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

東大情報基盤センター
活動状況

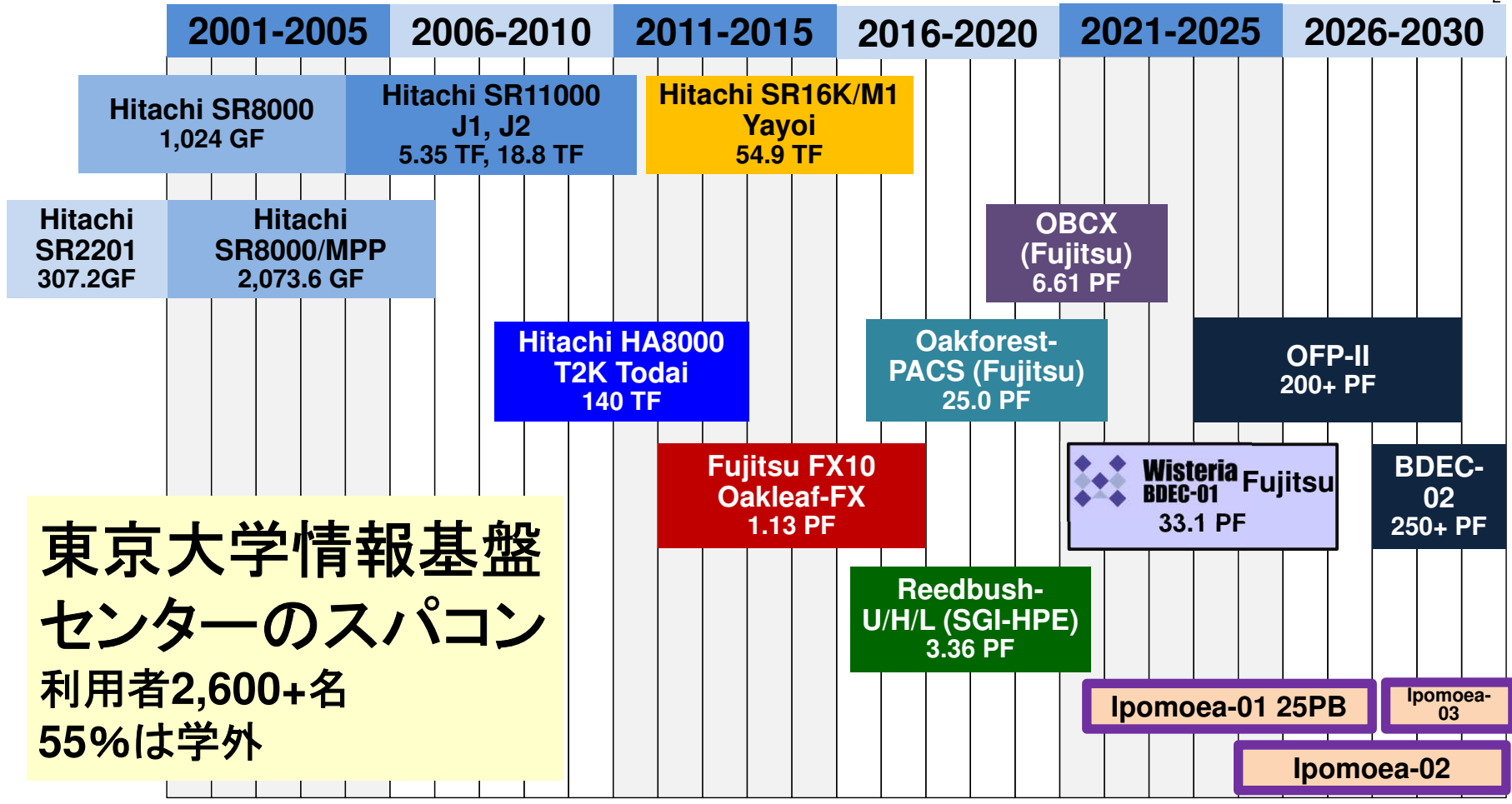
Wisteria
BDEC-01

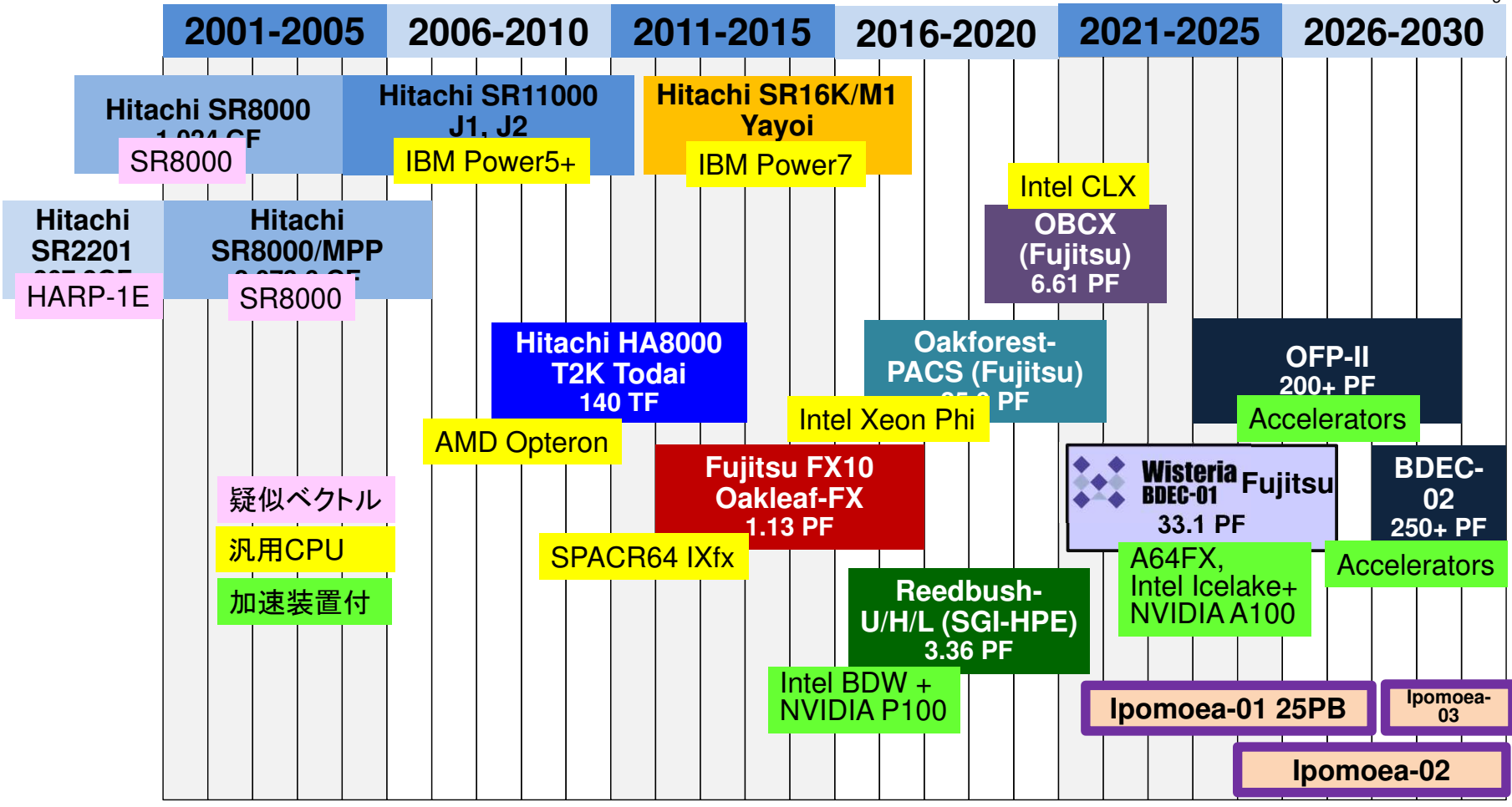
Hierarchical, Hybrid, Heterogeneous
h3-Open-BDEC
Big Data & Extreme Computing

