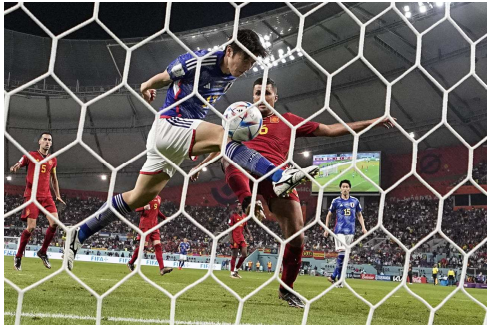
東京大学情報基盤センター
INFORMATION TECHNOLOGY CENTER, THE UNIVERSITY OF TOKYO

東京大学
THE UNIVERSITY OF TOKYO

PC Cluster Consortium



東大情報基盤センターの目指す『計算・データ・学習』の融合による革新的スーパーコンピューティング

Wisteria
BDEC-01

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

中島 研吾
東京大学情報基盤センター

PCCC22「HPCシステム技術の最前線」
2022年12月5日

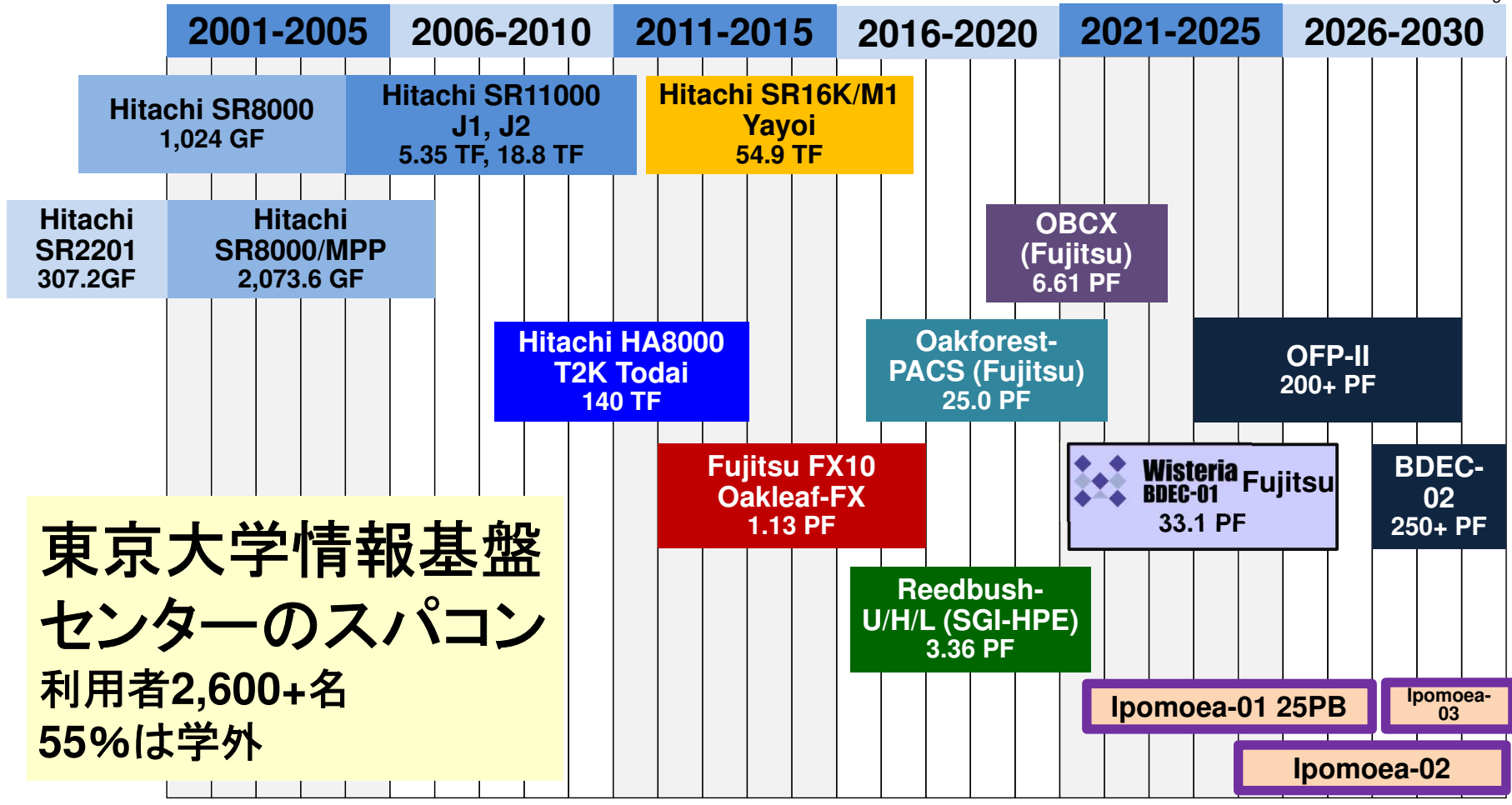
東京大学情報基盤センター

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

- 東京大学大型計算機センター（1965年）
- 東京大学情報基盤センター（1999年～）
  - 全国共同利用施設
  - 学際大規模情報基盤共同利用・共同研究拠点　中核拠点（2010年～）
  - 革新的ハイパフォーマンス・コンピューティング・インフラ(HPCI)　構成機関（2010年～）
  - 最先端共同HPC基盤施設（JCAHPC）（2013年～）
    - 筑波大学計算科学研究センター・東大情報基盤センター：OFP

JCAHPC

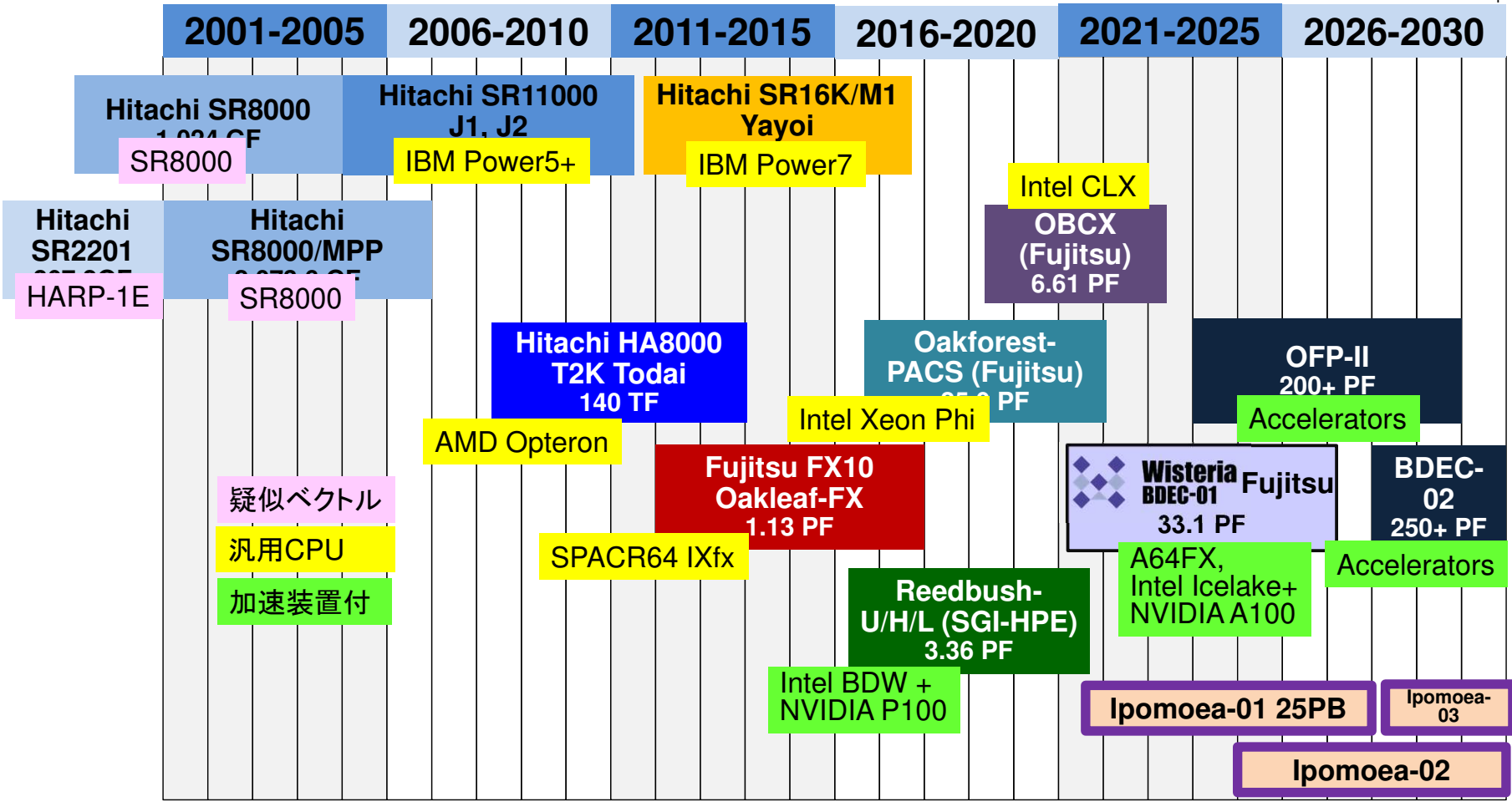
- 2022年11月現在
  - 2式のスパコンシステムを運用
    - Oakbridge-CX（OBCX）：2023年9月末運用終了
    - Wisteria/BDEC-01（「計算・データ・学習」融合スーパーコンピュータシステム）：2021年5月運用開始
  - 大規模共通ストレージ「Ipomoea-01」：2022年1月運用開始
  - データ活用社会創成プラットフォーム（mdx）：2021年3月設置

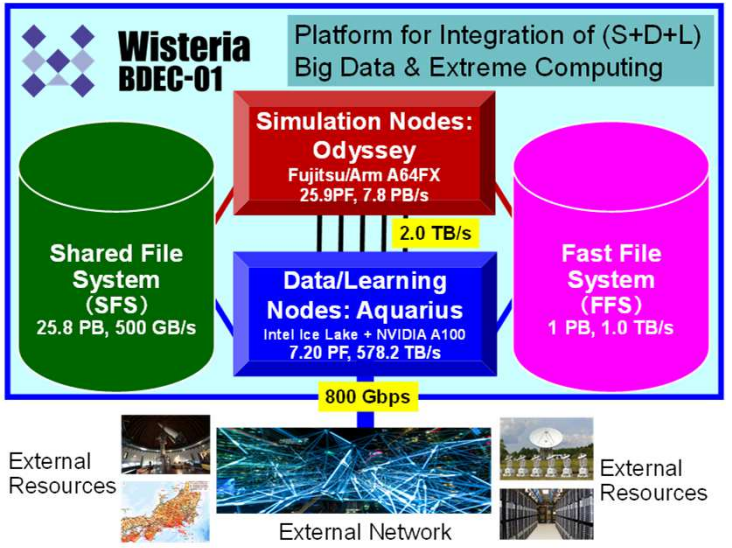


東京大学情報基盤センターのスパコン

利用者2,600+名

55%は学外


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 200+ PF
Accelerators
疑似ベクトル
汎用CPU
加速装置付
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
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))

- データ解析・シミュレーション融合スーパーコンピュータ
- 2016年7月～2021年11月末
- 東大ITC初のGPUクラスタ，ピーク性能3.36 PF（Reedbush-H/L）

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

- JCAHPC（筑波大CCS・東大ITC），2016年10月～2022年3月末
- 25 PF, #39 in 58th TOP 500 (November 2021)

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

- 2019年7月～2023年9月末（予定）
- 6.61 PF, #129 in 60th TOP500（November 2022）

Wisteria/BDEC-01（Fujitsu）

- シミュレーションノード群（Odyssey）：A64FX (#23)
- データ・学習ノード群（Aquarius）：Intel Icelake + NVIDIA A100 (#125)
- 33.1 PF, 2021年5月14日運用開始
- 「計算・データ・学習（S+D+L）」融合のためのプラットフォーム
- 革新的ソフトウェア基盤「h3-Open-BDEC」
  （科研費基盤（S）2019年度～2023年度）

h3-Open-BDEC

Reedbush

Oakforest-PACS

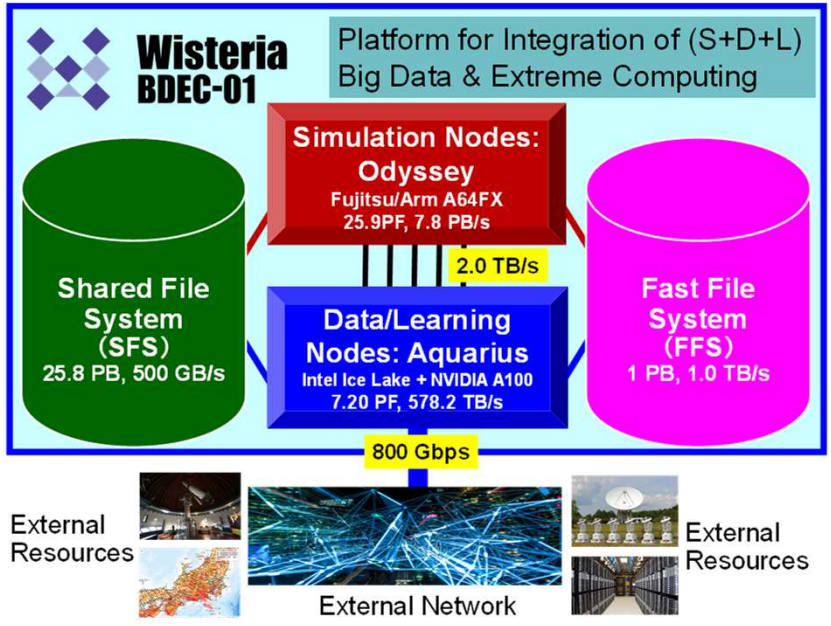
Oakbridge-CX

SC22における諸ランキング（2022年11月）

SC22 Dallas, TX | hpc accelerates.

