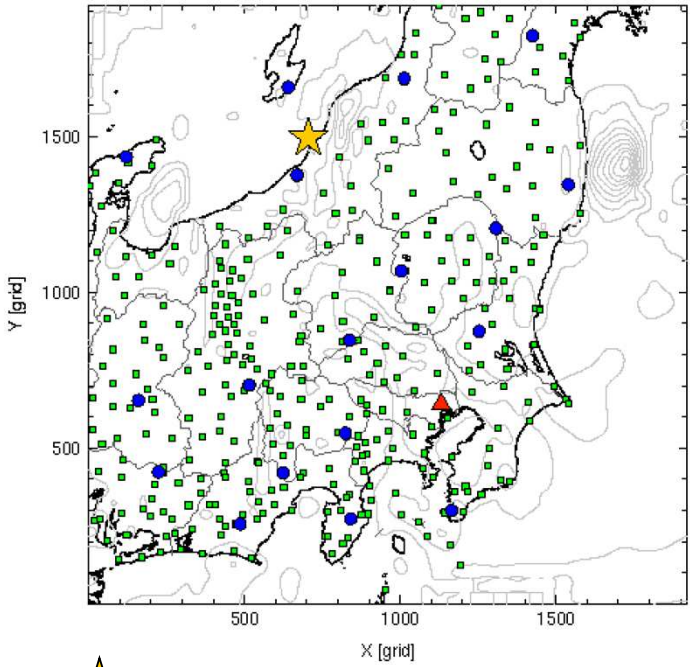
■ Hi-net (Short Period) 349 pts

● F-net (Broadband) 18 pts



Off Niigata 2007 Mw6.6 Earthquake

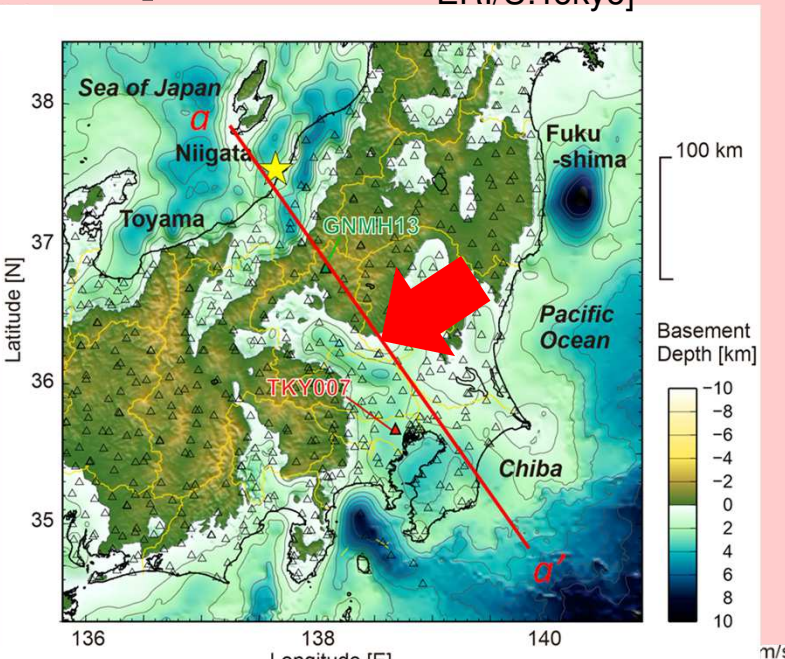
[c/o Prof. T. Furumura, ERI/U.Tokyo]

