



東京大学スーパーコンピュータの 企業による利用制度のご紹介

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/>

問合せ先: uketsuke@cc.u-tokyo.ac.jp

- 東大情報基盤センターについて
- サービス概要
- 企業による利用制度
 - 制度概略と利用提案
 - 申込・審査・成果報告
 - スパコン利用への道
 - ①お試しアカウント付き並列プログラミング講習会
 - ②トライアルユース制度
 - ③ 企業利用制度
- **スーパーコンピュータシステム概要**
 - **Reedbush**
 - **Oakforest-PACS**
 - **Oakbridge-CX**
 - **Wisteria/BDEC-01**
 - 運用
- ご質問は

- 東大情報基盤センターの概要
- **Wisteria/BDEC-01**
 - 導入の背景
 - システム概要
- Wisteria/BDEC-01: 利用プログラム
 - 一般, 企業, 萌芽課題, JHPCN, HPCI, 大規模HPCチャレンジ
 - 負担金, リソースグループ等
 - 講習会, GPUミニキャンプ, JCAHPCセミナー
- Wisteria/BDEC-01: 利用例
 - h3-Open-BDEC
 - 予備的計算事例
- 今後の動向等
 - 「共通ストレージ」について

各種ベンチマーク

- TOP 500 (Linpack, HPL)
 - 連立一次方程式ソルバー(直接法), 計算速度(FLOPS値)
 - 規則的な密行列: 連続メモリアクセス
 - 計算性能
- HPCG
 - 連立一次方程式ソルバー(反復法), 計算速度(FLOPS値)
 - 有限要素法から得られる疎行列(ゼロが多い)
 - 不連続メモリアクセス
 - 実アプリケーションに近い
 - メモリアクセス性能, 通信性能
- Green 500
 - HPL (TOP500) 実行時のFLOPS/W値

June 2021 (ISC-HPC 2021 Digital)の諸ランキング

System	TOP500	Green500	HPCG	Graph500	HPL-AI
Oakforest-PACS	32	58	23	-	-
Oakbridge-CX	97	55	67	-	-
Wisteria/BDEC-01 (Odyssey)	13 (柏市で第位)	21	9	3	7
Wisteria/BDEC-01 (Aquarius)	93	10 (IceLake+ A100として は最高位)	53	-	-

57th TOP500 List (June, 2021)

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

<http://www.top500.org/>

	Site	Computer/Year Vendor	Cores	R _{max} (TFLOPS)	R _{peak} (TFLOPS)	Power (kW)
1	<u>Fugaku, 2020, Japan</u> 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
2	<u>Summit, 2018, USA</u> 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	148,600	200,795	10,096
3	<u>Sierra, 2018, USA</u> DOE/NNSA/LLNL	IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR InfiniBand	1,572,480	94,640	125,712	7,438
4	<u>Sunway TaihuLight, 2016, China</u> National Supercomputing Center in Wuxi	Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway	10,649,600	93,015	125,436	15,371
5	<u>Perlmutter, 2021, USA</u> DOE/NERSC/LBNL	HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10	706,304	64,590	89,794	2,528
6	<u>Selene, 2020, USA</u> NVIDIA	NVIDIA DGX A100 SuperPOD, AMD EPYC 7742 64C 2.25GHz, NVIDIA GA100, Mellanox Infiniband HDR	555,520	63,460	79,215.0	2,646
7	<u>Tianhe-2A, 2018, China</u> 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	61,445	100,679	18,482
8	<u>JUWELS Booster Module, 2020, Germany</u> Julich (FZJ)	Bull Sequana XH2000, AMD EPYC 7402 24c 2.8GHz, NVIDIA A100, Mellanox InfiniBand HDR	449,280	44,120	70,980	1,764
9	<u>HPC5, 2020, Italy</u> Eni S.p.A.	Dell C4140, Xeon Gold 6252 24c 2.1GHz, NVIDIA Volta GV100, Mellanox InfiniBand HDR	669,760	35,450	51,720	2,252
10	<u>Frontera, 2019, USA</u> Texas Advanced Computing Center	Dell C6420, Xeon Platinum 8280 28c 2.7GHz, Mellanox InfiniBand HDR	448,448	23,516	38,746	
12	<u>ABCI 2.0, 2021, Japan</u> AIST	Fujitsu PRIMERGY GX2570 M6, Xeon Platinum 8360Y 36C 2.4GHz, NVIDIA A100 SXM4 40 GB, InfiniBand HDR	504,000	22,208	54,341	1,600
13	<u>Wisteria/BDEC-01 (Odyssey), 2021, Japan</u> ITC, University of Tokyo	Fujitsu PRIMEHPC FX1000, A64FX 48C 2.2GHz, Tofu interconnect D	368,640	22,121	25,952	1,468

