



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



「計算・データ・学習」融合とポストコロナ時代の 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

2022年6月24日

PCクラスタワークショップ in 神戸2022
「クラウドとHPC」

中島 研吾

「地球シミュレータ」に
~~翻弄された~~ 導かれた人生

- 大学卒業後(工学部航空学科)
- 三菱総合研究所(1985-1999)
 - 原子力関連諸業務
 - テキサス大学(Dept. Aerospace Eng. & Eng. Mechanics)(1991-1993)
- 高度情報科学技術研究機構(RIST)(1999-2004)
 - GeoFEMやってみました
 - 博士(工学)(東京大学大学院)(2003)
- 東京大学
 - **理学系研究科地球惑星科学専攻(2004-2008) COE特任教員**
 - 情報基盤センター(2008-)
- 海洋研究開発機構(非常勤)(2008-2011)
- 理化学研究所
 - 計算科学研究機構(非常勤)(2009-2018)
 - 計算科学研究センター・副センター長(2018-)
- 数値流体力学, 計算力学, 数値線形代数, 並列アルゴリズム

クラウドとHPC (1/2)

- 「Grid」のことを久々に思い出す
- 昔昔, 「Grid vs. スパコン」というような議論が盛んだった時期がある
- 事業仕分けのときも「1PFのシステム10個揃えたら?」というような人が居た
- 今も「クラウド vs. HPC」という図式で捉え, 「バッチジョブのスパコンは時代遅れ」と言いたがる人もある程度はいる。ただ今回はそういう場ではないはず。
- 少し前の話をふりかえってみることに...

今日, 話しの主題の一つにしたいことが書かれていた

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 Today
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

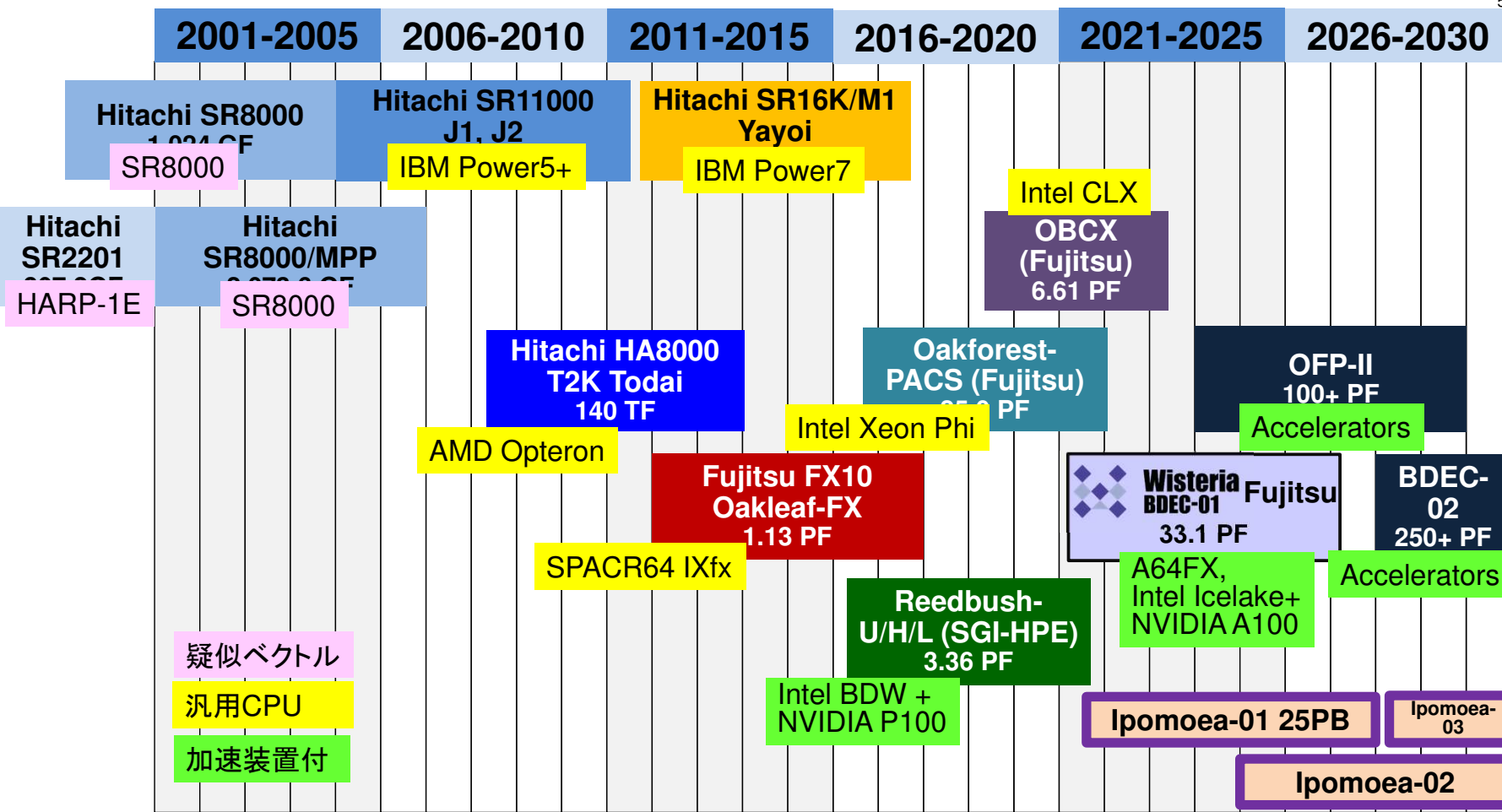
Reedbush-
U/H/L (SGI-HPE)
3.36 PF

Ipomoea-01 25PB

Ipomoea-
03

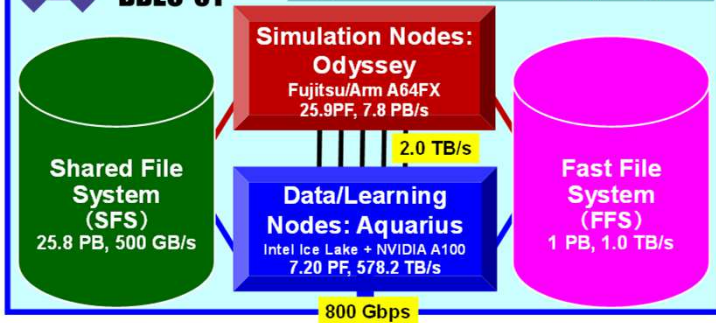
Ipomoea-02

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





Platform for Integration of (S+D+L)
Big Data & Extreme Computing



External Resources



External Network



External Resources



Wisteria
BDEC-01



Wisteria
BDEC-01

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クラスター, ピーク性能3.36 PF

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6月末 (予定)
- 6.61 PF, #119 in 59th TOP500 (June 2022)



Wisteria/BDEC-01 (Fujitsu)

- シミュレーションノード群 (Odyssey) : A64FX (#20)
- データ・学習ノード群 (Aquarius) : Intel Icelake+NVIDIA A100) (#115)
- 33.1 PF, #20 in 59th TOP 500, 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)



Reedbush



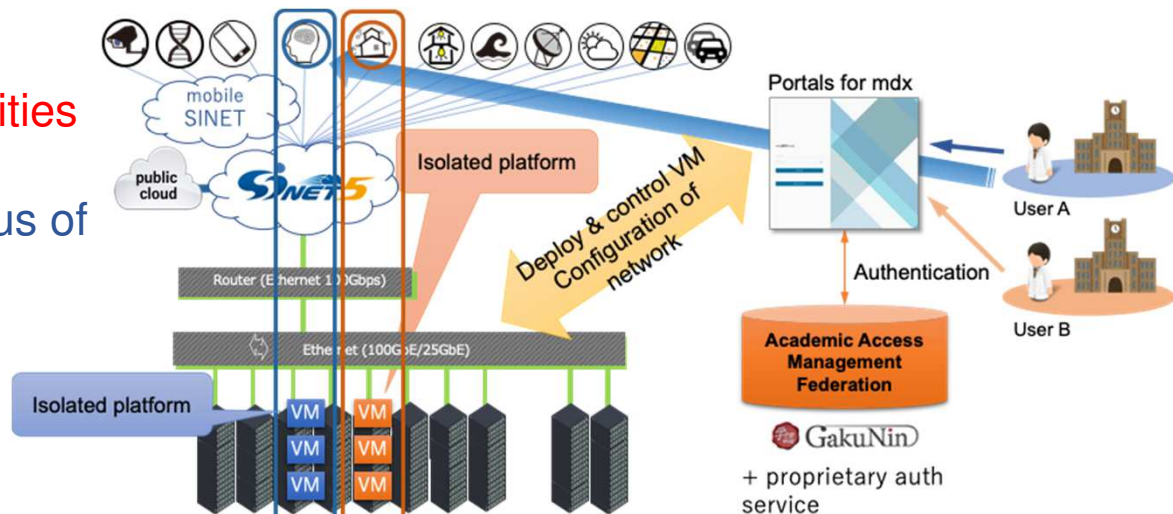
Oakforest-PACS



Oakbridge-CX

“Data Platform”

- High-performance virtualization environment focusing on leveraging data with security like “Cloud” using same HW as supercomputers.
- National joint usage system jointly operated by 9 universities and 2 research institutes
- Located at Kashiwa II campus of U.Tokyo



June 6, 2022



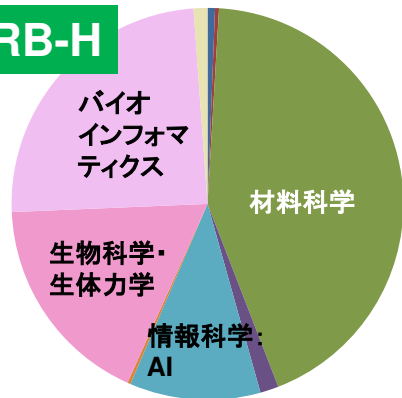
クラウドとHPC (2/2)

- クラウド＝「やりたいこと」を「やりたい時」にできる環境
 - 昨日のAmazonさんの話し
- それぞれの計算機は得意な計算をやる
 - ヘテロジニアスな環境になりうる

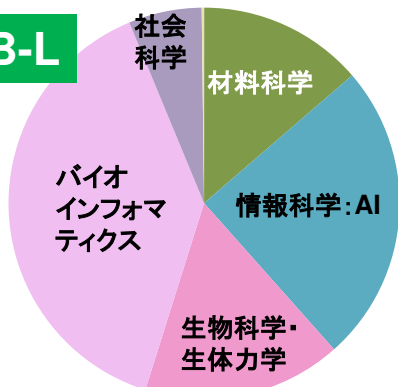
2021年度分野別 ■ 汎用CPU, ■ GPU

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

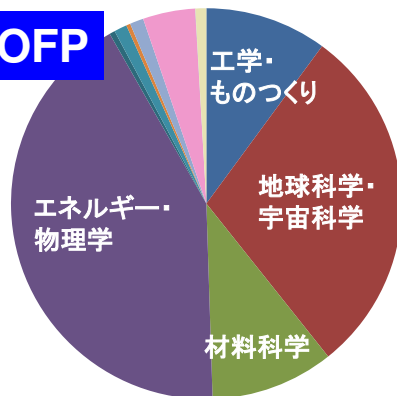
RB-H



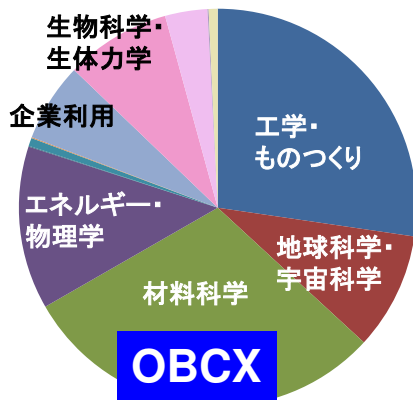
RB-L



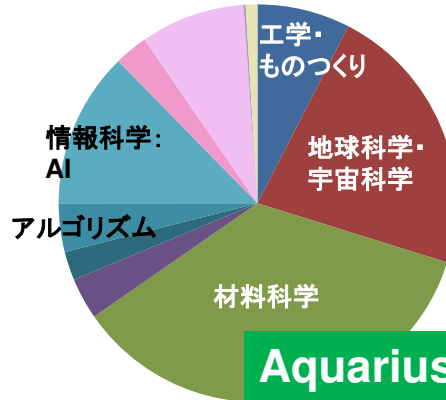
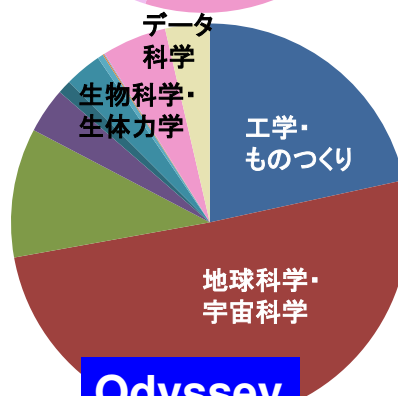
OFP