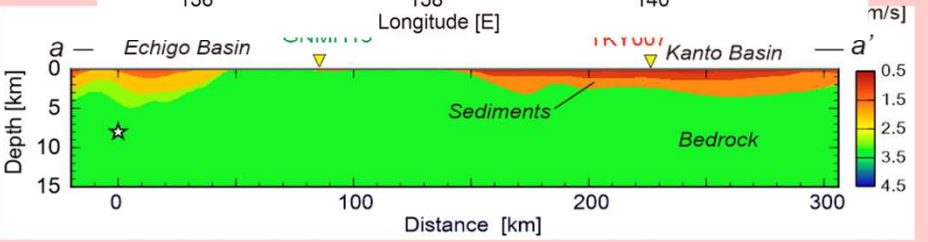


Epicenter

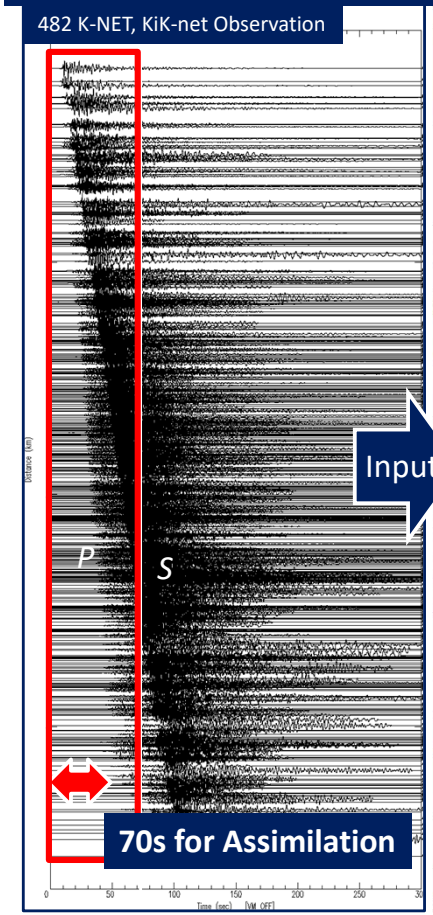
Hi-net (Short Period) 349 pts

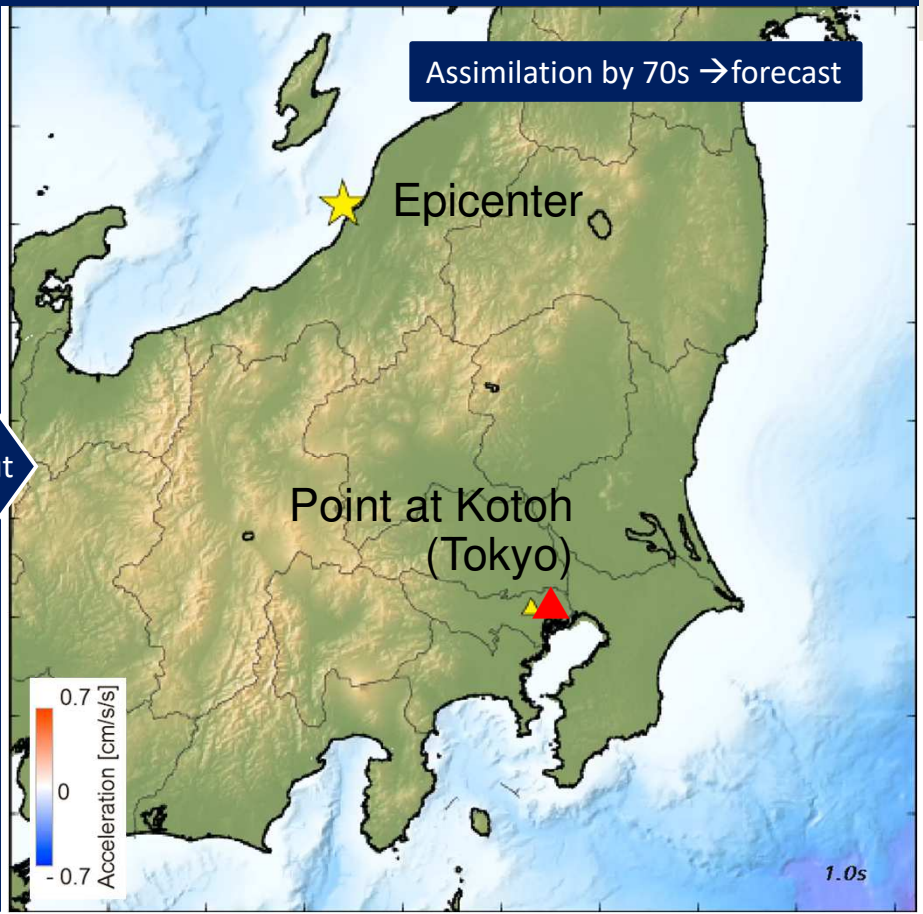
F-net (Broadband) 18 pts



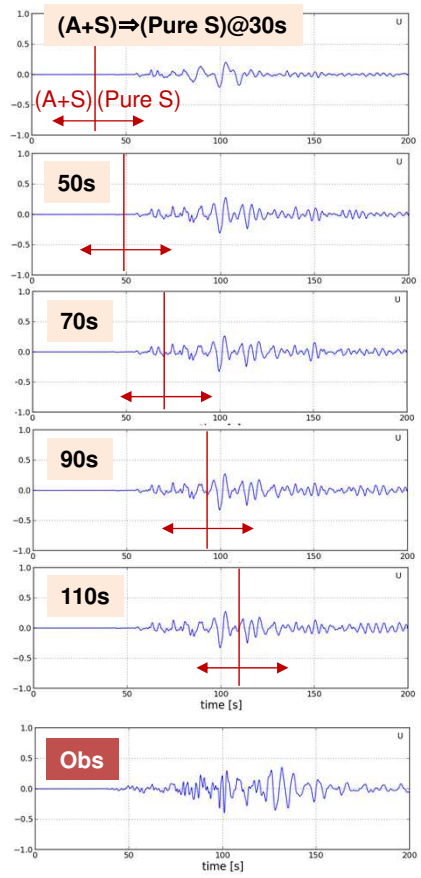


Data Assimilation + Pure Simulation/Forecast





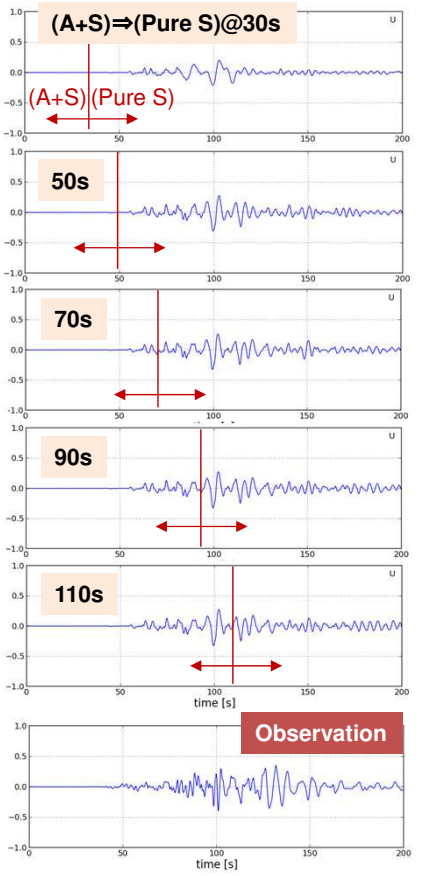
Results at Kotoh ▲ (N.KOTH)
N 35° 37.0'
E 139° 46.9'



Results: Off Niigata 2007 Mw6.6 Earthquake）

- (A+S)
  - Data assimilation is done using real-time observations, therefore this procedure cannot go ahead of real-time
  - Considering the overhead by preprocessing such as filtering, it is good to be able to calculate in about half the time of the actual phenomenon
- (Pure S)
  - 1/10 time of the actual phenomenon is required
  - Switching at 50 sec. from (A+s) to (Pure S)
  - If the subsequent 50 sec. can be computed in 5 sec., it is possible to predict the time when the peak wave will arrive in Tokyo, which is about 250km away from the epicenter (approx. 100 sec. after the occurrence of the earthquake)

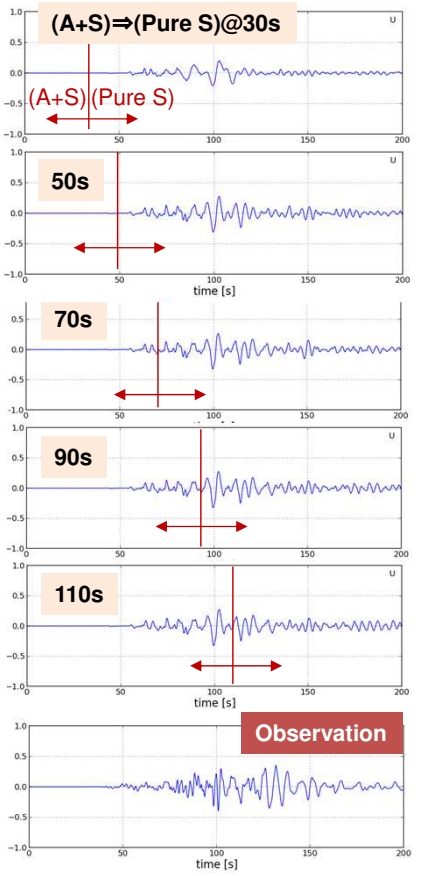
Koto, Tokyo ▲ (N.KOTH)
N 35° 37.0'
E 139° 46.9'



Results: Off Niigata 2007 Mw6.6 Earthquake）

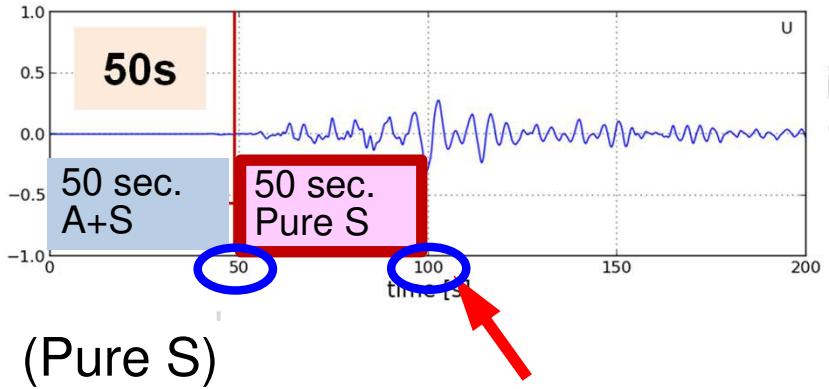
- (A+S)
  - Data assimilation is done using real-time observations, therefore this procedure cannot go ahead of real-time
  - Considering the overhead by preprocessing such as filtering, it is good to be able to calculate in about half the time of the actual phenomenon
- (Pure S)
  - 1/10 time of the actual phenomenon is required
  - Switching at 50 sec. from (A+s) to (Pure S)
  - If the subsequent 50 sec. can be computed in 5 sec., it is possible to predict the time when the peak wave will arrive in Tokyo, which is about 250km away from the epicenter (approx. 100 sec. after the occurrence of the earthquake)

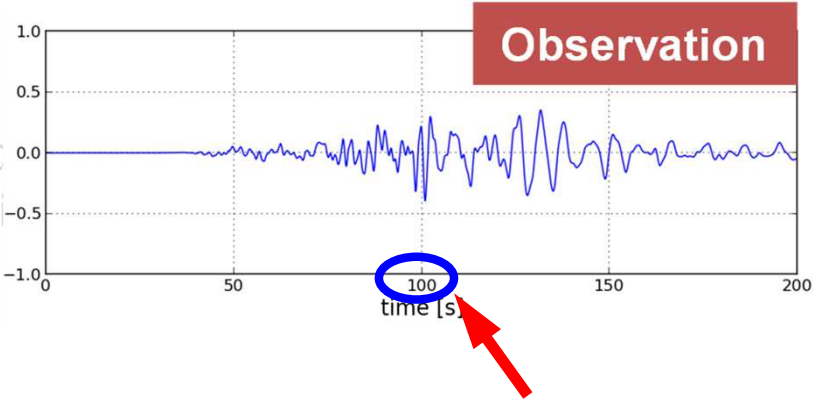
Koto, Tokyo ▲ (N.KOTH)
N 35° 37.0'
E 139° 46.9'



Results: Off Niigata 2007 Mw6.6 Earthquake)

Koto, Tokyo ▲ (N.KOTH)
N 35° 37.0'
E 139° 46.9'

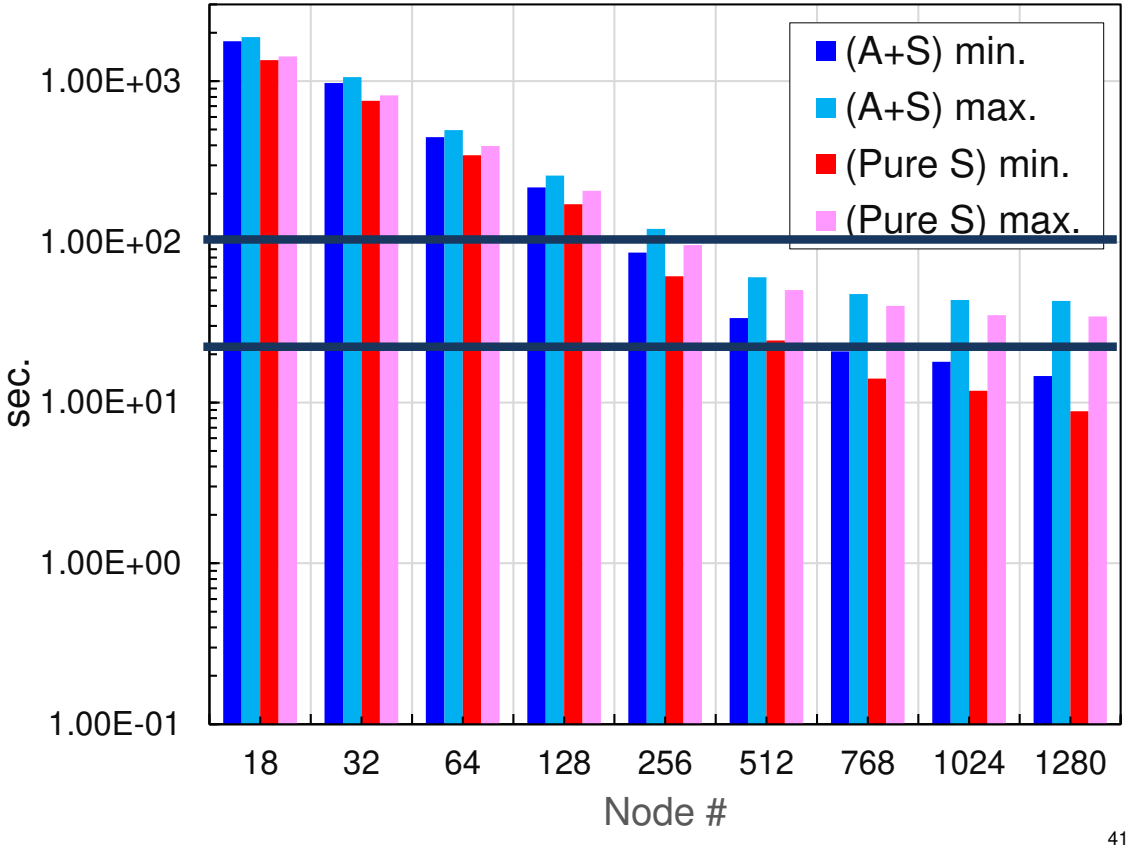




- (Pure S)
  - 1/10 time of the actual phenomenon is required
  - Switching at 50 sec. from (A+s) to (Pure S)
  - If the subsequent 50 sec. can be computed in 5 sec., it is possible to predict the time when the peak wave will arrive in Tokyo, which is about 250km away from the epicenter (approx. 100 sec. after the occurrence of the earthquake)