HPCG Ranking (June, 2021)

	Computer	Cores	HPL Rmax (Pflop/s)	TOP500 Rank	HPCG (Pflop/s)
1	Fugaku	7,630,848	442.010	1	16.004
2	Summit	2,414,592	148.600	2	2.926
3	Perlmutter	706,304	64.590	5	1.905
4	Sierra	1,572,480	94.640	3	1.796
5	Selene	555,520	63.460	6	1.622
6	JUWELS Booster Module	449,280	44.120	8	1.275
7	Dammam-7	672,520	22.400	11	0.881
8	HPC5	669,760	35.450	9	0.860
9	Wisteria/BDEC-01 (Odyssey)	368,640	22.121	13	0.817
10	Earth Simulator -SX-Aurora TSUBASA	43,776	9.990	39	0.747
23	Oakforest-PACS	556,104	13.555	32	0.385

Green 500 Ranking (June, 2021)

<http://www.top500.org/>

	TOP 500 Rank	System	Accelerator	Cores	HPL Rmax (Pflop/s)	Power (kW)	GFLOPS/W
1	335	MN-3, Preferred Networks, Japan	MN-Core	1,664	1.822	61.36	*29.70
2	22	HiPerGator AI	NVIDIA A100	138,880	17.200	582.63	*29.52
3	100	Wilkes-3, U. Cambridge, UK	NVIDIA A100	44,800	4.124	146.53	28.14
4	36	MeluXina - Accelerator Module	NVIDIA A100	99,200	10.520	390.25	*26.96
5	214	NVIDIA DGX SuperPOD, USA	NVIDIA A100	19,840	2.356	90	26.195
6	5	Perlmutter	NVIDIA A100	706,304	64.590	2,528	25.550
7	8	JUWELS Booster Module, Germany	NVIDIA A100	449,280	44.120	1,764	25.01
8	43	JURECA Data Centric Module	NVIDIA A100	105,840	9.330	384.10	*24.29
9	189	Spartan2, France	NVIDIA A100	23,040	2.566	106	24.262
10	93	Wisteria/BDEC-01 (Aquarius), Fujitsu, Japan	NVIDIA A100	42,120	4.425	183.93	24.06
20	1	Fugaku		7,630,848	442.01	29.899	15.418
21	13	Wisteria/BDEC-01 (Odyssey), Fujitsu, Japan		368,640	22.121	1,468.00	15.069

Graph500 BFS, June 2021

グラフ探索問題による性能評価、頂点数 = 2^{Scale}

	Site	Computer/ Vendor	Cores	Scale	GTEPS
1	<u>Fugaku, 2020, Japan</u> R-CCS, RIKEN	Fujitsu PRIMEHPC FX1000, Fujitsu A64FX 48C 2.2GHz, Tofu-D	7,630,848	41	102956
2	<u>Sunway TaihuLight, 2016, China</u> National Supercomputing Center in Wuxi	Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway	10,599,680	40	23755.7
3	<u>Wisteria/BDEC-01 (Odyssey), 2021, Japan</u> ITC, University of Tokyo	Fujitsu PRIMEHPC FX1000, A64FX 48C 2.2GHz, Tofu interconnect D	368,640	37	16118
4	<u>Toki-Sora, 2021, Japan</u> JAXA	Fujitsu PRIMEHPC FX1000, A64FX 48C 2.2GHz, Tofu interconnect D	276,480	36	10813
5	<u>Summit (CPU Only), 2018, USA</u> 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	40	7665.7
6	<u>SuperMUC, 2018, Germany</u> Leibniz	Lenovo ThinkSystem SD650, Xeon Platinum 8174 24C 3.1GHz, Intel Omni-Path	196,608	39	6279.47
7	<u>Lise, 2021, Germany</u> ZIB	Bull Intel Cluster Intel Xeon Platinum 9242 48C 2.3GHz Intel Omni-Path	121,920	38	5423.94
8	<u>Cori - 1024 haswell partition, 2017, USA</u> NERSC	Cray XC40	32,768	37	2562.16
9	<u>Tianhe-2 (MilkyWay-2), 2013, China</u>	TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P	196,608	36	2061.48
10	<u>MeluXina - Cluster Module, 2021, Luxembourg</u>	Bull Sequana XH2000	49,152	36	1527.39