- 工学・ものづくり
- 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



Odyssey



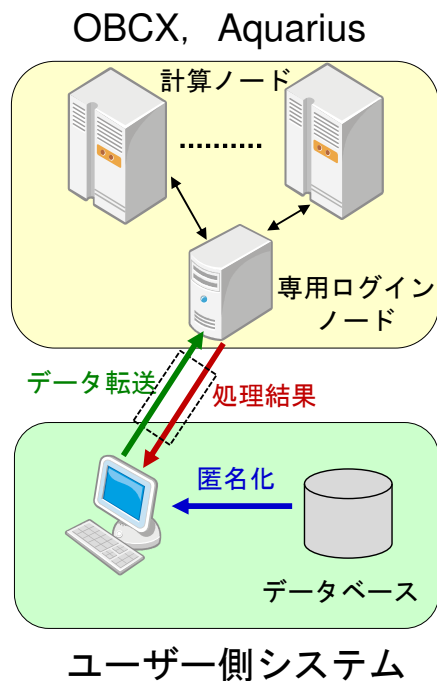
地球科学・宇宙科学分野ではOFP ⇒ Wisteria/BDEC-01への移行が順調に進んでいる

クラウドとHPC (2/2)

- クラウド＝「やりたいこと」を「やりたい時」にできる環境
 - 昨日のAmazonさんの話し
- それぞれの計算機は得意な計算をやる
 - ヘテロジニアスな環境になりうる
- Wisteria/BDEC-01における経験をもとに、「スーパーコンピューティング」の視点から、東大センターが研究機関、資源提供機関(企業利用含む)として、「クラウド+HPC」にどのような貢献ができるか、について語る
 - 多様な計算機資源の利用
 - 大量の計算機資源の確保
 - セキュリティ, ストレージ, ネットワーク...普通の話はしない(できない)
 - HPC≡Computational Science & Engineering (CSE), ではあるが, D(Data), L(Learning)も含む

現状：ノード固定：OBCX, Aquarius

- ノード占有, 個別なカスタマイズが可能
 - 高度な機密性
 - 専用ログインノード, 専用ストレージ, 専用回線
 - 個別ソフトウェア, 商用コードインストール
 - 「長時間専有希望」原則 $\times \Rightarrow$ 増えつつある, MLユーザー
- 審査(ヒアリング)あり, 負担金は通常の1.5倍**
- T2K東大(2008年)より実施
 - GOSAT「いぶき」観測データ自動処理(国立環境研)
 - 医療画像処理(東大病院)
- Aquarius: 8GPU/ノード
 - ノード固定(8GPU)
 - GPU専有(1,2,4GPU)(審査無し)
 - 特定のノードのGPUを利用, どのGPUを使うかは変動
 - OdysseyはHWの構成上「ノード固定」はない $\Rightarrow$ トークン消費量1.50倍の優先キューがある
- mdx: よりフレキシブルに「ノード固定」実現**



(シミュレーション(計算)+データ+学習)融合(S+D+L)

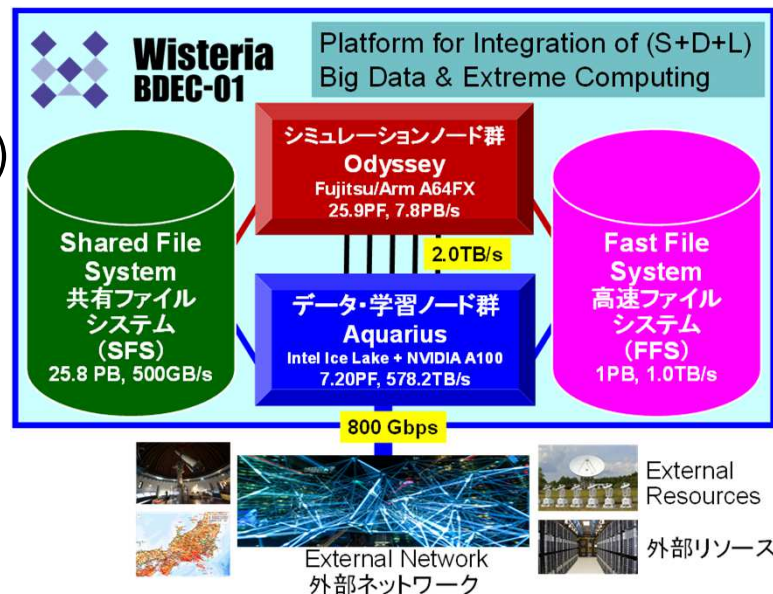
- 東大情報基盤センターでは、2015年頃から「(S+D+L)融合」の重要性に注目し、それを実現するためのハードウェア、ソフトウェア、アプリケーション、アルゴリズムに関する研究開発を開始
 - BDEC計画(Big Data & Extreme Computing)
 - 「データ+学習」による、より高度な「シミュレーション」
 - AI for HPC, AI for Science, Digital Twin
 - 地球科学関連では自然な発想(すでに実施されている)
- 2021年5月に運用を開始した「Wisteria/BDEC-01」は「BDEC計画」の1号機
 - Reedbush, Oakbridge-CX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

- 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), ~360m²
- 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

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)



**Wisteria
BDEC-01**

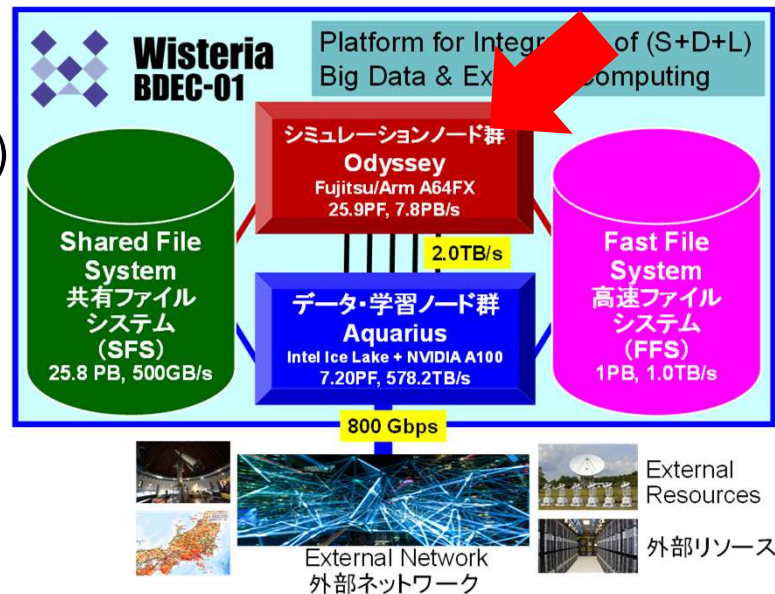
Wisteria/BDEC-01

- 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), ~360m²
- 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

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)

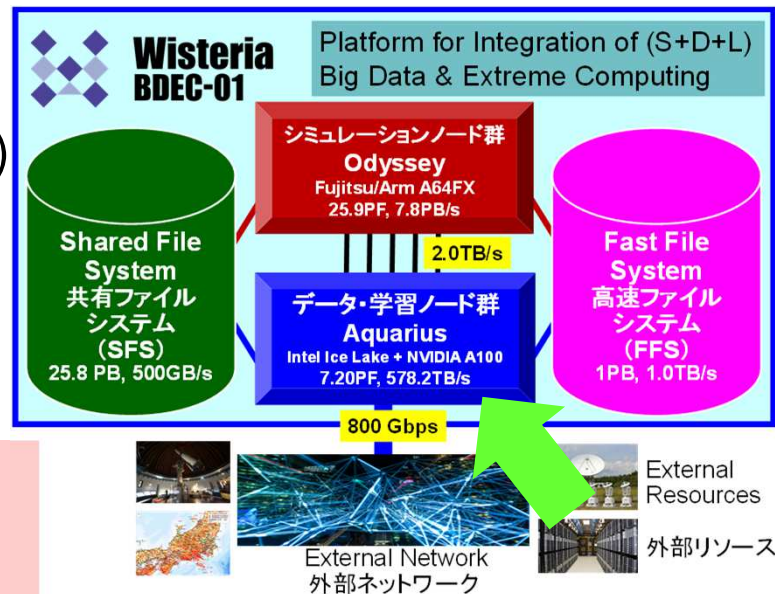


**Wisteria
BDEC-01**

Wisteria/BDEC-01

- 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), ~360m²
- 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

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)



**Wisteria
BDEC-01**

Simulation Nodes Odyssey

25.9 PF, 7.8 PB/s

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

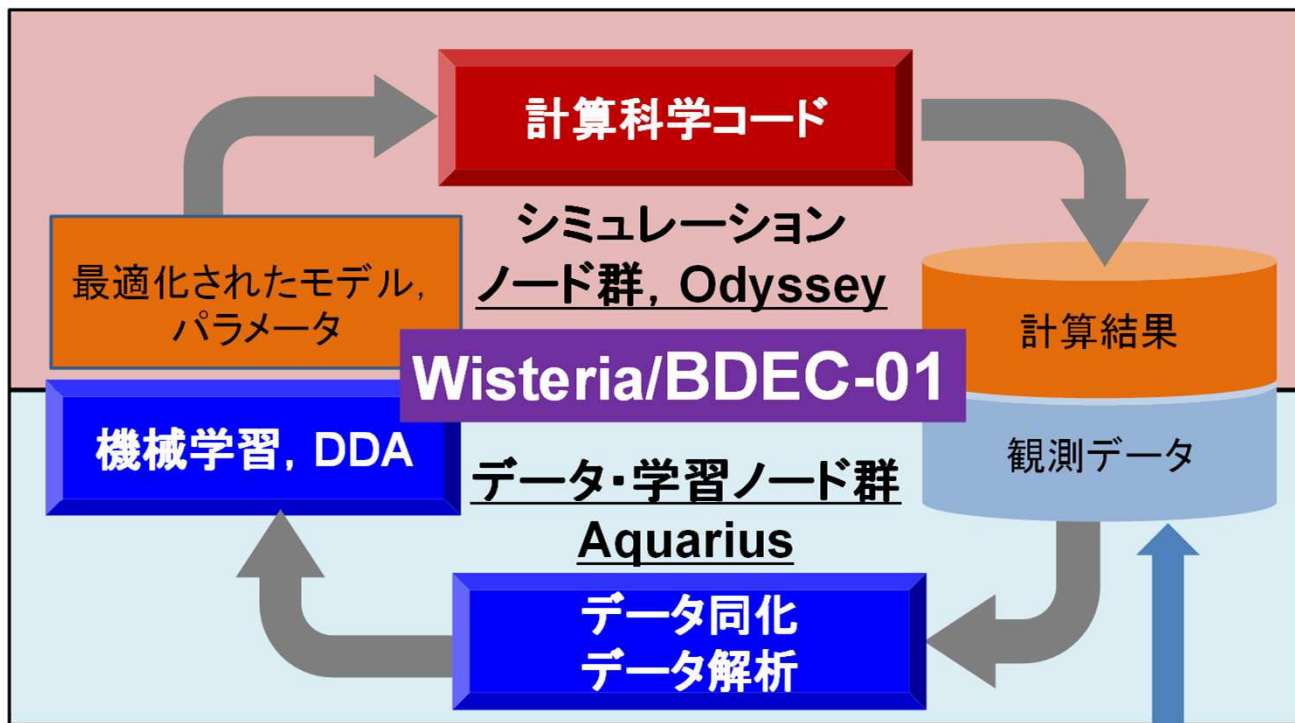
Shared File
System
(SFS)
25.8 PB,
0.50 TB/s