Computation Time for 200 sec. Phenomenon

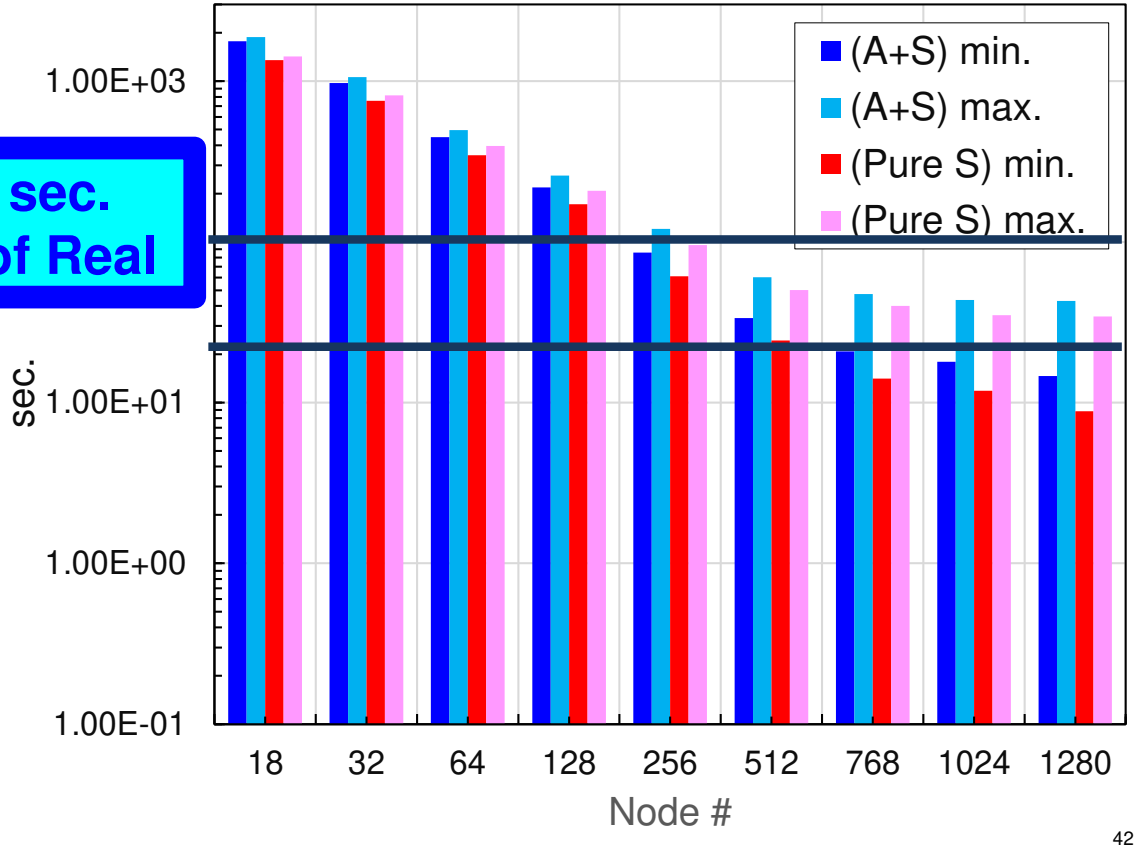
- **Communications for I/O are included**
  - **min.: Comm. excluded**
  - **max.: Comm. Included**
- (A+S)
  - Computation in 100 sec. (Half of 200 sec.)
  - 300-400 nodes
- (Pure S)
  - Computation in 20 sec. (1/10 of 200 sec.)
  - 1,000+ nodes



Computation Time for 200 sec. Phenomenon

- Communications for I/O are included
  - min.: Comm. excl
  - max.: Comm. Incl
- (A+S)
  - Computation in 100 sec. (Half of 200 sec.)
  - 300-400 nodes
- (Pure S)
  - Computation in 20 sec. (1/10 of 200 sec.)
  - 1,000+ nodes