中島 研吾
東京大学情報基盤センター

PCCC-OSSSワークショップ
2023年3月30日

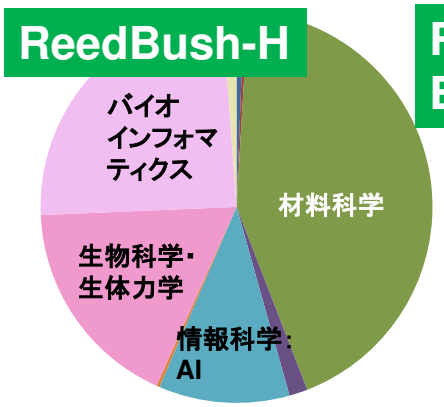
PC Cluster Consortium

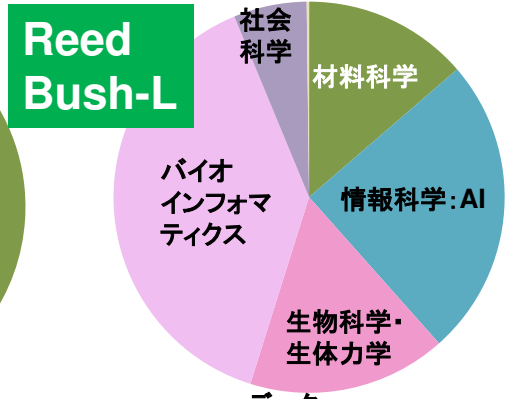

2001-2005
2006-2010
2011-2015
2016-2020
2021-2025
2026-2030
Hitachi SR8000
1,024 GF
Hitachi SR11000
J1, J2
5.35 TF, 18.8 TF
Hitachi SR16K/M1
Yayoi
54.9 TF
Hitachi
SR2201
307.2GF
Hitachi
SR8000/MPP
2,073.6 GF
OBCX
(Fujitsu)
6.61 PF
Hitachi HA8000
T2K Todai
140 TF
Oakforest-
PACS (Fujitsu)
25.0 PF
OFP-II
200+ 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

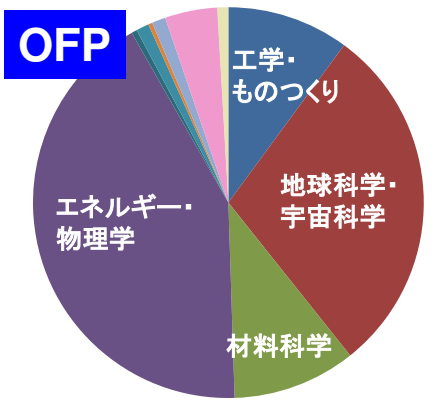

2001-2005
2006-2010
2011-2015
2016-2020
2021-2025
2026-2030
Hitachi SR8000
SR8000
Hitachi SR11000 J1, J2
IBM Power5+
Hitachi SR16K/M1 Yayoi
IBM Power7
Intel CLX
OBCX (Fujitsu) 6.61 PF
Hitachi SR2201
HARP-1E
Hitachi SR8000/MPP
SR8000
Hitachi HA8000 T2K Todai 140 TF
AMD Opteron
Oakforest-PACS (Fujitsu)
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

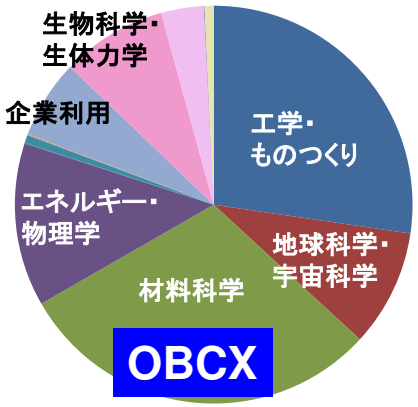
2021年度分野別（東大） ■汎用CPU，■GPU

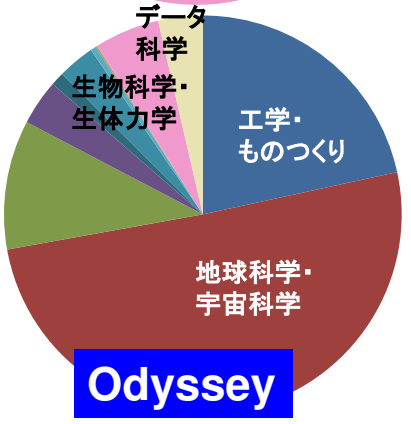
Odyssey，Aquariusは8月以降，RB-H，RB-Lは11月末時点

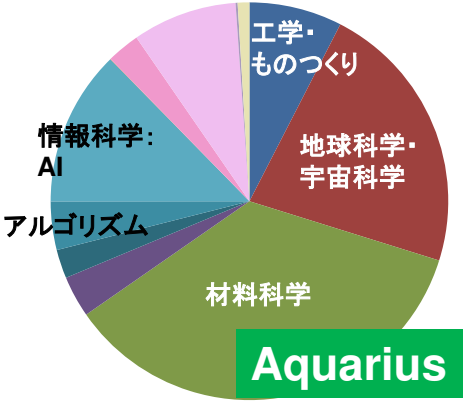






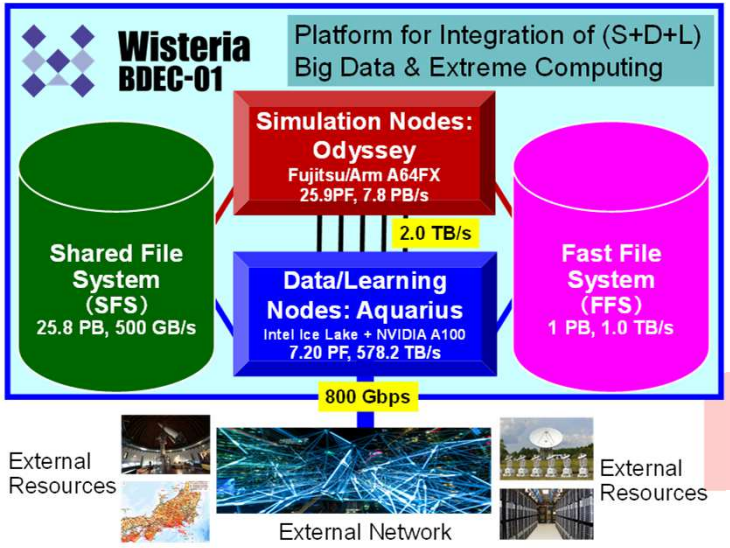






- 工学・ものつくり
- 地球科学・宇宙科学
- 材料科学
- エネルギー・物理学
- 情報科学：システム
- 情報科学：アルゴリズム
- 情報科学：AI
- 教育
- 産業利用
- 生物科学・生体力学
- バイオインフォマティクス
- 社会科学・経済学
- データ科学・データ同化

