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

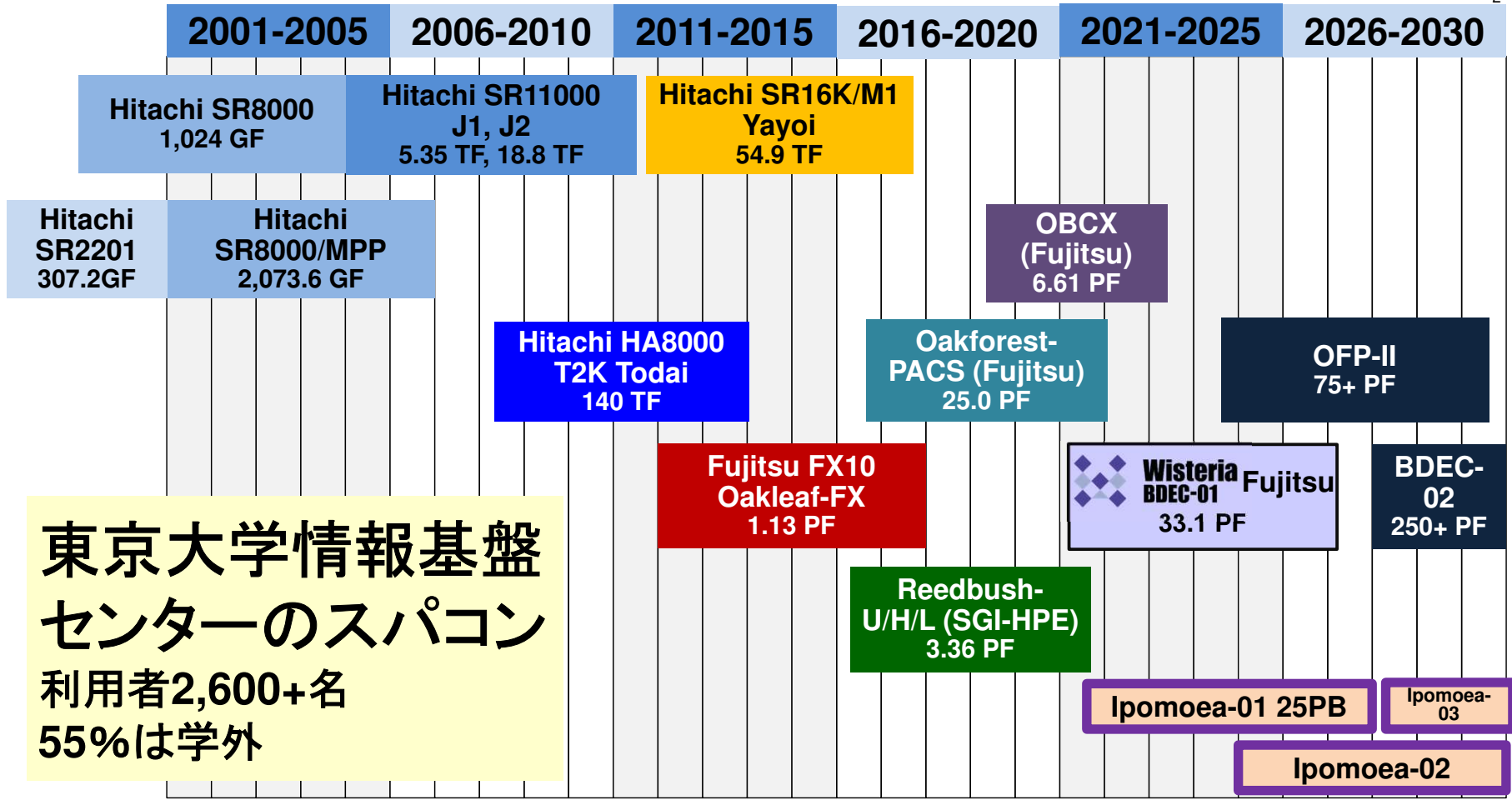
東京大学
THE UNIVERSITY OF TOKYO

Wisteria/BDEC-01

「計算＋データ＋学習」融合へ向けて

Wisteria
BDEC-01

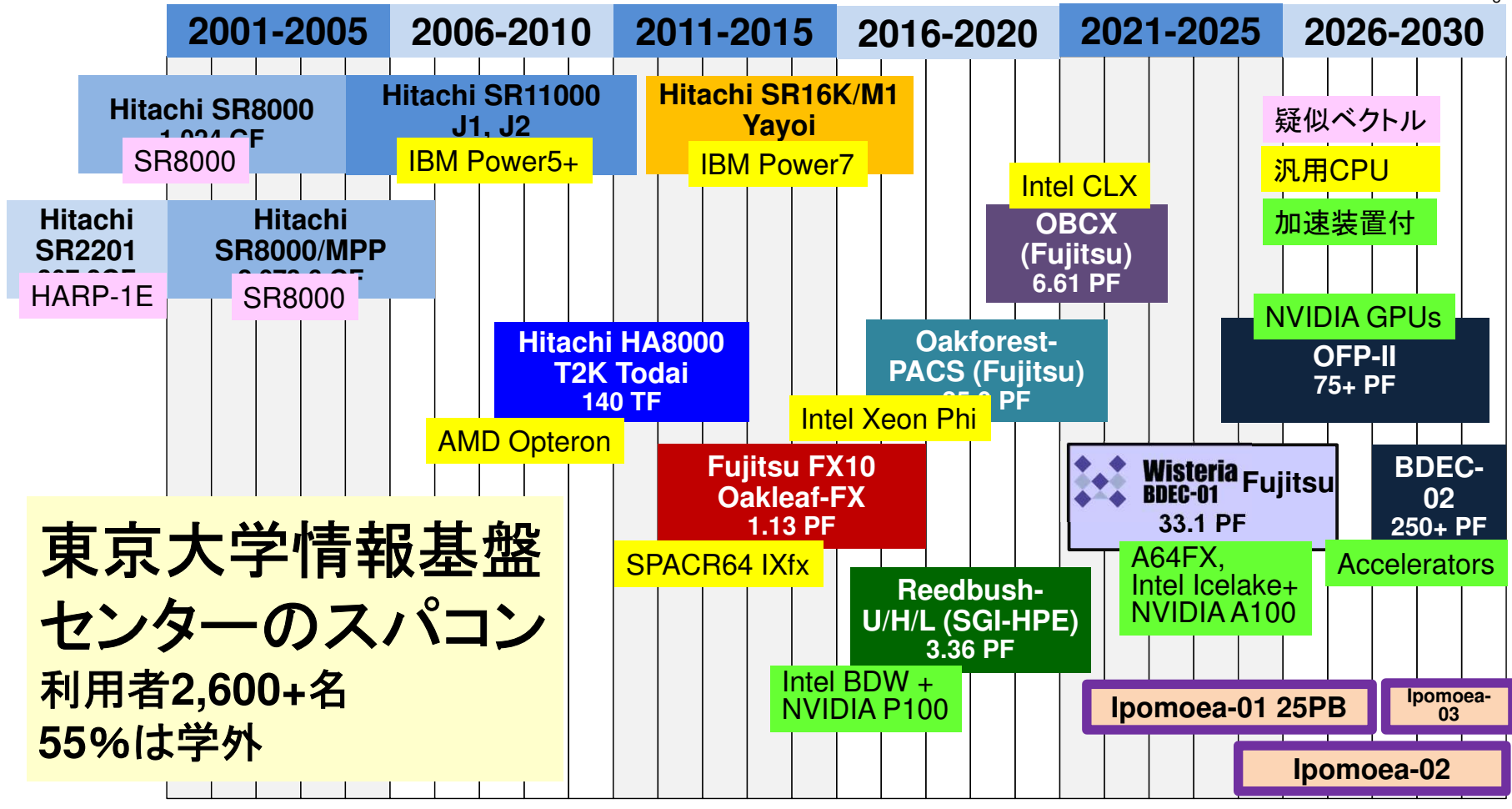
東京大学情報基盤センター
スーパーコンピューティング研究部門
http://www.cc.u-tokyo.ac.jp/
問合先：uketsuke@cc.u-tokyo.ac.jp



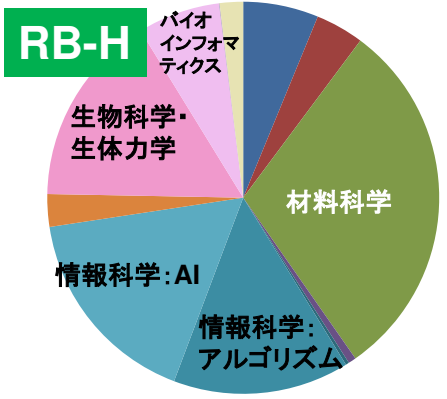
東京大学情報基盤センターのスパコン

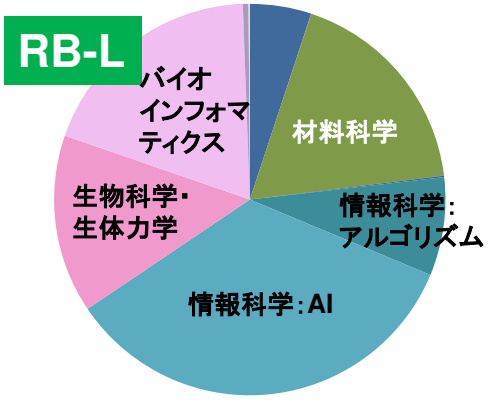
利用者2,600+名

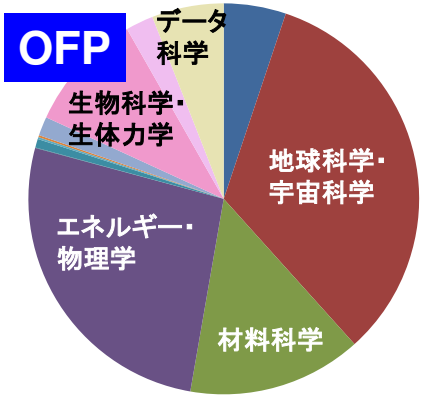
55%は学外

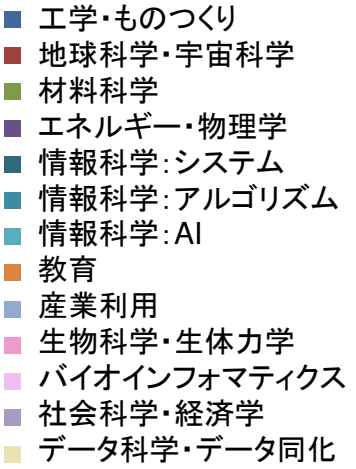

2001-2005
2006-2010
2011-2015
2016-2020
2021-2025
2026-2030
Hitachi SR8000
SR8000
Hitachi SR11000 J1, J2
IBM Power5+
Hitachi SR16K/M1 Yayoi
IBM Power7
疑似ベクトル
汎用CPU
加速装置付
Hitachi SR2201
HARP-1E
Hitachi SR8000/MPP
SR8000
Intel CLX
OBCX (Fujitsu)
6.61 PF
NVIDIA GPUs
OFP-II
75+ PF
Hitachi HA8000 T2K Todai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