HPL-AI (June 2021)

$R_{\max}$: Performance of Linpack (TFLOPS)

低精度演算利用による性能向上

	Site	Computer/Year Vendor	Cores	HPL-AI (EFLOPS)	Top500	HPL $R_{\max}$ (TFLOPS)	Speedup
1	<u>Fugaku, 2020, Japan</u> R-CCS, RIKEN	Fujitsu PRIMEHPC FX1000, Fujitsu A64FX 48C 2.2GHz, Tofu-D	7,630,848	2.0	1	442,010 (= 442.0 PF)	4.5
2	<u>Summit, 2018, USA</u> 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	1.15	2	148,600	7.7
3	<u>Selene, 2020, USA</u> NVIDIA	NVIDIA DGX A100 SuperPOD, AMD EPYC 7742 64C 2.25GHz, NVIDIA GA100, Mellanox Infiniband HDR	555,520	0.63	6	63,460	9.9
4	<u>Perlmutter, 2021, USA</u> DOE/NERSC/LBNL	HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10	761,856	0.59	5	64,590	9.1
5	<u>JUWELS Booster Module, 2020, Germany</u> Julich (FZJ)	Bull Sequana XH2000, AMD EPYC 7402 24c 2.8GHz, NVIDIA A100, Mellanox InfiniBand HDR	449,280	0.47	8	44,120	10
6	<u>HiPerGator, 2021, USA</u> NVIDIA	NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Infiniband HDR	138,880	0.17	23	63,460	9.9
7	<u>Wisteria/BDEC-01 (Odyssey), 2021, Japan</u> ITC, University of Tokyo	Fujitsu PRIMEHPC FX1000, A64FX 48C 2.2GHz, Tofu interconnect D	368,640	0.10	13	22,121	4.5
8	<u>Berzelius, 2021, Sweden</u>	NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, A100, Infiniband HDR, Atos	59,520	0.05	84	44,120	9.9
9	<u>Flow, 2020, Japan</u> ITC, Nagoya.U.	Fujitsu PRIMEHPC FX1000, A64FX 48C 2.2GHz, Tofu interconnect D	110,592	0.03	87	6,617	4.3
10	<u>MTS GROM, 2021, Russia</u>	NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100 40GB, Infiniband	19,840	0.015	245	2,258	7

GFLOPS(ピーク性能)当たり利用負担(円):電気代 GFLOPS/W(Green 500)