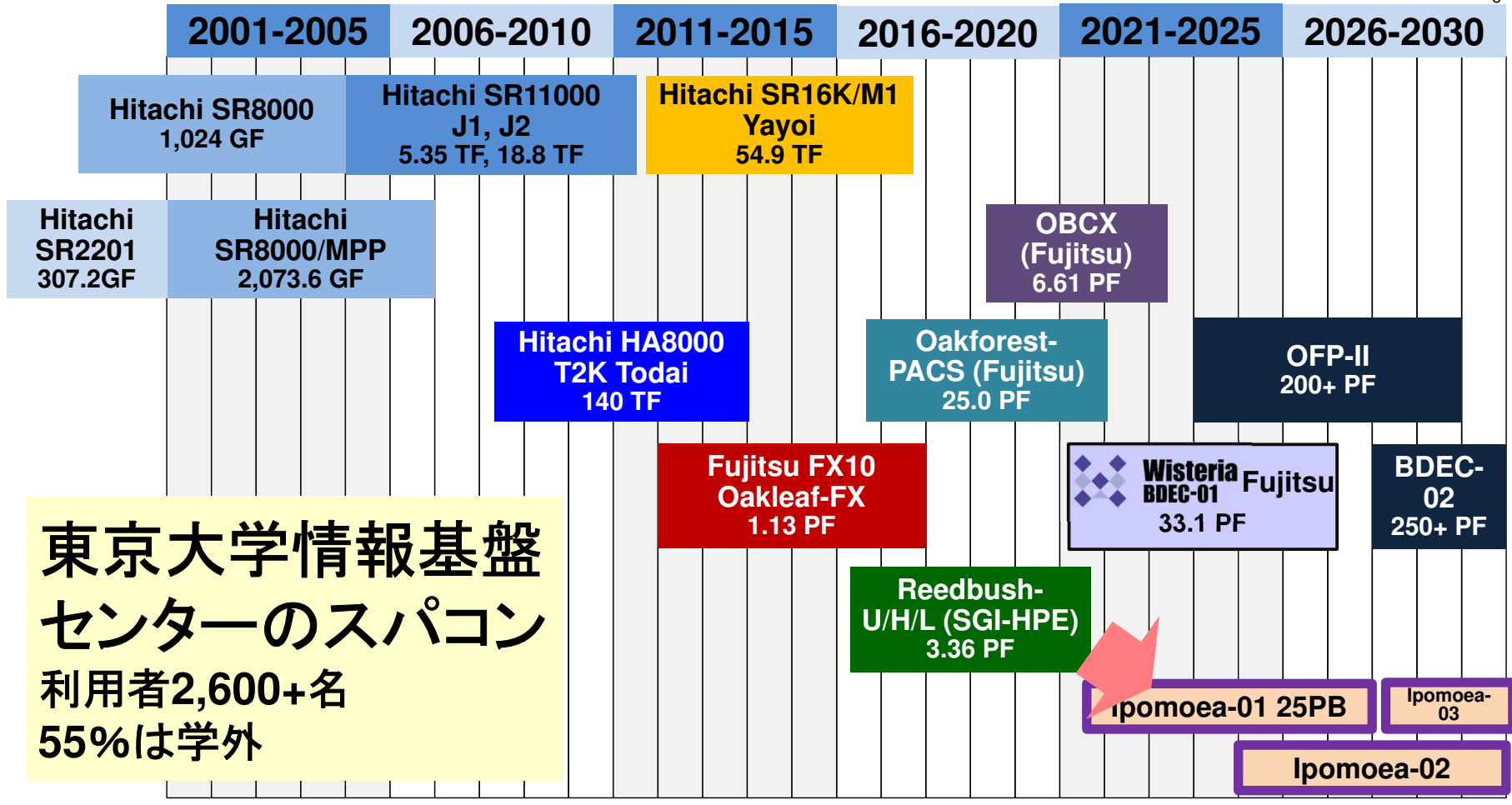
	Odyssey	Aquarius
TOP 500	23	125
Green 500	45	28
HPCG	12	68
Graph 500 BFS	4	-
HPL-MxP (HPL-AI)	10*	-

*) ISC 2022 (June 2022)



GFLOPS（ピーク性能）当たり利用負担（円）：電気代
GFLOPS/W（Green 500）（2023年度から値上げ予定）

System	JPY/GFLOPS Small is Good	GFLOPS/W Large is Good
Oakleaf-FX/Oakbridge-FX (Fujitsu) (Fujitsu SPARC64 IXfx)	125	**0.866**
Reedbush-U (HPE)　(Intel Xeon Broadwell (BDW))	61.9	**2.310**
Reedbush-H (HPE)　(Intel BDW+NVIDIA P100x2/node)	15.9	**8.575**
Reedbush-L (HPE)　(Intel BDW+NVIDIA P100x4/node)	13.4	**10.167**
Oakforest-PACS (Fujitsu)　(Intel Xeon Phi/KNL)	16.5	**4.986**
Oakbridge-CX (Fujitsu)　(Intel Xeon Cascade Lake)	20.7	**5.076**
Wisteria-Odyssey（Fujitsu/Arm A64FX）	17.8	**15.069**
Wisteria-Aquarius（Intel Xeon Ice Lake + NVIDIA A100x8）	9.00	**24.058**



東京大学情報基盤センターのスパコン

利用者2,600+名

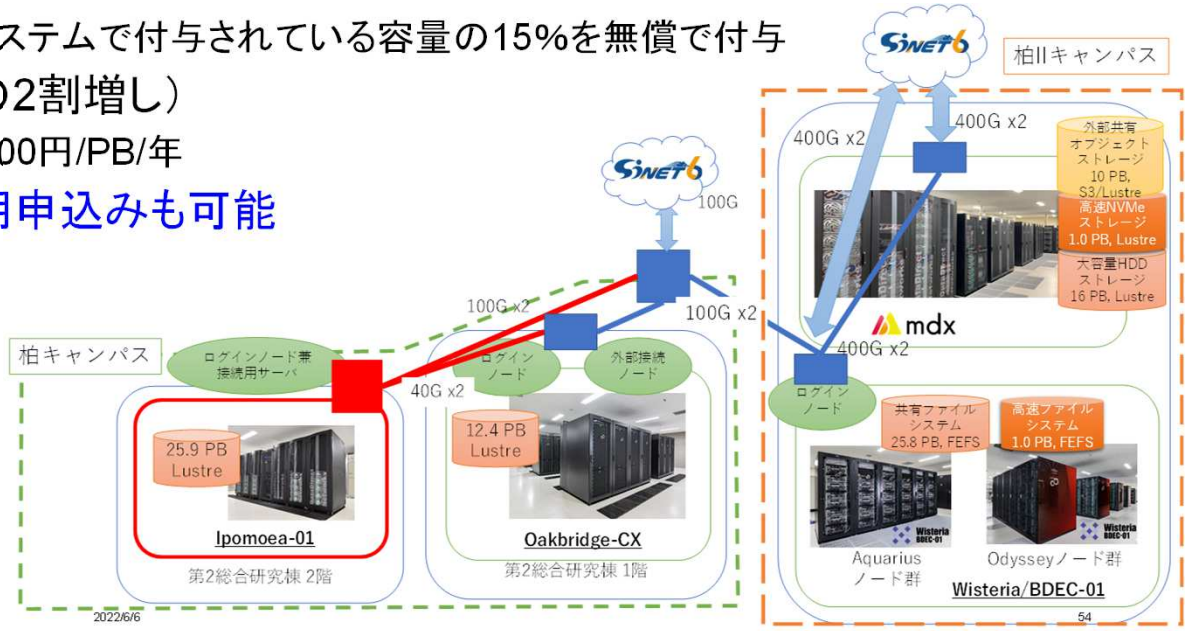
55%は学外

大規模共通ストレージシステム「Ipomoea」

- スーパーコンピュータの処理能力の向上に伴い，扱うデータ量も増加の一途
- 東大センターでは従来ストレージは各システムに附属して導入され，各システムのストレージは独立
- このような状況（注：ストレージがシステム毎に独立）は利用者に多大な不便を強いることになり，東大センターの全システムからアクセス可能な共通ストレージの導入が強く求められていた
- 各システムからアクセスできる「大規模共通ストレージ（Ipomoea）」導入決定
  - OFP運用終了が契機
  - 1システムを約5–6 年使用し，約3年ごとに新しいストレージシステム（25+PB）を導入し，入れ替えることを想定している

- 2022年1月運用開始・6月より一般に公開，25+PB，富士通製
  - 2022 年5月末までにOFPのLustre 領域の必要ファイルの移行完了
- 割当容量
  - 東大センターのシステムに利用者番号（教育利用，講習会除く）を有する場合
    - 各利用者ごとに5TB
    - 各グループごとに登録システムで付与されている容量の15%を無償で付与
  - 追加負担金（企業はこの2割増し）
    - 7,200円/TB/年，2,100,000円/PB/年
  - Ipomoea-01のみの利用申込みも可能

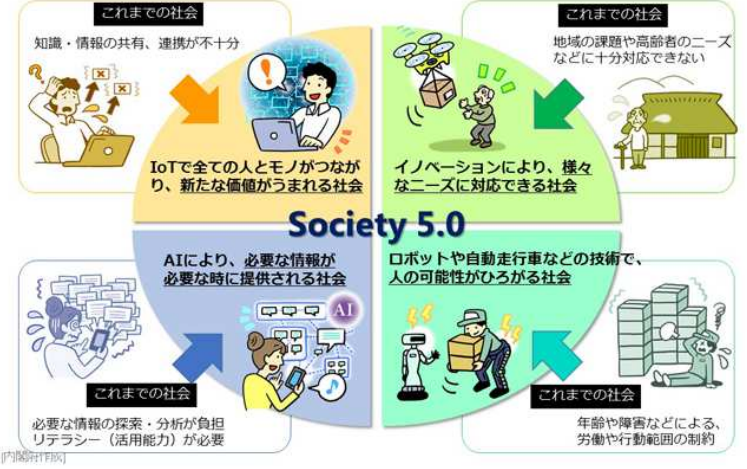
Ipomoea-01



mdx データ活用社会創成プラットフォーム

超スマート社会 Society 5.0： AI, IoT, ビッグデータなどの革新技術を社会全体に活用し、サイバー空間（仮想）とフィジカル空間（現実）を高度に融合させた社会

持続可能な開発目標(SDGs)：経済発展と社会的課題解決の両立



用途に応じてオンデマンドで短時間に構築・拡張・融合できるデータ収集・集積・解析機能を提供するプラットホーム

3 本柱

1 高性能計算環境によるデータ科学と計算科学の融合

データ科学、計算科学の手法を融合し、さらに国内最高の計算環境を用いて他に無い高精度の予測を行えるようにする

2 SINETを活かしたリアルタイム収集・集積・解析環境の動的な構築

遠隔地のセンサーやストレージ、データプラットフォームの計算資源、ストレージをつないで、リアルタイムに入力から出力を得られるアプリケーションごとの収集・集積・解析環境を、使いたいときに即時に構築するSINETモバイル基盤によりセンサー等のデータを安定してセキュアにつなぐ

3 異種データ・異種知識の融合活用の推進

様々な分野のデータ保持者、解析者、利用者が産学にまたがって連携するコミュニティーを形成し、新たな価値創造につなげる。

mdx データ活用社会創成プラットフォーム

- データ利活用・セキュリティを重視した
  クラウド型の高性能仮想化環境
- 9大学2研究所が共同運営し、全国共同利用

ネットワーク
I2つのネットワーク 外部接続ネットワーク
SINET6と400G x2で接続
内部高速ネットワーク RDMA
ストレージ

汎用CPUノード
Intel IceLake x2ソケット **x368ノード**
理論ピーク性能(FP64): **2.1PFLOPS**
総メモリバンド幅: 150.7 TB/s

GPU 演算加速ノード
Intel IceLake x2ソケット+NVIDIA A100 x8 **x40ノード**
理論ピーク性能(FP64): **6.4PFLOPS**
理論ピーク性能(FP16): **100.7PFLOPS**
総メモリバンド幅: 496.3 TB/s