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



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）

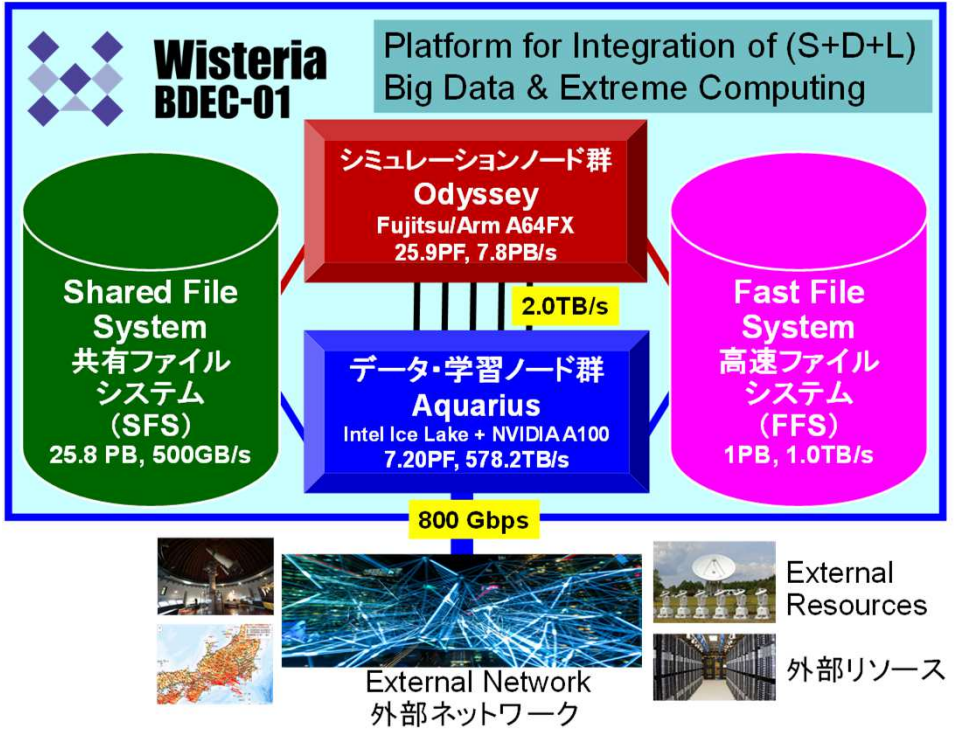
h3-Open-BDEC
Hierarchical, Hybrid, Heterogeneous
Big Data & Extreme Computing

Reedbush

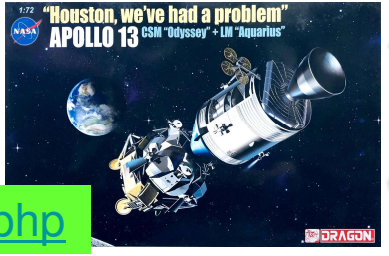
Oakforest-PACS

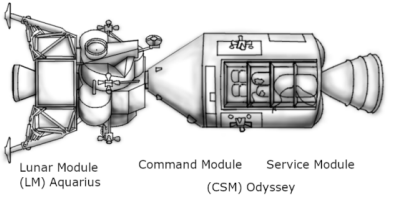
Oakbridge-CX

Wisteria/BDEC-01
(S+D+L)融合プラットフォーム



- Wisteria（紫藤）
  - 手賀沼（柏市）に伝わる「藤姫伝説」
- Odyssey
  - アポロ13号・司令船（Command Module, CM）のコールサイン
- Aquarius
  - アポロ13号・月着陸船（Lunar Module, LM）のコールサイン
- 人類と地球を護る





https://www.cc.u-tokyo.ac.jp/public/pr/pr-wisteria.php

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）

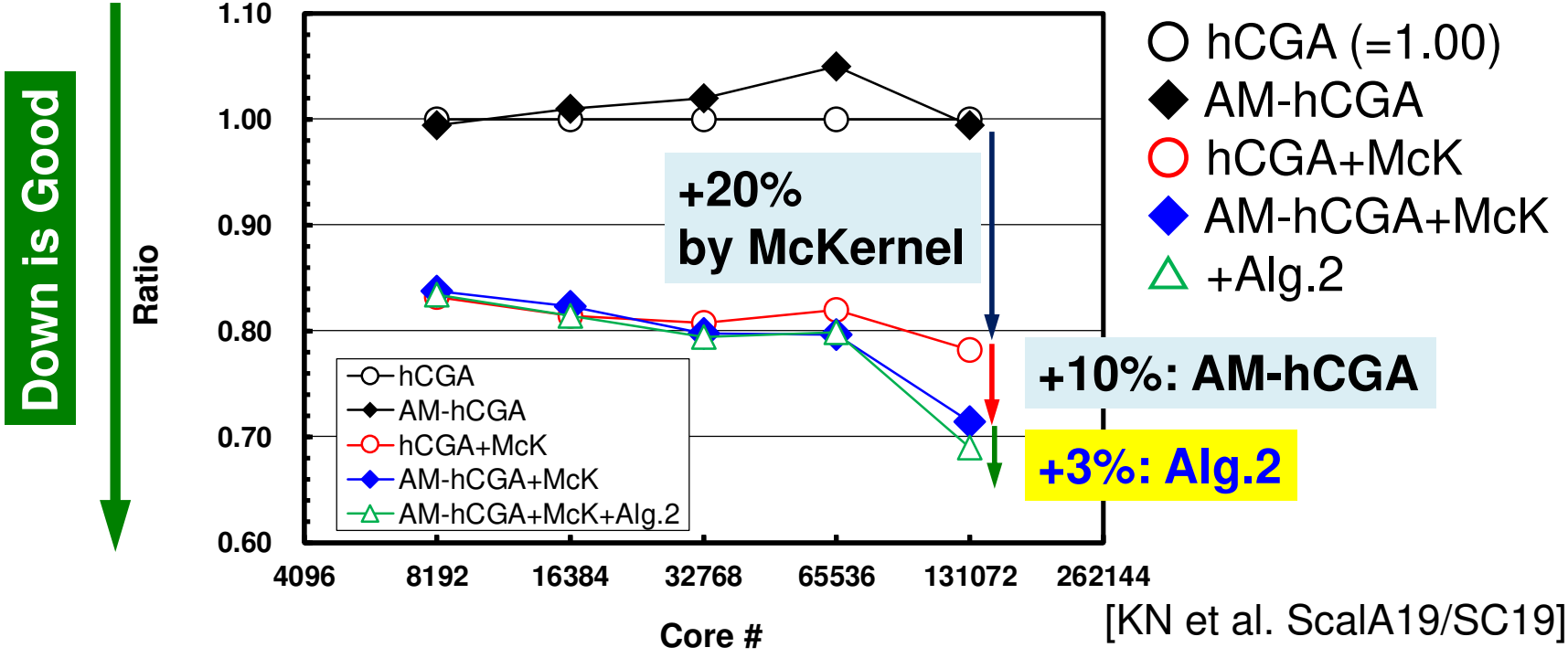
System	JPY/GFLOPS Small is Good	GFLOPS/W Large is Good
Oakleaf-FX/Oakbridge-FX (Fujitsu) (Fujitsu SPARC64 IXfx)	125	**0.866**
Reedbush-U (HPE)　(Intel Xeon Broadwell (BDW))	61.9	**2.310**
Reedbush-H (HPE)　(Intel BDW+NVIDIA P100x2/node)	15.9	**8.575**
Reedbush-L (HPE)　(Intel BDW+NVIDIA P100x4/node)	13.4	**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**