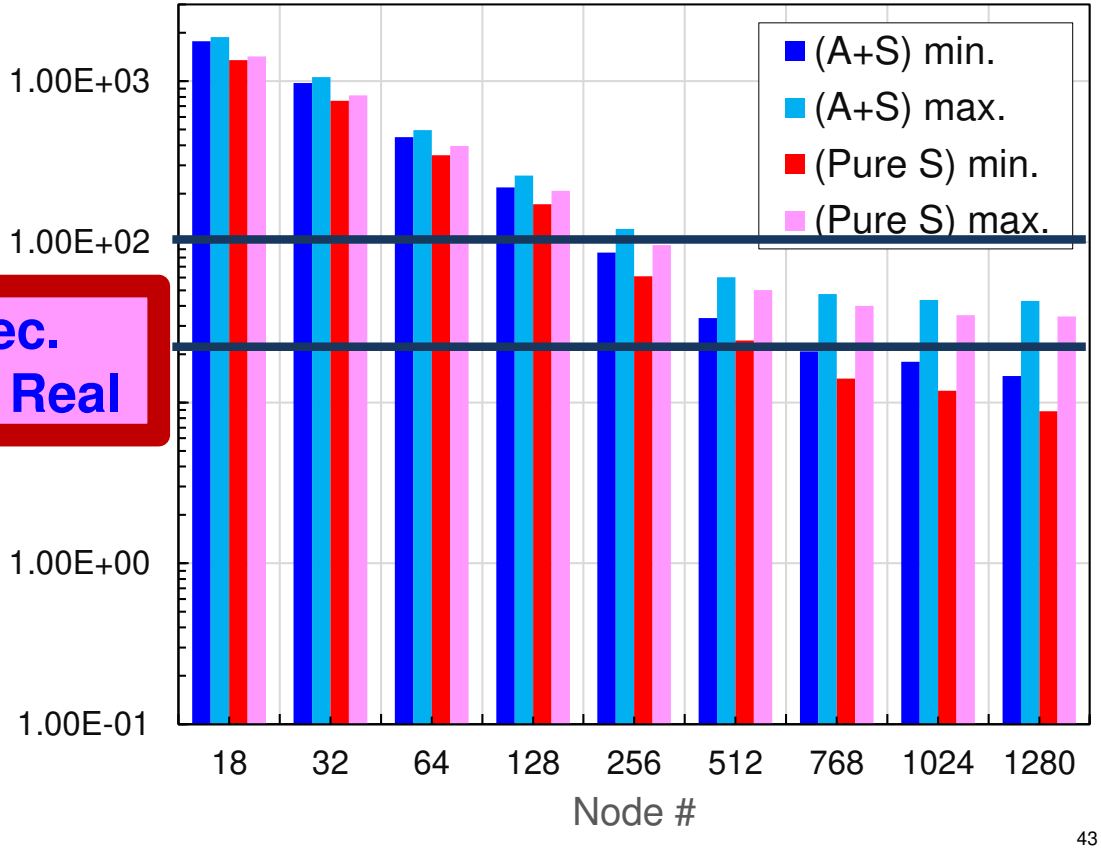
100 sec.
50% of Real



Computation Time for 200 sec. Phenomenon

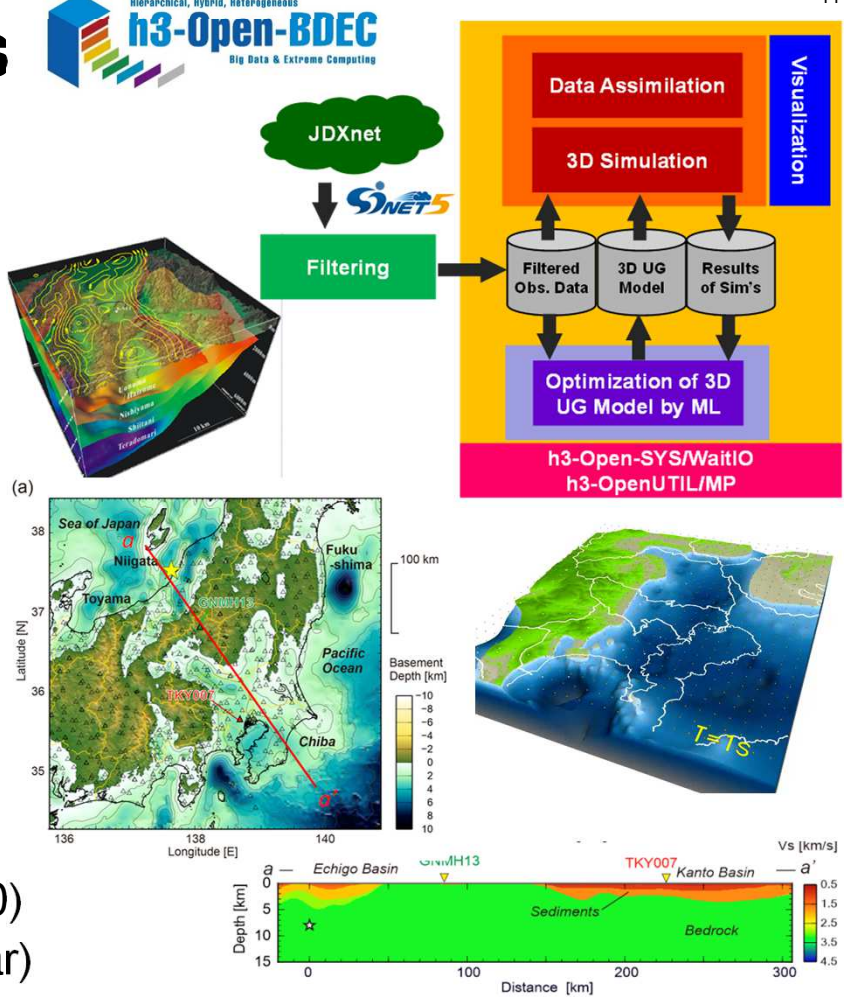
- Communications for I/O are included
  - min.: Comm. excluded
  - max.: Comm. Included
- (A+S)
  - Computation in 10
    (Half of 200 sec.)
  - 300-400 nodes
- (Pure S)
  - Computation in 20 sec.
    (1/10 of 200 sec.)
  - 1,000+ nodes

20 sec.
10% of Real



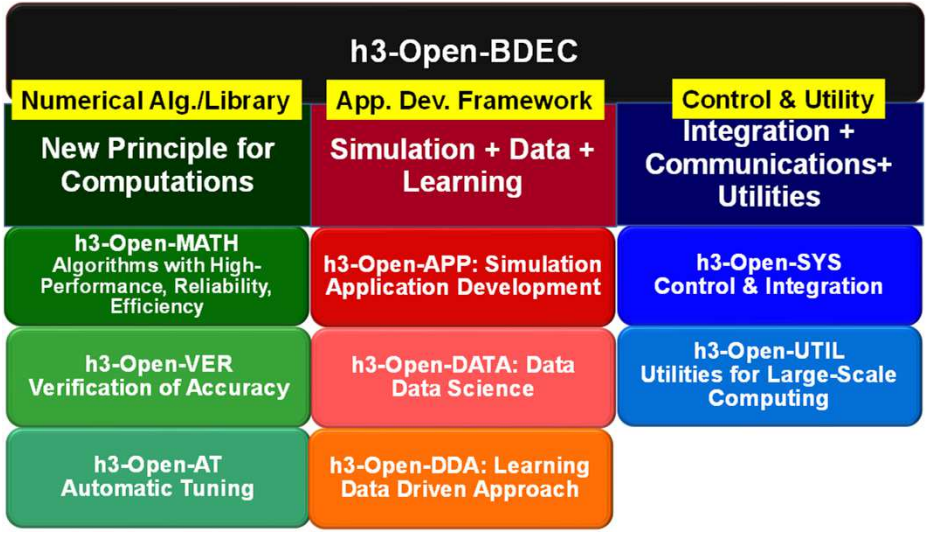
Future Directions towards Integration of (S+D+L)

- Accurate Prediction of Seismic Wave Propagation with Real-Time Data Observation/Assimilation
  - Emergency Info. for Safer Evacuation
- 3D Underground Model
  - Heterogeneous, Observation is difficult
  - Inversion analyses of seismic waves are important for prediction of structure of underground model
  - ML may be utilized for acceleration of this prediction based on analyses of small earthquakes in normal time (e.q. Mw < 3.0)
  - More sophisticated DA method (e.g. 4DVar)



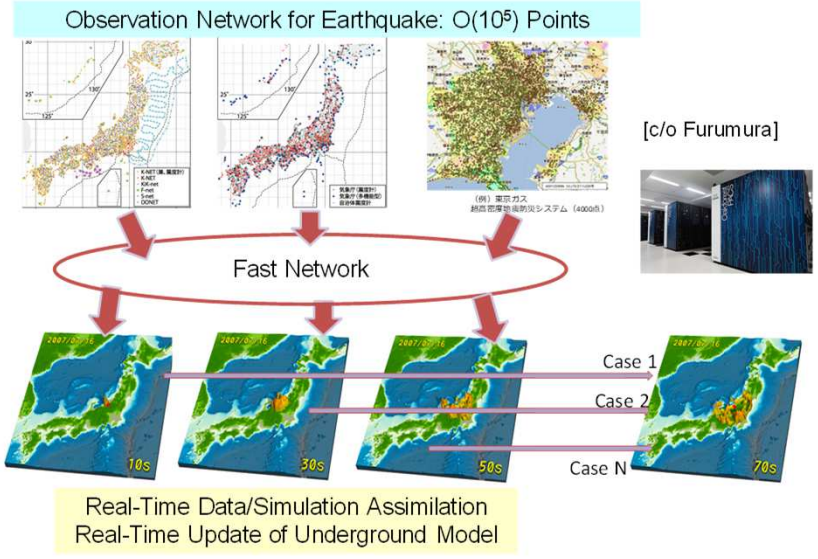
h3-Open-BDEC Innovative Software Platform for Integration of (S+D+L) on the BDEC System, such as Wisteria/BDEC-01

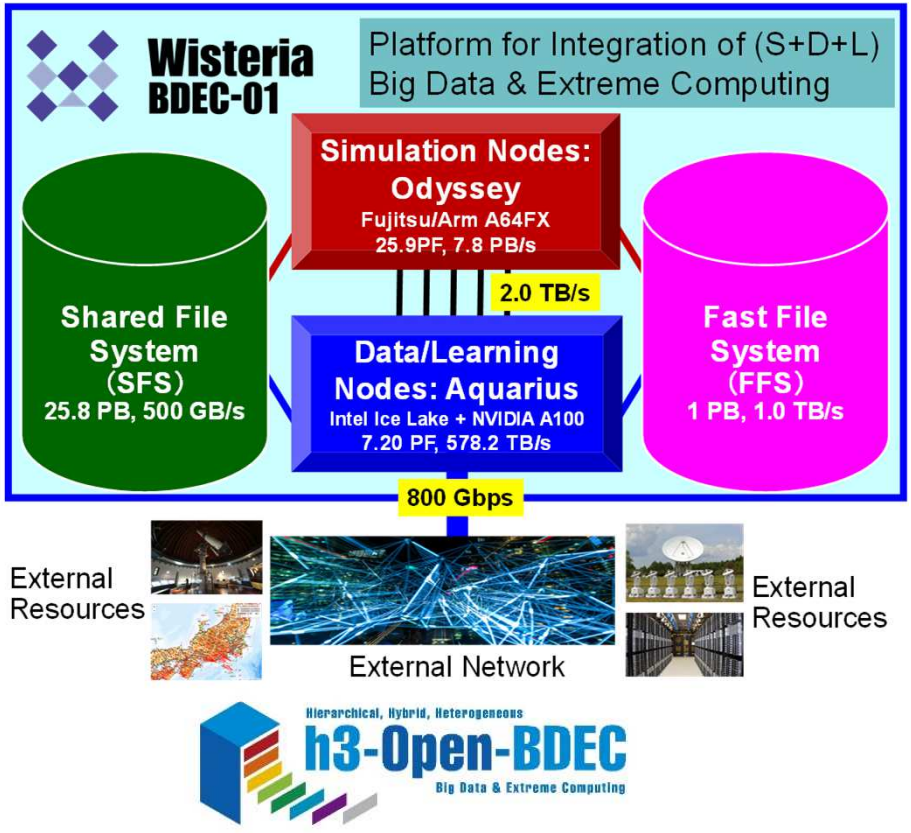
- "Three" Innovations
  - New Principles for Numerical Analysis by Adaptive Precision, Automatic Tuning & Accuracy Verification
  - Hierarchical Data Driven Approach (*h*DDA) based on Machine Learning
  - Software & Utilities for Heterogenous Environment, such as Wisteria/BDEC-01