高速 NVMe ストレージ
Lustre Filesystem
1.0 PB (NVMe SSD)
252 GByte/sec

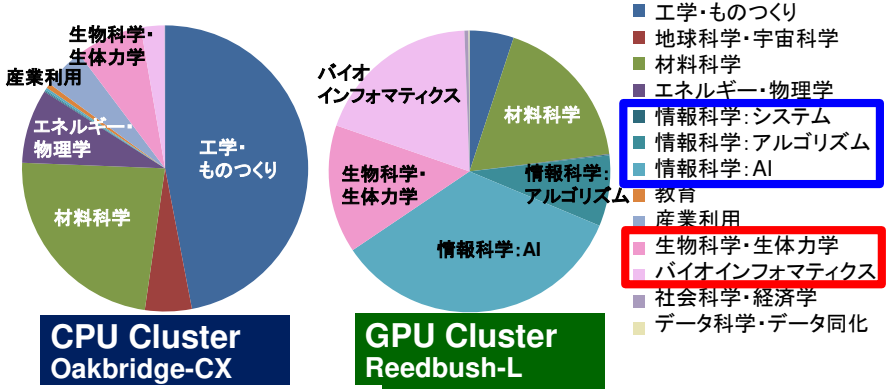
大容量HDDストレージ
Lustre Filesystem
16.3 PB (HDD)
157.5 GByte/sec

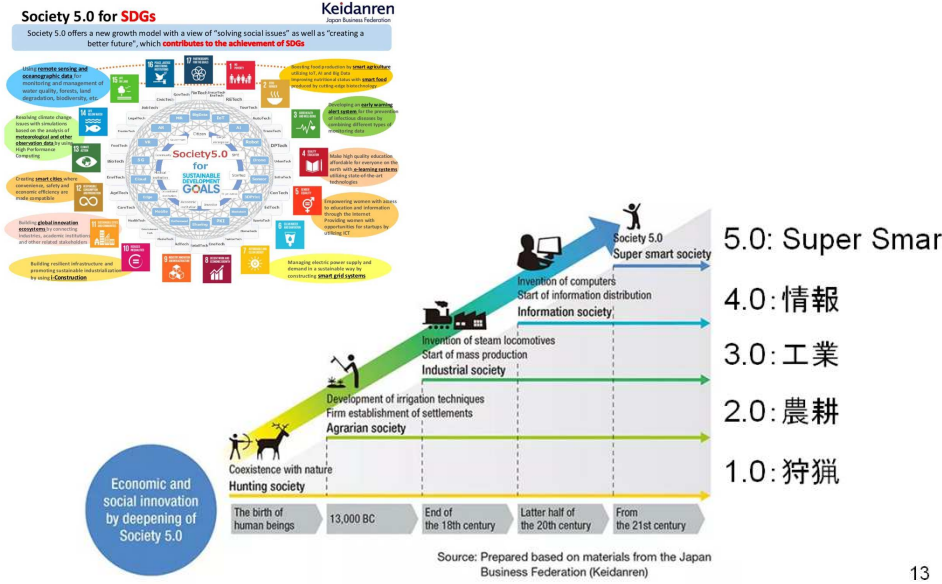
外部共有オブジェクトストレージ
S3 Data Service
10.3 PB (HDD)
63.0 GByte/sec

スーパーコンピューティングの今後

- ワークロードの多様化
  - 計算科学，計算工学：Simulations
  - 大規模データ解析
  - AI，機械学習
- **（シミュレーション（計算）＋データ＋学習）融合**

  ⇒Society 5.0実現に有効
  - **フィジカル空間とサイバー空間の融合**
    - **S:シミュレーション（計算）（Simulation）**
    - **D:データ（Data）**
    - **L：学習（Learning）**
  - **Simulation + Data + Learning = S+D+L**





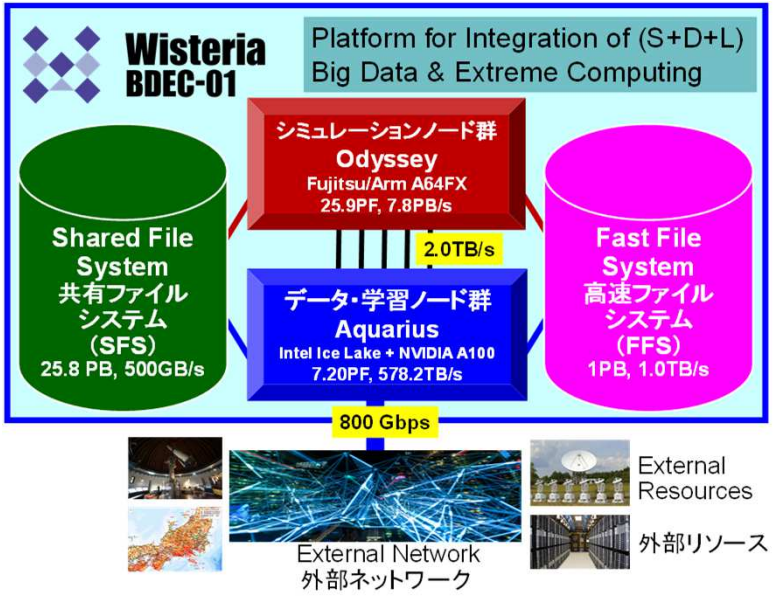
（シミュレーション（計算）＋データ＋学習）融合（S+D+L）

- 東大情報基盤センターでは，2015年頃から「（S+D+L）融合」の重要性に注目し，それを実現するためのハードウェア，ソフトウェア，アプリケーション，アルゴリズムに関する研究開発を開始
  - BDEC計画（Big Data & Extreme Computing）
  - 「データ＋学習」による，より高度な「シミュレーション」
    - AI for HPC, AI for Science
  - 地球科学関連では自然な発想（すでに実施されている）
- 2021年5月に運用を開始した「Wisteria/BDEC-01」は「BDEC計画」の1号機
  - Reedbush，Oakbridge-CXは「BDEC」のプロトタイプと位置づけられる
  - 「計算・データ・学習（S+D+L）」融合を実現する，世界でも初めてのプラットフォーム

Wisteria/BDEC-01

BDEC:「計算・データ・学習（S+D+L）」融合のためのプラットフォーム（Big Data & Extreme Computing）

- 2021年5月14日運用開始
  - 東京大学柏Ⅱキャンパス
- 33.1 PF, 8.38 PB/sec. , **富士通製**
  - ~4.5 MVA（空調込み）, ~360m^2
- Hierarchical, Hybrid, Heterogeneous (h3)
- **2種類のノード群**
  - **シミュレーションノード群（S, SIM）：Odyssey**
    - **従来のスパコン**
    - **Fujitsu PRIMEHPC FX1000 (A64FX), 25.9 PF**
      - **7,680ノード（368,640 コア）, 20ラック, Tofu-D**
  - **データ・学習ノード群（D/L, DL）：Aquarius**
    - **データ解析，機械学習**
    - **Intel Xeon Ice Lake + NVIDIA A100, 7.2 PF**
      - **45ノード（Ice Lake：90基，A100：360基）, IB-HDR**
    - **一部は外部リソース（ストレージ，サーバー，センサーネットワーク他）に直接接続**
  - ファイルシステム：共有（大容量）＋高速

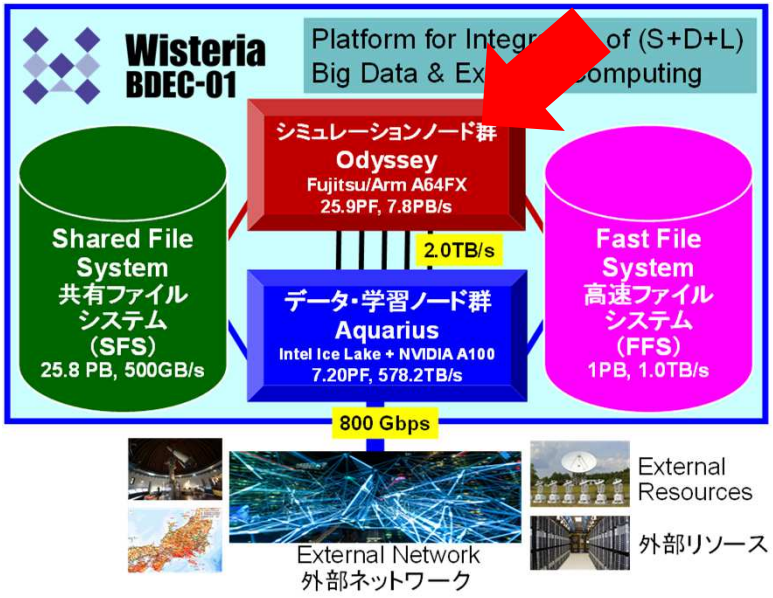


Wisteria
BDEC-01

Wisteria/BDEC-01

BDEC:「計算・データ・学習（S+D+L）」融合のためのプラットフォーム（Big Data & Extreme Computing）

- 2021年5月14日運用開始
  - 東京大学柏Ⅱキャンパス
- 33.1 PF, 8.38 PB/sec. , **富士通製**
  - ~4.5 MVA（空調込み）, ~360m^2
- Hierarchical, Hybrid, Heterogeneous (h3)
- **2種類のノード群**
  - **シミュレーションノード群（S, SIM）：Odyssey**
    - **従来のスパコン**
    - **Fujitsu PRIMEHPC FX1000 (A64FX), 25.9 PF**
      - **7,680ノード（368,640 コア）, 20ラック, Tofu-D**
  - **データ・学習ノード群（D/L, DL）：Aquarius**
    - **データ解析，機械学習**
    - **Intel Xeon Ice Lake + NVIDIA A100, 7.2 PF**
      - **45ノード（Ice Lake：90基，A100：360基）, IB-HDR**
    - **一部は外部リソース（ストレージ，サーバー，センサーネットワーク他）に直接接続**
  - ファイルシステム：共有（大容量）＋高速

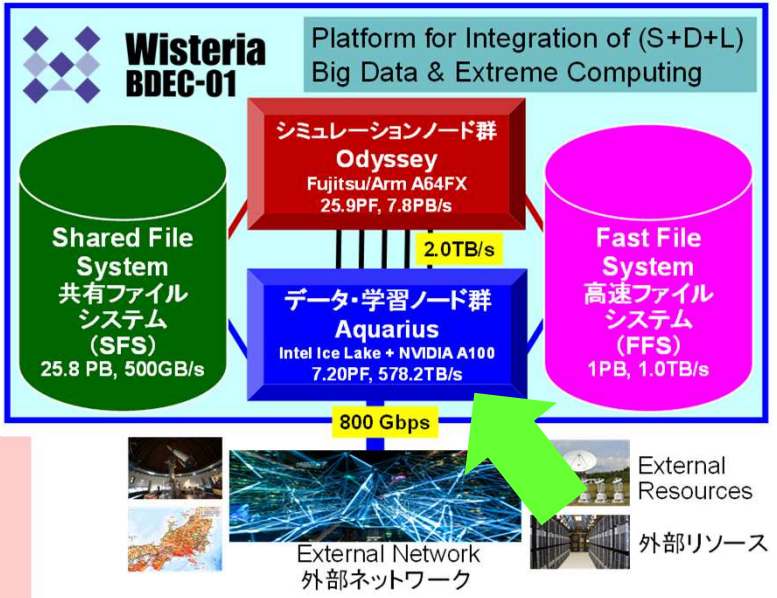


Wisteria
BDEC-01

Wisteria/BDEC-01

BDEC:「計算・データ・学習（S+D+L）」融合のためのプラットフォーム（Big Data & Extreme Computing）

- 2021年5月14日運用開始
  - 東京大学柏Ⅱキャンパス
- 33.1 PF, 8.38 PB/sec.，**富士通製**
  - ~4.5 MVA（空調込み），~360m^2
- Hierarchical, Hybrid, Heterogeneous (h3)
- **2種類のノード群**
  - **シミュレーションノード群（S, SIM）：Odyssey**
    - **従来のスパコン**
    - **Fujitsu PRIMEHPC FX1000 (A64FX), 25.9 PF**
      - **7,680ノード（368,640 コア），20ラック，Tofu-D**
  - **データ・学習ノード群（D/L, DL）：Aquarius**
    - **データ解析，機械学習**
    - **Intel Xeon Ice Lake + NVIDIA A100, 7.2 PF**
      - **45ノード（Ice Lake：90基，A100：360基），IB-HDR**
    - **一部は外部リソース（ストレージ，サーバー，センサーネットワーク他）に直接接続**
  - ファイルシステム：共有（大容量）＋高速