2022年のハイライト

- MATLAB導入（2022年3月）
- OFP運用終了（3月末）
  - 第11回JCAHPCセミナー（OFP運用終了記念シンポジウム）「ありがとうOFP:京から富岳への狭間で咲いた大輪の花」 2022年5月27日（金），ハイブリッド開催
    - https://www.jcahpc.jp/event/seminar11.html
- Ipomoea-01（共通ストレージ）運用開始
- Wistera/BDEC-01
  - Odyssey-Aquarius連携運用開始
  - mdxとの連携も試行的に実施
- OFP後継機（OFP-II）：JCAHPC
  - GPUを中心としたシステムに決定
  - アプリケーション移行作業開始
  - 調達開始，2024年10月以降運用開始目標
- 電気料金高騰⇒負担金1.50倍（ストレージ関連は据置）

“Tiny” Cases: More Siginicant Improvement by IHK/McKernel and AM-*h*CGA on OFP

+20% by IHK/McKernel, +10% by AM-*h*CGA, +3% by Alg.2

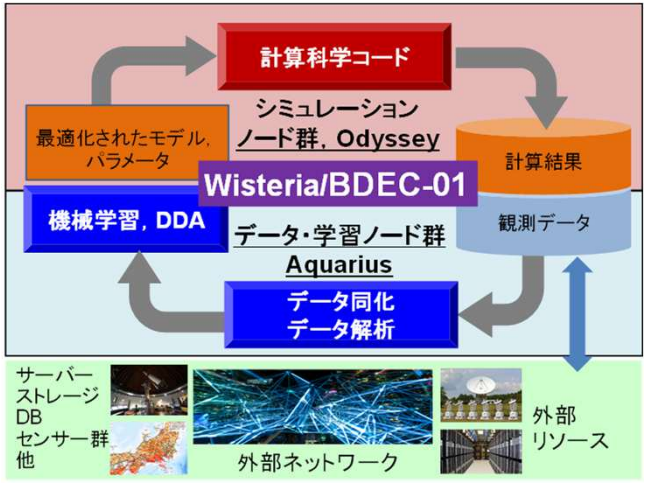


[KN et al. ScalA19/SC19]

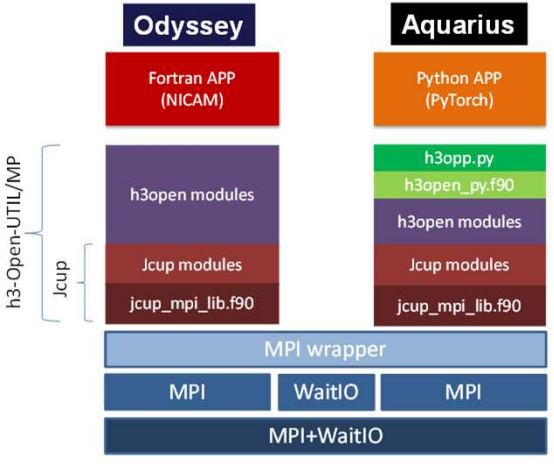
MATLABの導入「S+D+L」融合，AI for HPCの実現

Wisteria BDEC-01

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

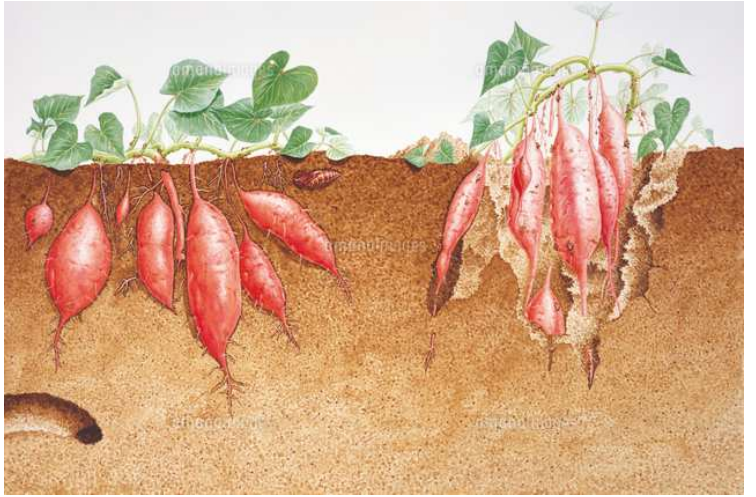


- 2022年3月からOBCX，Aquariusで利用可能
- MATLAB
  - ✓ 多様な機能
  - ✓ ユーザーのプログラムからの関数呼び出し重視⇒データ解析，機械学習系の豊富な機能⇒高度化
  - ✓ MATLABはAquarius（データ・学習ノード群）でのみ稼働するが，h3-Open-BDECと連携させて，Odyssey（シミュレーションノード群）上で実施する大規模シミュレーションのパラメータ最適化に適用する⇒「S+D+L」融合，AI for HPC
- h3-Open-BDECは様々な環境で動作⇒MATLABと組み合わせた使用による普及