Data/Learning Nodes Aquarius

7.20 PF, 578.2 TB/s



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

h3-Open-BDEC

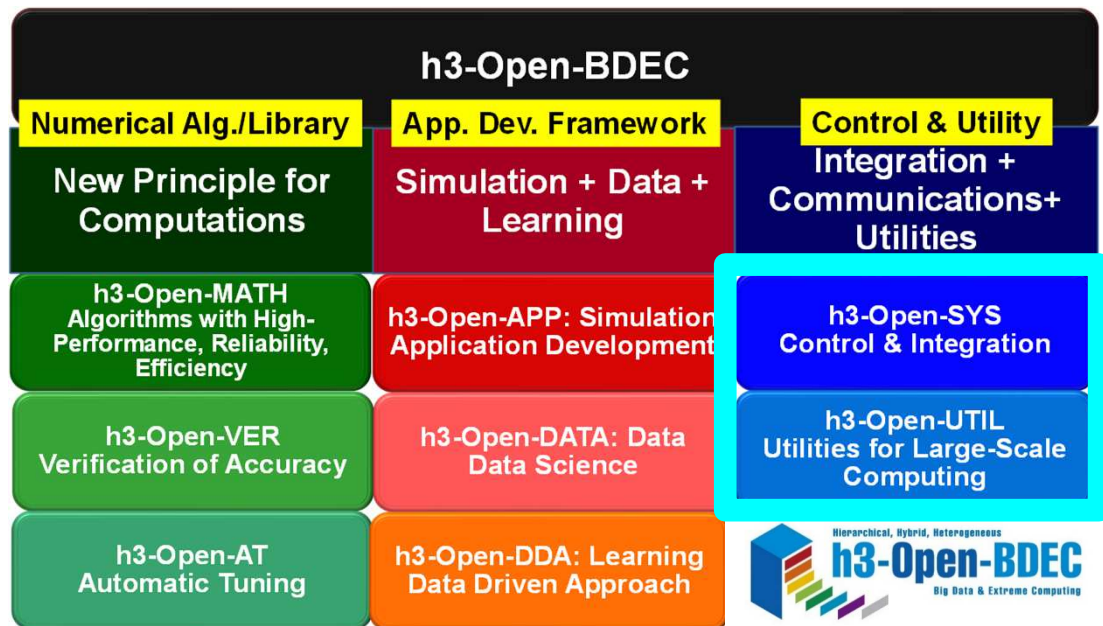
「計算＋データ＋学習」融合を実現する革新的ソフトウェア基盤
科研費基盤研究(S)(2019年度～23年度, 代表: 中島研吾)

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

Hierarchical,
Hybrid,
Heterogeneous

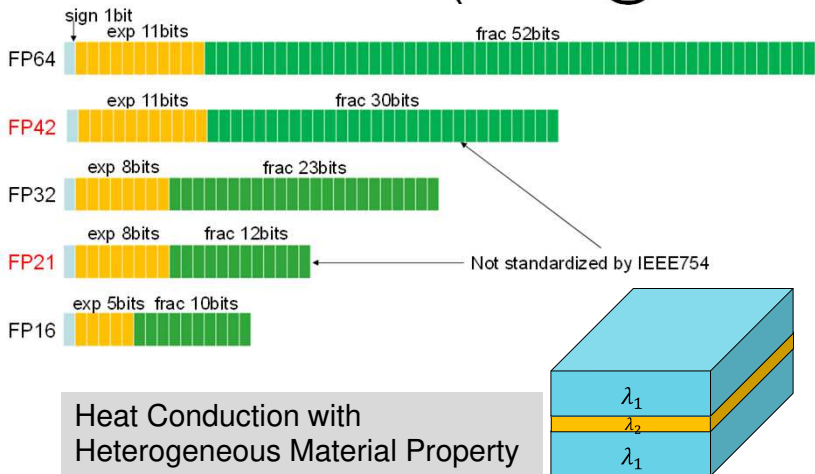
Big Data &
Extreme
Computing

- ① 新計算原理に基づく革新的数値解法
- ② 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(第3の柱)

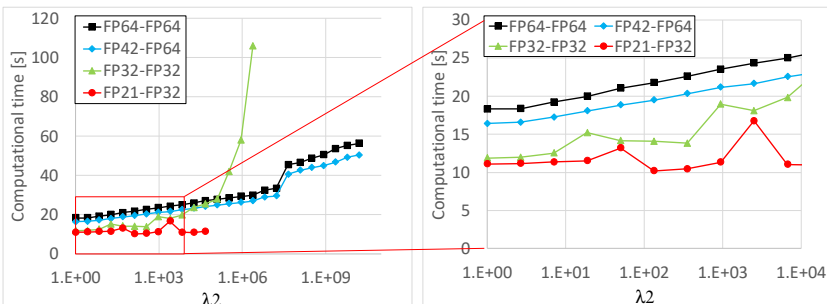


Adaptive Precision Computing with FP42/FP21

Masatoshi Kawai (kawai@cc.u-tokyo.ac.jp)



Heat Conduction with
Heterogeneous Material Property



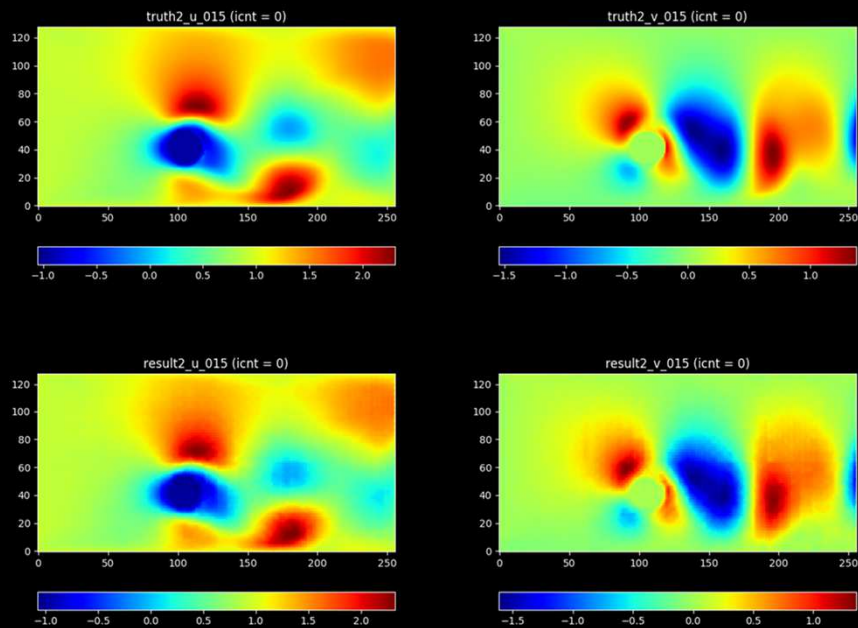
Computation Time for ICCG Solver
Various Types of Precisions on Intel Xeon Cascadelake

In recent years, the usefulness of low-precision floating-point representation has been studied in various fields such as machine learning. Low accuracy can be expected to have effects such as shortening calculation time and reducing power consumption. For example, in an application with a memory bandwidth bottleneck, the effect of reducing the calculation time by reducing the amount of memory transfer is significant. However, in fields such as iterative methods, it is common to use FP64 because the calculation accuracy strongly affects the convergence, and there are few application examples of low-precision arithmetic. This study investigates the applicability of low-precision representation to the Krylov subspace and stationary iterative methods. In this research, we focus on the FP32, FP16, and FP42, FP21, which are not standardized by IEEE754.

Developed method has been evaluated for ICCG solver, which solves linear equations derived from 3D FVM code for steady-state head conduction with heterogeneous material property ($\lambda_1=10^0$, $\lambda_2=10^0 \sim 10^9$). Generally, computation with lower precision (e.g. FP32-FP32, FP21-FP32) becomes unstable, if condition number of the coefficient matrix is larger (λ_2 is larger), FP21-FP32 provides the best performance if λ_2 is up to 10^4 . (“FP21-FP32” means “matrices are in FP21, and vectors are in FP32”).

Prediction of CFD Simulation by Deep Learning

Takashi Shimokawabe (shimokawabe@cc.u-tokyo.ac.jp)



Computational fluid dynamics (CFD) is widely used in science and engineering. However, since CFD simulations require a large number of grid points and particles for these calculations, these kinds of simulations demand a large amount of computational resources such as supercomputers. Recently, deep learning has attracted attention as a surrogate method for obtaining calculation results by CFD simulation approximately at high speed. We are working on a project to develop a parallelization method to make it possible to apply the surrogate method based on the deep learning to large scale geometry. Unlike the model parallel computing, the method we are currently developing predicts large-scale steady flow simulation results by dividing the input geometry into multiple parts and applying a single small neural network to each part in parallel. This method is developed based on considering the characteristics of CFD simulation and the consistency of the boundary condition of each divided subdomain. By using the physical values on the adjacent subdomains as boundary conditions, applying deep learning to each subdomain can predict simulation results consistently in the entire computational domain. It is possible to predict the simulation results in about 36.9 seconds by the developed method, compared to about 286.4 seconds by the conventional numerical method. In addition to this, we are also attempting to develop a method for fast prediction of time evolution calculations using deep learning.

Comparison of the flow velocity results obtained by the conventional simulation (upper) and the prediction of these results by deep learning (lower)

h3-Open-BDEC

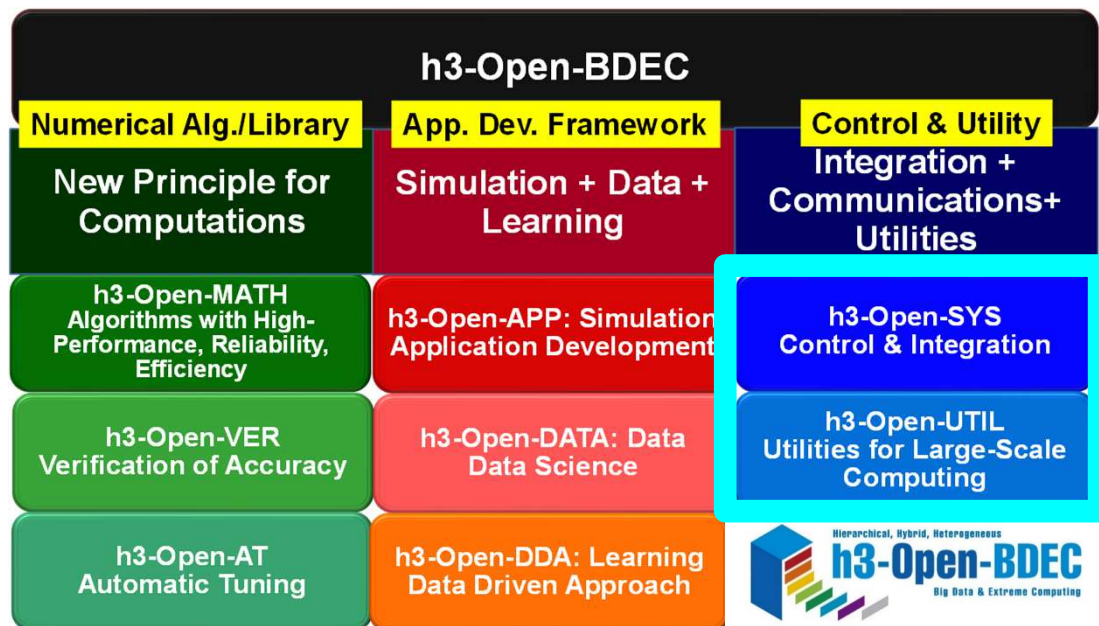
「計算＋データ＋学習」融合を実現する革新的ソフトウェア基盤
科研費基盤研究(S)(2019年度～23年度, 代表: 中島研吾)