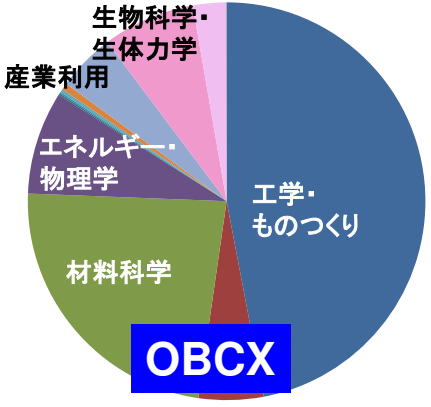
2020年度分野別 ■汎用CPU, ■GPU







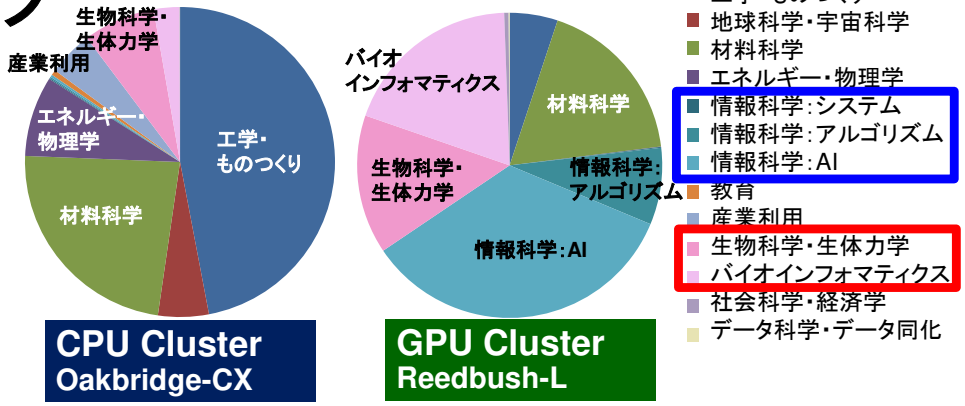


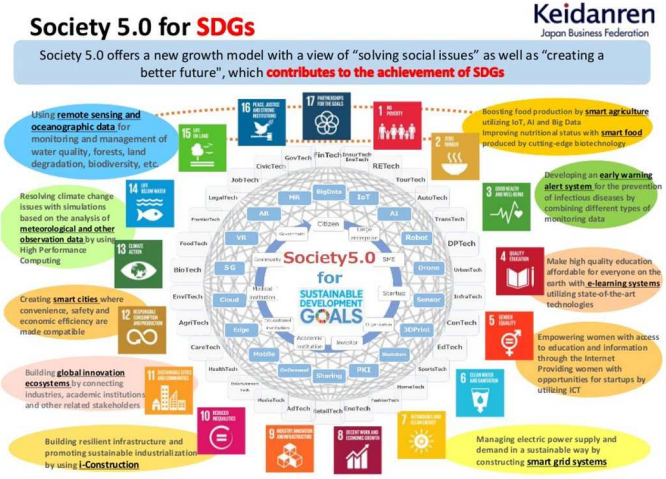


- 工学・ものつくり
- 地球科学・宇宙科学
- 材料科学

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，2015年頃から取り組み
  - 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





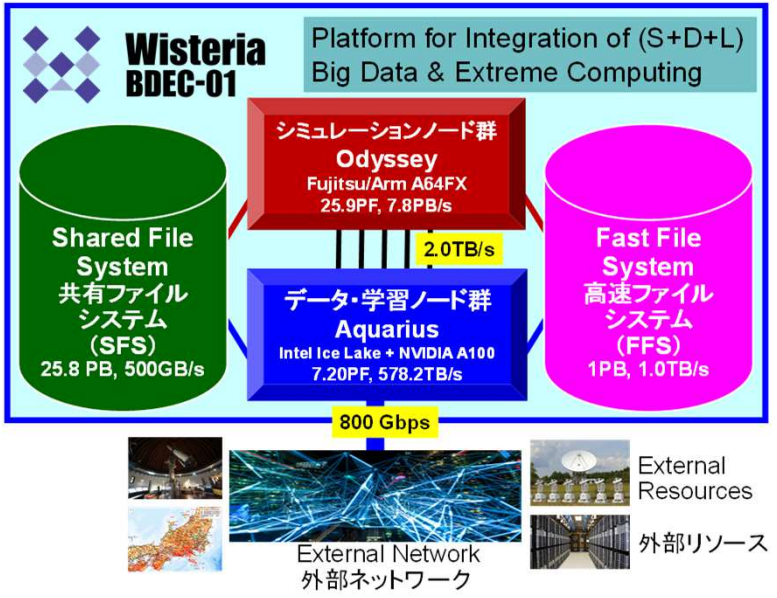
（シミュレーション（計算）＋データ＋学習）融合（S+D+L）

- 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
  - 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

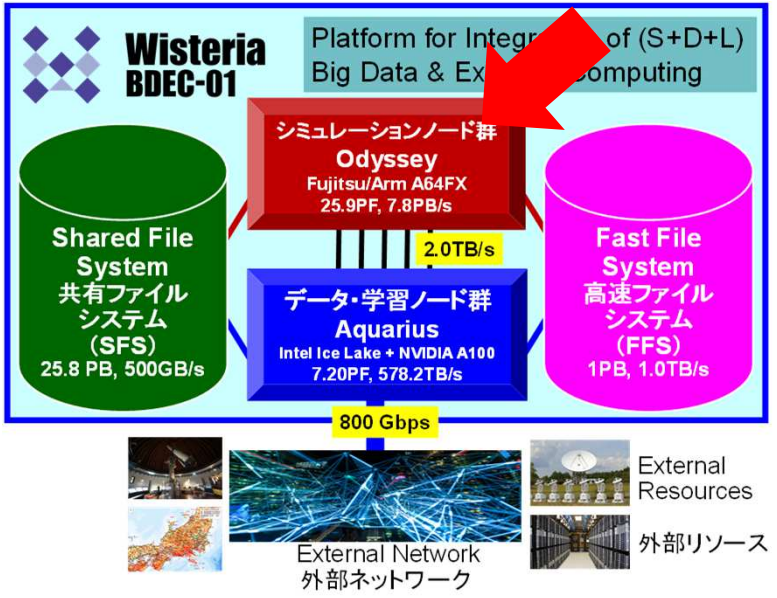


Wisteria
BDEC-01

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

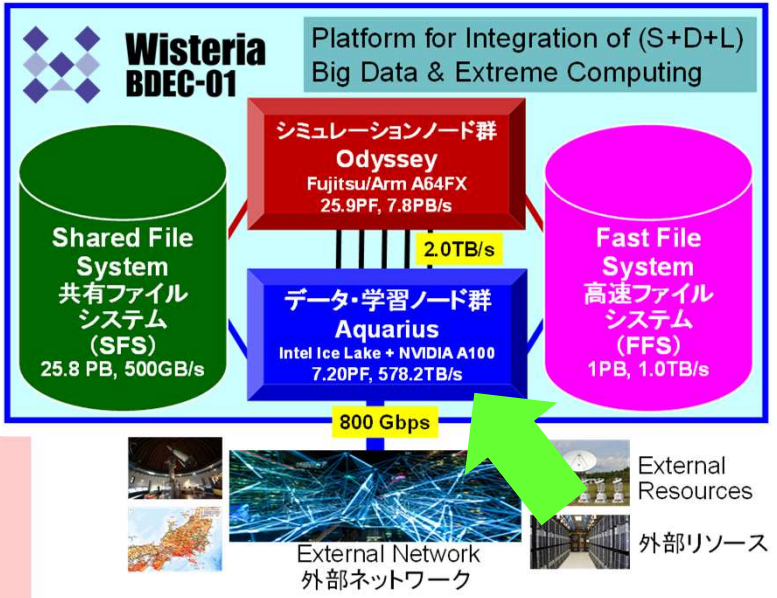


Wisteria
BDEC-01

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

システム構成図

シミュレーションノード：7,680ノード (総理論演算性能 25.9 PFLOPS、総メモリバンド幅 7.8 PB/s)
データ･学習ノード：45ノード (総理論演算性能 7.2 PFLOPS、総メモリバンド幅 578.2 TB/s)

外部接続ルータ (Ethernet)

・外部ネットワーク
・柏キャンパス
・mdxシステム

シミュレーションノード群

FUJITSU Supercomputer
PRIMEHPC FX1000 x 7,680ノード (20ラック)
総理論演算性能(倍精度)：25.9PFLOPS
総メモリ容量：240TiB
総メモリバンド幅：7.8PB/s

シミュレーションノード間ネットワーク
(TofuインターコネクトD)
バイセクションバンド幅：13TB/s

データ･学習ノード群

FUJITSU Server
PRIMERGY GX2570 M6 x 45ノード
(ノードあたり Intel Xeon Platinum 8360Y
Processor (2.40GHz, 36コア) x2)
(ノードあたり NVIDIA A100 x 8基)
総理論演算性能(倍精度)：7.2PFLOPS
総メモリ容量：36.5TiB,総メモリバンド幅：578.2TB/s

ログインノード

FUJITSU Server
PRIMERGY RX2530 M5 x 20ノード
総理論演算性能(倍精度)：96TFLOPS
総メモリ容量：7.5TiB

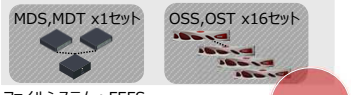
ログインノード共有ファイルシステム

FUJITSU Server
PRIMERGY RX2530 M5 x 2
ETERNUS DX60 S5 x 1
有効利用容量：14.4TB

データ･学習ノード間ネットワーク 兼 ノード群結合ネットワーク (InfiniBand EDR/HDR)

ライフネットワーク/管理ネットワーク (Ethernet)

高速ファイルシステム



ファイルシステム：FEFS
ストレージデータ転送速度：1.0TB/s
MDS：PRIMERGY RX2530 M5 x 2
MDT：ETERNUS AF250 S3 x 1
OSS, OST：2VM/CM, DDN SFA400NVXE x 16

共有ファイルシステム

ファイルシステム：FEFS
ストレージデータ転送速度：0.5TB/s
MDS：PRIMERGY RX2530 M5 x 4
MDT：ETERNUS AF250 S3 x 1
OSS, OST：1VM/CM, DDN SFA7990XE x 16

管理サーバ群

FUJITSU Server
PRIMERGY RX2530 M5
x 13
(ジョブ管理x5、運用管理x2、
認証x2、Webポータルx2、
セキュリティログx2)