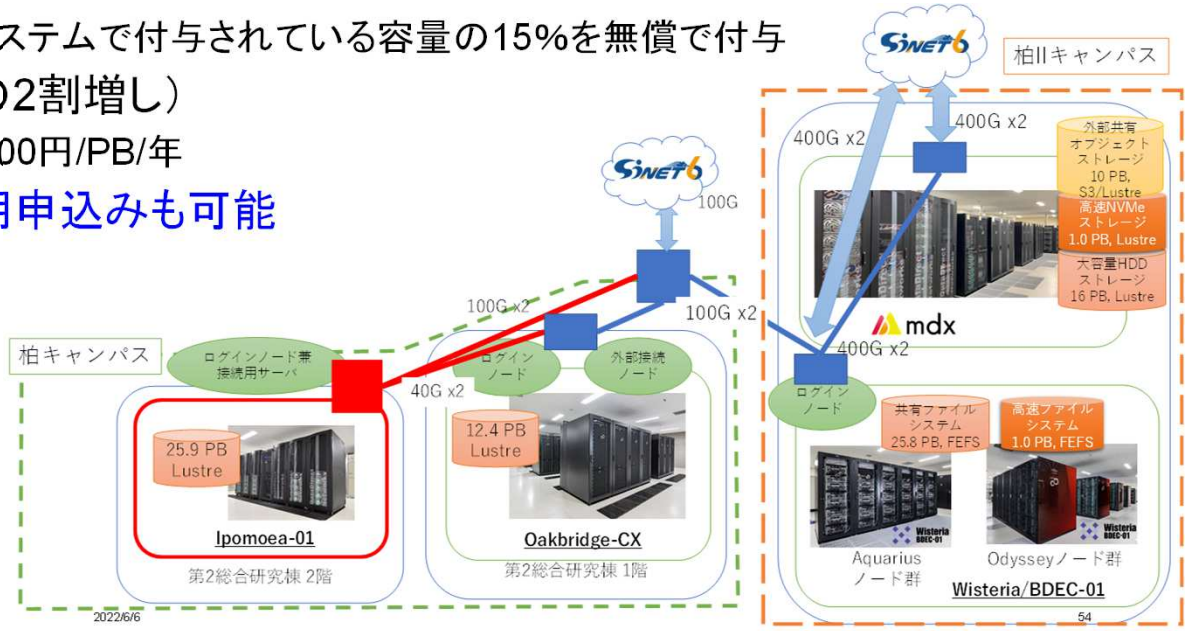
大規模共通ストレージシステム「Ipomoea」

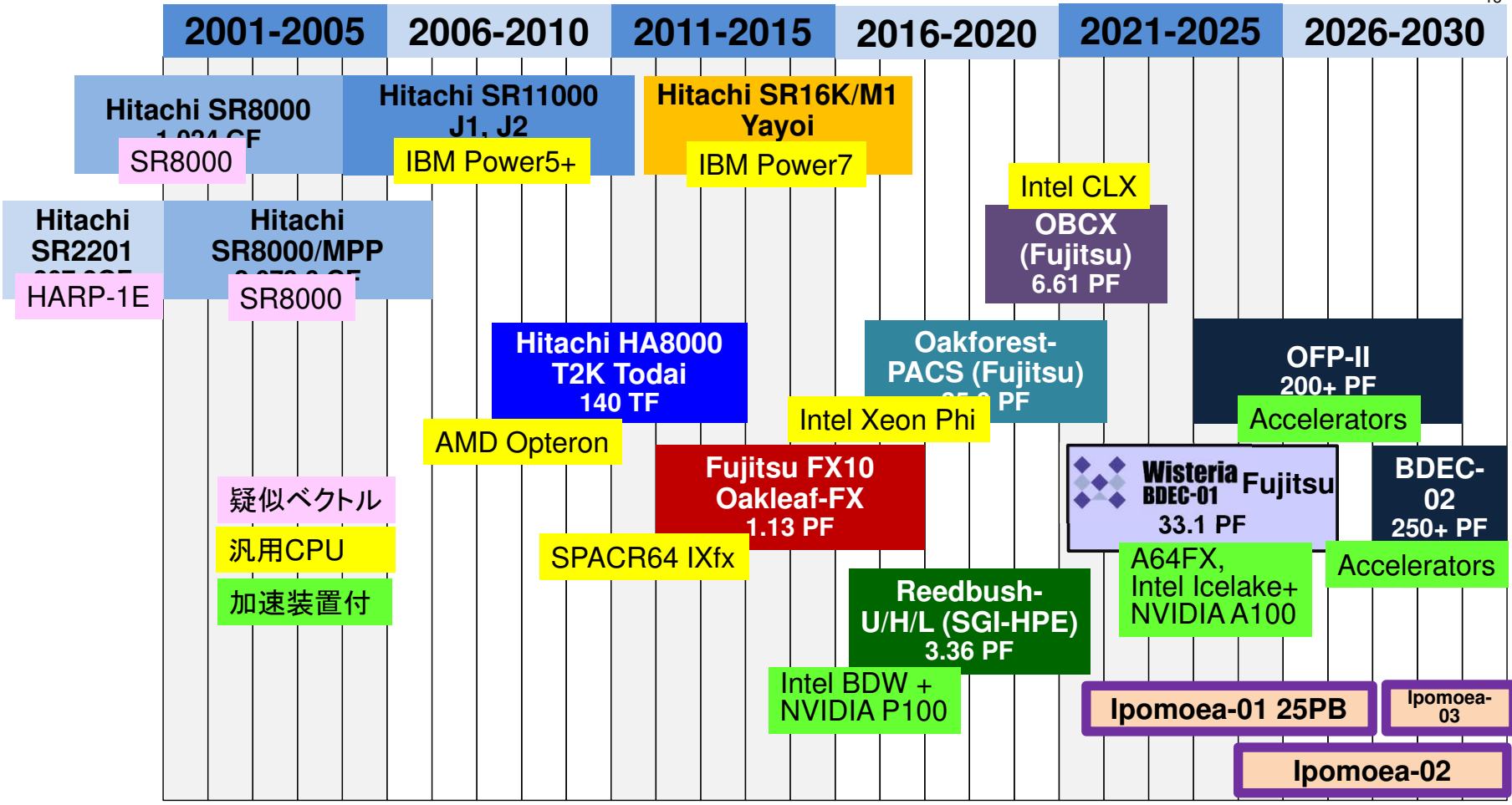
- スーパーコンピュータの処理能力の向上に伴い，扱うデータ量も増加の一途
- 東大センターでは従来ストレージは各システムに附属して導入され，各システムのストレージは独立
- このような状況（注：ストレージがシステム毎に独立）は利用者に多大な不便を強いることになり，東大センターの全システムからアクセス可能な共通ストレージの導入が強く求められていた
- 各システムからアクセスできる「大規模共通ストレージ（Ipomoea）」導入決定
  - OFP運用終了が契機
  - 1システムを約5–6 年使用し，約3年ごとに新しいストレージシステム（25+PB）を導入し，入れ替えることを想定している



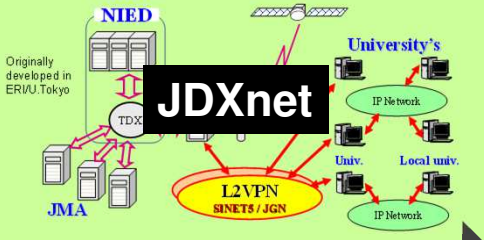
- 2022年1月運用開始・6月より一般に公開，25+PB，富士通製
  - 2022 年5月末までにOFPのLustre 領域の必要ファイルの移行完了
- 割当容量
  - 東大センターのシステムに利用者番号（教育利用，講習会除く）を有する場合
    - 各利用者ごとに5TB
    - 各グループごとに登録システムで付与されている容量の15%を無償で付与
  - 追加負担金（企業はこの2割増し）
    - 7,200円/TB/年，2,100,000円/PB/年
  - Ipomoea-01のみの利用申込みも可能