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

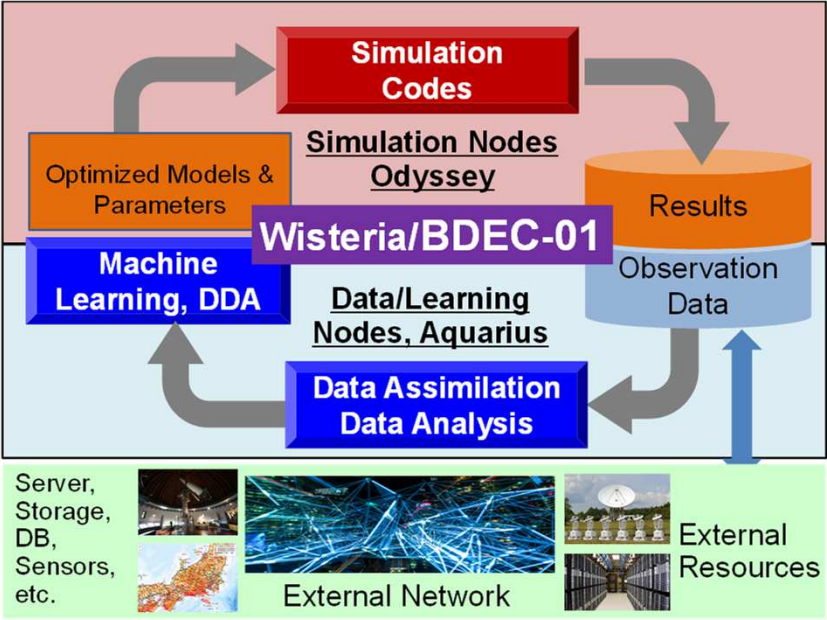
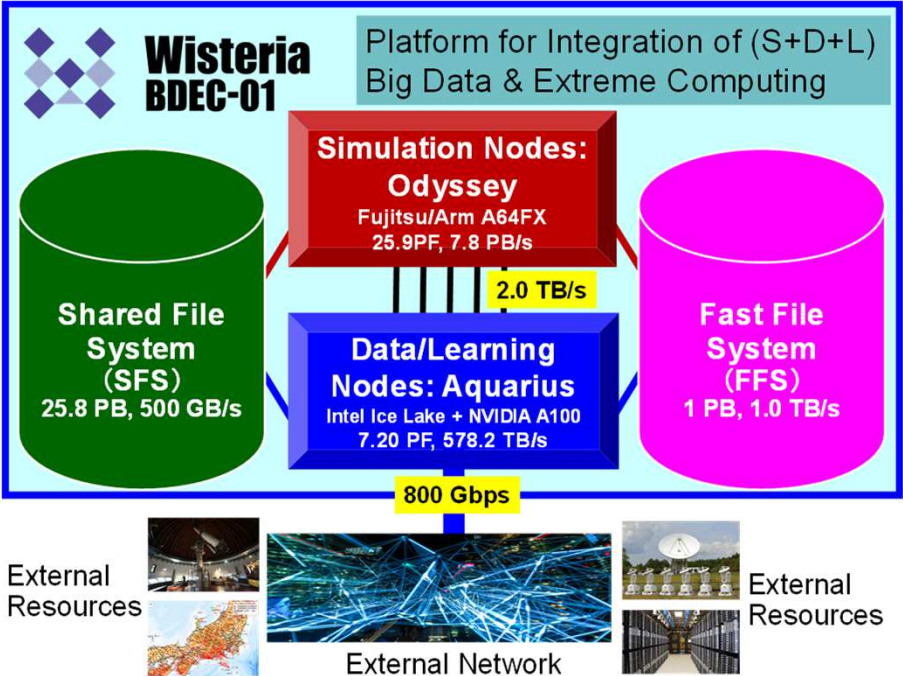
Hierarchical,
Hybrid,
Heterogeneous

Big Data &
Extreme
Computing

- ① 新計算原理に基づく革新的数値解法
- ② 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(第3の柱)



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



AI for HPC, AI for Science の実現へ向けて



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

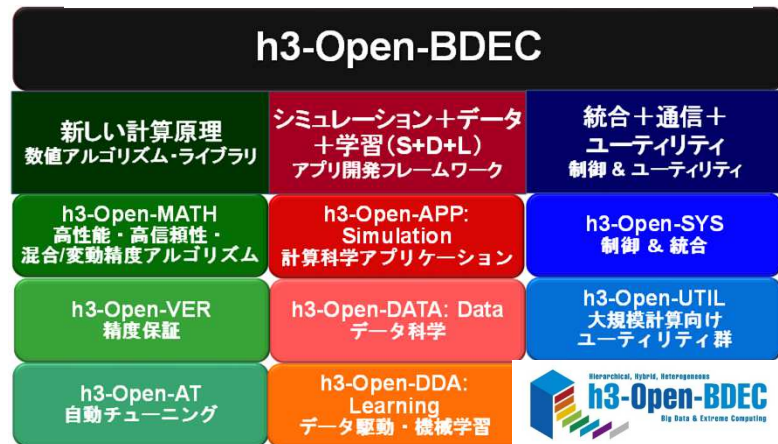
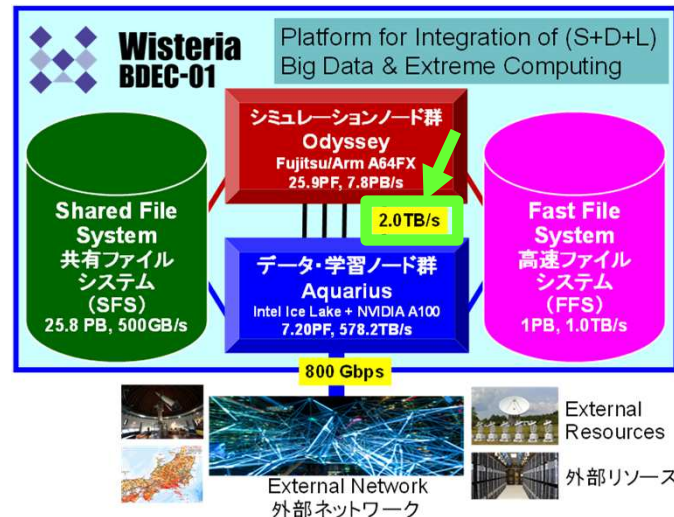
ソフトウェア開発

– 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)

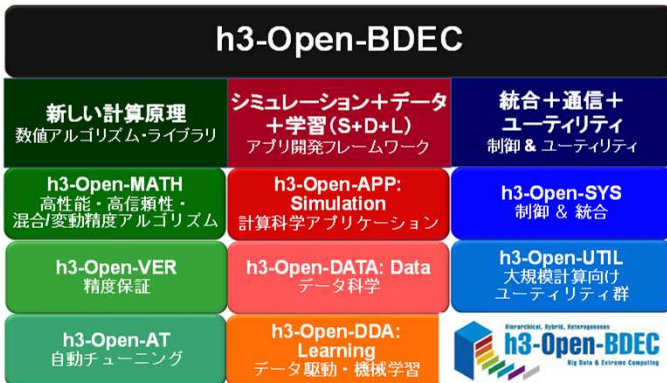
• 高速ファイルシステム (FFS) 経由連携 (WaitIO-File)



h3-Open-SYS/WaitIO

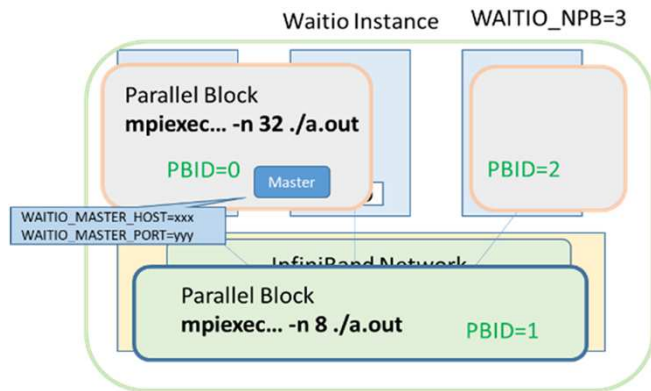
データ受け渡しライブラリ〔松葉, 2020〕
〔住元他, HPC-181, 2021〕

- 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

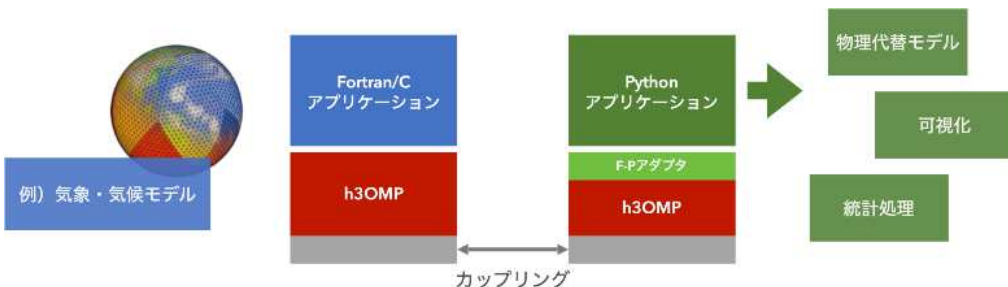


[Sumimoto et al. 2021]

「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。

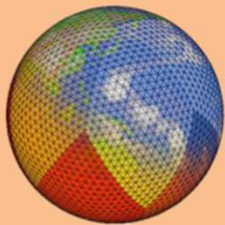


Fortran/CアプリとPythonアプリの連成計算の模式図
〔八代・荒川 2020〕

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



ARM: A64FX



A huge amount of
simulation data
output

HPC App
(Fortran)

h3o-U/MP

IceLake+A100

Analysis/ML
App
(Python)

F<->P adapter

h3o-U/MP

Surrogate
Model

Visualiztion

Statistics

Coupling

IB-EDR



**Wisteria
BDEC-01**

Odyssey



**Wisteria
BDEC-01**

Aquarius

h3-Open-UTIL/MP・

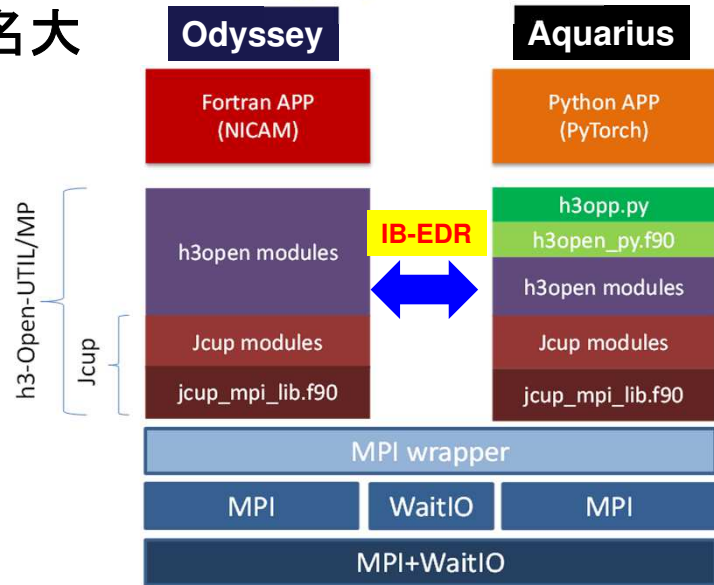
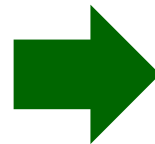
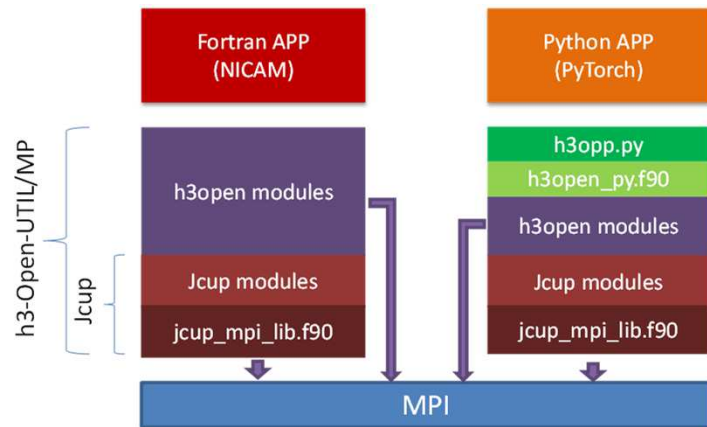
h3-Open-SYS/WaitIO-Socket連携