Wisteria
BDEC-01

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

計算科学コード

シミュレーション
ノード群，Odyssey

最適化されたモデル，
パラメータ

計算結果

Wisteria/BDEC-01

機械学習，DDA

観測データ

データ・学習ノード群
Aquarius

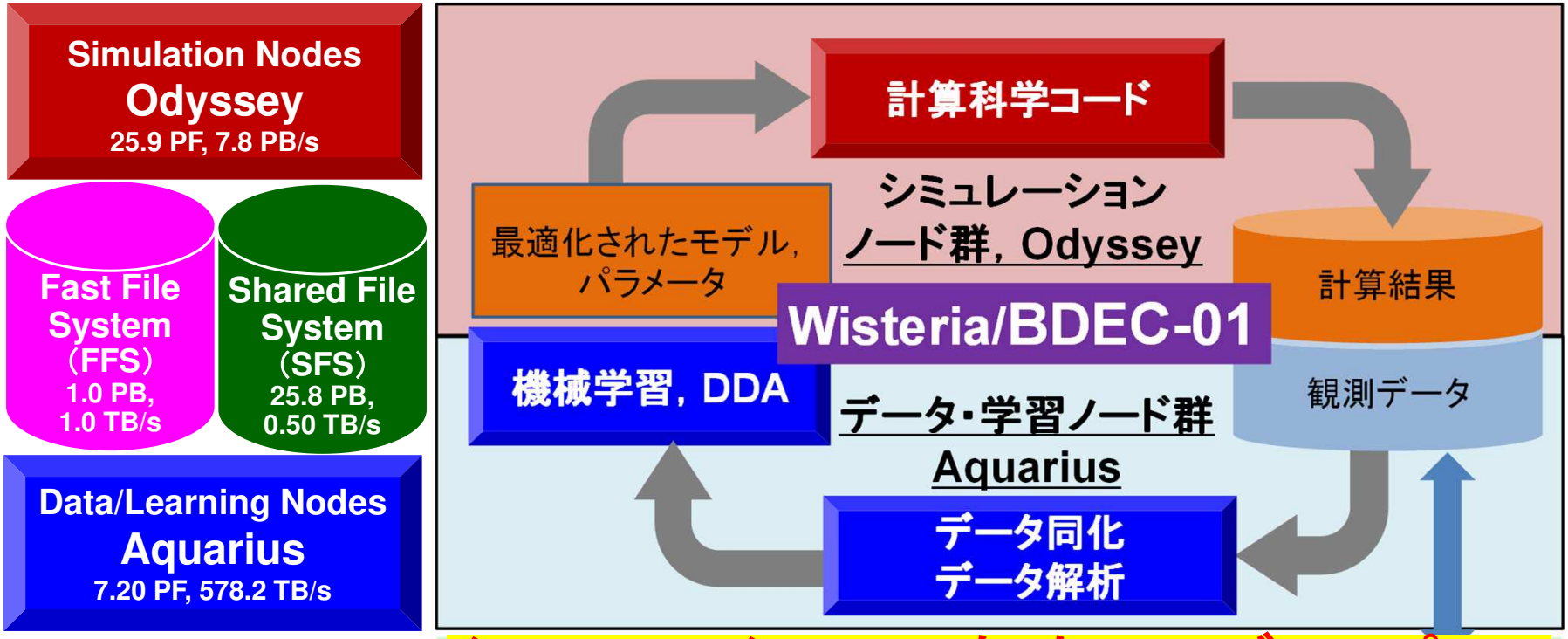
データ同化
データ解析

Wisteria
BDEC-01

サーバー
ストレージ
DB
センサー群
他

外部ネットワーク

外部
リソース



Wisteria
BDEC-01

シミュレーションのためのモデル・パラメータのデータ解析，AI/機械学習による最適化（S+D+L）

h3-Open-BDEC

「計算＋データ＋学習」融合を実現する革新的ソフトウェア基盤
科研費基盤研究（S）（2019年度～23年度，代表：中島研吾）
https://h3-open-bdec.cc.u-tokyo.ac.jp/

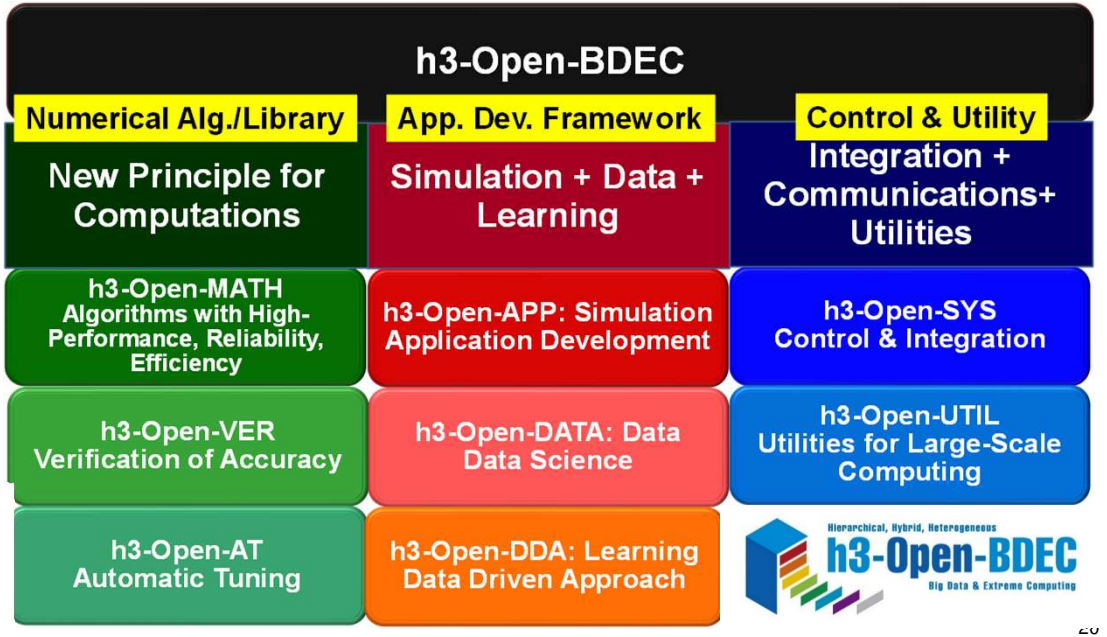
Hierarchical, Hybrid, Heterogeneous

Big Data & Extreme Computing

① 変動精度演算・精度保証・自動チューニングによる新計算原理に基づく革新的数値解法

② 階層型データ駆動アプローチ等に基づく革新的機械学習手法

③ ヘテロジニアス環境（e.g. Wisteria/BDEC-01）におけるソフトウェア，ユーティリティ群



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

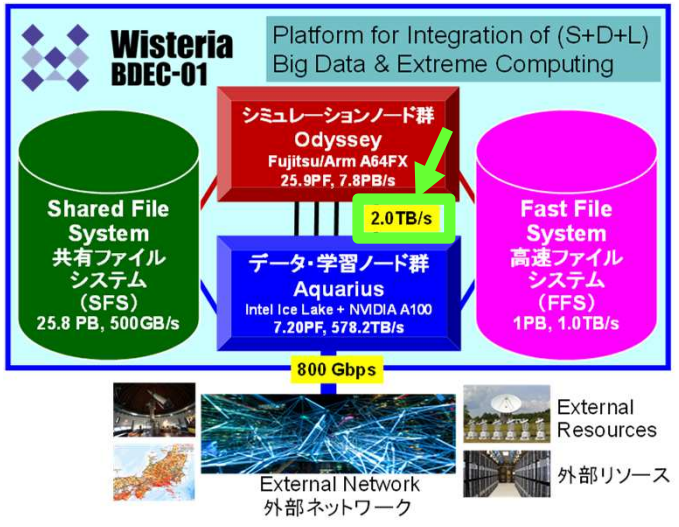
Hierarchical, Hybrid, Heterogeneous
h3-Open-BDEC
Big Data & Extreme Computing

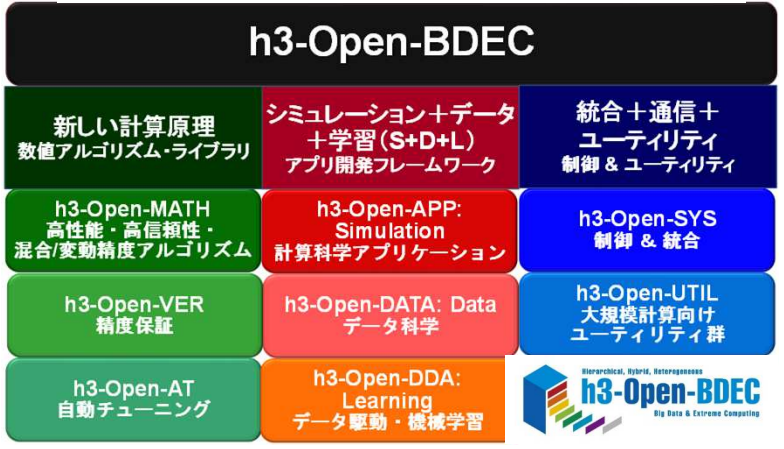
AI for HPC, AI for Science の実現へ向けて

Wisteria BDEC-01

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

- Odyssey-Aquarius連携
  - MPIによる通信は不可
    - O-Aを跨いでMPIプログラムは動かない
  - Odyssey-Aquarius間はInfiniband-EDR（2TB/sec）で結合されている
- ソフトウェア開発
  - 高機能カプラー：h3-Open-UTIL/MP
  - O-A間通信：h3-Open-SYS/WaitIO
    - IB-EDR経由（WaitIO-Socket）
    - 高速ファイルシステム（FFS）経由連携（WaitIO-File）

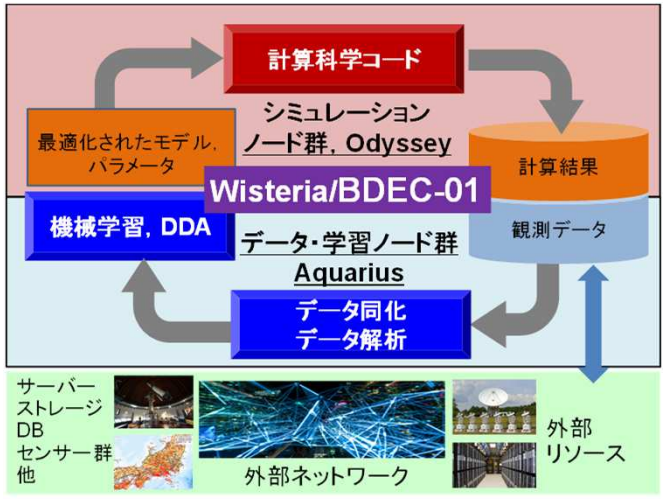


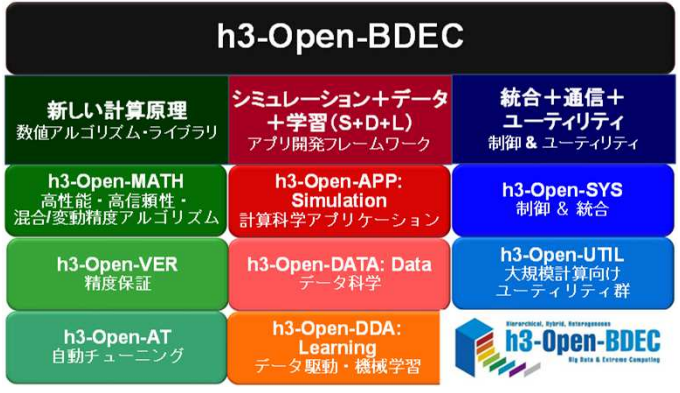


h3-Open-SYS/WaitIO

データ受け渡しライブラリ〔松葉，2020〕〔住元他，HPC-181，2021〕

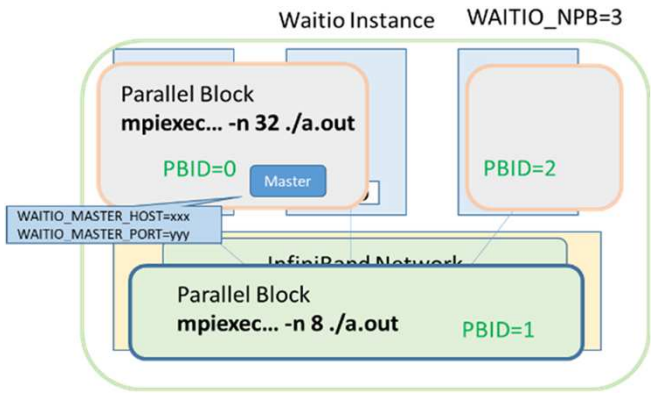
- ヘテロジニアス環境下での異なるコンポーネント間ファイル経由連携ライブラリとして考案
- 機能
  - ✓ Odyssey～Aquarius間連携
    - IB-EDR経由通信（WaitIO-Socket）
    - ファイル経由（WaitIO-File）
  - ✓ 外部からのデータ取得（観測データ等）
  - ✓ 読み込み・書き出しの同期
- API：C/C++，Fortranから呼び出し可能
  - ✓ MPIライクなインタフェースを提供