Ipomoea-01




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
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
疑似ベクトル
汎用CPU
加速装置付
Reedbush-U/H/L (SGI-HPE) 3.36 PF
Intel BDW + NVIDIA P100
Ipomoea-01 25PB
Ipomoea-03
Ipomoea-02

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

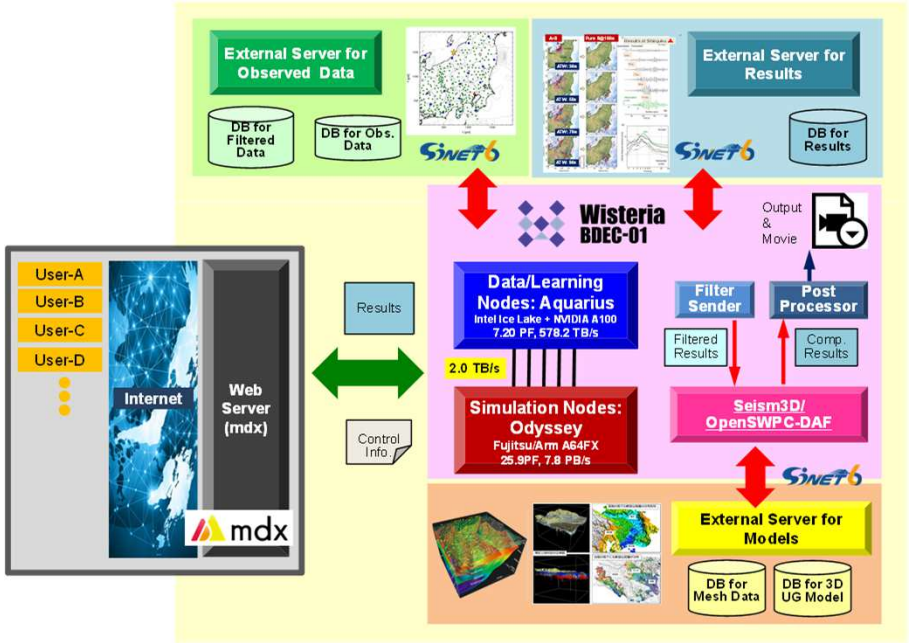
Seism3D/ OpenSWPC-DAF

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

データ活用社会創成プラットフォーム「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
- 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 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

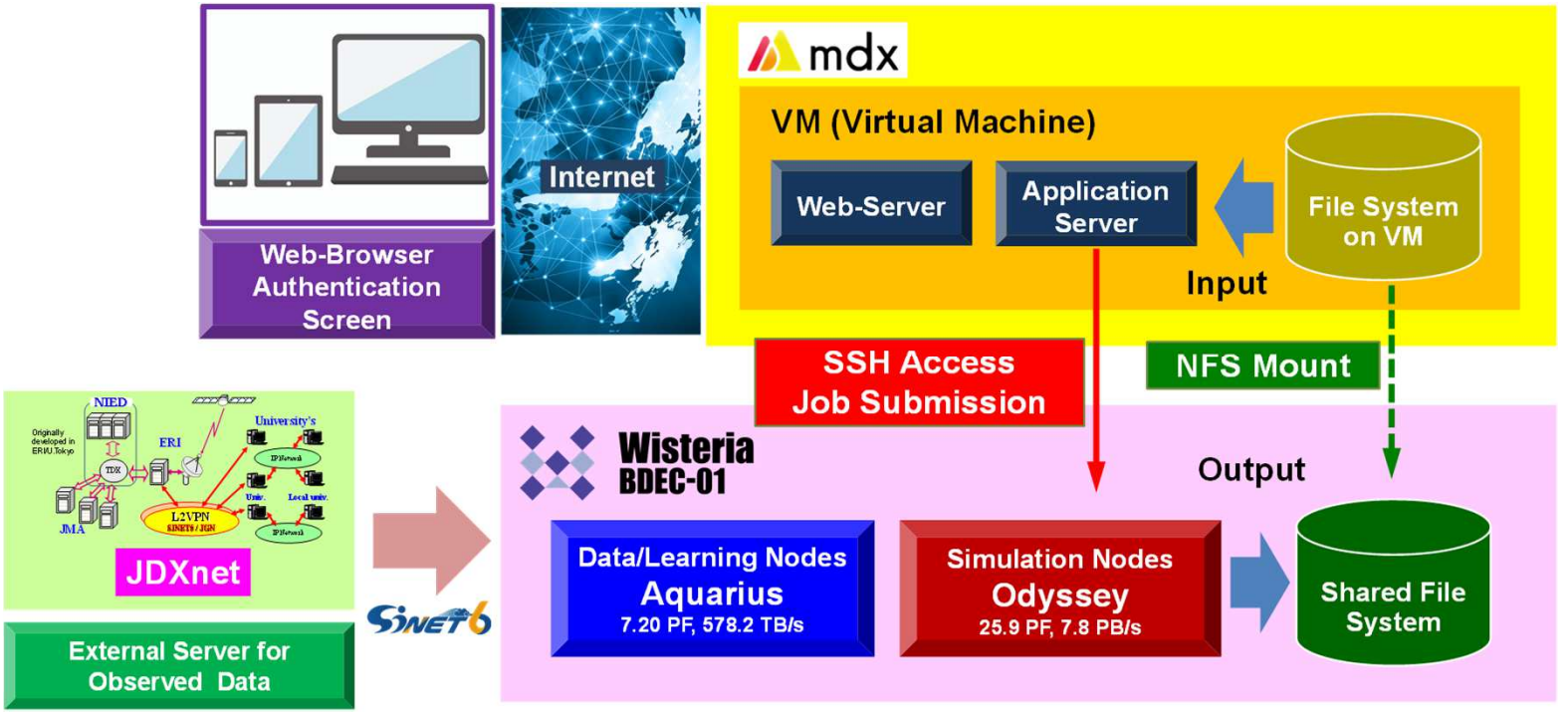
GPU 演算加速ノード
Intel IceLake x2ソケット+NVIDIA A100 x8 x40ノード
理論ピーク性能(FP64): 6.4PFLOPS
理論ピーク性能(FP16): 100.7PFLOPS
総メモリバンド幅: 496.3 TB/s

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

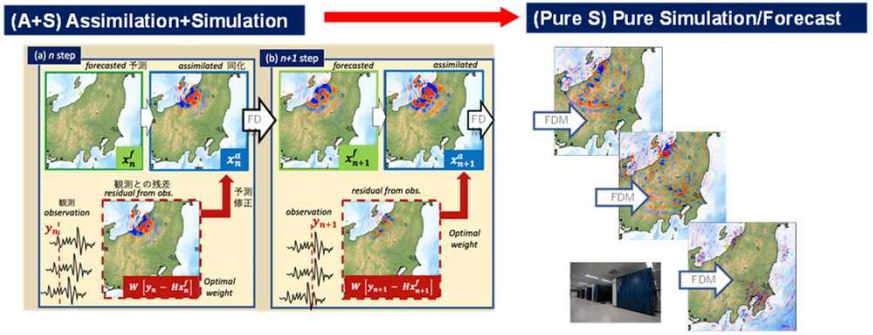
大容量HDDストレージ
Lustre Filesystem
16.3 PB (HDD)
157.5 GByte/sec