Computing on Wisteria/BDEC-01

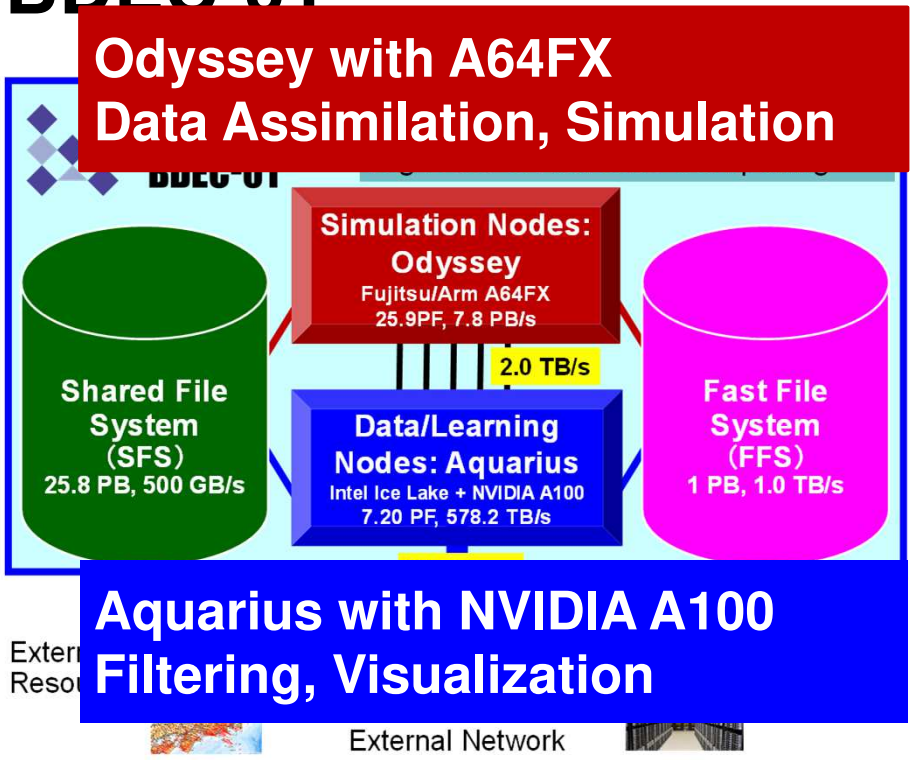
- **Wisteria/BDEC-01**
  - **Aquarius (GPU: NVIDIA A100)**
    - **Filtering, ML, Visualization**
  - **Odyssey (CPU: A64FX)**
    - **Data Assimilation, Simulation**





Computing on Wisteria/BDEC-01

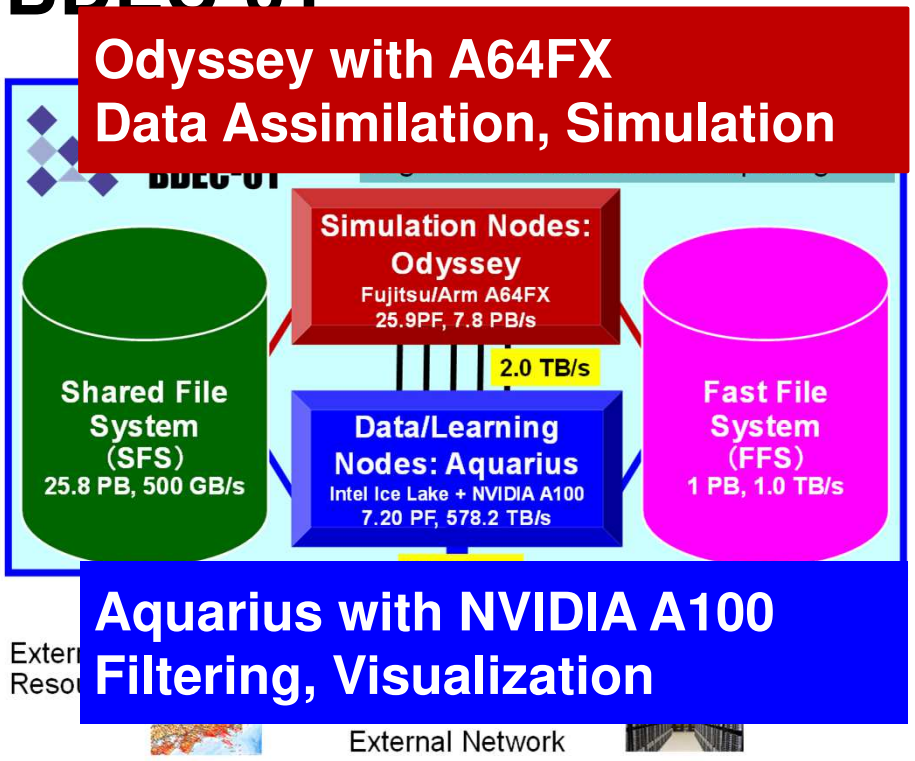
- **Wisteria/BDEC-01**
  - **Aquarius (GPU: NVIDIA A100)**
    - **Filtering, ML, Visualization**
  - **Odyssey (CPU: A64FX)**
    - **Data Assimilation, Simulation**
- Combining Odyssey-Aquarius
  - Single MPI Job over O-A is impossible



Hierarchical, Hybrid, Heterogeneous
h3-Open-BDEC
Big Data & Extreme Computing

Computing on Wisteria/BDEC-01

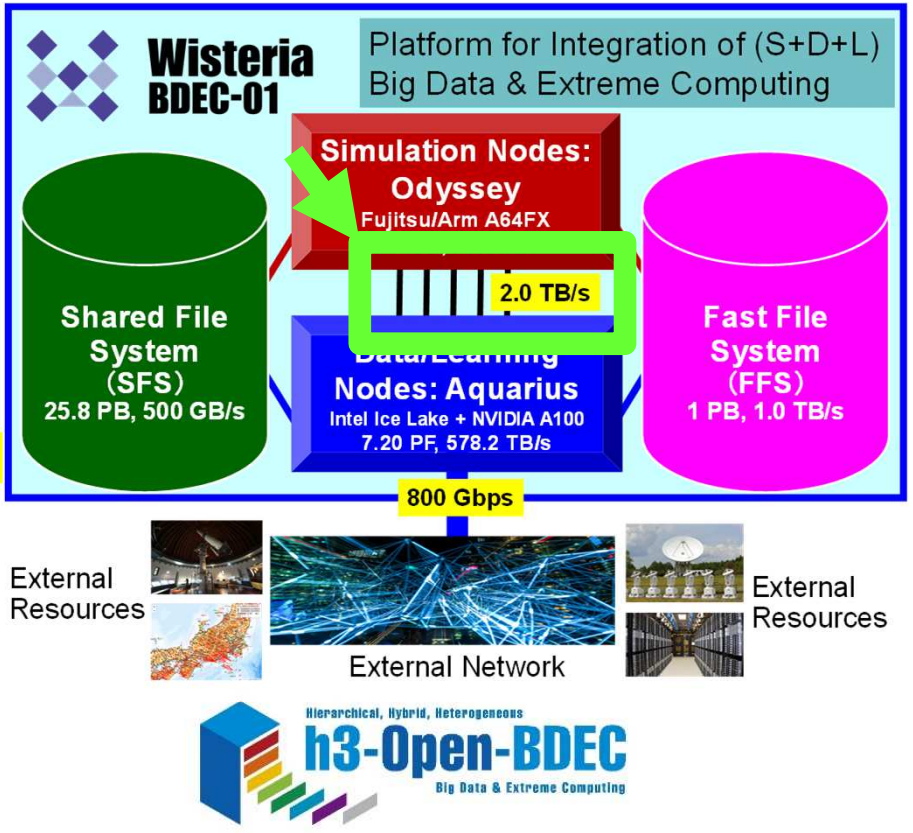
- **Wisteria/BDEC-01**
  - **Aquarius (GPU: NVIDIA A100)**
    - **Filtering, ML, Visualization**
  - **Odyssey (CPU: A64FX)**
    - **Data Assimilation, Simulation**
- Combining Odyssey-Aquarius
  - Single MPI Job over O-A is impossible



Hierarchical, Hybrid, Heterogeneous
h3-Open-BDEC
Big Data & Extreme Computing

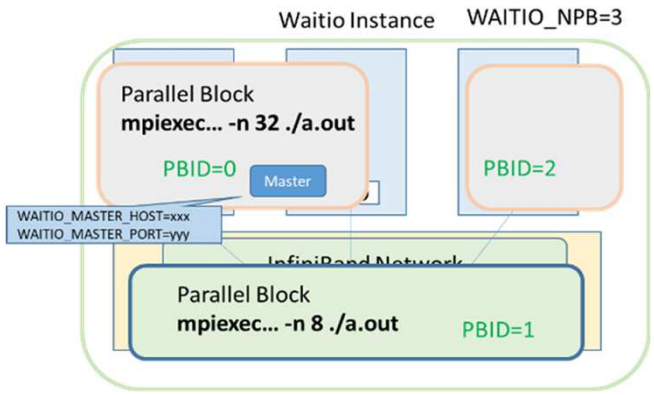
Computing on Wisteria/BDEC-01

- Wisteria/BDEC-01
  - Aquarius (GPU: NVIDIA A100)
    - Filtering, ML, Visualization
  - Odyssey (CPU: A64FX)
    - Data Assimilation, Simulation
- Combining Odyssey-Aquarius
  - Single MPI Job over O-A is impossible
  - Actually, O-A are connected through IB-EDR with 2TB/sec.
  - h3-Open-SYS/WaitIO-Socket
    - Library for Inter-Process Communication through IB-EDR with MPI-like interface
  - h3-Open-UTIL/MP
    - Multiphysics Coupler