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

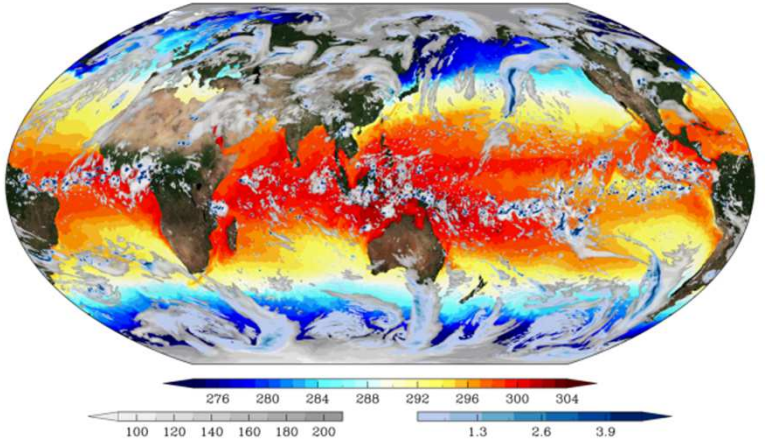
WaitIO API	Description
waitio_isend	Non-Blocking Send
waitio_irecv	Non-Blocking Receive
waitio_wait	Termination of waitio_isend/irecv
waitio_init	Initialization of WaitIO
waitio_get_nprocs	Process # for each PB (Parallel Block)
waitio_create_group waitio_create_group_wranks	Creating communication groups 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

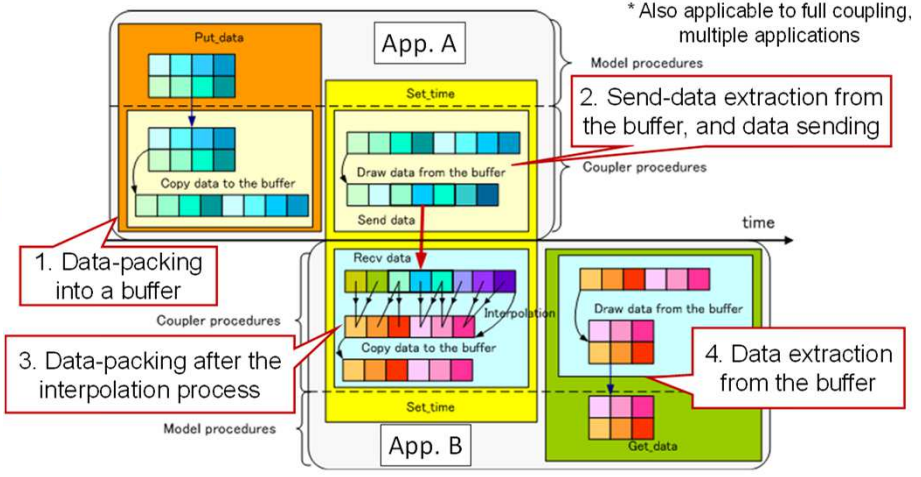


[Sumimoto et al. 2021]

連成シミュレーションのためのカプラー〔荒川，八代〕

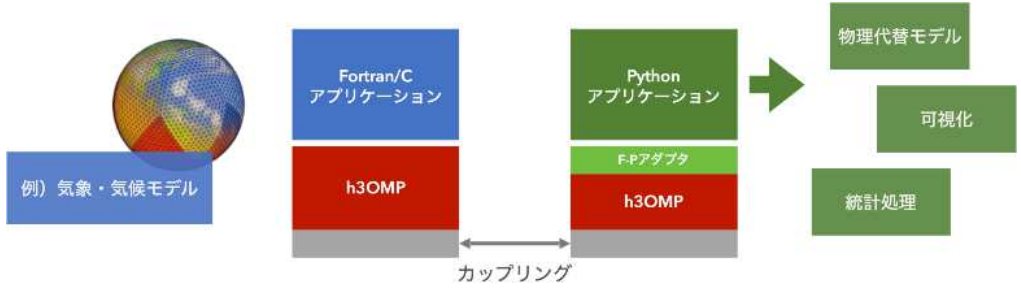
- 従来のカプラー（Coupler）：ppOpen-MATH/MP
  - 複数（通常2つ：大気（NICAM）＋海洋（COCO））のアプリケーションの弱連成（Weak Coupling）をサポート
  - 各アプリケーションは1種類の計算をやる





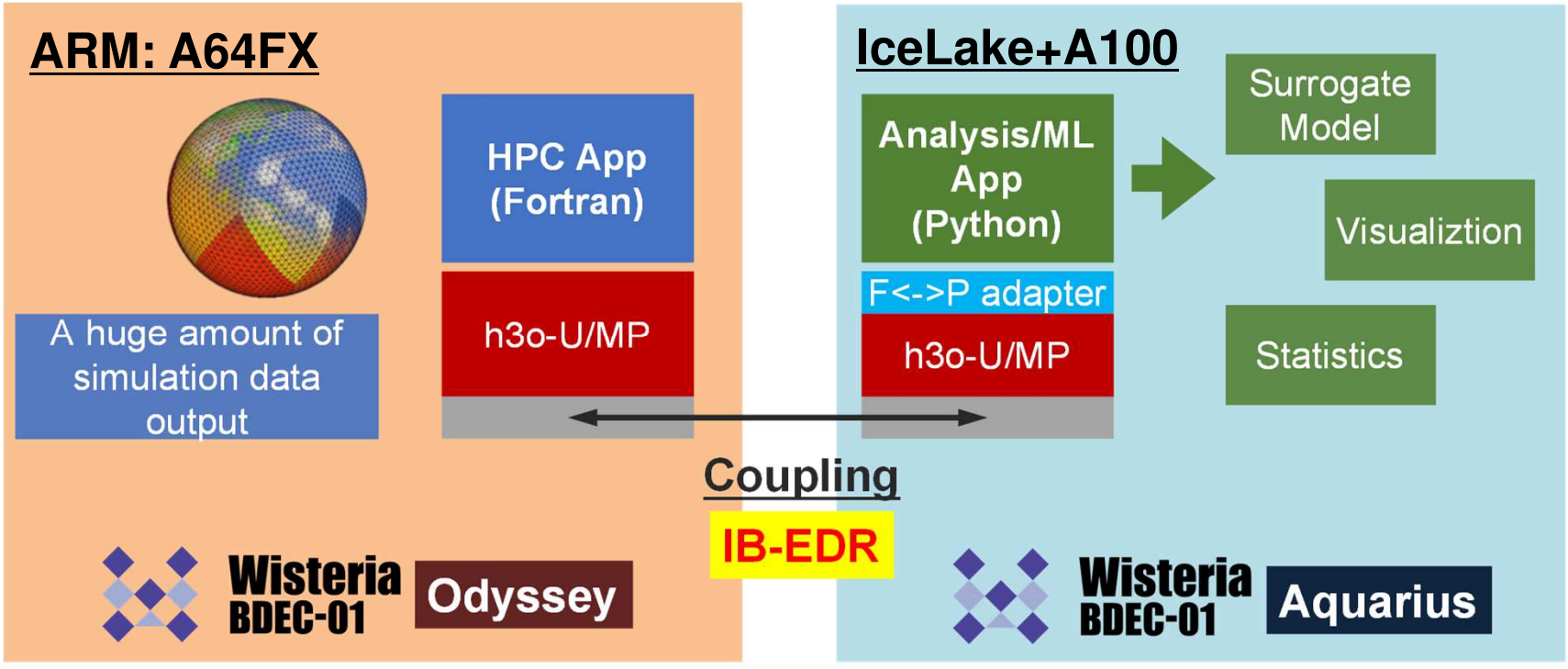
「計算＋データ＋学習」融合を支援する多機能カプラーh3-Open-UTIL/MP

- 異なる物理モデル連成のアンサンブル実行を支援・統合するための機能
  - MPI通信、時刻同期、格子系間マッピング等の管理機能の他、従来のカプラーには無い、複数の弱連成結合シミュレーションのアンサンブル実行、片側のモデルのみをアンサンブル実行する多対1の弱連成結合が可能
  - スパコン上で、全地球大気海洋連成シミュレーションによって動作検証済み
- Fortran/Cコード（物理モデル）とPythonコードの弱連成を実現する機能
  - FortranやCで記述されたプログラム同士の連成計算に限って開発を行ってきたカプラーを、Pythonによって記述されたAI・機械学習、可視化処理系のワークロードからも活用できるよう機能拡充。



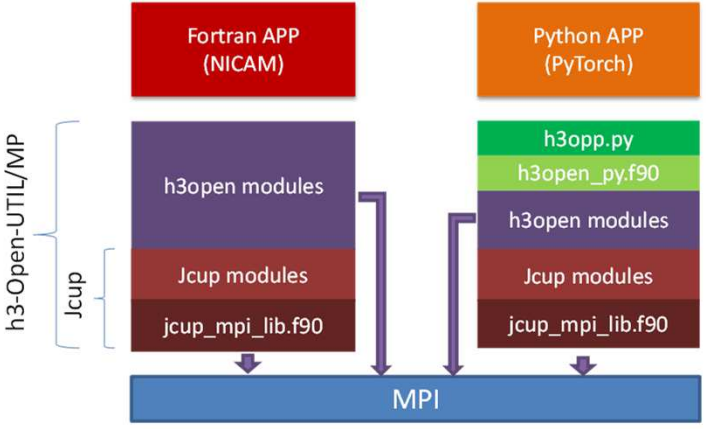
Fortran/CアプリとPythonアプリの連成計算の模式図
〔八代・荒川　2020〕

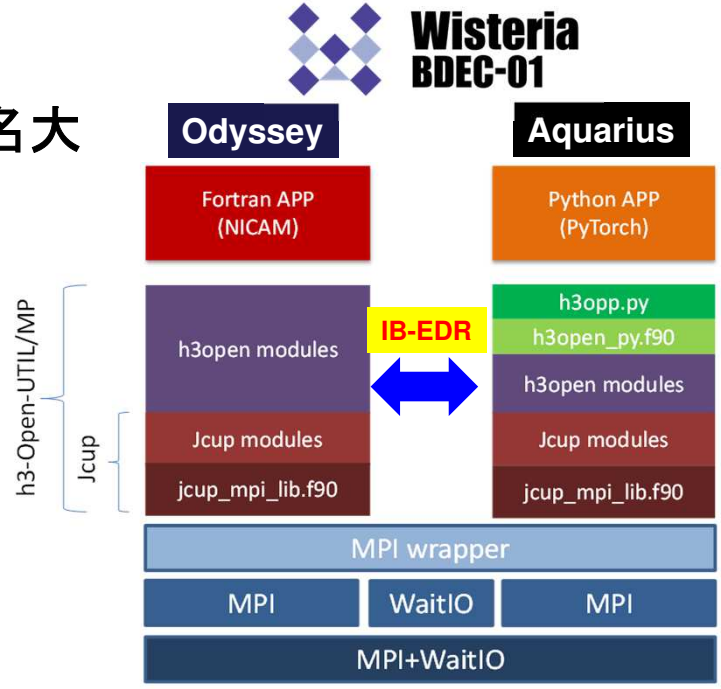
h3-Open-UTIL/MP (h3o-U/MP) + h3-Open-SYS/WaitIO-Socket



h3-Open-UTIL/MP・h3-Open-SYS/WaitIO-Socket連携2022年6月から利用可能

2022年度はFS経由のWaitIO-File整備：名大





2021年4月：MPI通信可能な環境を前提

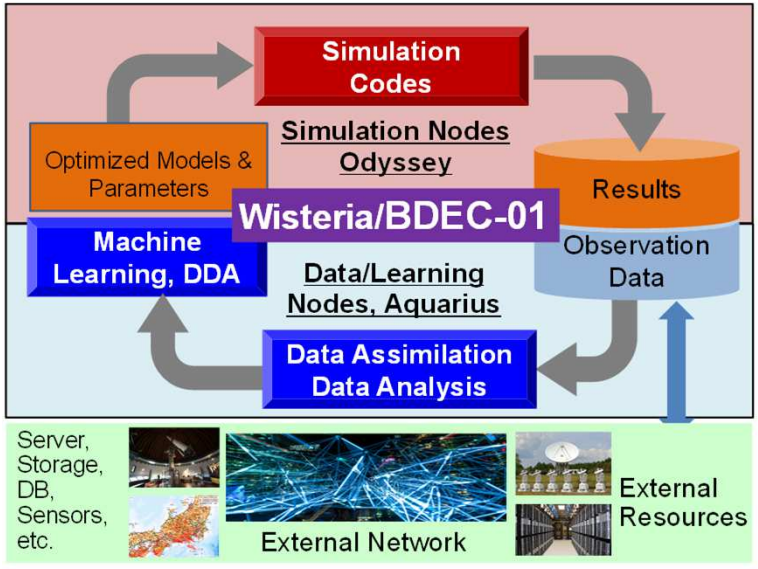
2022年6月：Coupler＋WaitIO

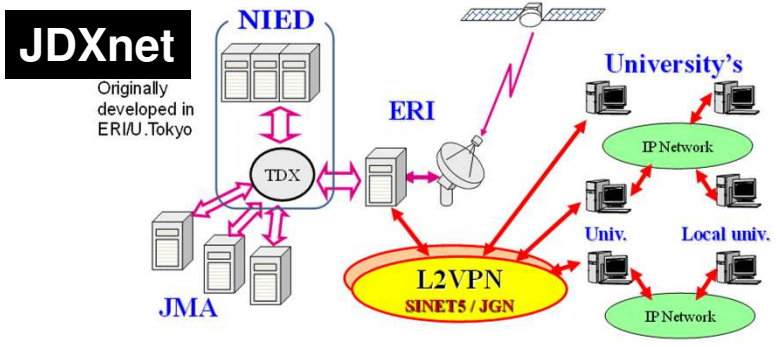
解説記事：h3-Open-UTIL/MP・h3-Open-SYS/WaitIO-Socket