管理サーバ群ネットワークストレージ

FUJITSU Server
PRIMERGY RX2530 M5 x 2
ETERNUS DX100 S5 x 1
有効利用容量：420TB

Wisteria
BDEC-01

項目		Wisteria-O (Odyssey)	Wisteria-A (Aquarius)
総理論演算性能		25.9 PFLOPS	7.2 PFLOPS
総ノード数		7,680	45
総主記憶容量		240.0 TiB	36.5 TiB
ネットワークトポロジー		6 次元メッシュ / トーラス	Full-bisection Fat Tree
インターコネクト		TofuインターコネクトD	InfiniBand HDR(200Gbps) x 4
共有ファイルシステム	システム名	FEFS（Fujitsu Exabyte File System）	
	サーバ(OSS)	DDN SFA7990XE	
	サーバ(OSS)数	16	
	ストレージ容量	25.8 PB	
	ストレージデータ転送速度	504 GB/s	
高速ファイルシステム	システム名	FEFS（Fujitsu Exabyte File System）	
	サーバ(OSS)	DDN SFA400NVXE	
	サーバ(OSS)数	16	
	ストレージ容量	1.0 PB	
	ストレージデータ転送速度	1.0 TB/s	

項目		Wisteria-O (Odyssey)	Wisteria-A (Aquarius)
マシン名		FUJITSU Supercomputer PRIMEHPC FX1000	FUJITSU Server PRIMERGY GX2570 M6
CPU	プロセッサ名	A64FX	Intel Xeon Platinum 8360Y （開発コード名：Ice Lake）
	プロセッサ数 (コア数)	1 (48+アシスタントコア2 or 4)	2 (36+36)
	周波数	2.2 GHz	2.4 GHz
	理論演算性能	3.3792 TFLOPS	5.53 TFLOPS
	メモリ容量	32 GB	512 GiB
	メモリ帯域幅	1,024 GB/s	409.6 GB/s
GPU	プロセッサ名	-	NVIDIA A100
	SM数（単体）		108
	メモリ容量（単体）		40 GB
	メモリ帯域幅（単体）		1,555 GB/s
	理論演算性能（単体）		19.5 TFLOPS
	搭載数		8
	CPU-GPU間接続		PCI Express Gen4 x 16レーン (1レーンあたり片方向32 GB/s)
	GPU間接続		NVLink x 12本 (1本あたり片方向25GB/s)

ソフトウェア群

項目	Wisteria-O（Odyssey）	Wisteria-A（Aquarius）
OS	Red Hat Enterprise Linux 8 (aarch64)	Red Hat Enterprise Linux 8 (x86_64)
コンパイラ	GNU コンパイラ	GNU コンパイラ
	富士通社製 コンパイラ (Fortran77/90/95/2003/2008、C、C++)	Intel コンパイラ(Fortran77/90/95/2003/2008、C、C++) NVIDIA HPC SDK (Fortran77/90/95/2003/2008、C、C++、OpenACC 2.7) NVIDIA CUDA SDK (CUDA C、CUDA C++)
メッセージ通信ライブラリ	富士通社製MPI	Intel MPI、Open MPI

項目	Wisteria-O（Odyssey）	Wisteria-A（Aquarius）
ライブラリ	SuperLU、SuperLU MT、SuperLU DIST、METIS、MT-METIS、ParMETIS、Scotch、PT-Scotch、PETSc、Trillinos、FFTW、GNU Scientific Library、NetCDF、Parallel netCDF、HDF5、Parallel HDF5、CMake、Miniconda、Xabclib、ppOpen-HPC、MassiveThreads、Boost C++、mpiJava	
	富士通社製ライブラリ(BLAS、CBLAS、LAPACK、ScaLAPACK)	Intel社製ライブラリ(MKL)(BLAS、CBLAS、LAPACK、ScaLAPACK）、cuBLAS、cuSPARSE、cuFFT、MAGMA、cuDNN、NCCL
アプリケーション	OpenFOAM、ABINIT-MP、PHASE、FrontFlow/blue、FrontISTR、REVOCAP-Coupler、REVOCAP-Refiner、OpenMX、MODYLAS、GROMACS、BLAST、R packages、bioconductor、BioPerl、BioRuby、BWA、GATK、SAMtools、Quantum ESPRESSO、Xcrypt、ROOT、Geant4、LAMMPS、CP2K、NWChem、DeepVariant、Paraview、VisIt、POV-Ray、TensorFlow、Chainer、PyTorch、Keras、Horovod、MXNet	
		Theano
フリーソフトウェア	autoconf、automake、bash、bzip2、cvs、emacs、findutils、gawk、gdb、make、grep、gnuplot、gzip、less、m4、python、perl、ruby、screen、sed、subversion、tar、tcsh、tcl、vim、zsh、git など	
		Globus Toolkit、Gfarm、FUSE
コンテナ仮想化	Singularity Community Edition	

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

Wisteria BDEC-01

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

データ・学習ノード群
Aquarius

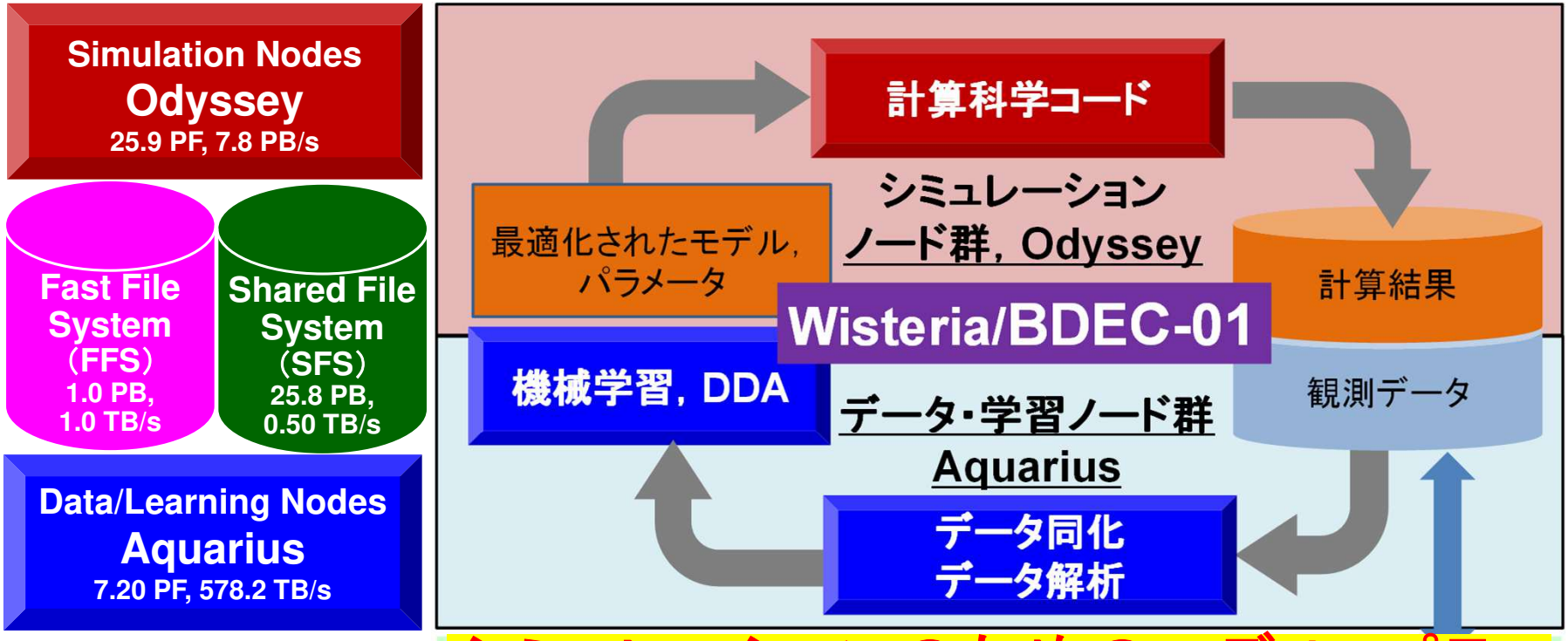
観測データ

データ同化
データ解析

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）

61st TOP500 List (June, 2023)

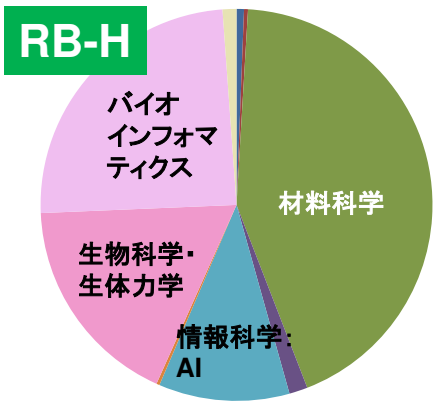
R_{max}: Performance of Linpack (TFLOPS)
R_{peak}: Peak Performance (TFLOPS), Power: kW

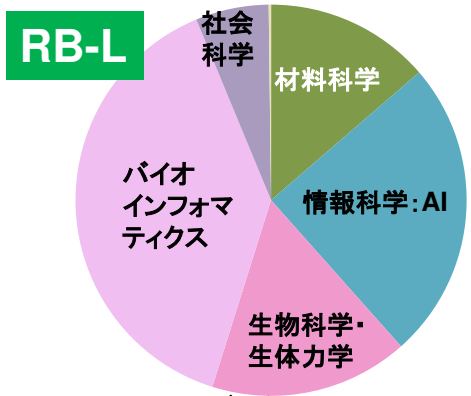
http://www.top500.org/

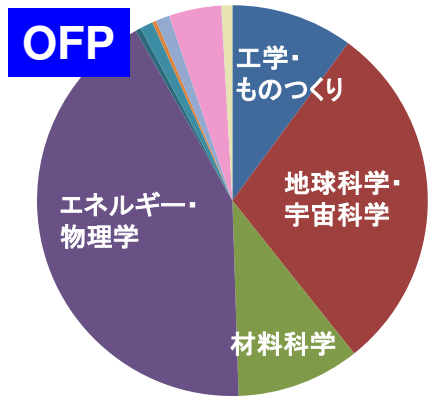
	Site	Computer/Year Vendor	Cores	R_{max} (PFLOPS)	R_{peak} (PFLOPS)	Power (kW)
1	**Frontier, 2022, USA** DOE/SC/Oak Ridge National Laboratory	HPE Cray EX235a, AMD Optimized 3rd Gen. EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11	8,699,904	1,194.00 (=1.194 EF)	1,679.82	22,703
2	**Fugaku, 2020, Japan** R-CCS, RIKEN	Fujitsu PRIMEHPC FX1000, Fujitsu A64FX 48C 2.2GHz, Tofu-D	7,630,848	442,010 (= 442.0 PF)	537,212.0	29,899
3	**LUMI, 2022, Finland** EuroHPC/CSC	HPE Cray EX235a, AMD Optimized 3rd Gen. EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11	2,220,288	309.10	428.70	6,016
4	**Leonard, 2022, Italy** EuroHPC/Cineca	BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64GB, Quad-rail NVIDIA HDR100	1,824,768	238.70	304.47	7,404
5	**Summit, 2018, USA** DOE/SC/Oak Ridge National Laboratory	IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR InfiniBand	2,414,592	148.60	200.79	10,096
6	**Sierra, 2018, USA** DOE/NNSA/LLNL	IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR InfiniBand	1,572,480	94.64	125.71	7,438
7	**Sunway TaihuLight, 2016, China** National Supercomputing Center in Wuxi	Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway	10,649,600	93.01	125.44	15,371
8	**Perlmutter, 2021, USA** DOE/NERSC/LBNL	HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10	761,856	70.87	93.75	2,528
9	**Selene, 2020, USA** NVIDIA	NVIDIA DGX A100 SuperPOD, AMD EPYC 7742 64C 2.25GHz, NVIDIA GA100, Mellanox Infiniband HDR	555,520	63.46	79.22	2,646
10	**Tianhe-2A, 2018, China** National Super Computer Center in Guangzhou	TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000	4,981,760	61.44	100.68	18,482
24	**ABCI 2.0, 2021, Japan** AIST	PRIMERGY GX2570 M6, Xeon Platinum 8360Y 36C 2.4GHz, NVIDIA A100 SXM4 40 GB, InfiniBand HDR	504,000	22.21	54.34	1,600
25	**Wisteria/BDEC-01 (Odyssey), 2021, Japan** ITC, University of Tokyo	PRIMEHPC FX1000, A64FX 48C 2.2GHz, Tofu interconnect D	368,640	22.12	25.95	1,468