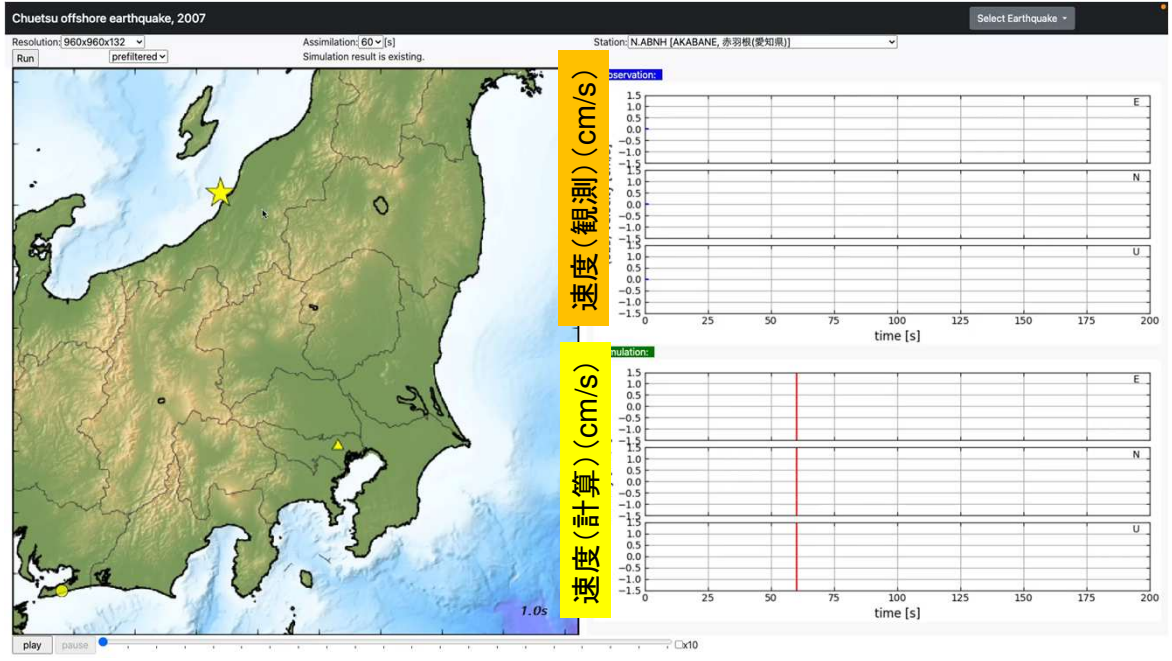
外部共有オブジェクトストレージ
S3 Data Service
10.3 PB (HDD)
63.0 GByte/sec

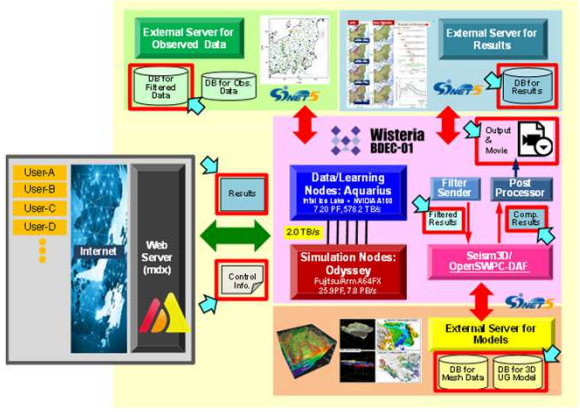
Web-based Simulation System for Outreach Activities



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



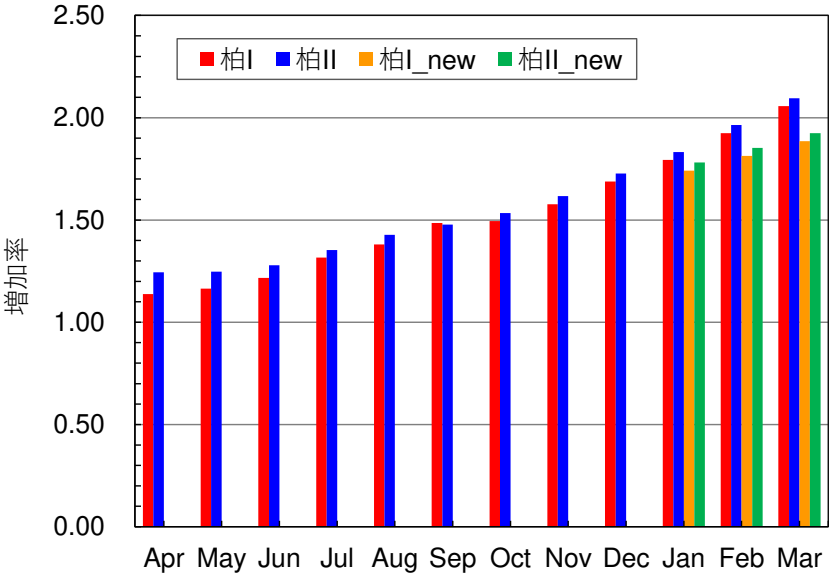




将来構想：OFP-II，Mercury

- 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（電気代も高騰）
- （S+D+L）融合路線は継続

System (Top/Green 500)	HW	GF/W
Henri (405,1)	NVIDIA H100	65.1
Frontier (1,6)	AMD MI250X	52.2
Leonardo (4,14)	NVIDIA A100	31.1
Fugaku (2, 43)	A64FX	15.4