API of h3-Open-SYS/WaitIO-Socket PB (Parallel Block): Each Application

WaitIO API	Description
waitio_isend	Non-Blocking Send
waitio_irecv	Non-Blocking Receive
waitio_wait	Termination of waitio_isend/irecv
waitio_init	Initialization of WaitIO
waitio_get_nprocs	Process # for each PB (Parallel Block)
waitio_create_group waitio_create_group_wranks	Creating communication groups among PB's
waitio_group_rank	Rank ID in the Group
waitio_group_size	Size of Each Group
waitio_pb_size	Size of the Entire PB
waitio_pb_rank	Rank ID of the Entire PB

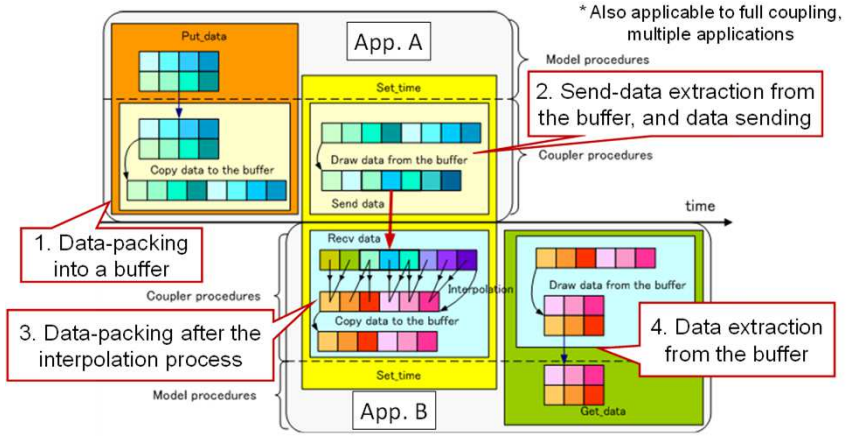


[Sumimoto et al. 2021]

h3-Open-UTIL/MP

Multilevel Coupler/Data Assimilation

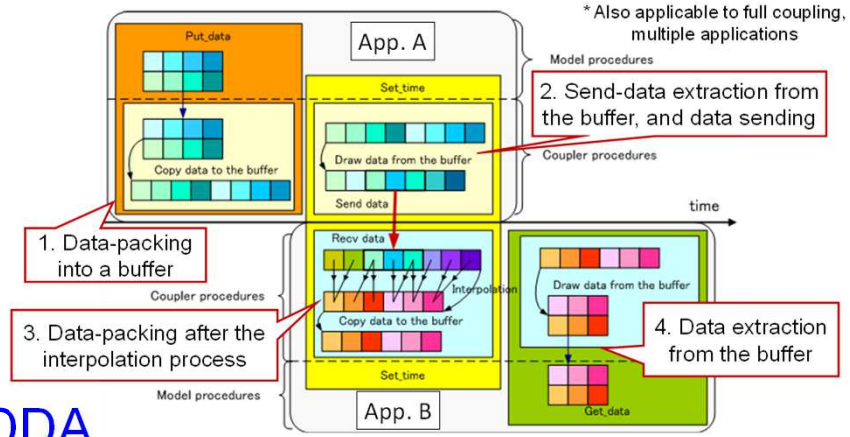
- Current Coupler: ppOpen-MATH/MP
  - Weak-Coupling of Multiple (usually two) Applications
    - Each application does a single computation



h3-Open-UTIL/MP

Multilevel Coupler/Data Assimilation

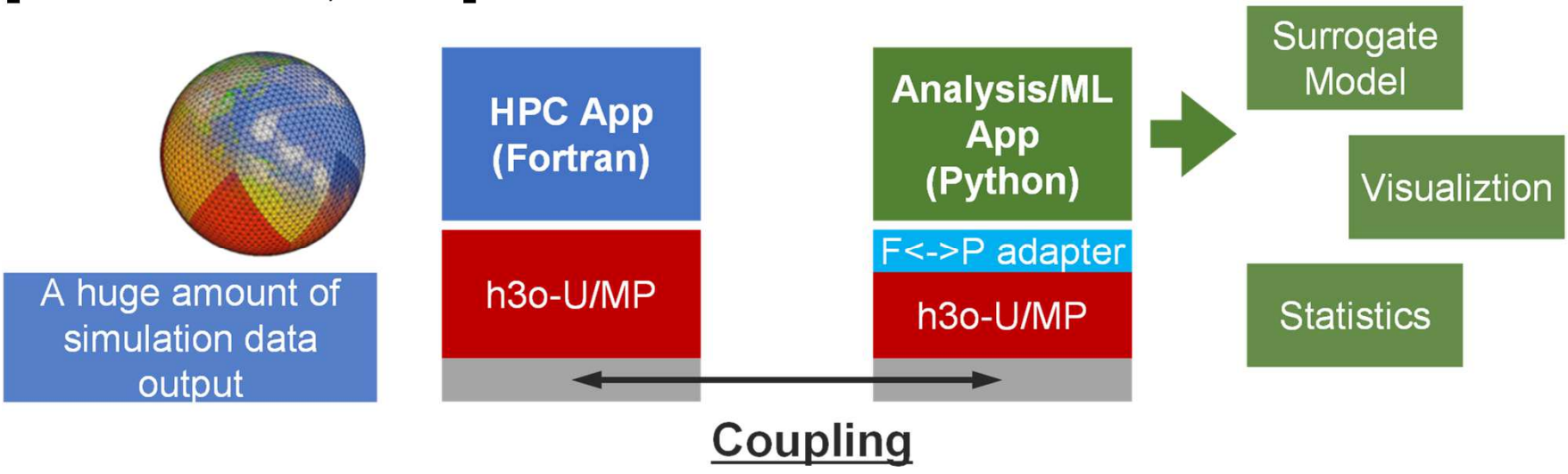
- Current Coupler: ppOpen-MATH/MP
  - Weak-Coupling of Multiple (usually two) Applications
    - Each application does a single computation
- h3-Open-UTIL/MP
  - Data Assimilation (Multiple Computations: Ensemble)
  - Assimilation of Computations with Different Resolutions
    - h3-Open-DATA, h3-Open-APP
  - Data Assimilation by Coupled Codes
    - e.g. Atmosphere-Ocean
- Data Assimilation: h3-Open-DATA
  - Karman Filter, Particle Karman Filter
  - LETKF
  - Adjoint Method
- Generation of Simplified Models in hDDA



h3-Open-UTIL/MP (h3o-U/MP)

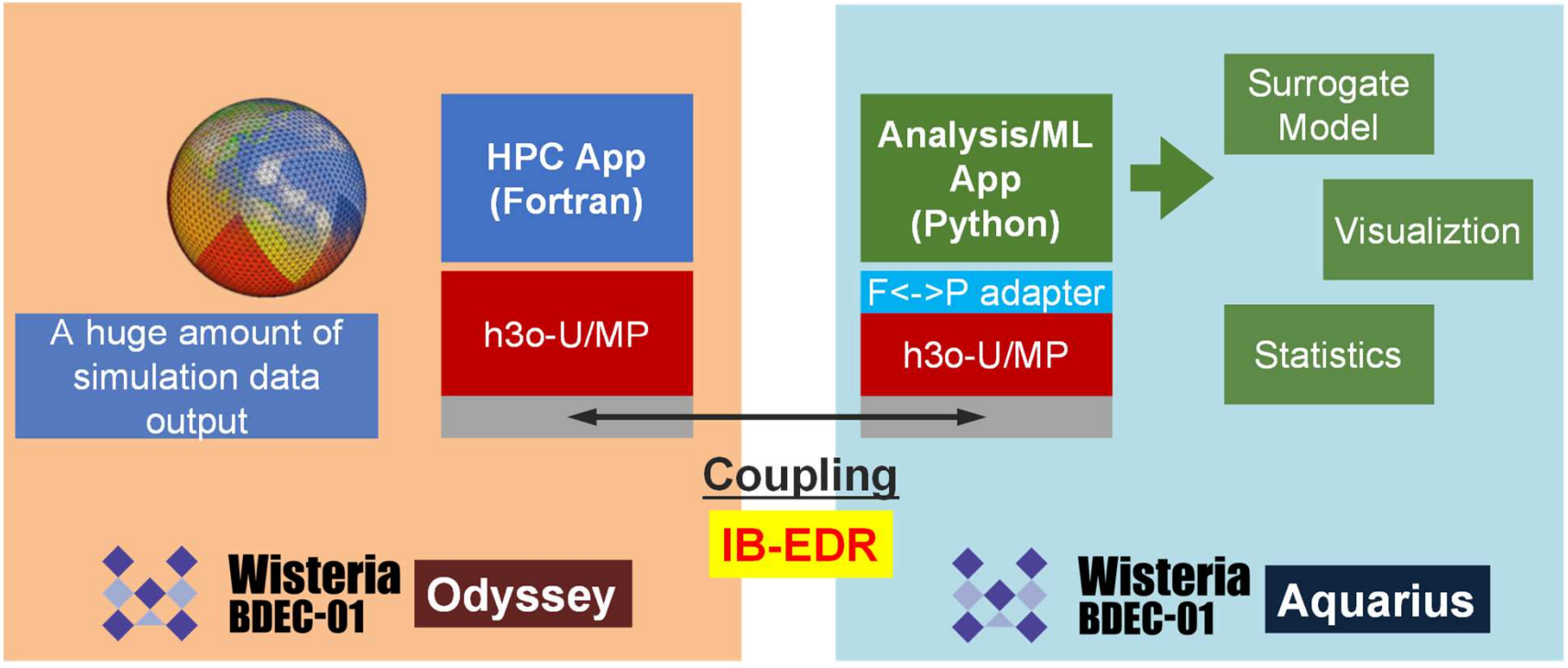
(HPC+AI) Coupling

[Dr. H. Yashiro, NIES]