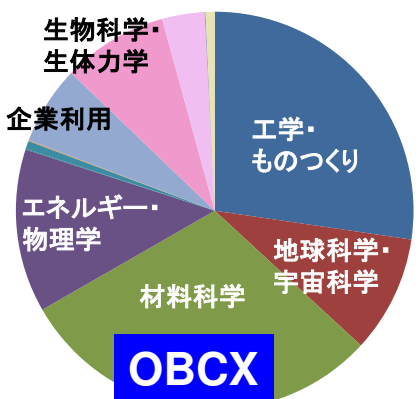
2021年度分野別 ■汎用CPU, ■GPU

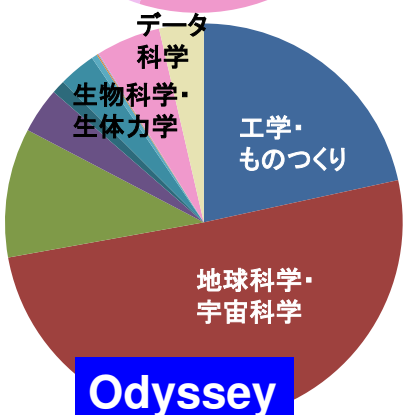
Odyssey, Aquariusは8月以降, RB-H, RB-Lは11月末時点

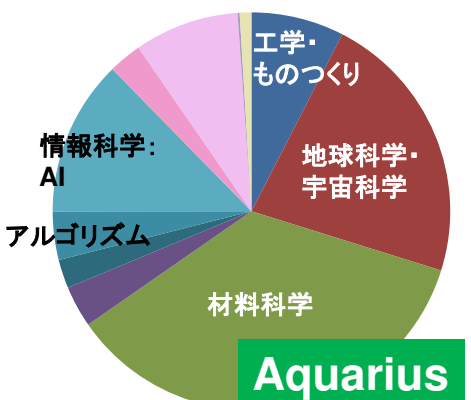








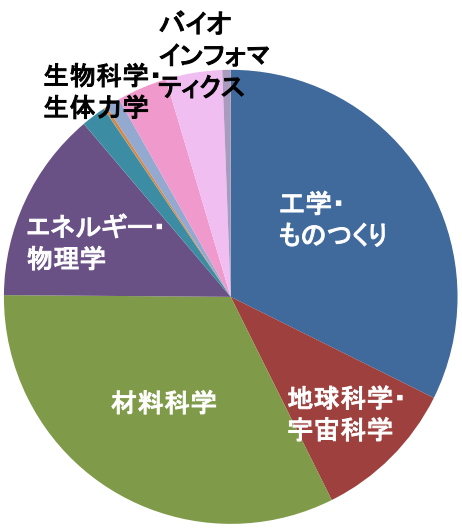




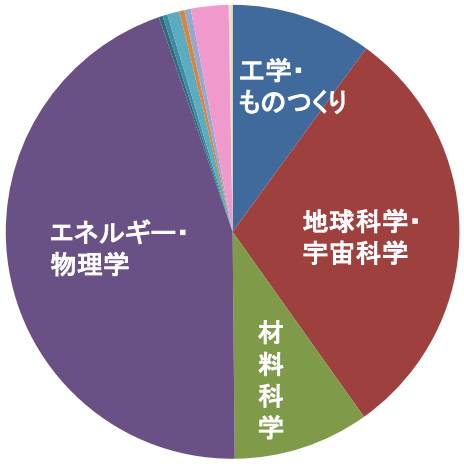
- 工学・ものつくり
- 地球科学・宇宙科学
- 材料科学
- エネルギー・物理学
- 情報科学:システム
- 情報科学:アルゴリズム
- 情報科学:AI
- 教育
- 産業利用
- 生物科学・生体力学
- バイオインフォマティクス
- 社会科学・経済学
- データ科学・データ同化

地球科学・宇宙科学分野ではOFP ⇒ Wisteria/BDEC-01への移行が順調に進んでいる

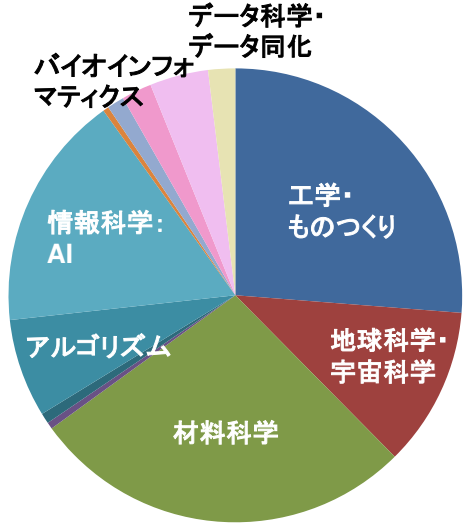
2022年度分野別 ■汎用CPU, ■GPU



OBCX
CascadeLake



Odyssey
A64FX

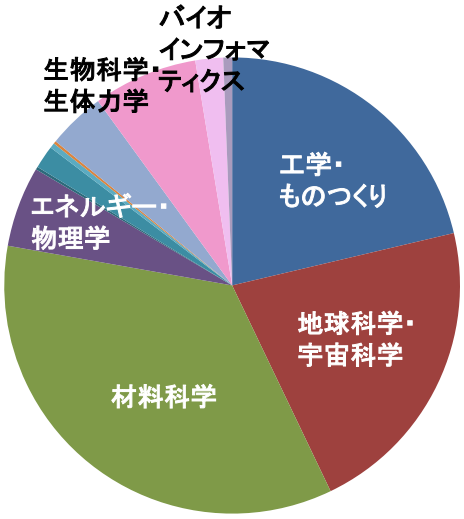


Aquarius
A100

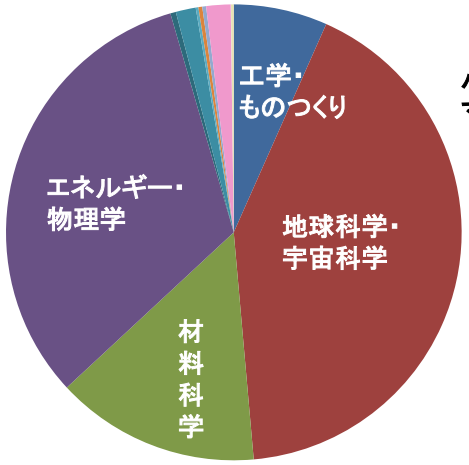
- 工学・ものつくり
- 地球科学・宇宙科学
- 材料科学
- エネルギー・物理学
- 情報科学:システム
- 情報科学:アルゴリズム
- 情報科学:AI
- 教育
- 産業利用
- 生物科学・生体力学
- バイオインフォマティクス
- 社会科学・経済学
- データ科学・データ同化

2023年度分野別(4月～9月末)

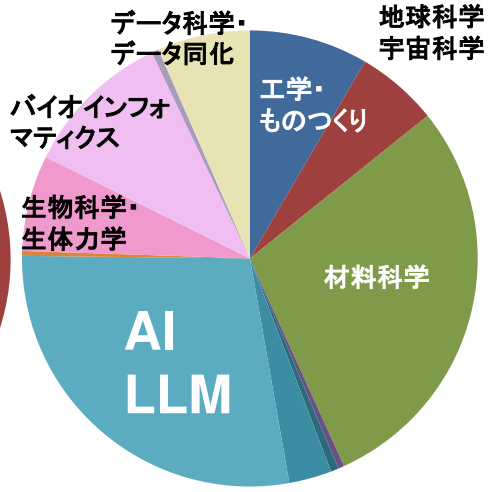
■汎用CPU, ■GPU



OBCX
CascadeLake
2023年9月末退役



Odyssey
A64FX



Aquarius
A100

- 工学・ものつくり
- 地球科学・宇宙科学
- 材料科学
- エネルギー・物理学
- 情報科学:システム
- 情報科学:アルゴリズム
- 情報科学:AI
- 教育
- 産業利用
- 生物科学・生体力学
- バイオインフォマティクス
- 社会科学・経済学
- データ科学・データ同化