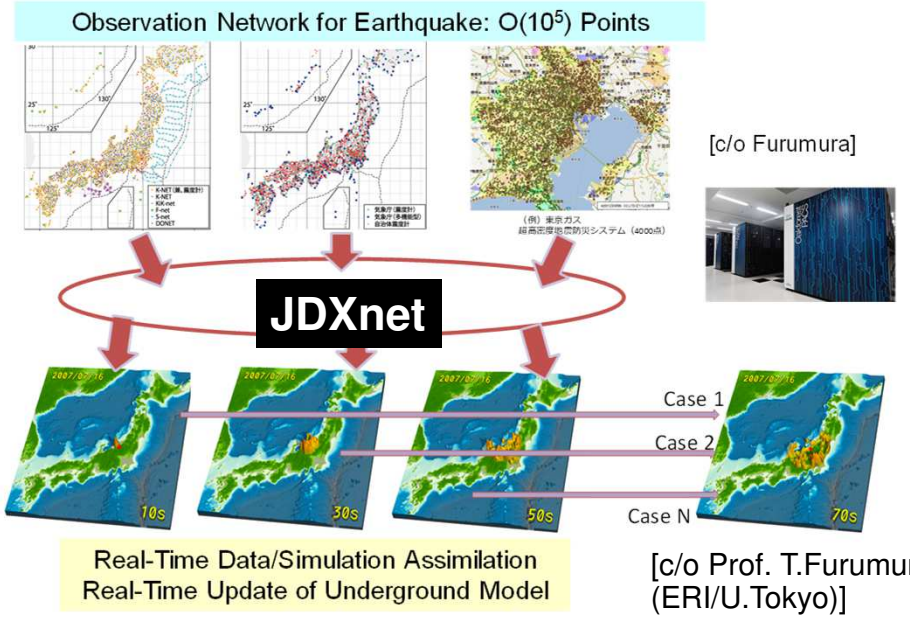
- h3-Open-UTIL/MP
  - https://www.cc.u-tokyo.ac.jp/public/VOL24/No3/13_202205-Wisteria-2.pdf
  - http://nkl.cc.u-tokyo.ac.jp/files/202207UtilMPfinal.pdf
- h3-Open-SYS/WaitIO-Socket
  - https://www.cc.u-tokyo.ac.jp/public/VOL24/No2/10_202203Wisteria-1.pdf
  - https://www.cc.u-tokyo.ac.jp/public/VOL24/No3/12_202205-Wisteria-1.pdf

リアルタイムデータ同化＋3D強震動シミュレーション融合

JDXnetによるリアルタイム観測データ活用

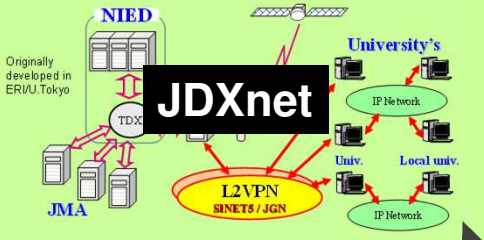






[c/o Prof. T.Furumura (ERI/U.Tokyo)]

長周期地震動シミュレーション＋観測データ同化



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

Filtered Results

WaitIO-Socket

Comp. Results

Seism3D/ OpenSWPC-DAF

Mesh Data

UG Model

Communications by WaitIO-Socket

[Kasai et al. 2021]

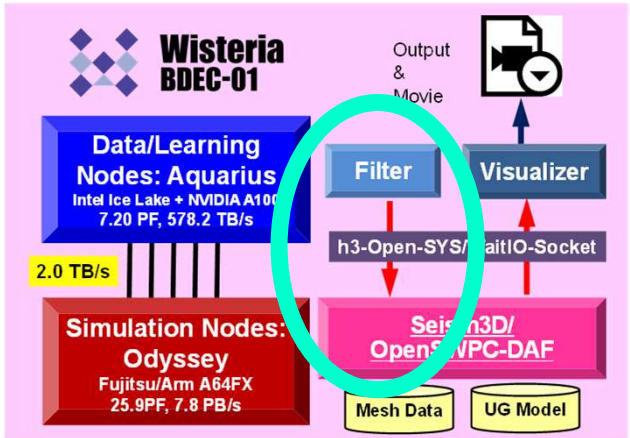
Aquarius: SEND

```
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
```

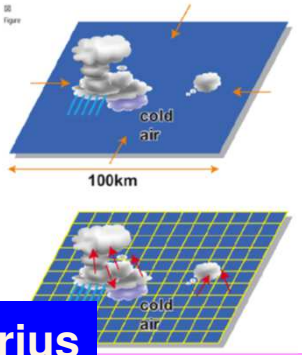
Odyssey: RECV

```
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,...)
```



Atmosphere-ML Coupling
[Yashiro (NIES), Arakawa (ClimTech/U.Tokyo)]

- Motivation of this experiment
  - Tow types of Atmospheric models: Cloud resolving VS Cloud parameterizing
  - Could resolving model is difficult to use for climate simulation
  - Parameterized model has many assumptions
  - Replacing low-resolution cloud processes calculation with ML!



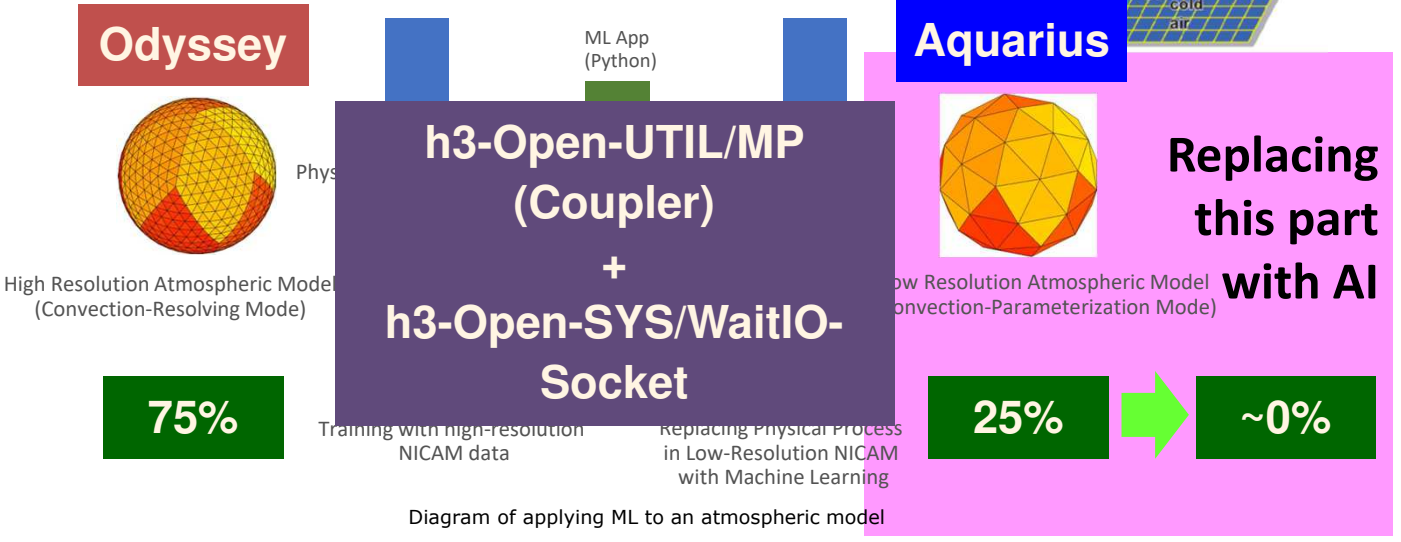
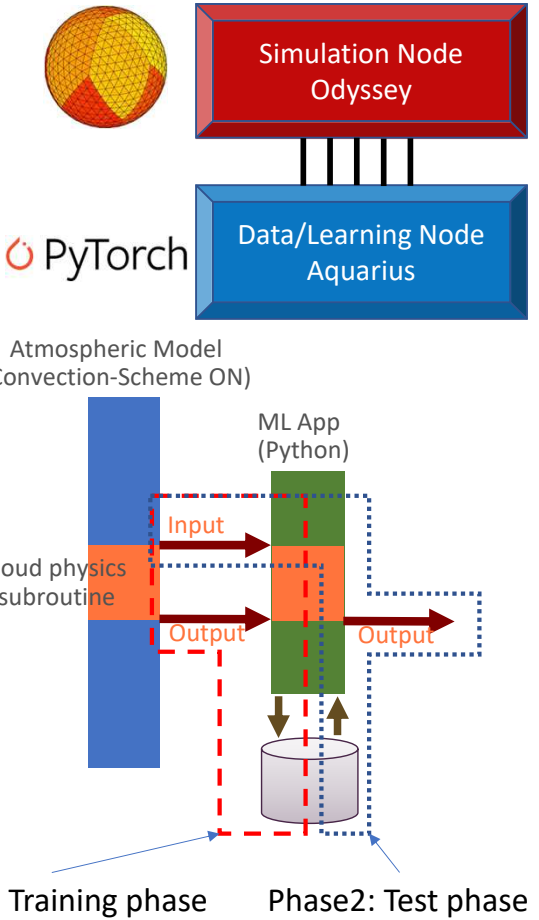


Diagram of applying ML to an atmospheric model

Experimental Design

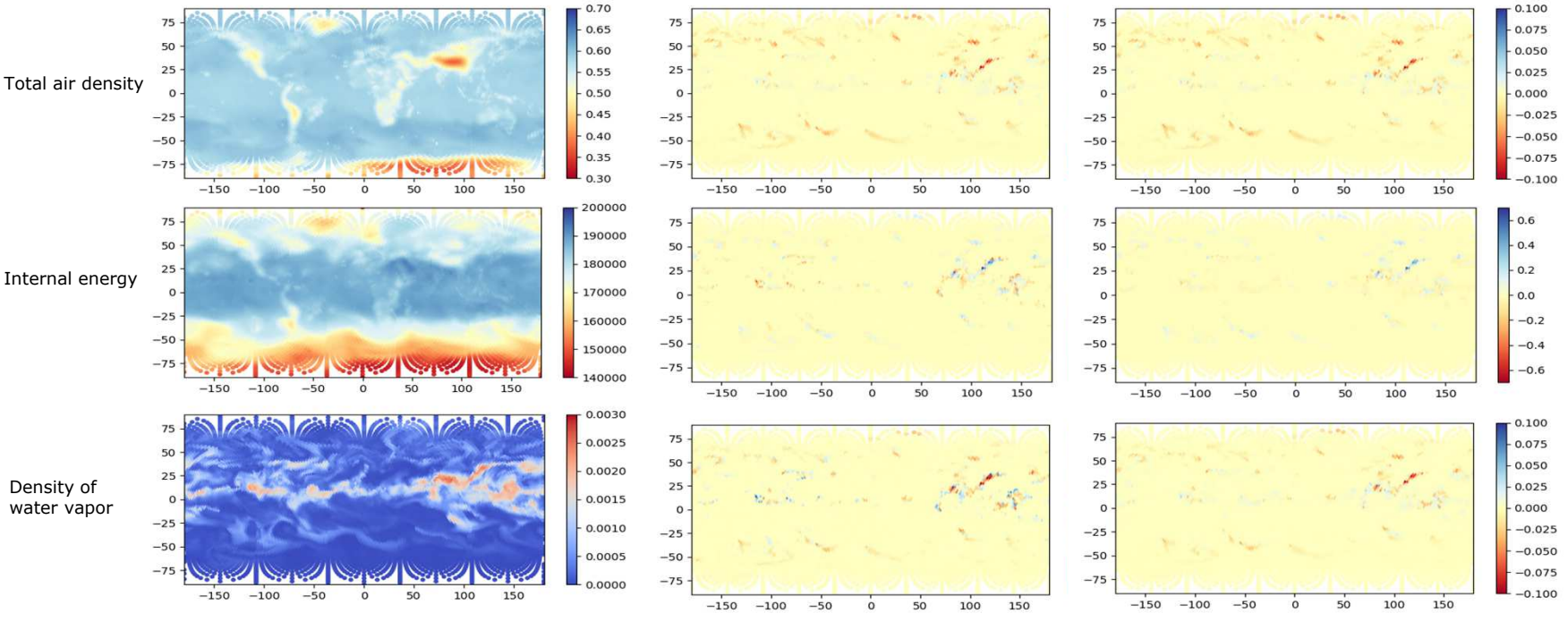
- Atmospheric model on Odyssey
  - NICAM : global non-hydrostatic model with an icosahedral grid
  - Resolution : horizontal : 10240, vertical : 78
- ML on Aquarius
  - Framework : PyTorch
  - Method : Three-Layer MLP
  - Resolution : horizontal : 10240, vertical : 78
- Experimental design
  - Phase1: 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