2022年6月から利用可能

2022年度はFS経由のWaitIO-File整備：名大

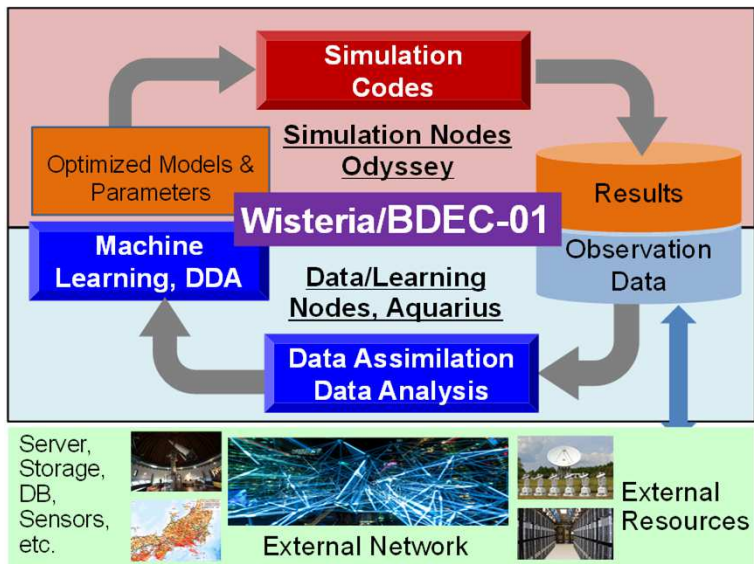


**Wisteria
BDEC-01**

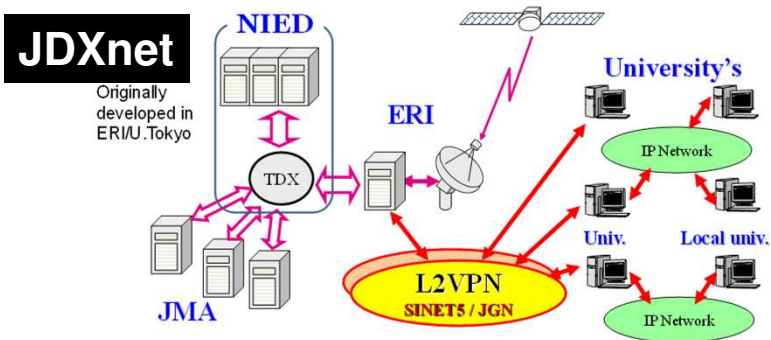
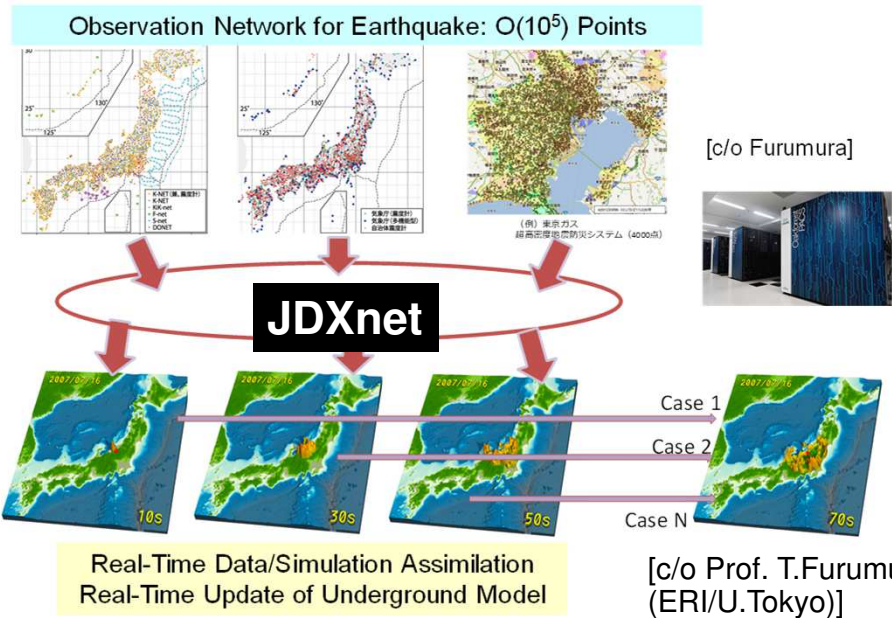


2021年4月：MPI通信可能な環境を前提

2022年6月：Coupler + WaitIO



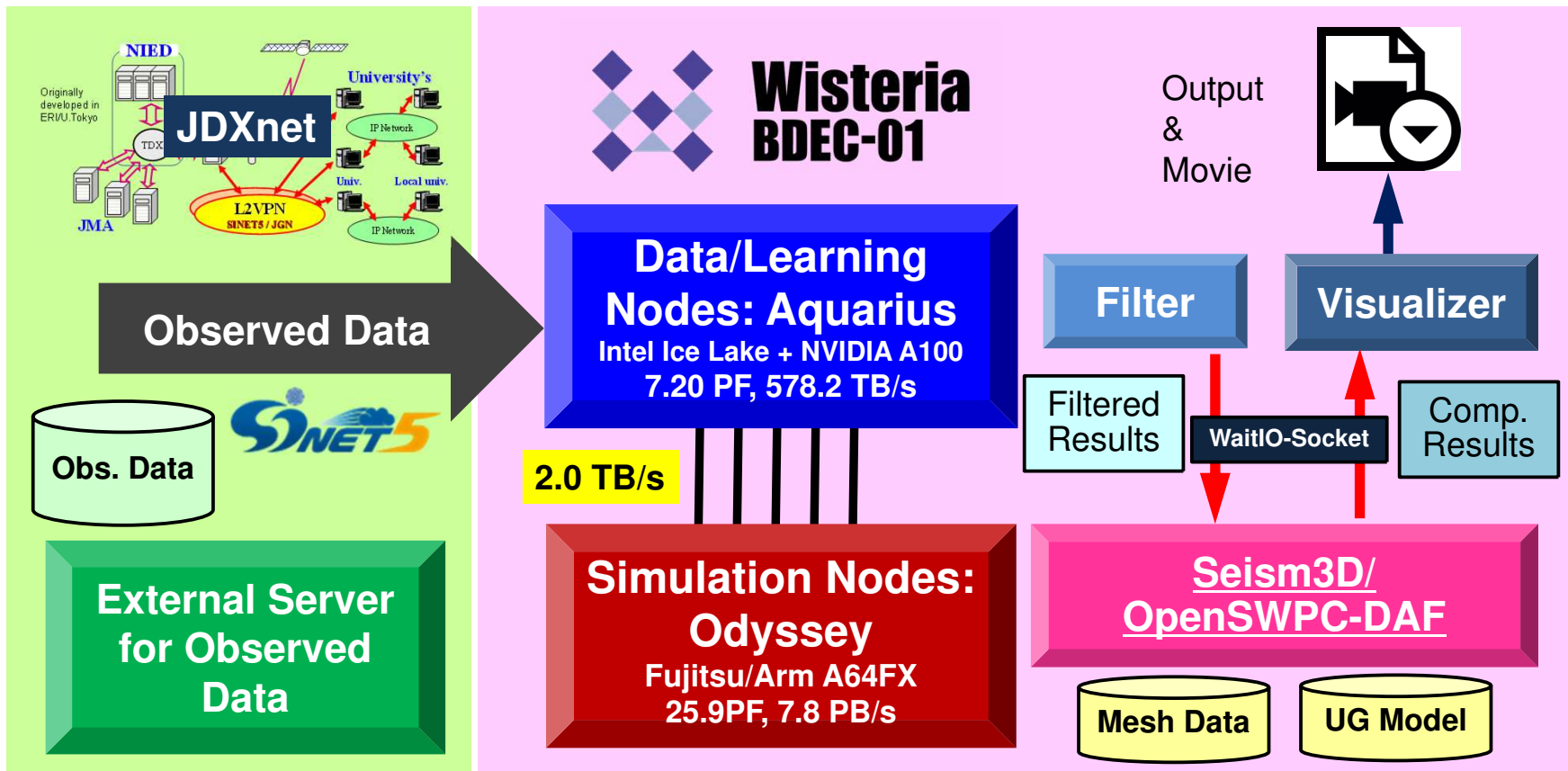
リアルタイムデータ同化＋ 3D強震動シミュレーション融合 JDXnetによるリアルタイム観測データ活用