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/

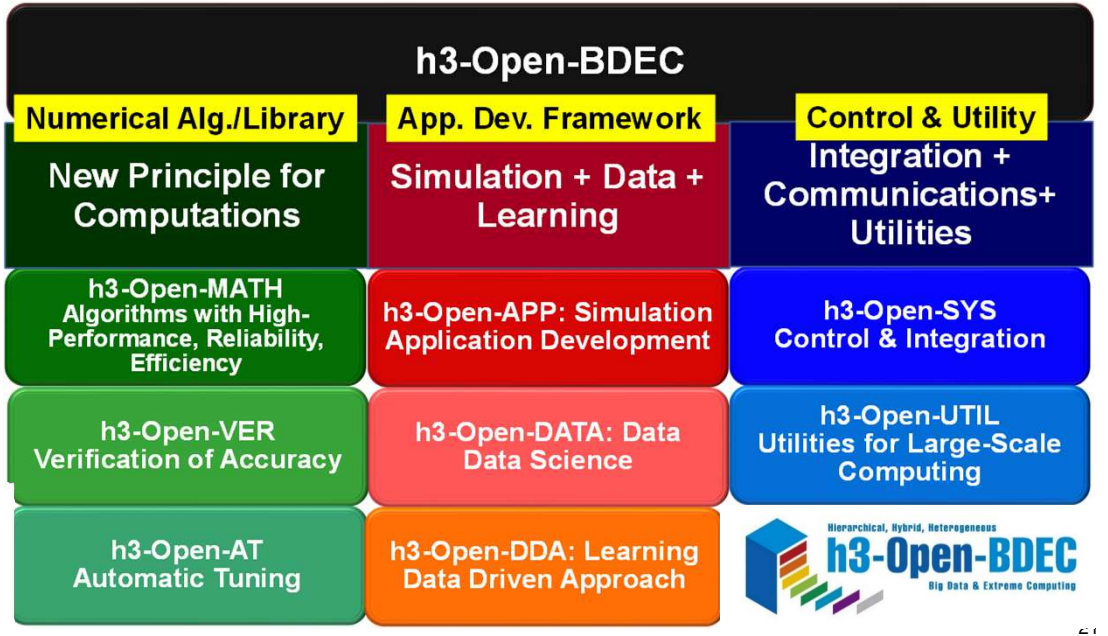
Hierarchical, Hybrid, Heterogeneous

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



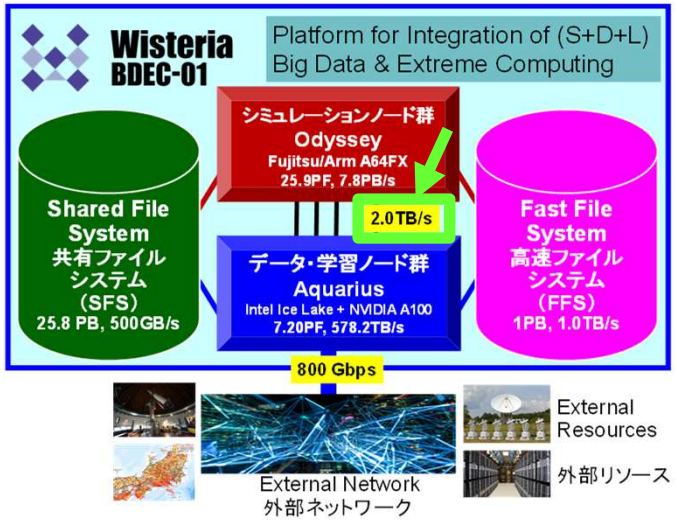
Hierarchical, Hybrid, Heterogeneous
h3-Open-BDEC
Big Data & Extreme Computing

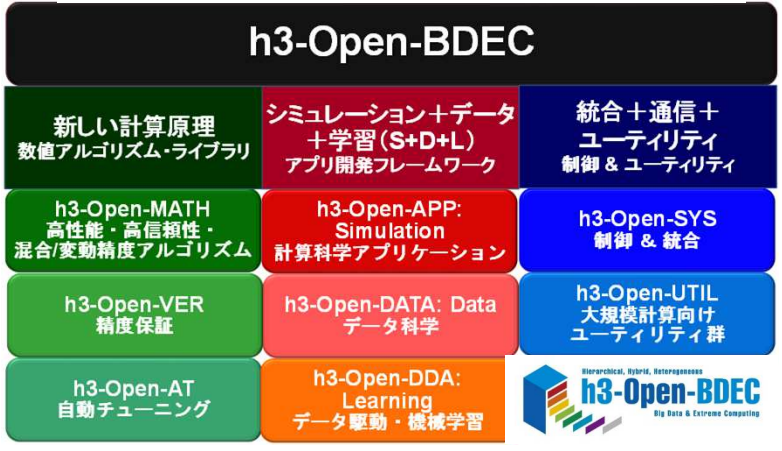
AI for HPC, AI for Science の実現へ向けて

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）

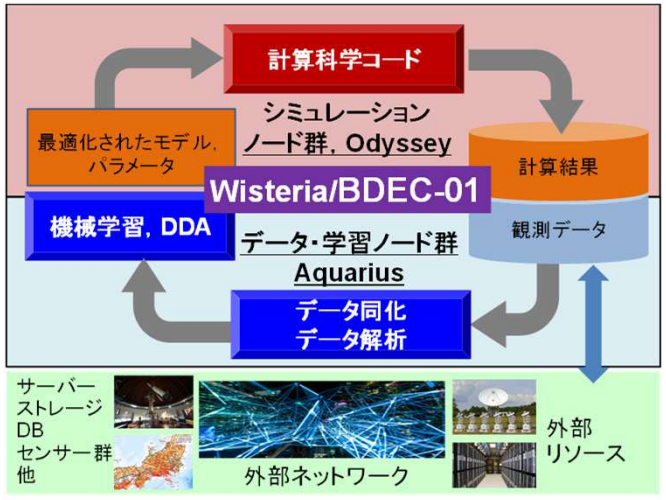


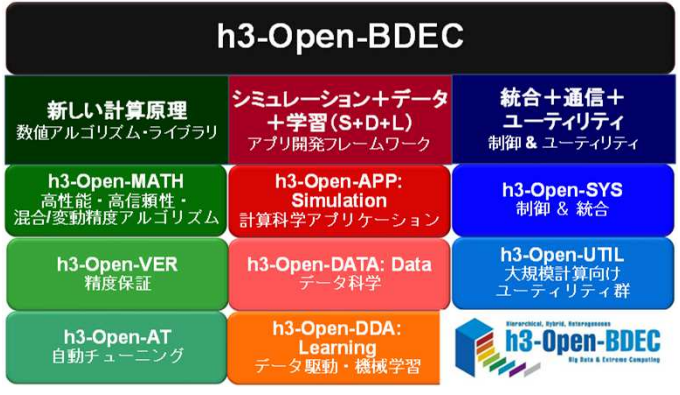


h3-Open-SYS/WaitIO

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

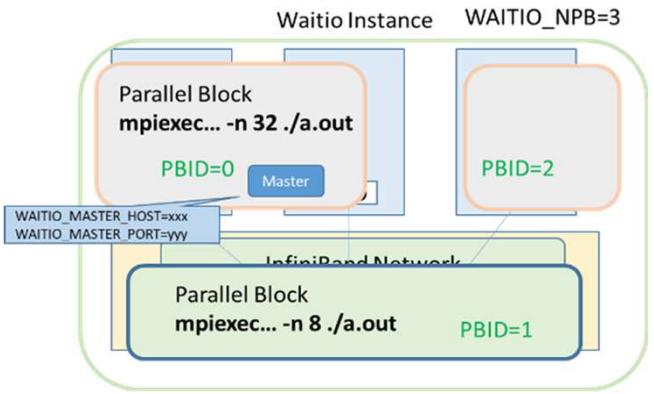
- ヘテロジニアス環境下での異なるコンポーネント間ファイル経由連携ライブラリとして考案
- 機能
  - ✓ Odyssey～Aquarius間連携
    - IB-EDR経由通信（WaitIO-Socket）
    - ファイル経由（WaitIO-File）
  - ✓ 外部からのデータ取得（観測データ等）
  - ✓ 読み込み・書き出しの同期
- API：C/C++，Fortranから呼び出し可能
  - ✓ MPIライクなインタフェースを提供





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

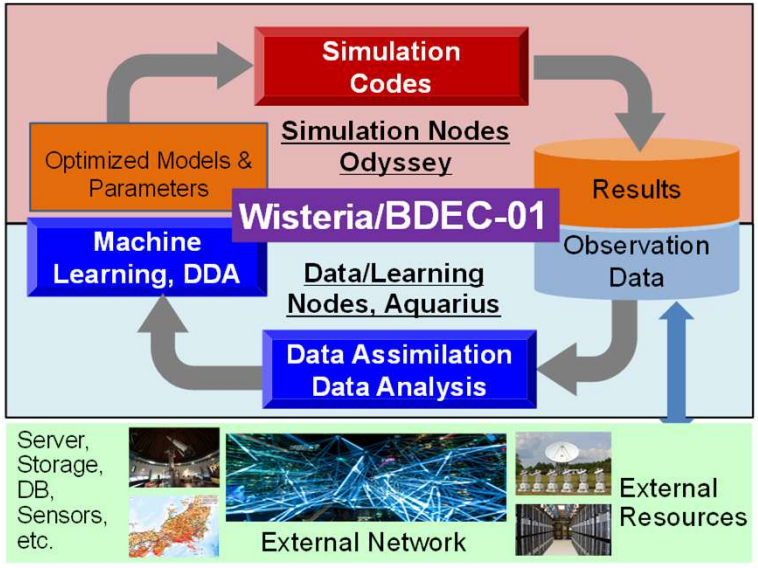
WaitIO API	Description
waitio_isend	Non-Blocking Send
waitio_irecv	Non-Blocking Receive
waitio_wait	Termination of waitio_isend/irecv
waitio_init	Initialization of WaitIO
waitio_get_nprocs	Process # for each PB (Parallel Block)
waitio_create_group waitio_create_group_wranks	Creating communication groups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

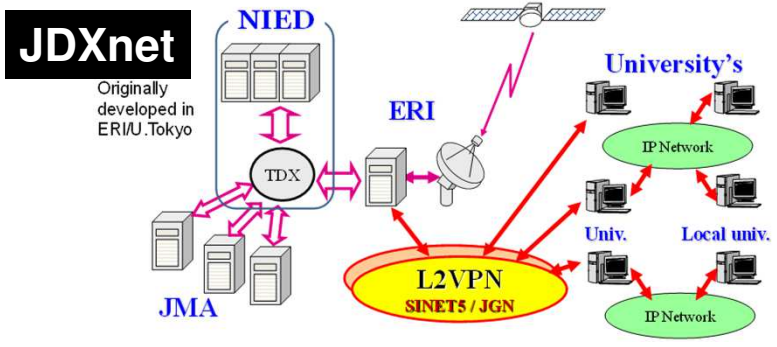


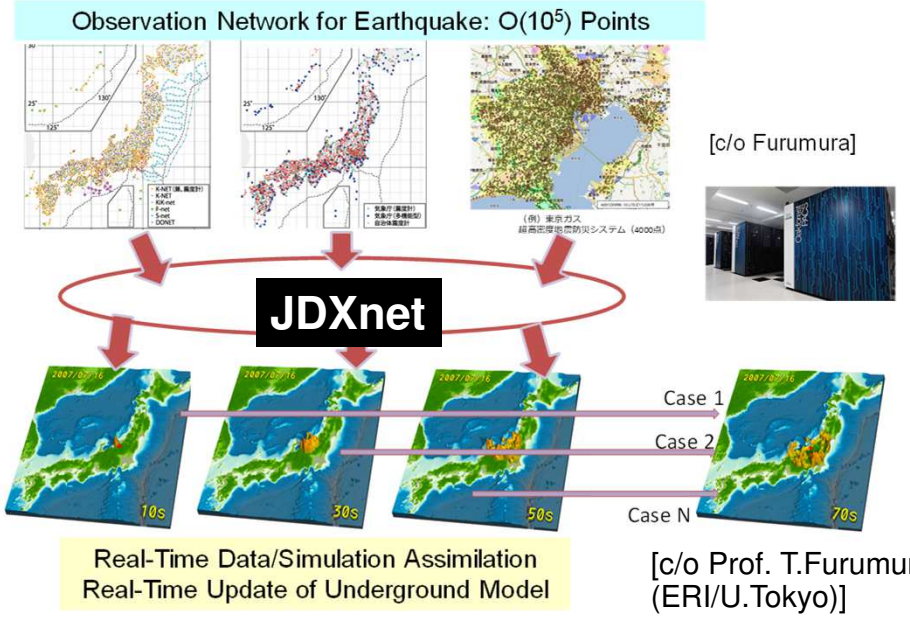
[Sumimoto et al. 2021]