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



Test calculation

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



Examples of Scripts [Sumimoto, Arakawa]

Odyssey for Simulation

```
#!/bin/bash
#PJM -N "test_waitio"
#PJM -L rscgrp=coupler-lec-o
#PJM -L node=10:noncont
#PJM --mpi proc=80
#PJM -L elapse=00:10:00
#PJM -g gt00
#PJM -j
#PJM -e err

module load fj
module load fjmpi
module load waitio

export WAITIO_MASTER_HOST=`hostname`
export WAITIO_MASTER_PORT=7100
export WAITIO_PBID=0
export WAITIO_NPB=2

hostname
waitio-serv-a64fx -d -m $WAITIO_MASTER_HOST

#mpiexec -oferr-proc errnicam -np 160 ./nicam
mpiexec -np 80 ./nicam
```

Aquarius for AI

```
#!/bin/bash
#PJM -N "test_waitio"
#PJM -L rscgrp=coupler-lec-a
#PJM -L node=1
#PJM --mpi proc=10
#PJM -L elapse=00:10:00
#PJM -g gt00
#PJM -j
#PJM -e err

module unload aquarius
module unload gcc ompi
module load intel
module load impi
module load waitio

export WAITIO_MASTER_HOST=`waitio-serv -c`
export WAITIO_MASTER_PORT=7100
export WAITIO_PBID=1
export WAITIO_NPB=2

module unload intel
module unload impi
module load gcc ompi

mpiexec -n 10 ./ada
```

Lessons Learned & Future

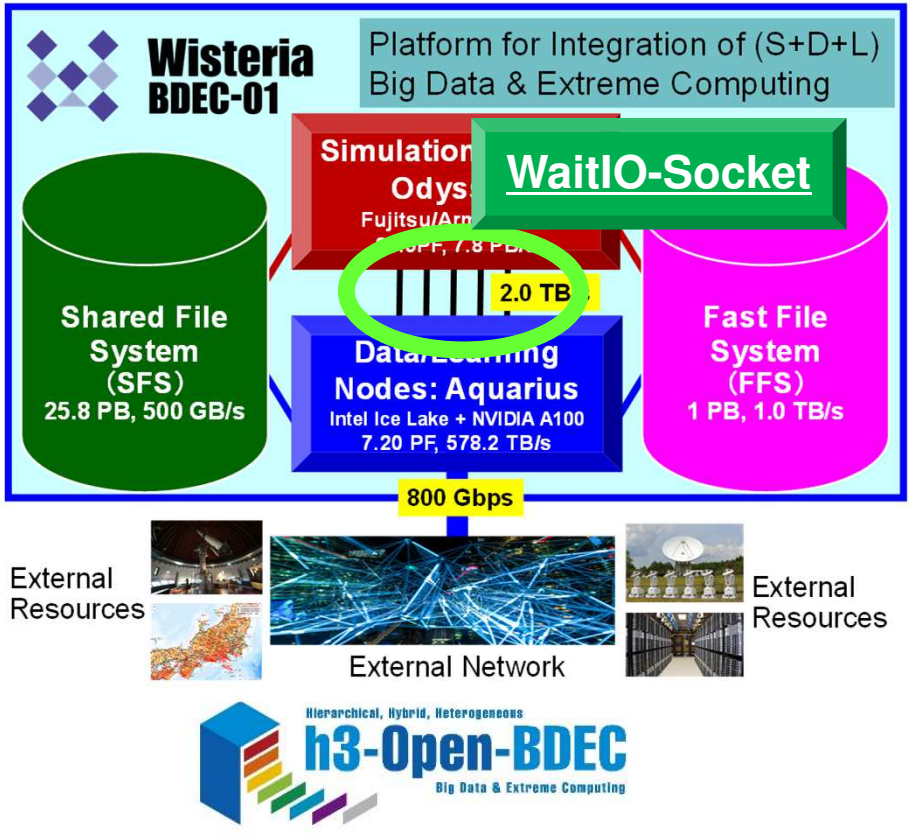
SC22 BoF 131

Disaggregated Heterogeneous Architectures

- **Software (h3-Open-BDEC: WaitIO, Coupler) enabled integration of (S+D+L) on Odyssey-Aquarius**
  - **WatioIO-Socket/File**
  - **Job Submission System**
- Policy for Operation
  - Current status is preliminary, Very few workloads for (S+D+L)
  - More flexible (& complicated) policy needed
- **Please contact us at nakajima@cc.u-tokyo.ac.jp if you are interested in installing and using our software on your systems !**
  - **to be installed at systems of Nagoya U. and Kyushu U.**
- Publication
  - Shinji Sumimoto, et al. PDCAT'22, Dec.7-9, 2022 (in press)
    - https://www.hpc.is.tohoku.ac.jp/pdcat2022/

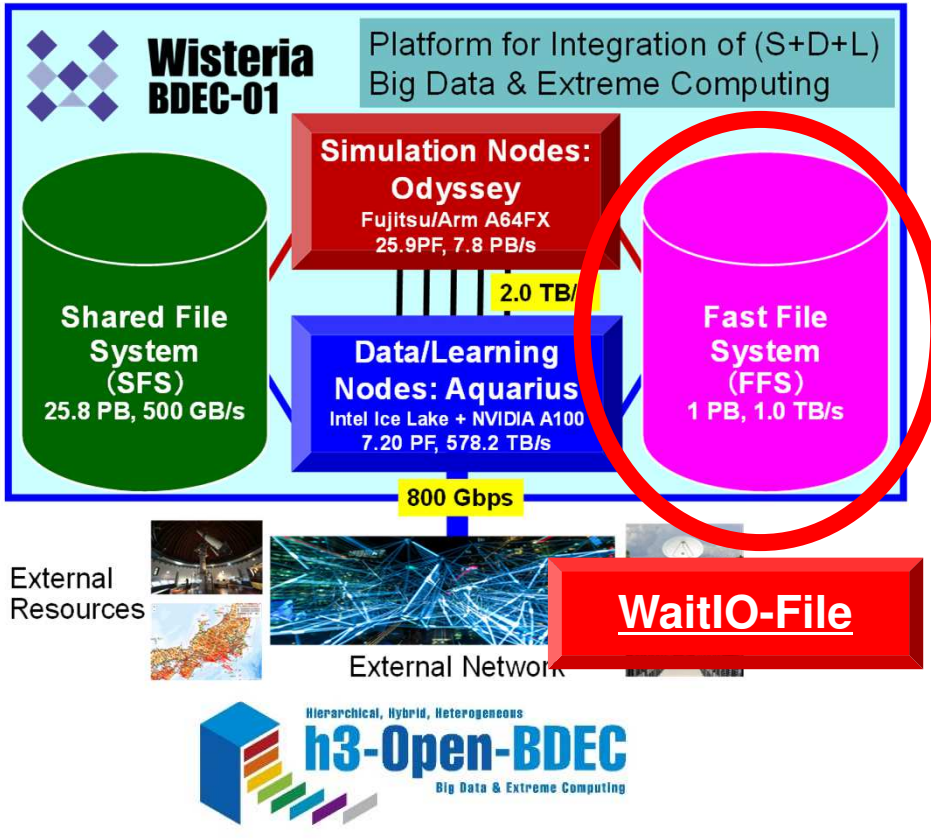
h3-Open-SYS/WaitIO-Socket

- Wisteria/BDEC-01
  - Aquarius (GPU: NVIDIA A100)
  - Odyssey (CPU: A64FX)
- Combining Odyssey-Aquarius
  - Single MPI Job over O-A is impossible
- **Connection between Odyssey-Aquarius**
  - **IB-EDR with 2TB/sec.**
  - **Fast File System**
  - **h3-Open-SYS/WaitIO-Socket**
    - **Library for Inter-Process Communication through IB-EDR with MPI-like interface**



h3-Open-SYS/WaitIO-File

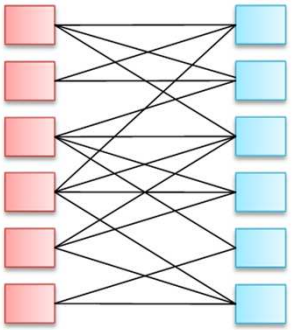
- Wisteria/BDEC-01
  - Aquarius (GPU: NVIDIA A100)
  - Odyssey (CPU: A64FX)
- Combining Odyssey-Aquarius
  - Single MPI Job over O-A is impossible
- **Connection between Odyssey-Aquarius**
  - **IB-EDR with 2TB/sec.**
  - **Fast File System**
  - **h3-Open-SYS/WaitIO-File**
    - **Library for Inter-Process Communication through FFS with MPI-like interface**

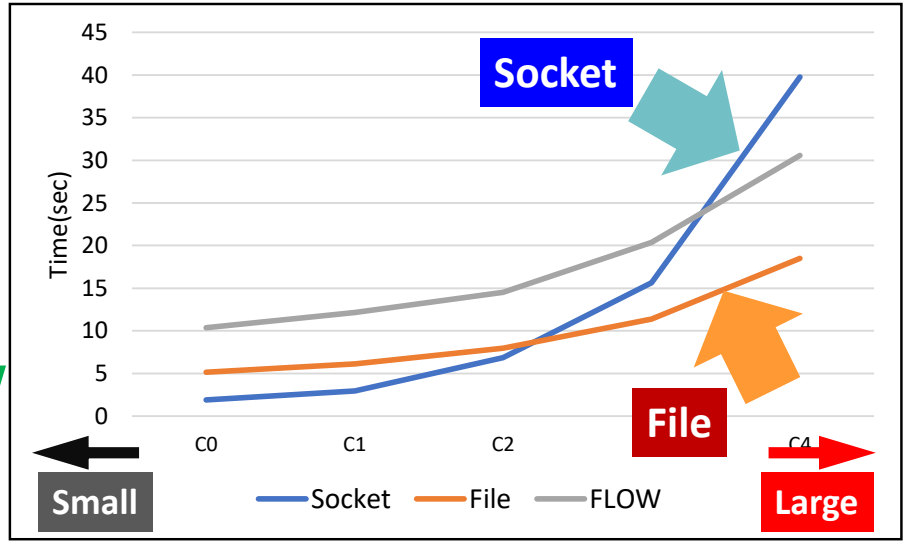


Preliminary Evaluation: WaitIO-Socket, -File

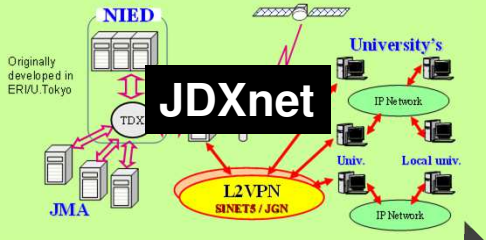
- Odyssey-Aquarius, Flow (Nagoya University)
- Test Cases
  - Model A: 160 proc's/40 nodes: A64FX (Odyssey, Flow I)
  - Model B: 20 proc's/1 node: Intel Xeon (Aquarius, Flow II)
    - Waitio_isend/irecv/wait
- **Results**
  - **"Socket" is better for small cases, "File" is better for large ones**
  - "Flow" with WaitIO-Files is 2x slower than Odyssey-Aquarius
    - Reasonable Number

Case	Grid #	Total Size
C0	81,920	26,214,400
C1	163,840	52,428,800
C2	327,680	104,857,600
C3	655,360	209,715,200
C4	1,310,720	419,430,400