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

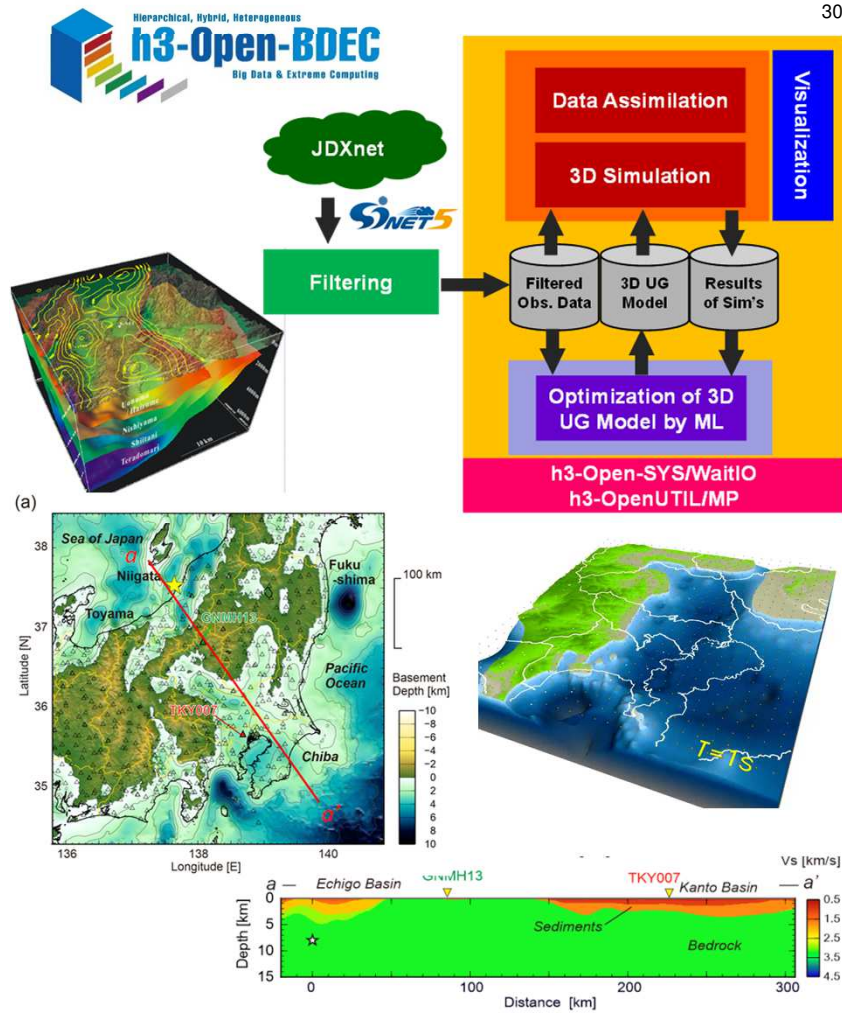
長周期地震動シミュレーション＋観測データ同化

他



(計算・データ・学習)融合へ 向けた(更なる)取り組み

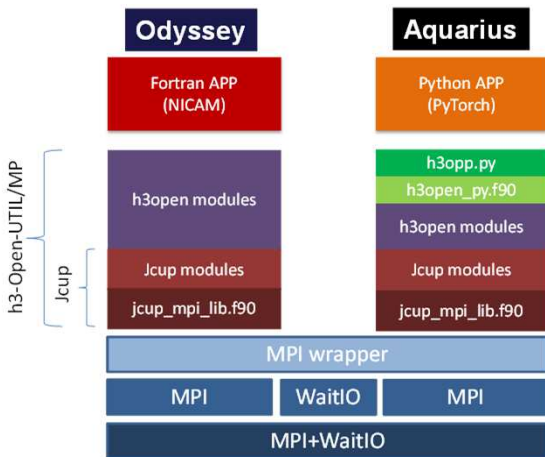
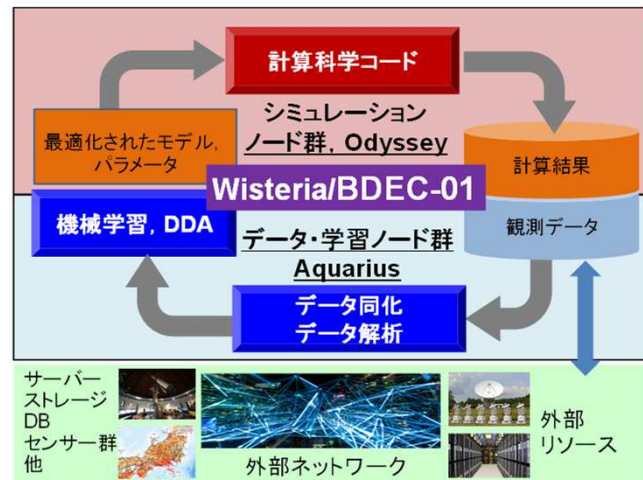
- 大地震時: 正確な地震波動伝播予測
 - 地震波観測データを同化して得られる変位分布を初期条件として入力
 - 100秒先の予測を10秒以内に計算
 - 的確な避難計画の策定, 震源は遠い場合
- 通常時: 地下構造モデル改良
 - 地下構造は複雑, 不均質, 実はよくわかっておらず, 大小の地震時の逆解析等によりモデルを少しずつ改良する, のが現状
 - 観測データ・三次元シミュレーション・データ解析を元に地下構造モデルの改良に使う
 - ✓ 例えばMw=3.0+の地震が起きた場合
 - ✓ 機械学習により精度高いモデル生成
 - ⇒ (S+D+L) 融合へ



MATLABの導入 「S+D+L」融合, AI for HPCの実現



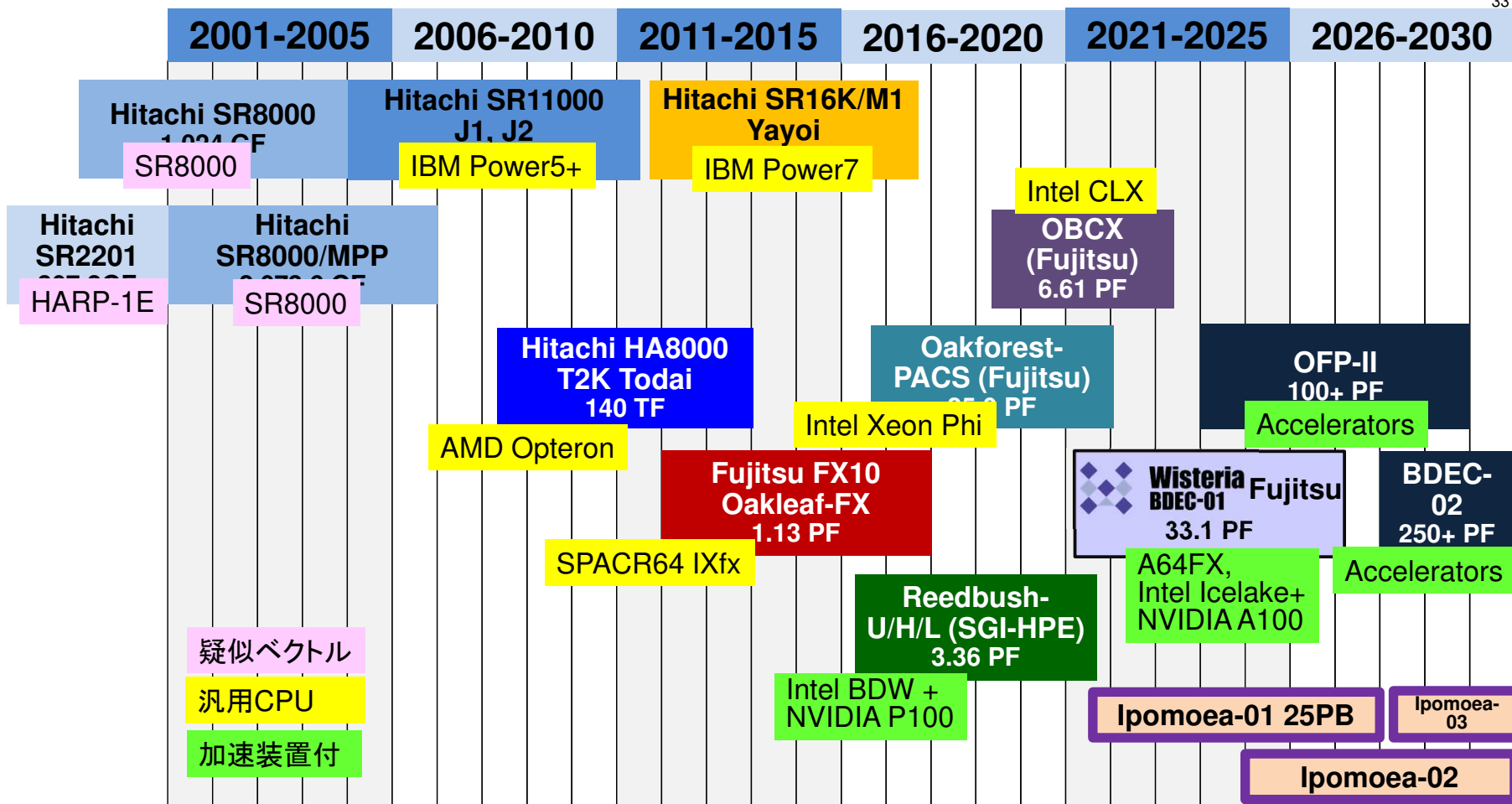
- 2022年3月からOBCX, Aquariusで利用可能
- MATLAB
 - ✓ 多様な機能
 - ✓ ユーザーのプログラムからの関数呼び出し重視⇒データ解析, 機械学習系の豊富な機能⇒高度化
 - ✓ MATLABはAquarius(データ・学習ノード群)でのみ稼働するが, h3-Open-BDECと連携させて, Odyssey(シミュレーションノード群)上で実施する大規模シミュレーションのパラメータ最適化に適用する⇒「S+D+L」融合, AI for HPC
- h3-Open-BDECは様々な環境で動作⇒MATLABと組み合わせた使用による普及



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





で？

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 Today
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

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

Reedbush-
U/H/L (SGI-HPE)
3.36 PF

Mercury

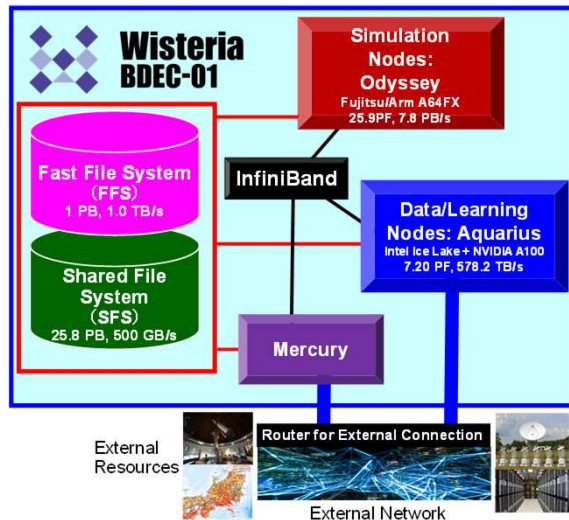
Ipomoea-01 25PB

Ipomoea-
03

Ipomoea-02

東大センター将来構想(1/2)

- 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6月～)**
 - GPUクラスタ
 - OFP-IIプロトタイプ (32+ノード, 128+GPU)
- **OFP-II (2024年4月)**
 - OFP後継機 (JCAHPC)
 - GPUクラスタ (Mercuryと同じ, もしくはその同型GPU), 100+PF
 - **CPU Only**のノード群もあり



JCAHPC



筑波大学
University of Tsukuba



東京大学
THE UNIVERSITY OF TOKYO

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 Today
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

Mercury

Ipomoea-01 25PB

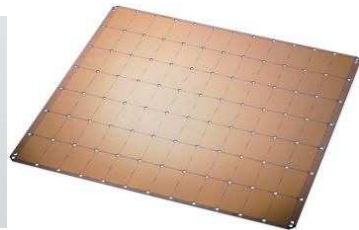
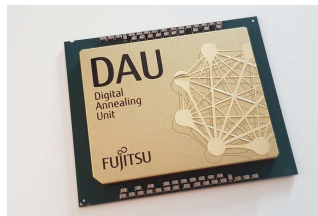
Ipomoea-
03

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

東大センター将来構想(2/2)

- BDEC-02(2027秋～2028春)
 - 「S+D+L」融合プラットフォーム
 - DPU, IPU, Quantum-Inspired Deviceによる(D+L)ワークロードサポートを検討, でも買うと高い！
 - ✓ Graphcore, SambaNova, Cerebras
 - ✓ PIUMA@Intel
 - ✓ Digital Annealer@Fujitsu
- 東大情基セ: Digital Annealer使用(2019年～)
 - 組み合わせ最適化問題への適用
- DPU/IPU/QID/DAとスパコンのクラウド経由連携
 - プログラミング環境, 通信ライブラリ整備必要
 - ✓ WaitIOの拡張: WaitIO-over-Cloud
 - 各社からは「OFP-IIの段階で行ける」との声がある
 - ✓ ALCF AI Testbed, NECさんのご発表



ALCF AI Testbeds

<https://www.alcf.anl.gov/alcf-ai-testbed>



Cerebras (CS-2)



SambaNova



Graphcore



Habana



Groq

➢ Infrastructure of next-generation machines with hardware accelerators customized for artificial intelligence (AI) applications.

➢ Provide a platform to evaluate usability and performance of machine learning based HPC applications running on these accelerators.

➢ The goal is to better understand how to integrate AI accelerators with ALCF's existing and upcoming supercomputers to accelerate science insights

eX3@Simula Research Laboratory, Norway

Experimental Infrastructure for Exploration of Exascale Computing

<https://www.ex3.simula.no/resources/>

simula

simula
+
GRAPHCORE

IN PARTNERSHIP WITH



- The eX3 infrastructure includes an IPU-POD, consisting of 64 Colossus Mk2 GC200 IPUUs with 1,472 cores each, making up a total of more than 94K cores.
- The total In-Processor Memory™ is more than 57 GB supported by a memory bandwidth of 47.5 TB/s per IPU. This IPU-POD is capable of 16 petaflops mixed precision peak performance. The IPU processor has been designed bottom-up for machine learning and AI workloads.