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

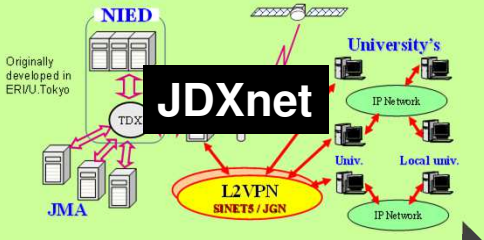






[c/o Prof. T.Furumura (ERI/U.Tokyo)]

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

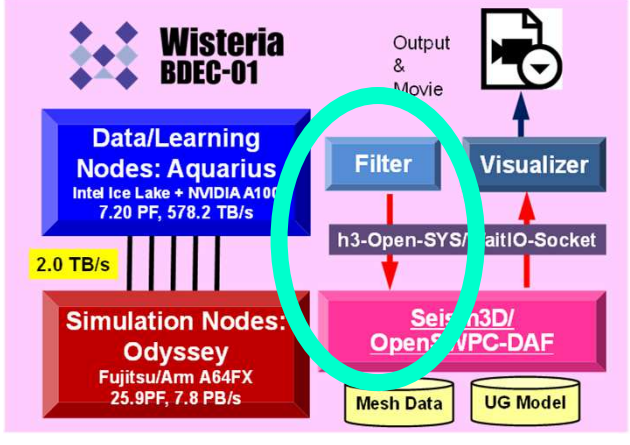
[Kasai et al. 2021]

Aquarius: SEND

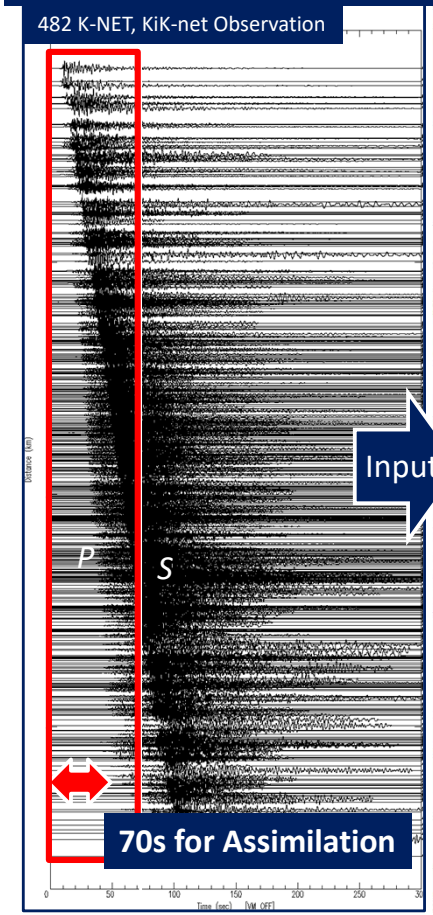
```
program dmy_filter
<省略: 型宣言等>
call mpi_init (ierr)
call mpi_comm_size (MPI_COMM_WORLD, nprocs, ierr)
call mpi_comm_rank (MPI_COMM_WORLD, myrank,  ierr)
call WAITIO_CREATE_UNIVERSE (WAITIO_COMM_UNIVERSE, ierr)

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
```

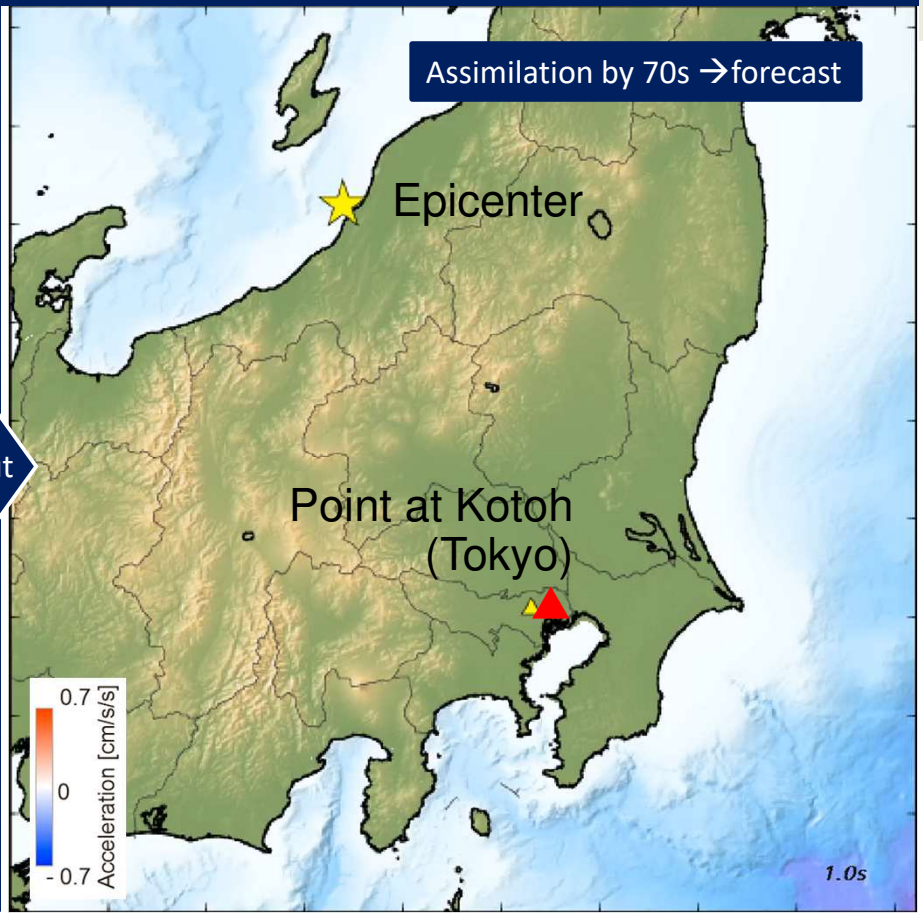
Odyssey: RECV

```
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,...)
```

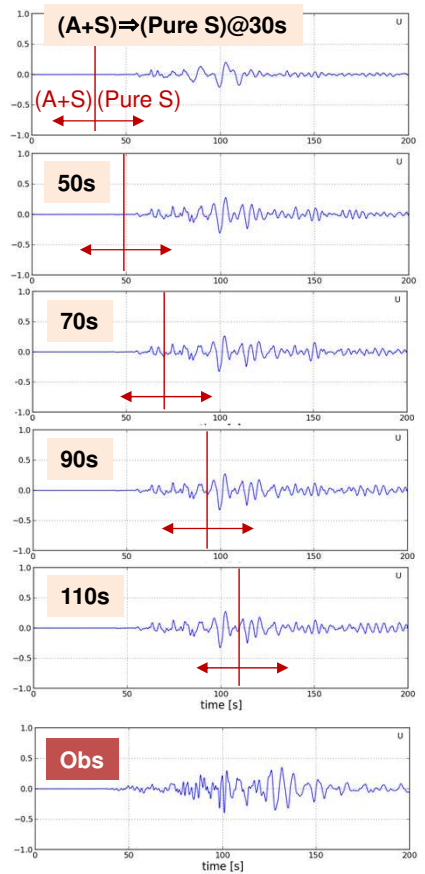


Data Assimilation + Pure Simulation/Forecast



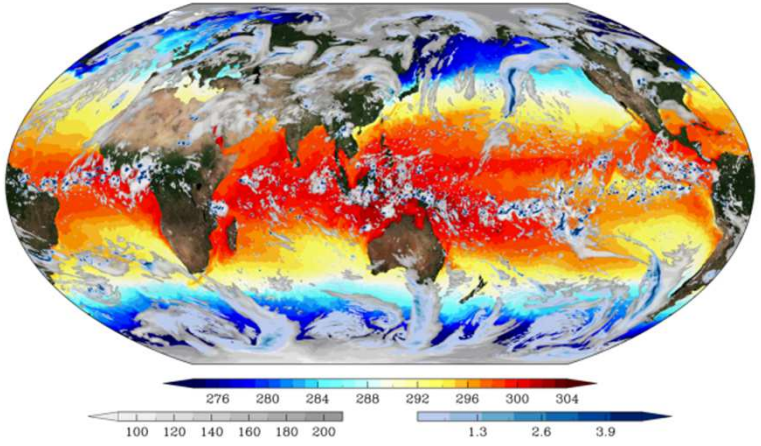


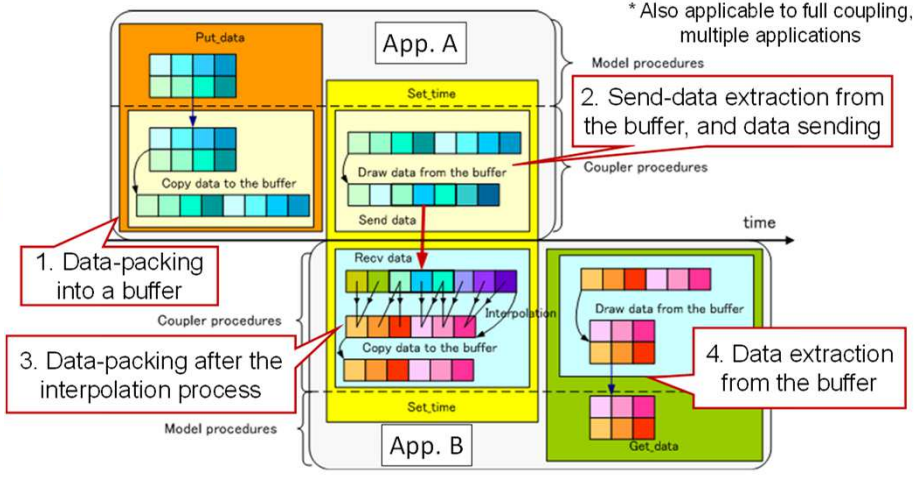
Results at Kotoh ▲ (N.KOTH)
N 35° 37.0'
E 139° 46.9'



連成シミュレーションのためのカプラー〔荒川，八代〕

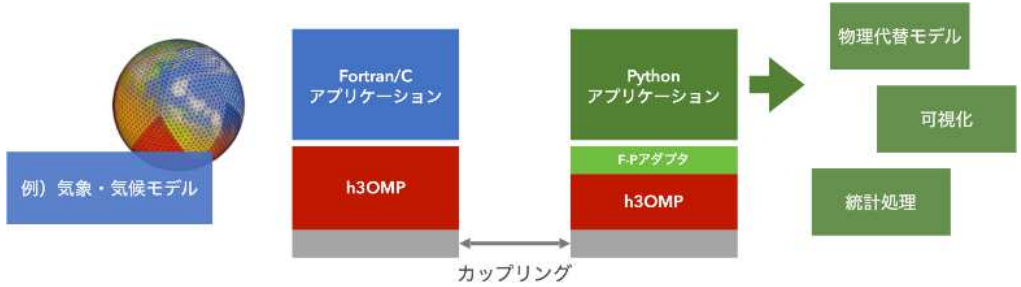
- 従来のカプラー（Coupler）：ppOpen-MATH/MP
  - 複数（通常2つ：大気（NICAM）＋海洋（COCO））のアプリケーションの弱連成（Weak Coupling）をサポート
  - 各アプリケーションは1種類の計算をやる