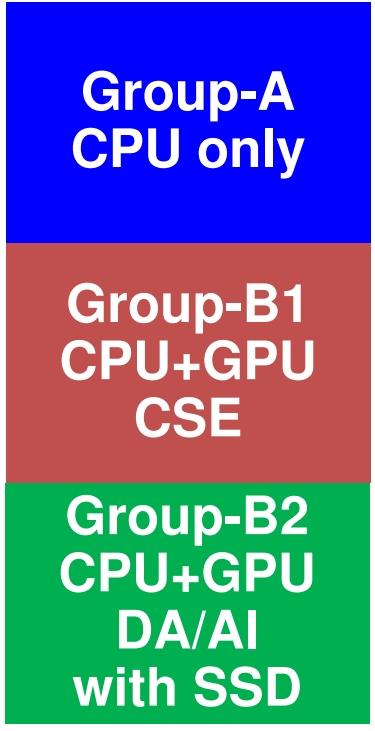
2022年度の電力単価（対2019年度比）

2022年11月推測

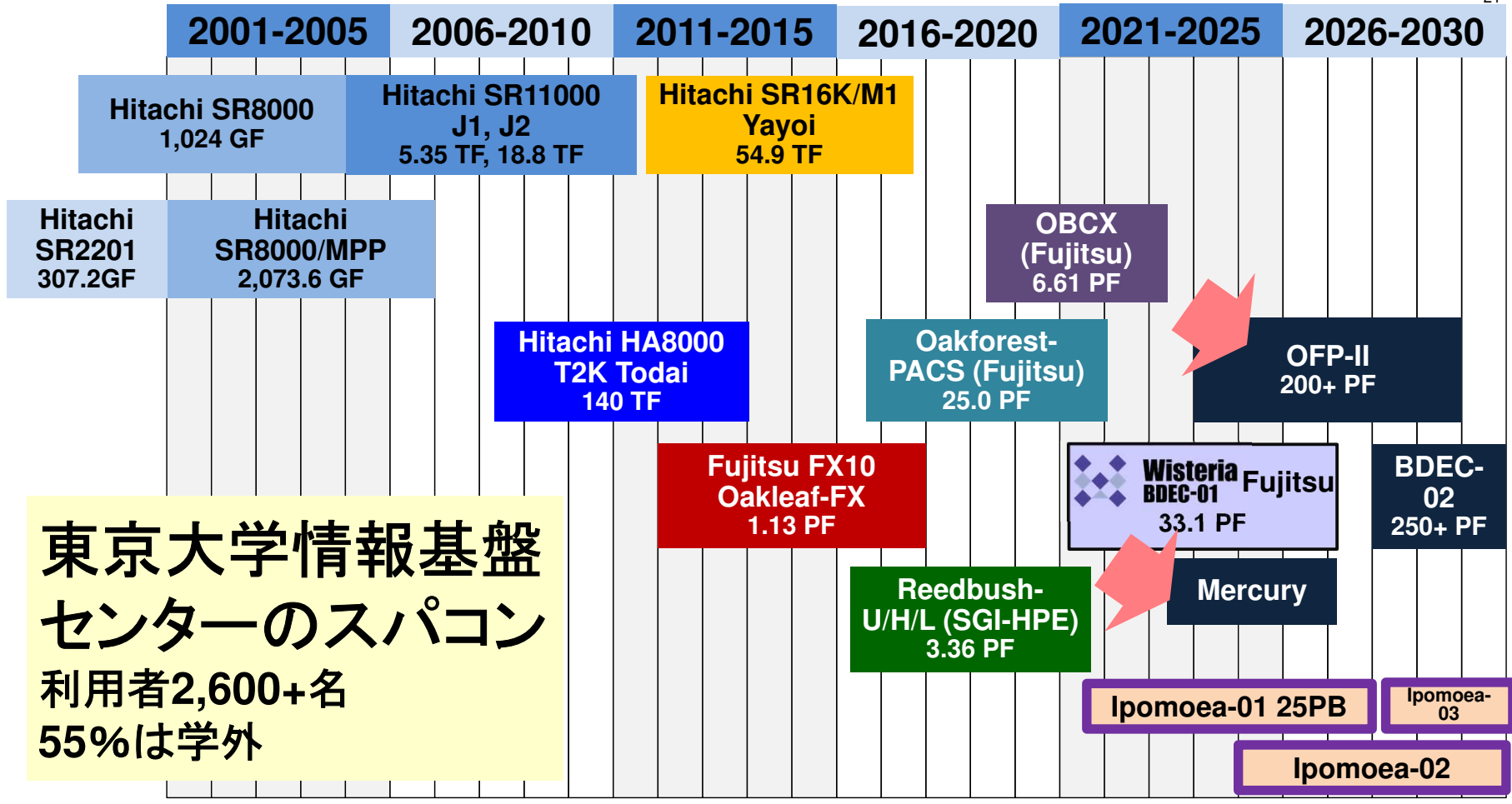
2022年12月推測

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（電気代も高騰）
- （S+D+L）融合路線は継続
- OFP-II（2024年4月以降）
  - OFP後継機（JCAHPC：筑波大学と共同），200+PF
  - GPUを主とするシステム
    - Group-A (Only CPU)，Group-B (CPU+GPU), : Group-A とGroup-BのCPUは異なる可能性あり
- Wisteria-Mercury（2023年11月～）
  - GPUクラスタ，OFP-IIプロトタイプ（同じ系統のGPU）
- **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**
  - **OpenACC, Stdpar**



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：NVIDIA
  - 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⇒2024年10月以降

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

Pre-Benchmarks

OFP-II及びMercury向けGPU決定

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

- 7種類，計算科学系（次頁）
  - CPU向け（OpenMP+MPI），GPU化済（CUDA/OpenACC）
  - 3つのカテゴリー（事実上は2つ：B,C），それぞれ2レベル
  - Mercury（2023秋），OFP-II（2024春（以降））に使用するHWを想定した性能推定

A	CPU向けコードのホストCPU上での最適化（Optional）	A-1	As-Is
		A-2	最適化
B	CPU向けコードのGPUへの移植	B-1	最低限の変更（OpenACC, OpenMP, Standard Language）
		B-2	フル最適化（OpenACC，CUDA）
C	OpenACC/CUDAによるGPU化済みコードの最適化	C-1	As-Is
		C-2	フル最適化

Seven Pre-Benchmarks

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

Name of the Code	Description	Lang.	Parallelization	GPU	Category
P3D	3-D Poisson's Equation by Finite Volume Method	C	OpenMP		A & B
GeoFEM/ICCG	Finite Element Method	Fortran	OpenMP, MPI		
H-Matrix	Hierarchical-Matrix calculation	Fortran	OpenMP, MPI		
QCD	Quantum-Chromo Dynamics simulation	Fortran	OpenMP, MPI	CUDA	C
N-Body	N-Body simulation using FDPS	C++	OpenMP, MPI	CUDA	
GROMACS	Molecular Dynamics simulation	C++	OpenMP, MPI	CUDA, HIP, SYCL	
SALMON	Ab-initio quantum-mechanical simulator for optics and nanoscience	Fortran	OpenMP, MPI	(OpenACC)	B

各カテゴリの概要・評価基準

- Category-B: (OpenMP+MPI)
  - B-1: 最低限の変更 (OpenMP, OpenACC, Standard Language etc.)
  - B-2: フル最適化（CUDA, OpenACC etc.）
  - 評価
    - B-2性能（絶対値）
    - （B-1）／（B-2）の性能比（できるだけ1に近い場合を高く評価）
- Category-C: (OpenACC, CUDA)，GPU化済み
  - C-1: As Is
  - C-2: フル最適化
    - C-2性能の絶対値で評価
- 性能・移植性の両者を評価
  - Performance & Portability
  - B-1においては特にPortability重視

ポーティング概要
Mercury & OFP-II

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

- サポート：基本的には各自によるポーティング
  - ミニキャンプ，講習会を頻繁に実施する
  - 相談会（毎月1回）
  - 移行ポータルサイト：https://www.cc.u-tokyo.ac.jp/supercomputer/gpu_porting.php
- 特別サポート：大口ユーザー（HPCI），コミュニティコード：16個
  - JCAHPC，GPUベンダーによるサーヴェイ，最適化
  - コード量が多い場合には外部業者に委託することも可，予算も準備してある
    - 全コード，一部，ミニアプリ
  - 2022年11月から徐々に開始
    - Slackによるコミュニケーション，一部会話は他グループとも共有
    - 商用コードを含むため，別途NDA締結の必要あり（対応済み）
- 基本的にOpenACCを推奨