Simulation Nodes Odyssey

25.9 PF, 7.8 PB/s

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

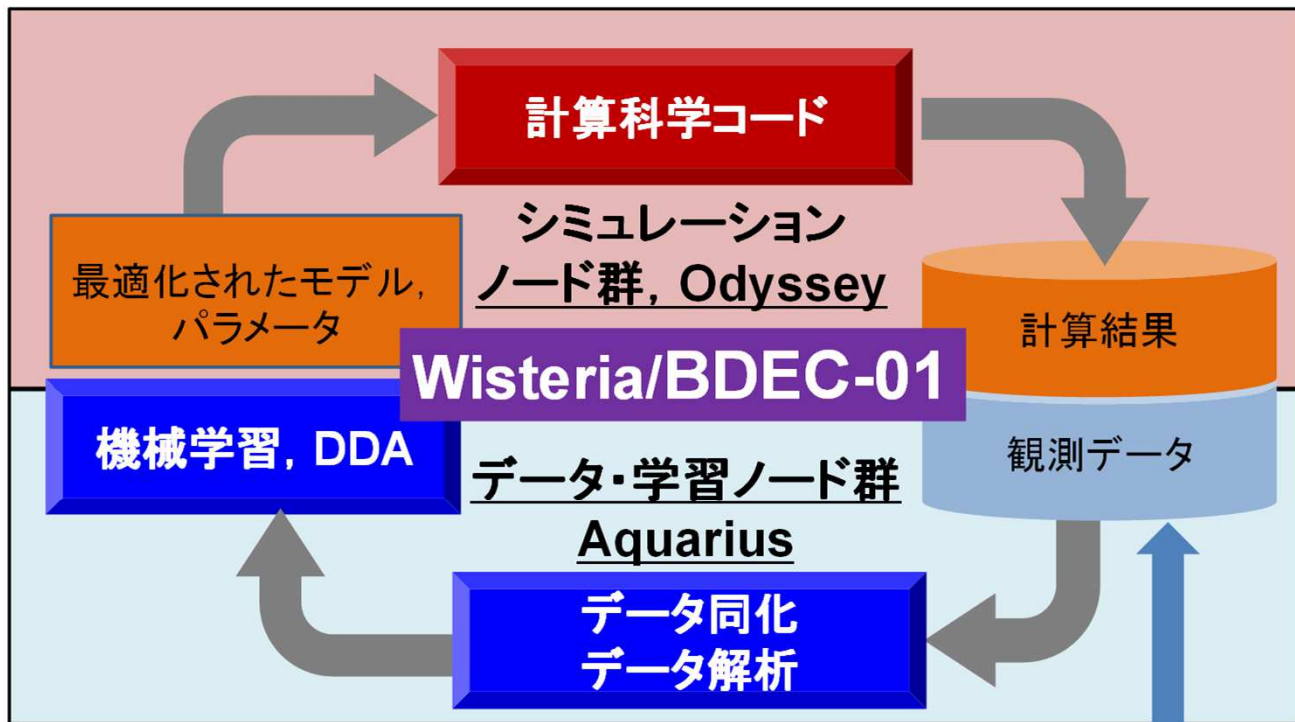
Shared File
System
(SFS)
25.8 PB,
0.50 TB/s

Data/Learning Nodes Aquarius

7.20 PF, 578.2 TB/s



Wisteria BDEC-01



サーバー
ストレージ
DB
センサー群
他

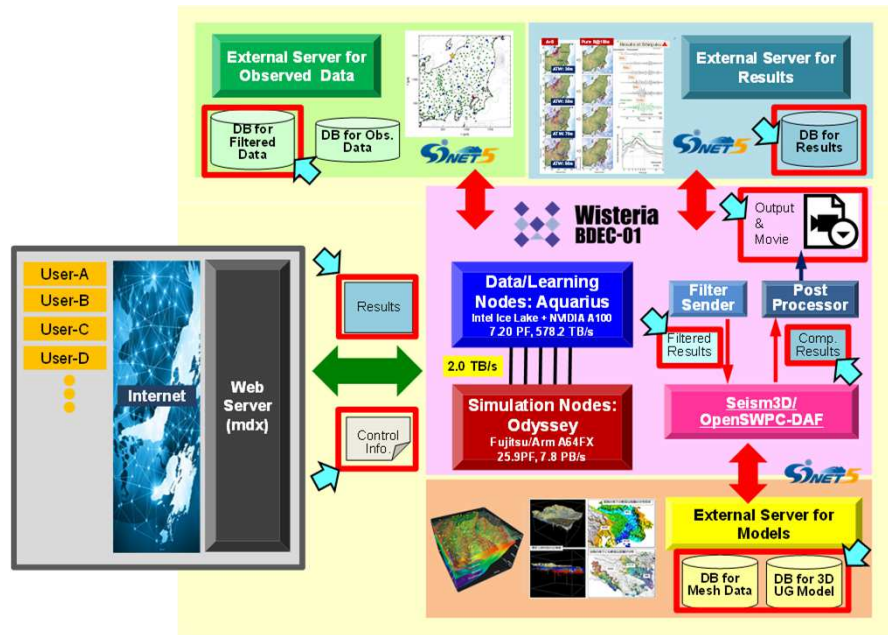


外部ネットワーク



外部
リソース

Webベース シミュレーション体験・ データ利活用システム



- 「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

実現象の10分の1の時間で計算: $O(10^3)$ ノード必要

- I/O用通信が含まれている

- min.: 通信含まず
- max.: 通信含む

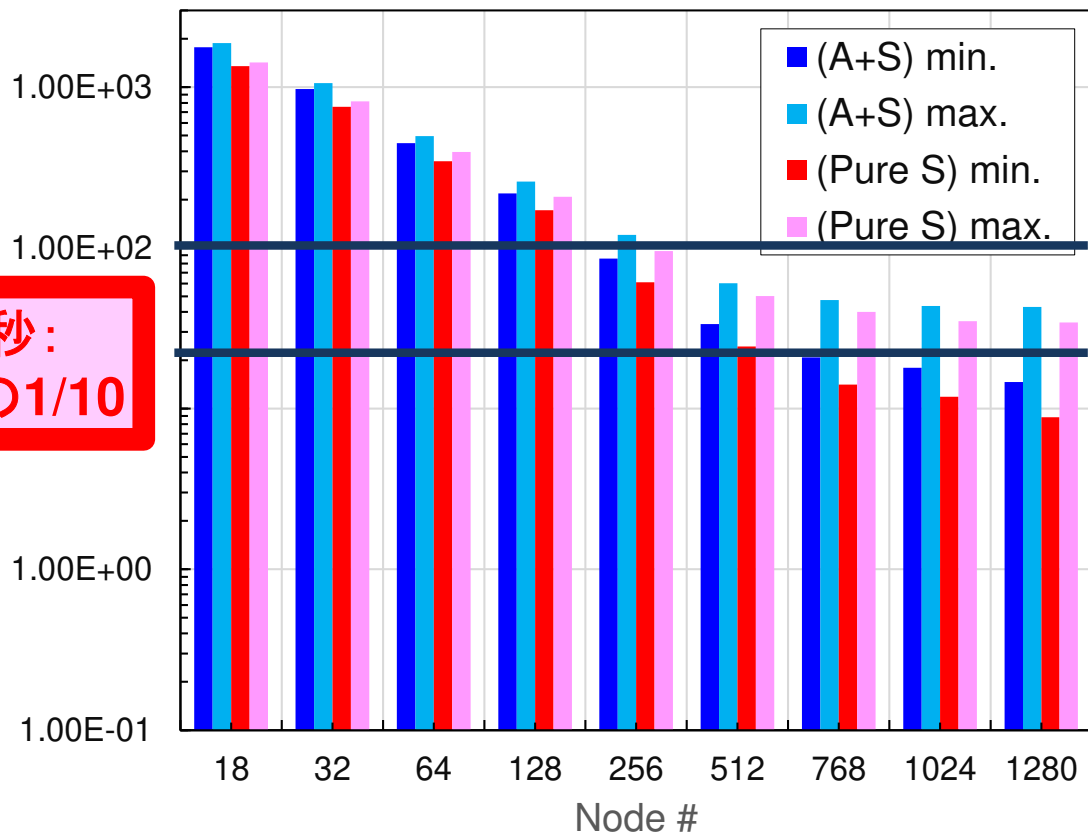
- (A+S)

- 現実の半分(100秒)で
できれば良い
- 300-400ノード程度

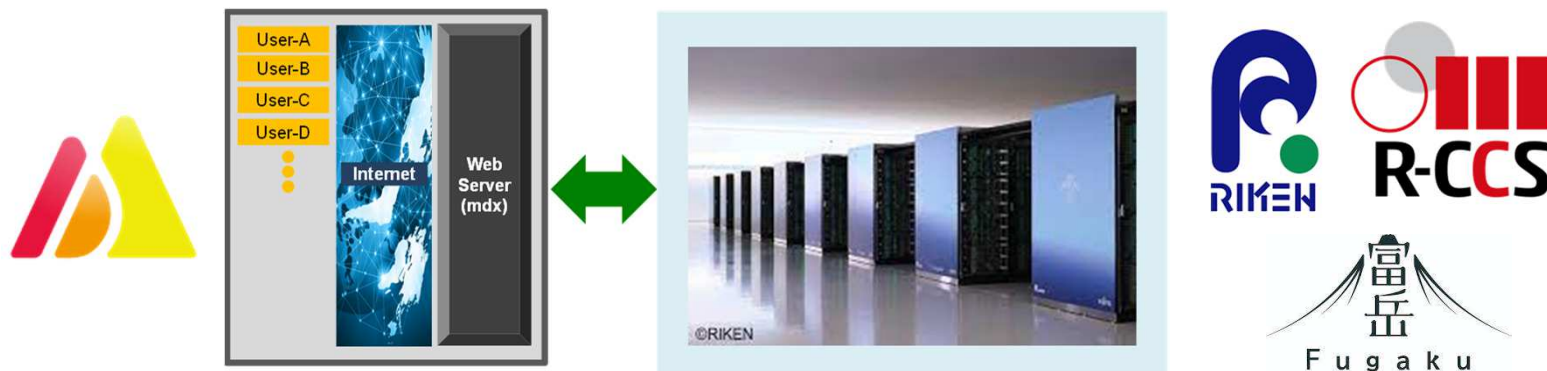
- (Pure S)

- 現実の10分の1(20秒)が求
められる
- 1,000ノード以上(OBCX)

20秒:
現実の1/10



理研R-CCS人材育成プログラムにおける シミュレーション・計算科学体験システム



- 計算科学体験システムとして理研R-CCSの人材育成への活用を検討中
 - ✓ 「富岳」のクラウド的利用, ウェブサーバーをmdxに置く
- 2022年度はまず大学生以上向けに「構造解析」を題材に作成
 - ✓ 小中高生向けに広報Gと連携して2023年度以降プログラム整備
 - ✓ 今後様々なアプリに拡張

で？

どうやって1,000ノード確保するの？

Urgent ComputingとResource Provisioning

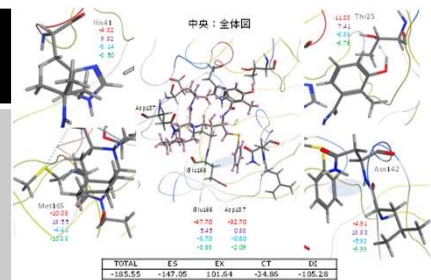
- 中島, 住元, 塙, Urgent Computingに向けたアプリケーション, 第178回 HPC研究会, 2021年3月
- Urgent Computingとは
 - 地震, 津波, 台風, 森林火災等大規模な自然災害等の緊急事態が生じた場合に, スーパーコンピュータのリソースを集中的に投入して, 迅速な被害想定と減災のための意志決定を実施
 - 東北大学サイバーサイエンスセンターではシステムレベルのチェックポイントニングを常時実施することにより, 大規模な地震, 津波等を想定したUrgent Computingに対応
- HPCI「COVID-19」課題
 - 東大: 全14課題のうち6課題(2020年度)
 - 数百ノード規模の資源を優先して割り当てる必要が生じた
 - ジョブの優先度を高めることで何とか対応
 - もっと柔軟にリソース割り当てができないものか, と考えた⇒研究としても成立(多分)