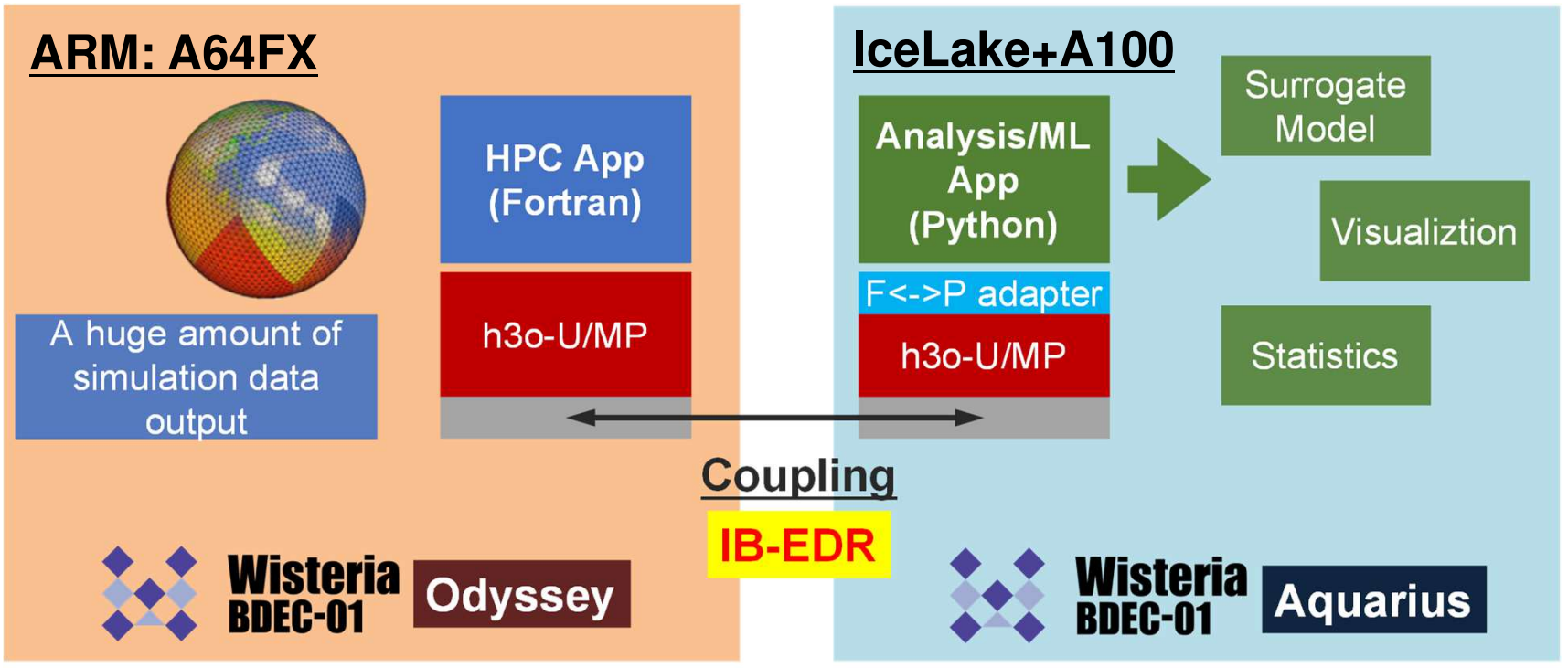
「計算＋データ＋学習」融合を支援する多機能カプラーh3-Open-UTIL/MP

- 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。



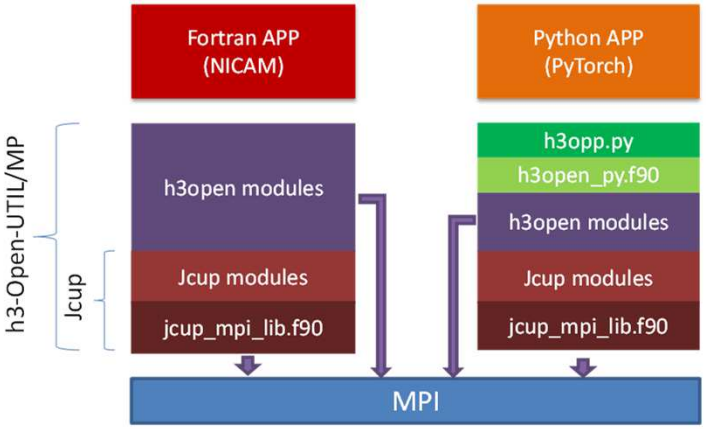
Fortran/CアプリとPythonアプリの連成計算の模式図
〔八代・荒川　2020〕

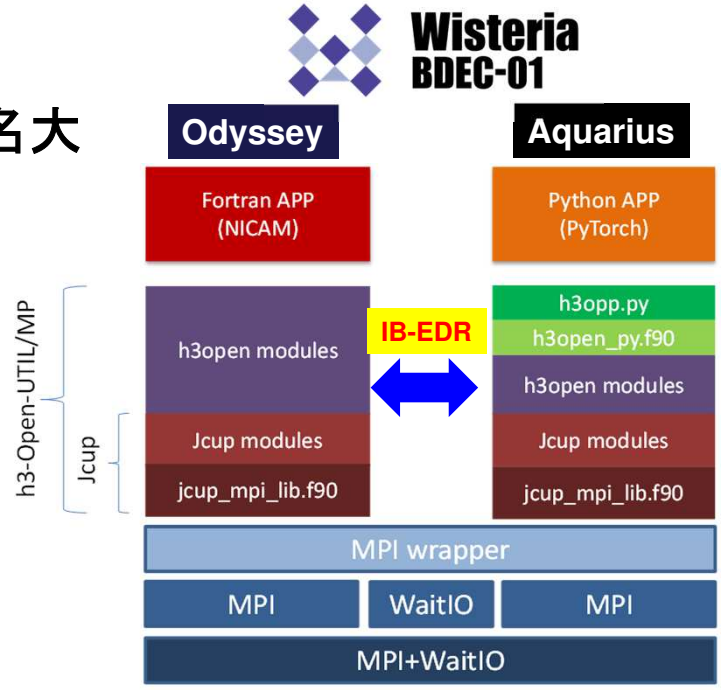
h3-Open-UTIL/MP (h3o-U/MP) + h3-Open-SYS/WaitIO-Socket



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



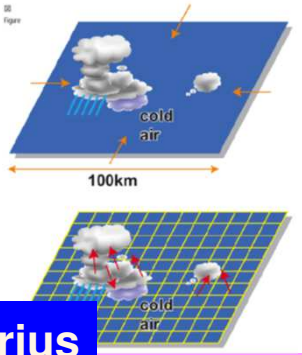


2021年4月：MPI通信可能な環境を前提

2022年6月：Coupler＋WaitIO

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!



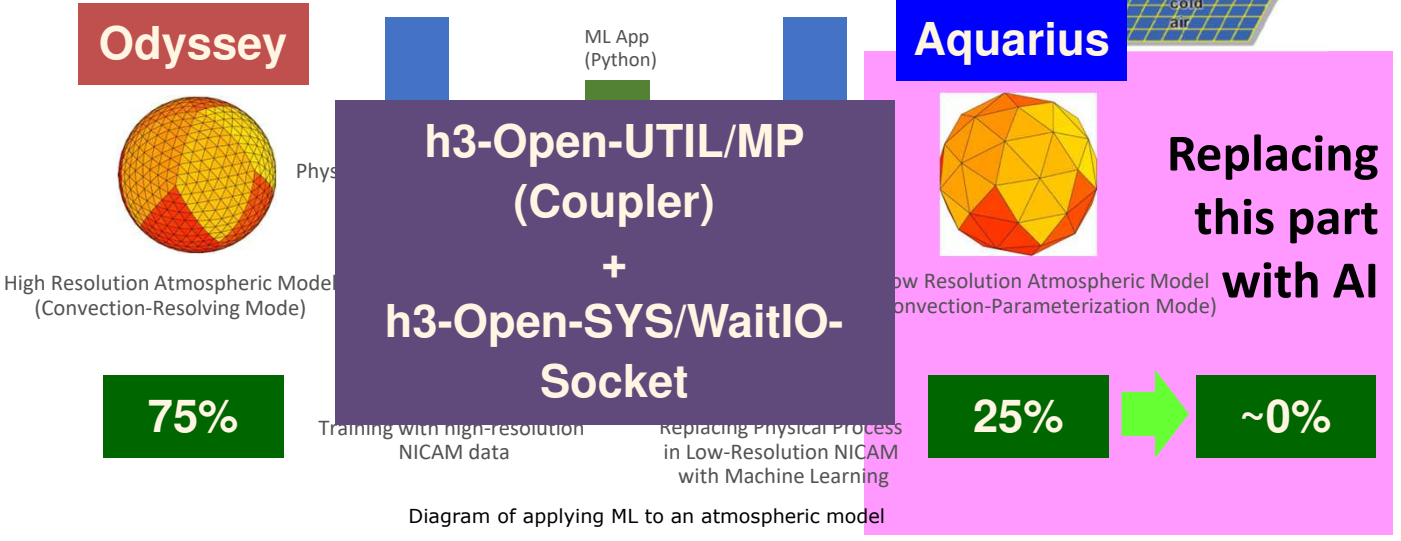
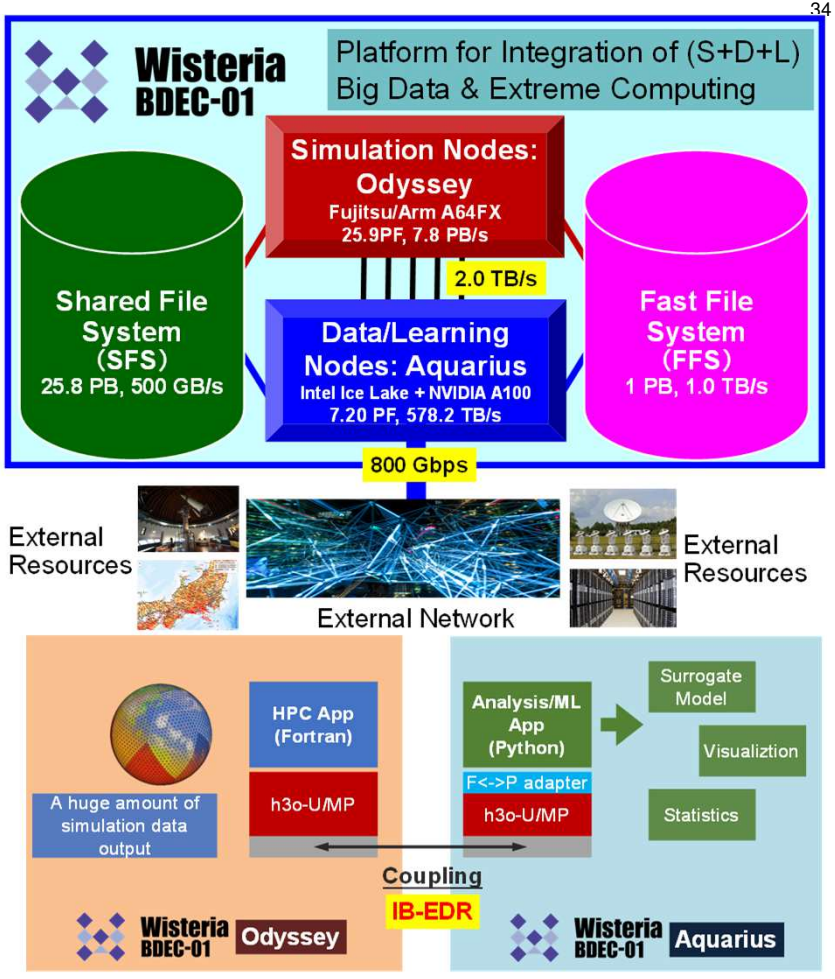