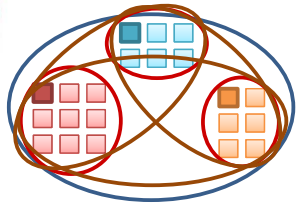


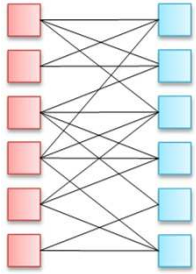
- Providing on-the-fly input/output/training data to the Analysis/ML tools
  - Easy to apply to existing HPC applications
  - Easy access to existing Python-based tools for AI/ML

h3-Open-UTIL/MP (h3o-U/MP) + h3-Open-SYS/WaitIO-Socket

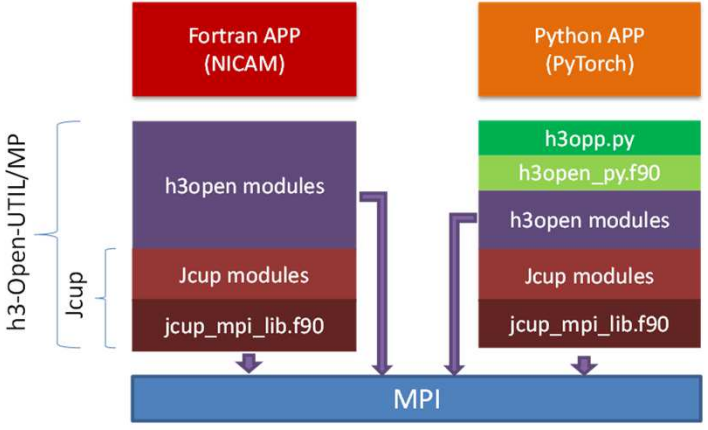


h3-Open-UTIL/MP + h3-Open-SYS/WaitIO-Socket



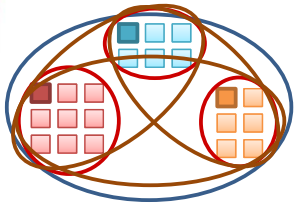


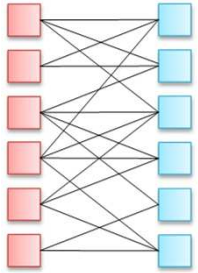
- Current Statue: Single MPI Job
- Direct Communication between Odyssey-Aquarius through IB-EDR by h3-Open-SYS/WaitIO, which provides MPI-like Interface



Current Status: Single MPI Job

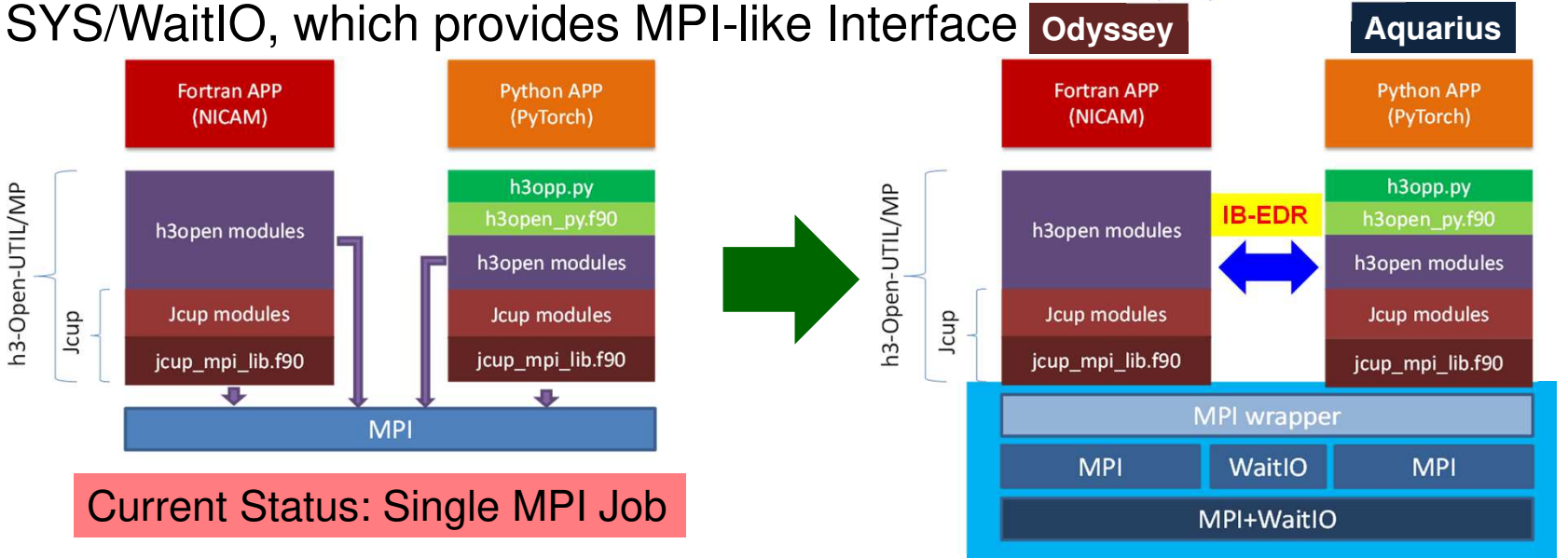
h3-Open-UTIL/MP + h3-Open-SYS/WaitIO-Socket





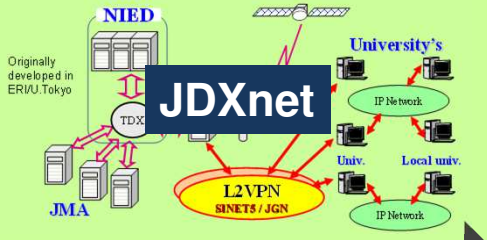
- Current Statue: Single MPI Job
- Direct Communication between Odyssey-Aquarius through IB-EDR by h3-Open-SYS/WaitIO, which provides MPI-like Interface

Wisteria BDEC-01



Current Status: Single MPI Job

System on Wisteria/BDEC-01 using WaitIO



Observed Data

Obs. Data

SINET5

External Server for Observed Data

Wisteria BDEC-01

Data/Learning Nodes: Aquarius
Intel Ice Lake + NVIDIA A100
7.20 PF, 578.2 TB/s

2.0 TB/s

Simulation Nodes: Odyssey
Fujitsu/Arm A64FX
25.9PF, 7.8 PB/s

Output & Movie

Filter

Visualizer

h3-Open-SYS/WaitIO-Socket

Seism3D/ OpenSWPC-DAF

Mesh Data

UG Model

Communications by WaitIO-Socket [Kasai et al. 2021]