「COVID-19」対応HPCI臨時公募課題

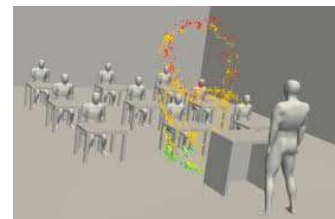
全14のうち6課題が東大システムを利用(2020年度)



課題名	代表者(所属)	使用システム
新型コロナウイルスの主要プロテアーゼに関するフラグメント分子軌道計算	望月 祐志 (立教大学)	Oakforest PACS
COVID-19治療の候補薬: chloroquine、hydroxychloroquine、azithromycinの催不整脈リスクの評価ならびにその低減策に関する研究	久田 俊明(株式会社UT-Heart研究所 / 東大)	
新型コロナウイルス表面のタンパク質動的構造予測	杉田 有治 (理化学研究所)	
計算機解析によるSARS-CoV-2増殖阻害化合物の探索	星野 忠次 (千葉大学)	Oakbridge CX
室内環境におけるウイルス飛沫感染の予測とその対策: 富岳大規模解析に向けたケーススタディ	坪倉 誠 (神戸大学)	
Spreading of polydisperse droplets in a turbulent puff of saturated exhaled air	Marco Edoardo Rosti (OIST)	



資料提供: 望月祐志教授(立教大学)



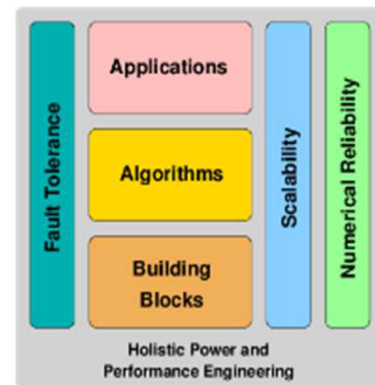
[資料提供: 坪倉誠教授(神戸大学)]

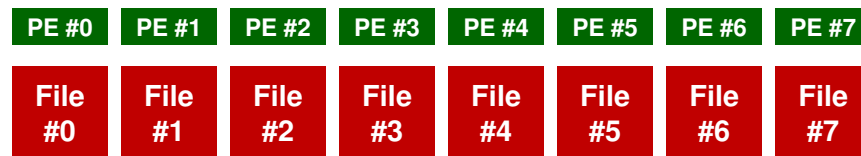
COVID-19時代のUrgent Computing

- 実行優先度と計算機資源消費率の異なる複数のジョブクラスを提供することも可能であるが、計算機リソースを予め優先度の高低によって分割しておく必要があり、計算機資源の効率的な利用が困難となる。
- 突発的な自然災害に迅速に対応することが中心であった従来のUrgent Computingでは、通常のジョブは全て一旦停止して、災害対応のワークロードに資源を集中して配分することが主要な課題であった
- COVID-19時代では、対応は長期間にわたる一方、創薬、変異種への対応など迅速な対応が求められるフェーズもあり、より柔軟な枠組みが必要
- アプリケーション、システムソフトウェアが備えるべき機能について検討
 - ULFM-MPI (User Level Failure Mitigation) によって、ジョブの実行途中で利用計算機資源量を変化させる場合を想定し、アプリケーションを試作、Oakbridge-CX上での評価を実施
 - Fault Tolerance, Fault Resilienceの研究成果を活用

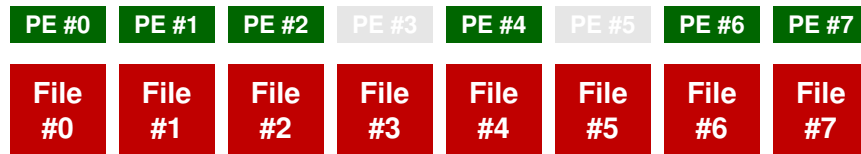
ULFM-MPIによる先行研究: 耐故障関連

- 本研究: ULFM-MPIによりMPIプロセス数を動的に変動
- User Level Failure Mitigation (ULFM)
 - MPIの拡張規格, または実装
 - 利用者がMPIにおけるプロセス故障を管理できる
 - 現在ULFMによって提供される関数群や定数群はMPI標準規格には含まれていないが, 既に実アプリケーションも含めた応用事例が多数ある
- CRAFT (Checkpoint-Restart and Automatic Fault Tolerance)
 - ULFM-MPIを利用したチェックポイントリスタートに基づく, 並列科学技術計算向けの耐故障性を有するライブラリ[Shahzad et al. 2019]
 - ESSEX-II (DFG/SPPEXA-II) で開発 (Equipping Sparse Solvers for Exascale) (Leading PI: Prof. Gerhard Wellein (FAU Erlangen-Nuremberg))
 - pFEM-CRAFT [Fukasawa et al. SIAM PP18] は, MPIを使用した並列有限要素法による非定常熱伝導アプリケーション (pHEAT-3Dと同じ) に対して, CRAFTを使用して, Spare Nodeを使用しない耐故障フレームワークを提案, 実装, 動作を検証

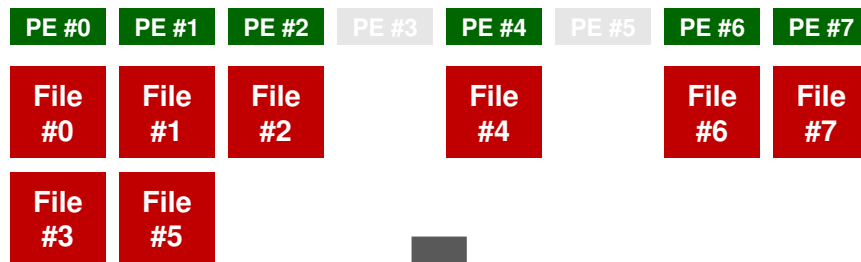




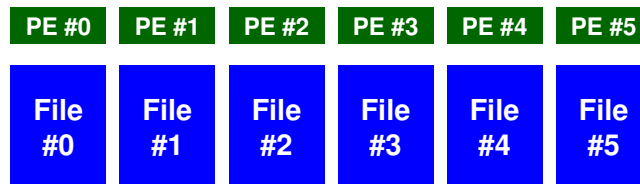
Starting
with 8 Proc's



#3 & #5 failed



Files of #3 & #5
are read by #0 &
#1 (Merging)



Repartitioning &
Data Migration

pFEM-CRAFT

Spare Node無しの

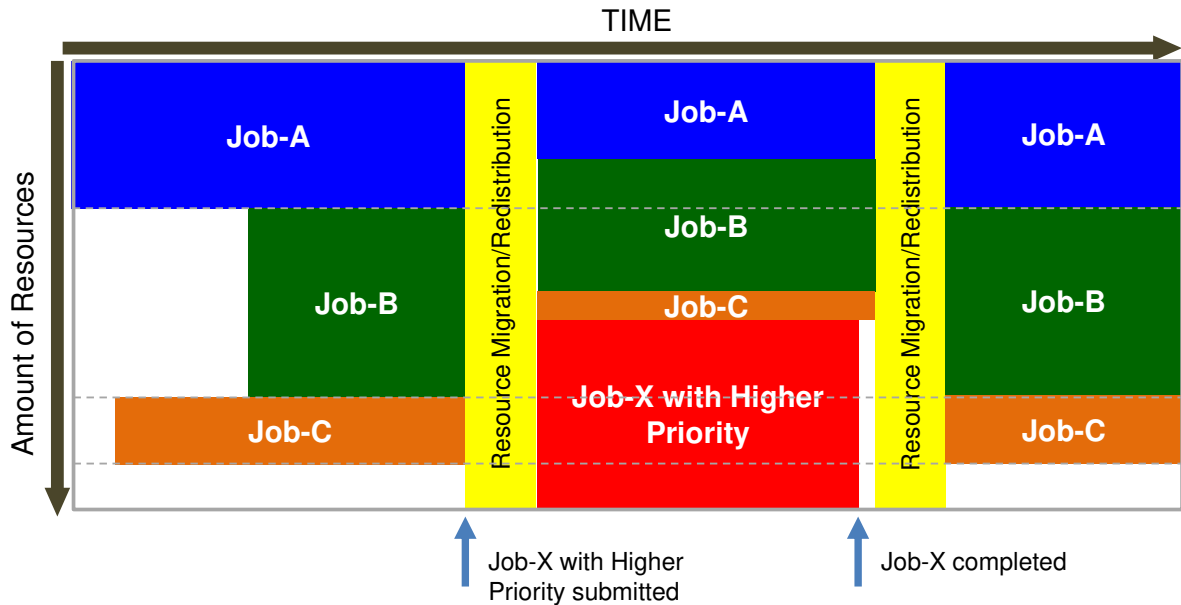
Fault Resilient機能付き

並列FEMアプリ

[Fukasawa, Shahzad, Nakajima,
Wellein, SIAM PP18]

提案手法

- ジョブ実行途中で利用計算機資源量を変化(減少)させる
- 一定のノード数で開始したジョブが、よりプライオリティの高いジョブに、利用している計算ノードの一部を供出する
 - Job-A, Job-B, Job-Cが実行中に、より優先度、緊急性の高いジョブJob-X投入
 - 資源のマイグレーション・再配分
 - Job-A, Job-B, Job-Cの一部のノードをJob-Xへ移動することによって、先行3ジョブを停止することなく、Job-Xを実行
 - Job-X終了後は元のノード数を回復
 - もどらない場合も可能



この部分の知見・将来展望

- ジョブの実行を止めることなく、より柔軟にUrgent Computingを実施することを目的として、アプリケーションの備えるべき機能を検討
- いくつかの機能(Repartitioning, Checkpointing, Data Migration)を付加することによって、有限要素法、差分法、有限体積法等についてはUrgent Computing向けの並列アプリケーションを少ない追加作業で開発できる見通しがついた。
 - ULFM-MPIも実装、動作確認ができた
 - Repartitioning/Data Migration/Checkpointingのライブラリ化
- 今後は関係各位との協力のもと、スーパーコンピュータの実運用への適用を継続して検討していく予定である。
 - 各ジョブ間の利用計算機資源の移動(減少, 増加), 様々な運用形態
 - 利用資源減少を受け入れるインセンティブ⇒ノード時間消費率(トークン)を減らす
 - アプリケーション書き換えの手間の削減

全体のまとめ

- 多様な計算機資源の利用
 - ✓「計算・データ・学習」融合推進
 - ✓ WaitIO over Cloud: 長いWaitが必要か...
 - ✓ WaitIOは中島の考え方を根本的に変えた
 - 翼がないと自分で飛べることの意義はわからない
- 大量の計算機資源の確保
 - ✓ Urgent Computing (Fault Resilience) の知見
- ソフトウェア開発を継続: 協力者求ム
- (追記) ストレージ
 - ✓ 「超大規模計算＝超大規模データ」ではない
 - Ipomoea-01のデータ移行量はさほどではなかった...
 - ✓ Digital Twin系では膨大になり得る: 宇宙物理の例
 - Odyssey10ノード×20hの計算を数千～数万ケース
 - 1ケース500GB級出力⇒ペタスケール
 - Gfarmによる広域共有⇒Wisteria/BDEC-01は良い!



Japan



WaitIO
over
Cloud

Norway



(データ同化+シミュレーション (A+S))から(Pureシミュレーション(Pure S))への移行

$$\begin{array}{lcl}
 \text{Assim.} & \text{Comp.} & \text{Residual} \\
 x_n^a = x_n^f + W(y_n - Hx_n^f) & & n: \text{Time Step} \\
 & & W: \text{Weighting Matrix} \\
 \\
 \text{Comp.} & \text{Assim.} & \\
 x_{n+1}^f = Fx_n^a & & F: \text{Wave Propagation simulation}
 \end{array}$$

(A+S) Assimilation+Simulation

(Pure S) Pure Simulation/Forecast