長周期地震動シミュレーション＋観測データ同化



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

Filtered Results

WaitIO-Socket

Comp. Results

Seism3D/ OpenSWPC-DAF

Mesh Data

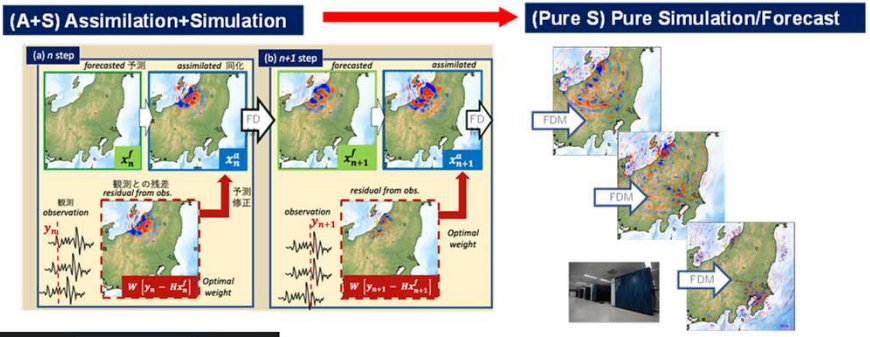
UG Model

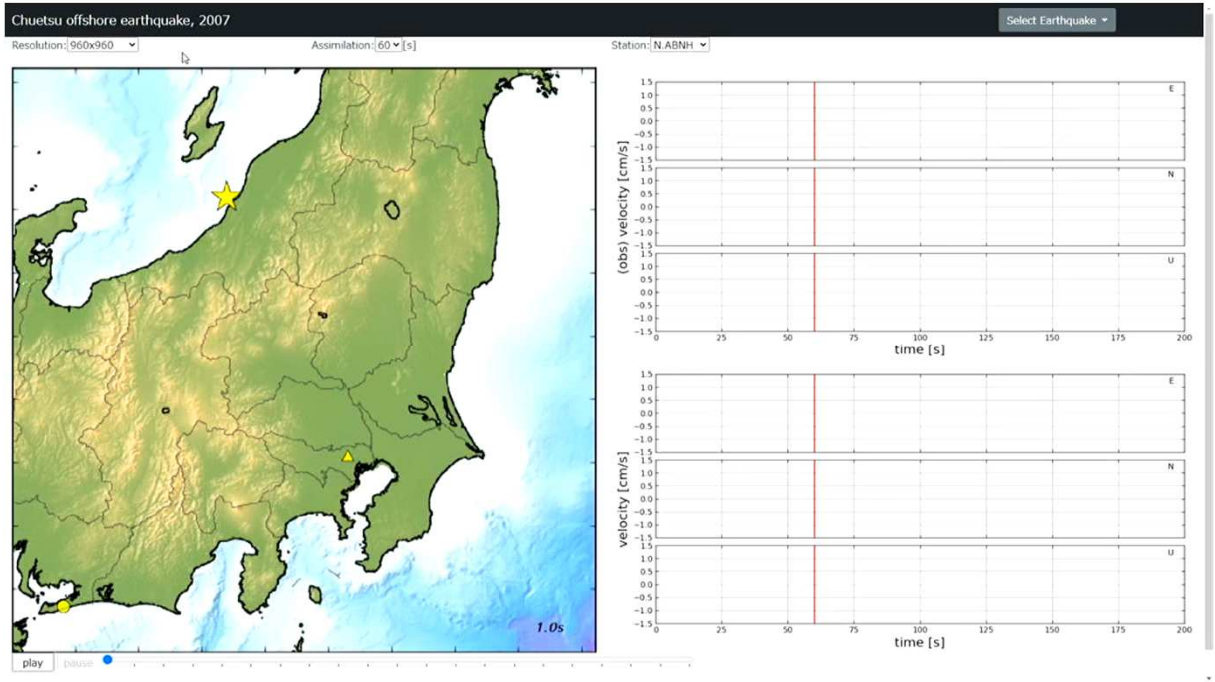
Webベース シミュレーション体験・データ利活用システム

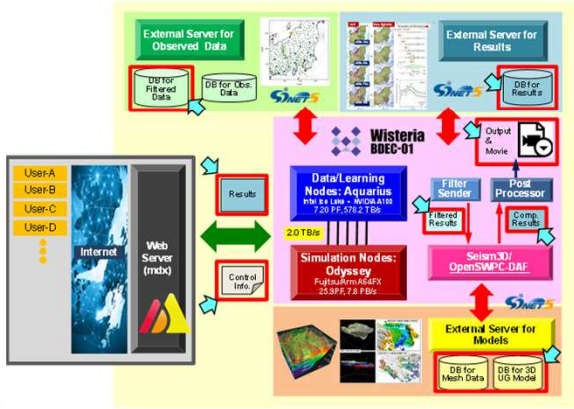
mdxとの連携事例

- 「3D長周期地震動＋リアルタイムデータ同化」融合シミュレーションシステムの「防災・減災」啓蒙・教育へ向けた利用・展開を図るため，Webベースのシミュレーション体験・データ利活用環境を構築（2022年度）
- 利用者はWeb Server（mdx上）にアクセスし，スパコン（Wisteria/BDEC-01）上でのシミュレーションの実施，計算結果，観測結果の可視化処理，表示等を行う。
- Web経由でデータ群をスパコン上で処理するフレームワークは様々なアプリケーションへの転用が可能

mdxからWeb経由で大規模シミュレーション・データ同化をインタラクティブに実行

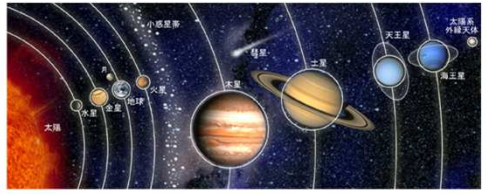


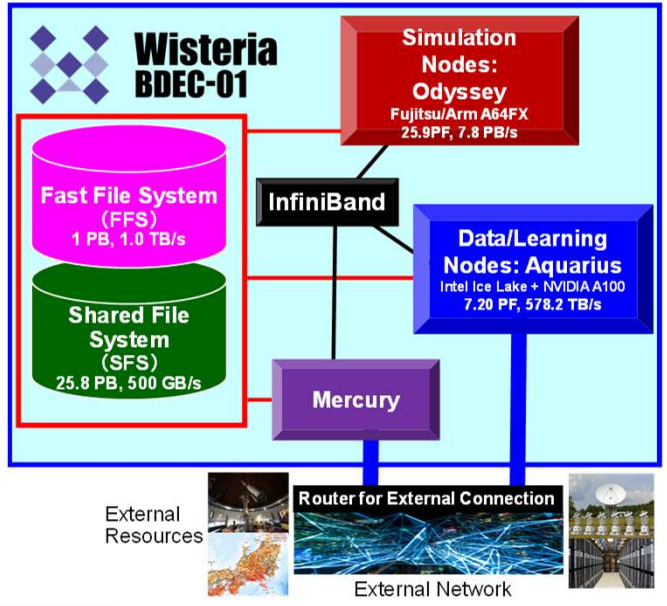




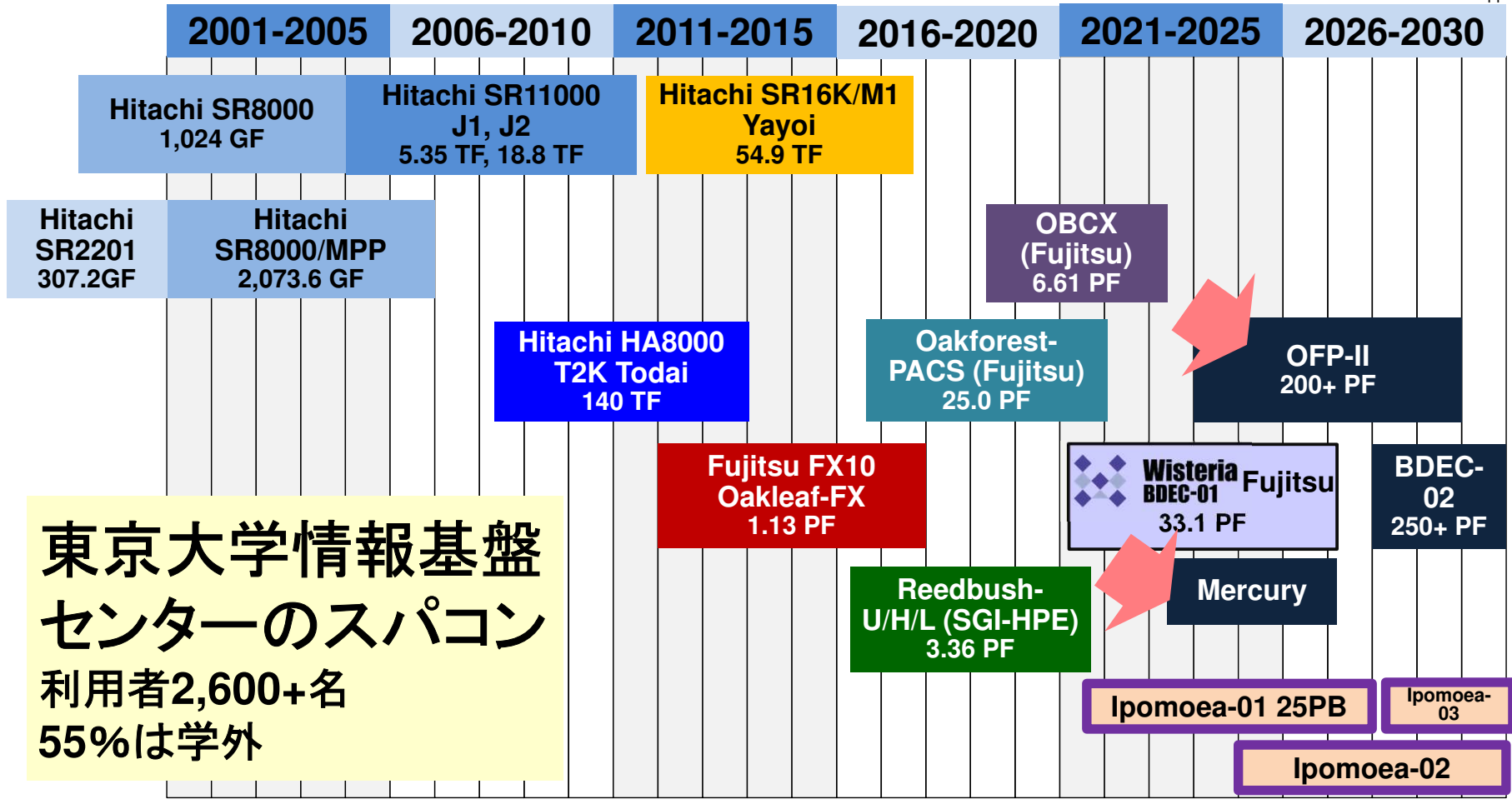
将来構想：OFP-II，Mercury

- スパコンへの性能要求，省電力，脱炭素化⇒演算加速器搭載は不可避
- Wisteria-Mercury（2023年11月～）
  - GPUクラスタ
  - OFP-IIプロトタイプ
- OFP-II（2024年4月）
  - OFP後継機（JCAHPC：筑波大学と共同），200+PF
  - GPUクラスタ
    - Mercuryと同じ，もしくはその後継GPU
  - **CPU Onlyノード群あり（GPUホストと異なる可能性）**
- **アプリケーションの移植が必要（3,000人）**
  - **講習会，GPUミニキャンプによる対応**





JCAHPC 筑波大学 University of Tsukuba 東京大学 THE UNIVERSITY OF TOKYO



東京大学情報基盤センターのスパコン

利用者2,600+名

55%は学外

スケジュール概要

Mercury & OFP-II

- GPU移行のための諸作業は遅くとも，2022年秋に始める必要がある
- それ以前にMercury・OFP-IIに搭載するGPU（両者は同じ）を決める必要あり
- 2022年2月～3月
  - プリベンチマークを各社に依頼（計算科学系7種類，Fortran，C）
- 2022年6月
  - GPUベンダーを決定
  - ポイント：性能，ポーティングのしやすさ，サポート体制，Fortranへの対応
- 2022年秋～
  - ポーティング開始，当初はAquariusをプラットフォームとして使用（多分12月頃）
  - OFP-II資料招請開始（2022年11月8日：導入説明会）
- 2023年秋～
  - Mercuryを使用した最適化，評価
- 2024年4月：OFP-II運用開始

MPI+OpenMPで並列化されたFortranプログラムのGPUへの移行手法（オンライン）

https://www.cc.u-tokyo.ac.jp/events/lectures/196/

- 趣旨
  - 本講習会ではOFP-IIへのプログラム移植に向け，既存のMPI+OpenMPでCPU向けに並列化されたFortranプログラムのGPU環境への移植手法を学ぶ
  - 受講料は無料
  - GPUへの移行は基本的にOpenMP（CPU向け）で並列化されたループを対象として実施
  - GPU向けのループ並列化の手段としてはOpenACC, OpenMP 5.x, do concurrent, CUDA Fortranなど様々であるため，それぞれの特徴や使い方を学ぶ
  - Wisteria/BDEC-01（Aquarius）を活用し，OpenMPでCPU向けに並列化された有限要素法のプログラムを題材とした，GPUへの移植の演習を行う
- **12月7日午後開催（オンライン）（申込締切：11月30日!!）**

GPUミニキャンプ（オンライン）（1/2）

https://www.cc.u-tokyo.ac.jp/events/lectures/197/

- ミニキャンプとは？
  - 参加者がコードやデータセットを持ち込み、CUDA、OpenACC、Deep Learning など、GPUに関連した課題に対して、メンターからの助言を受けながら、その課題解決に取り組みます。
- メンター
  - エヌビディア合同会社、東大情報基盤センターなどから参加し、参加者の課題解決にご協力します。

- **次回は12月12日～19日に実施（申込締切：11月30日!!）**
  - **各チームで実践中は、ベストエフォートでメンターがQ&A対応**
  - **年度内にもう1回，来年度は4回以上実施の予定**

GPUミニキャンプ（オンライン）（2/2）

https://www.cc.u-tokyo.ac.jp/events/lectures/197/

- 実施形式
  - Zoom と Slack を使ったオンライン形式
- 参加資格
  - 国公私立大学・高専の教員・学生・研究生
  - 研究機関研究員
  - 企業に所属する研究者・技術者（非営利目的に限る）
  - 作業に必要なコードおよびデータセット等をセンターに持ち込める方。
  - コマンドラインによるLinux上での作業やエディタ利用に支障のない方
- 2022年度は，企業利用に発展した事例もあり