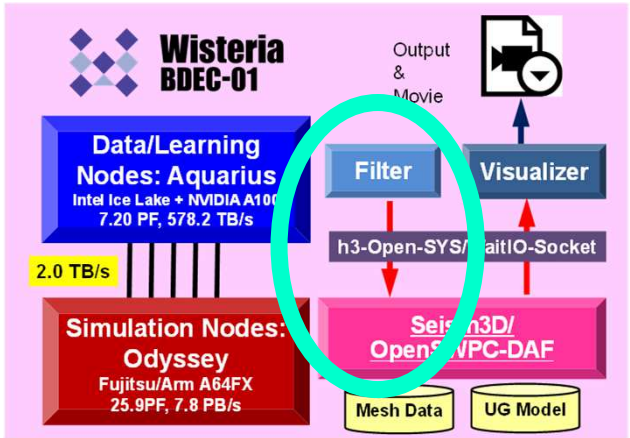
Aquarius: SEND

```
program dmy_filter
<省略: 型宣言等>
call mpi_init (ierr)
call mpi_comm_size (MPI_COMM_WORLD, nprocs, ierr)
call mpi_comm_rank (MPI_COMM_WORLD, myrank,  ierr)
call WAITIO_CREATE_UNIVERSE (WAITIO_COMM_UNIVERSE, ierr)

if (myrank==0) then
  open(100,file='./obsfile_list.txt', form='formatted', status='old', iostat=ierr)
  do i=1,300
    <省略: obsデータ読み込み処理>
    print *,"Send obs data ..... "
    call WAITIO_MPI_ISEND (NTMAX1_o, 1,           WAITIO_MPI_INTEGER,  2,1, WAITIO_COMM_UNIVERSE,req(1,1), ierr)
    call WAITIO_MPI_ISEND (DT_o,     1,           WAITIO_MPI_FLOAT,    2,2, WAITIO_COMM_UNIVERSE,req(1,2), ierr)
    call WAITIO_MPI_ISEND (NST_o,    1,           WAITIO_MPI_INTEGER,  2,3, WAITIO_COMM_UNIVERSE,req(1,3), ierr)
    call WAITIO_MPI_ISEND (AT_o,     1,           WAITIO_MPI_FLOAT,    2,4, WAITIO_COMM_UNIVERSE,req(1,4), ierr)
    call WAITIO_MPI_ISEND (T0_o,     1,           WAITIO_MPI_FLOAT,    2,5, WAITIO_COMM_UNIVERSE,req(1,5), ierr)
    call WAITIO_MPI_ISEND (ISO_X_o,  NSMAX,       WAITIO_MPI_INTEGER,  2,6, WAITIO_COMM_UNIVERSE,req(1,6), ierr)
    call WAITIO_MPI_ISEND (ISO_Y_o,  NSMAX,       WAITIO_MPI_INTEGER,  2,7, WAITIO_COMM_UNIVERSE,req(1,7), ierr)
    call WAITIO_MPI_ISEND (ISO_Z_o,  NSMAX,       WAITIO_MPI_INTEGER,  2,8, WAITIO_COMM_UNIVERSE,req(1,8), ierr)
    call WAITIO_MPI_ISEND (ISTX_o,   NST,         WAITIO_MPI_INTEGER,  2,9, WAITIO_COMM_UNIVERSE,req(1,9), ierr)
    call WAITIO_MPI_ISEND (ISTY_o,   NST,         WAITIO_MPI_INTEGER,  2,10,WAITIO_COMM_UNIVERSE,req(1,10),ierr)
    call WAITIO_MPI_ISEND (ISTZ_o,   NST,         WAITIO_MPI_INTEGER,  2,11,WAITIO_COMM_UNIVERSE,req(1,11),ierr)
    call WAITIO_MPI_ISEND (STC_o,    6*NST,       WAITIO_MPI_CHAR,     2,12,WAITIO_COMM_UNIVERSE,req(1,12),ierr)
    call WAITIO_MPI_ISEND (VxAll_obs,NST*NOBS_LEN,WAITIO_MPI_FLOAT,    2,13,WAITIO_COMM_UNIVERSE,req(1,13),ierr)
    call WAITIO_MPI_ISEND (VyAll_obs,NST*NOBS_LEN,WAITIO_MPI_FLOAT,    2,14,WAITIO_COMM_UNIVERSE,req(1,14),ierr)
    call WAITIO_MPI_ISEND (VzAll_obs,NST*NOBS_LEN,WAITIO_MPI_FLOAT,    2,15,WAITIO_COMM_UNIVERSE,req(1,15),ierr)
    call WAITIO_MPI_WAITALL (15,req, status, ierr)
    call sleep(1)
  enddo
  close (100)
endif
call WAITIO_FINALIZE (ierr)
call mpi_finalize (ierr)
end
```

Odyssey: RECV

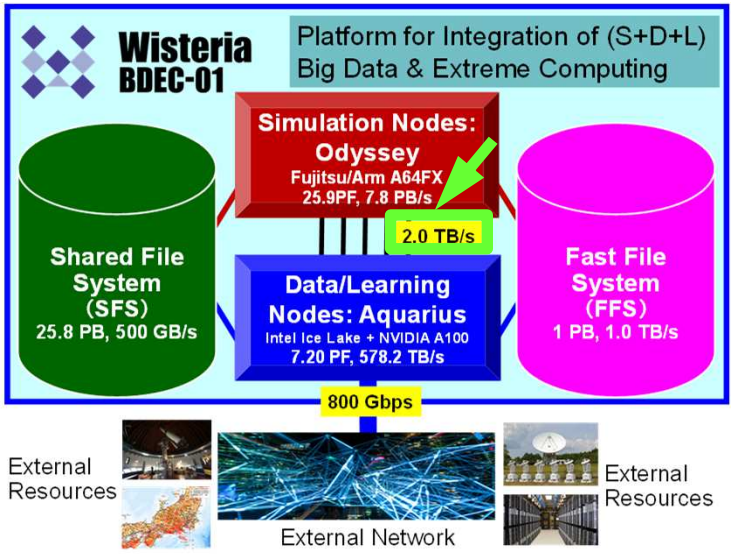
```
call WAITIO_MPI_IRECV (NTMAX1_o, 1,           WAITIO_MPI_INTEGER, 0,1, WAITIO_COMM_UNIVERSE,...)
call WAITIO_MPI_IRECV (DT_o,     1,           WAITIO_MPI_FLOAT,   0,2, WAITIO_COMM_UNIVERSE,...)
call WAITIO_MPI_IRECV (NST_o,    1,           WAITIO_MPI_INTEGER, 0,3, WAITIO_COMM_UNIVERSE,...)
call WAITIO_MPI_IRECV (AT_o,     1,           WAITIO_MPI_FLOAT,   0,4, WAITIO_COMM_UNIVERSE,...)
call WAITIO_MPI_IRECV (T0_o,     1,           WAITIO_MPI_FLOAT,   0,5, WAITIO_COMM_UNIVERSE,...)
call WAITIO_MPI_IRECV (ISO_X_o,  NSMAX,       WAITIO_MPI_INTEGER, 0,6, WAITIO_COMM_UNIVERSE,...)
call WAITIO_MPI_IRECV (ISO_Y_o,  NSMAX,       WAITIO_MPI_INTEGER, 0,7, WAITIO_COMM_UNIVERSE,...)
call WAITIO_MPI_IRECV (ISO_Z_o,  NSMAX,       WAITIO_MPI_INTEGER, 0,8, WAITIO_COMM_UNIVERSE,...)
call WAITIO_MPI_IRECV (ISTX_o,   NST,         WAITIO_MPI_INTEGER, 0,9, WAITIO_COMM_UNIVERSE,...)
call WAITIO_MPI_IRECV (ISTY_o,   NST,         WAITIO_MPI_INTEGER, 0,10,WAITIO_COMM_UNIVERSE,...)
call WAITIO_MPI_IRECV (ISTZ_o,   NST,         WAITIO_MPI_INTEGER, 0,11,WAITIO_COMM_UNIVERSE,...)
call WAITIO_MPI_IRECV (STC_o,    6*NST,       WAITIO_MPI_CHAR,    0,12,WAITIO_COMM_UNIVERSE,...)
call WAITIO_MPI_IRECV (VxAll_obs,NST*NOBS_LEN,WAITIO_MPI_FLOAT,   0,13,WAITIO_COMM_UNIVERSE,...)
call WAITIO_MPI_IRECV (VyAll_obs,NST*NOBS_LEN,WAITIO_MPI_FLOAT,   0,14,WAITIO_COMM_UNIVERSE,...)
call WAITIO_MPI_IRECV (VzAll_obs,NST*NOBS_LEN,WAITIO_MPI_FLOAT,   0,15,WAITIO_COMM_UNIVERSE,...)
```

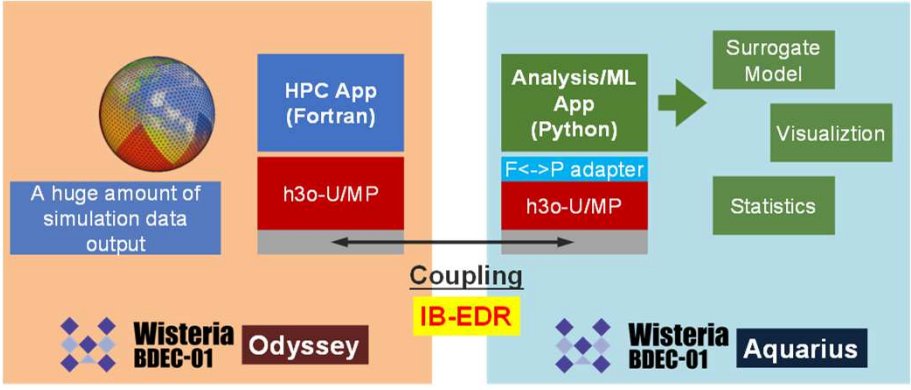


Schedule for Public Use

Collaborations are Welcome !!

- h3-Open-SYS/WaitIO-Socket
  - June 2022, O-A Direct Communication by MPI-like Interface
- h3-Open-SYS/WaitIO-File
  - Via File System, After October 2022
- h3-Open-UTIL/MP（HPC+Python）
  - June 2022 on Odyssey only (Single MPI)
- h3-Open-UTIL/MP+h3-Open-SYS/WaitIO-Socket via IB-EDR
  - June 2022





Wisteria BDEC-01

Hierarchical, Hybrid, Heterogeneous h3-Open-BDEC Big Data & Extreme Computing

Summary

- Earthquake Simulation/Real-Time Data Assimilation
  - On-Going Works for Real-Time Forecast/Assimilation
  - Preliminary Works on OBCX
  - Switching to Wisteria/BDEC-01
- Future Works
  - Improvement of the Simulation Method
  - Improvement of Underground/Subsurface Model by ML (Machine Learning)
  - More sophisticated algorithms for data assimilation (e.g. 4DVar, Ensemble 4DVar, 4DEnVar etc.)
  - Implementation/Optimization towards Real-Time System
- https://h3-open-bdec.cc.u-tokyo.ac.jp/