Diagram of applying ML to an atmospheric model

Odyssey-Aquarius連携

- 総ノード数
  - Odyssey：7,680ノード，やや空いている
  - Aquarius: 45ノード, 360 GPUs，混雑
- Aquariusのうち1ノードを（S+D+L）融合型ワークロード向けにリザーブ
  - Odyssey，Aquariusそれぞれに対する2つのジョブスクリプトをサブミットする必要がある
  - 両ジョブがリソースを確保⇒実行開始
- より柔軟な仕組みを整える必要あり
  - このようなシステム，運用例は世界的に見ても例がほとんどない
- O-A連携ワークロードを考えている場合はご相談ください



ジョブスクリプト例 [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
```

技術的な特徴など

Wisteria
BDEC-01

- Odyssey
  - SVE (Scalable Vector Extension)
    - Armv8-A命令セットアーキテクチャをスーパーコンピュータ向けに拡張
  - FP16
  - 機械学習・AIワークロードへの適用
- Aquarius
  - HPC・計算科学への適用
  - CPU：Intel Xeon Ice Lake
    - 3rd Generation Intel Xeon Scalable Processors
    - 推論，単独での利用は難しいが
  - GPU：NVIDIA A100 Tensor Core
    - Tensor Core + Tensor Float [TF32]
- Odyssey-Aquarius
  - InfiniBand-EDR，ファイルシステム（高速・共有）

スパコン利用にあたっての指針(1/3)
Odyssey，Aquarius

- 基本的には，自作コード，オープンソースの利用を前提
  - OpenFOAM(流体)
    - Odyssey
    - 今野雅博士(客員研究員):OpenFOAM関連チュートリアル
  - FrontISTR，FrontFlow，ABINIT(東大生研)
  - ppOpen-HPC，h3-Open-BDEC(東大センター)
- 商用コード
  - Altair HyperWorks(汎用CAEコード)
    - https://www.altairjp.co.jp/hyperworks/
    - Aquarius(一部)
    - 国内大学教職員・学生のみ利用可能
    - 研究機関，企業の場合は別途ライセンス取得が必要
  - MATLAB(2022年3月から利用可能)
    - Aquarius

スパコン利用にあたっての指針(2/3)
Odyssey，Aquarius

- 計算科学・大規模シミュレーション(S)
- データ科学(D)
- 機械学習・AI(L)
- 「S+D+L」融合

- 全てのシステム(Odyssey，Aquarius)がそれぞれの項目に対応可能
  - Aquarius(データ・学習ノード)でもシミュレーションはできる
- データ科学(D)，機械学習・AI(L)
  - コンテナ仮想化(Singularity)により対応

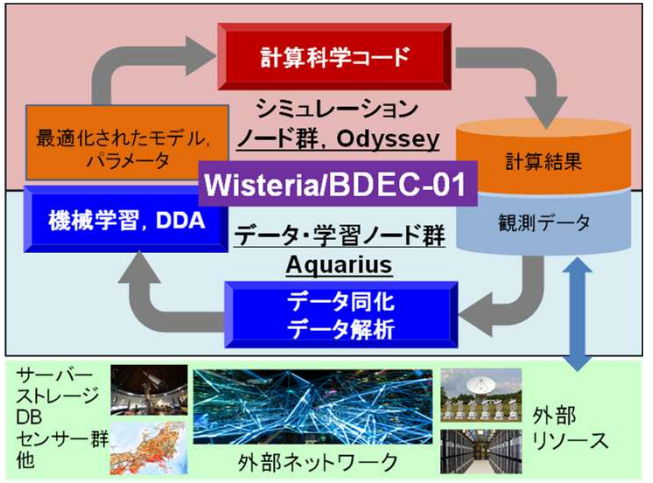
スパコン利用にあたっての指針(3/3)
Odyssey，Aquarius

	Odyssey	Aquarius	O+A
計算科学	◎	◎	-
データ科学	◎	◎	-
機械学習・AI	○	◎	-
「S+D+L」融合	○	◎	◎
商用コード・MATLAB等	×～△	○	-
その他の特徴	• A64FX(Arm) • チューニング必須 • FP16 • 商用コードへの対応がやや遅れている	• CPU(Ice Lake):高い推論性能 • GPU(A100):Tensor Core + Tensor Float [TF32] • 超大規模シミュレーションには不向き	• O-A連携についてはソフトウェア開発(h3-Open-BDEC，WaitIO)，応相談

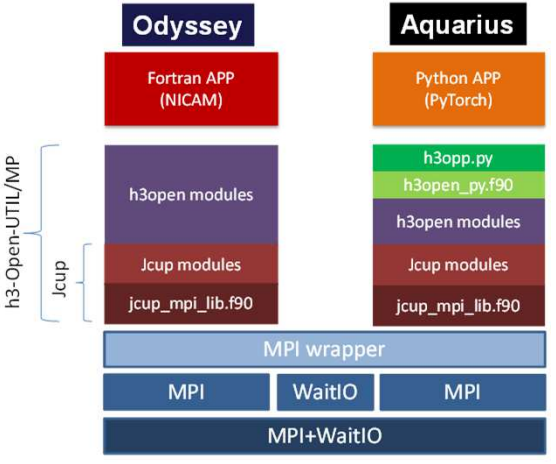
MATLABの導入「S+D+L」融合，AI for HPCの実現

Wisteria BDEC-01

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



- MATLAB
  - ✓ 多様な機能
  - ✓ ユーザーのプログラムからの関数呼び出し重視⇒データ解析，機械学習系の豊富な機能⇒高度化
  - ✓ MATLABはAquarius（データ・学習ノード群）でのみ稼働するが，h3-Open-BDECと連携させて，Odyssey（シミュレーションノード群）上で実施する大規模シミュレーションのパラメータ最適化に適用する⇒「S+D+L」融合，AI for HPC
- h3-Open-BDECは様々な環境で動作⇒MATLABと組み合わせた使用による普及
- 量子・HPCハイブリッド



更に詳細な情報

- A64FX（富士通）
  - https://www.fujitsu.com/jp/products/computing/servers/supercomputer/a64fx/
  - https://old.hotchips.org/hc30/2conf/2.13_Fujitsu_HC30.Fujitsu.Yoshida.rev1.2.pdf
- FUJITSU PRIMEHPC FX1000
  - https://www.fujitsu.com/jp/products/computing/servers/supercomputer/
- 3rd Gen Intel Xeon Scalable
  - https://www.intel.com/content/www/us/en/newsroom/news/3rd-gen-intel-xeon-scalable-video.html#gs.zb3u0m
  - https://www.intel.com/content/www/us/en/newsroom/news/3rd-gen-xeon-scalable-processors.html#gs.zb4d00
  - https://www.hotchips.org/assets/program/conference/day1/HotChips2020_Server_Processors_Intel_Irma_ICX-CPU-final3.pdf
- NVIDIA A100 TENSORコア　GPU
  - https://www.nvidia.com/ja-jp/data-center/a100/
  - https://www.hotchips.org/assets/program/conference/day1/HotChips2020_GPU_NVIDIA_Choquette_v01.pdf

参考リンク(ビデオ)

Wisteria
BDEC-01

- Wisteria/BDEC-01利用説明会
  - https://www.youtube.com/watch?v=1bbZVO6-UQg
- h3-Open-BDEC:プロジェクトHP(工事中)
  - https://h3-open-bdec.cc.u-tokyo.ac.jp/
- Wisteria/BDEC-01 & h3-Open-BDEC紹介講演(日本語)
  - https://www.youtube.com/watch?v=CsJ_9aGNXCg
  - https://www.pccluster.org/ja/event/pccc20/exhibition/itc-u-tokyo.html
- Wisteria/BDEC-01 & h3-Open-BDEC紹介講演(英語)
  - https://www.youtube.com/watch?v=jX51NF2LniE

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

- 住元真司，荒川隆，坂口吉生，松葉浩也，八代尚，塙敏博，中島研吾，WaitIO-Socket：異種システム上の複数MPIプログラムを結合する通信ライブラリの試作，情報処理学会研究報告（2021-HPC-181-07），2021
- h3-Open-SYS/WaitIO-Socket，h3-Open-UTIL/MP概要：
  https://www.dropbox.com/s/k1nd0p98p5cbdeg/KN_HPC182x.pdf?dl=0
- 住元真司他：Wistera/BDEC-01利用事例（3）データ受け渡しライブラリh3-Open-SYS/WaitIO（1/2）
  https://www.cc.u-tokyo.ac.jp/public/VOL24/No2/10_202203Wisteria-1.pdf
- 住元真司他：Wistera/BDEC-01利用事例（4）データ受け渡しライブラリh3-Open-SYS/WaitIO（2/2）
  https://www.cc.u-tokyo.ac.jp/public/VOL24/No3/12_202205-Wisteria-1.pdf
- 住元真司，荒川隆，坂口吉生，松葉浩也，八代尚，大島聡史，塙敏博，中島研吾，WaitIO-Hybrid:共有ファイルシステムとSocketを併用可能なシステム間通信ライブラリ，情報処理学会研究報告（2022-HPC-187-06），2022
- 荒川隆他：Wistera/BDEC-01利用事例（5）マルチプログラム連成ライブラリh3-Open-UTIL/MP（1/2）
  https://www.cc.u-tokyo.ac.jp/public/VOL24/No3/13_202205-Wisteria-2.pdf
- 荒川隆他：Wistera/BDEC-01利用事例（6）マルチプログラム連成ライブラリh3-Open-UTIL/MP（2/2）
  https://www.cc.u-tokyo.ac.jp/public/VOL24/No4/09_202207-Wisteria-1.pdf