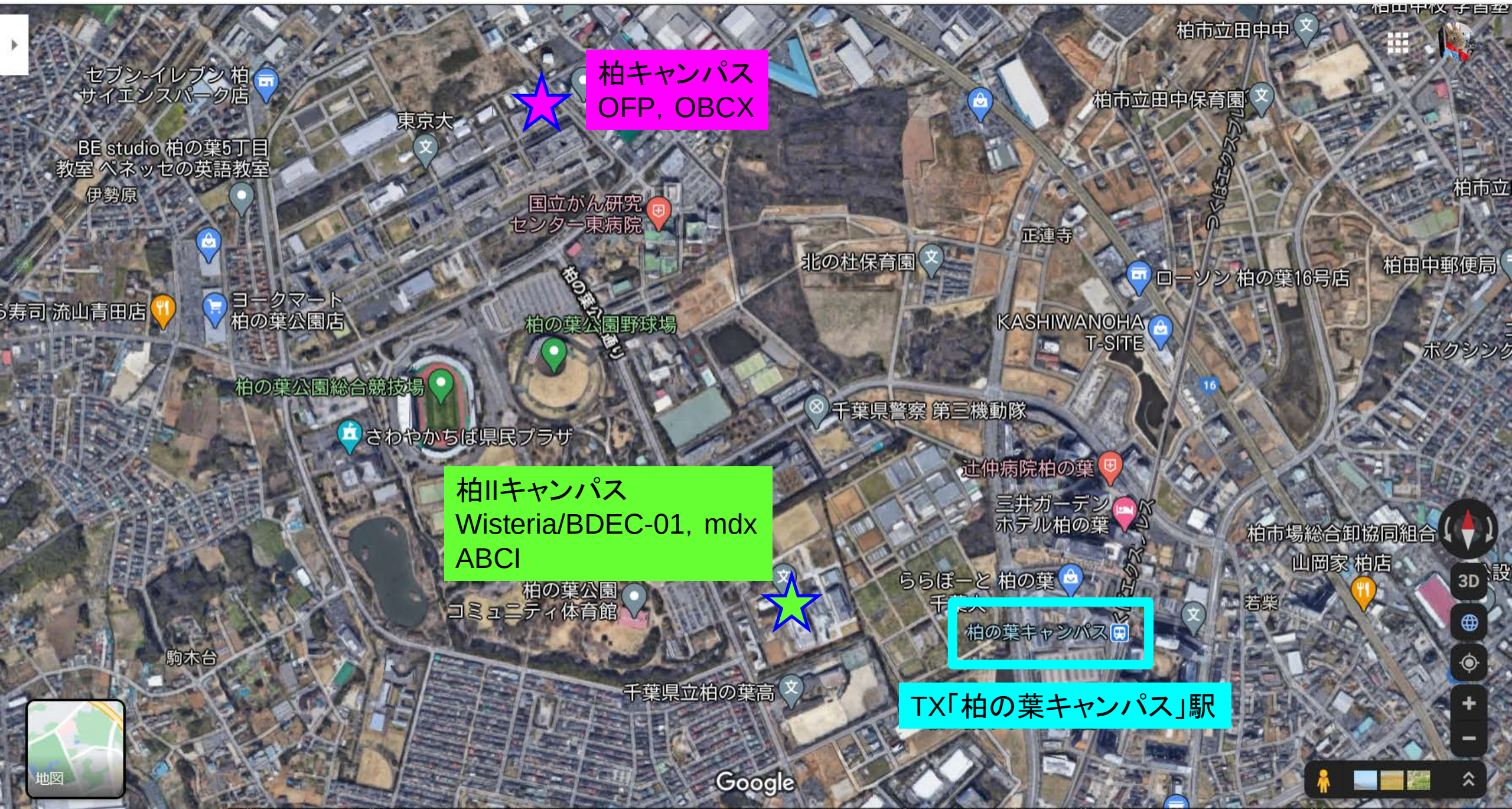
***1: Fugaku, *2 NVIDIA DGX SuperPOD, TOP/Green 500@Nov. 2020**

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	14.66^{*1}
Wisteria-Aquarius(Intel Xeon Ice Lake + NVIDIA A100x8)	9.00	26.20^{*2}

3システム: 利用者2,600+, 学外55+%

- Reedbush (HPE, Intel BDW + NVIDIA P100 (Pascal)) (本郷)
 - 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
 - 3.36 PF, 2016年7月～ 2021年11月末(予定)
 - Reedbush-U (CPU only, 2020年6月30日で退役)
 - Reedbush-H (2GPU's/n), Reedbush-L (4GPU's)
 - 東大ITC初GPUクラスタ (2017年3月より), DDN IME (Burst Buffer)
- Oakforest-PACS (OFP) (富士通, Intel Xeon Phi (KNL))
 - JCAHPC (筑波大CCS & 東大ITC), 2022年3月末退役予定
 - 25 PF, TOP500で32位(日本5位)(2021年6月)
 - Intel Omni-Path アーキテクチャ(OPA), DDN IME (Burst Buffer)
- Oakbridge-CX (OBCX) (富士通, Intel Xeon Platinum 8280)
 - 大規模超並列スーパーコンピュータシステム
 - 6.61 PF, 2019年7月 ～ 2023年6月, TOP500で97位(2021年6月)
 - 全1,368ノードの内128ノードにSSDを搭載, OPA





柏キャンパス
OFP, OBCX

柏IIキャンパス
Wisteria/BDEC-01, mdx
ABCI

TX「柏の葉キャンパス」駅

最先端共同HPC 基盤施設 JCAHPC

Joint Center for Advanced High Performance Computing

- 平成25年3月、筑波大学と東京大学は「計算科学・工学及びその推進のための計算機科学・工学の発展に資するための連携・協力推進に関する協定」を締結
 - 筑波大学計算科学研究センター
 - 東京大学情報基盤センター
- 東京大学柏キャンパスの東京大学情報基盤センター内に、両機関の教職員が中心となって設計するスーパーコンピュータシステムを設置し、最先端の大規模高性能計算基盤を構築・運営するための組織
 - <http://jcahpc.jp>



東京大学
THE UNIVERSITY OF TOKYO



筑波大学
University of Tsukuba



Oakforest-PACS (OFP)

Top500: 世界第32位

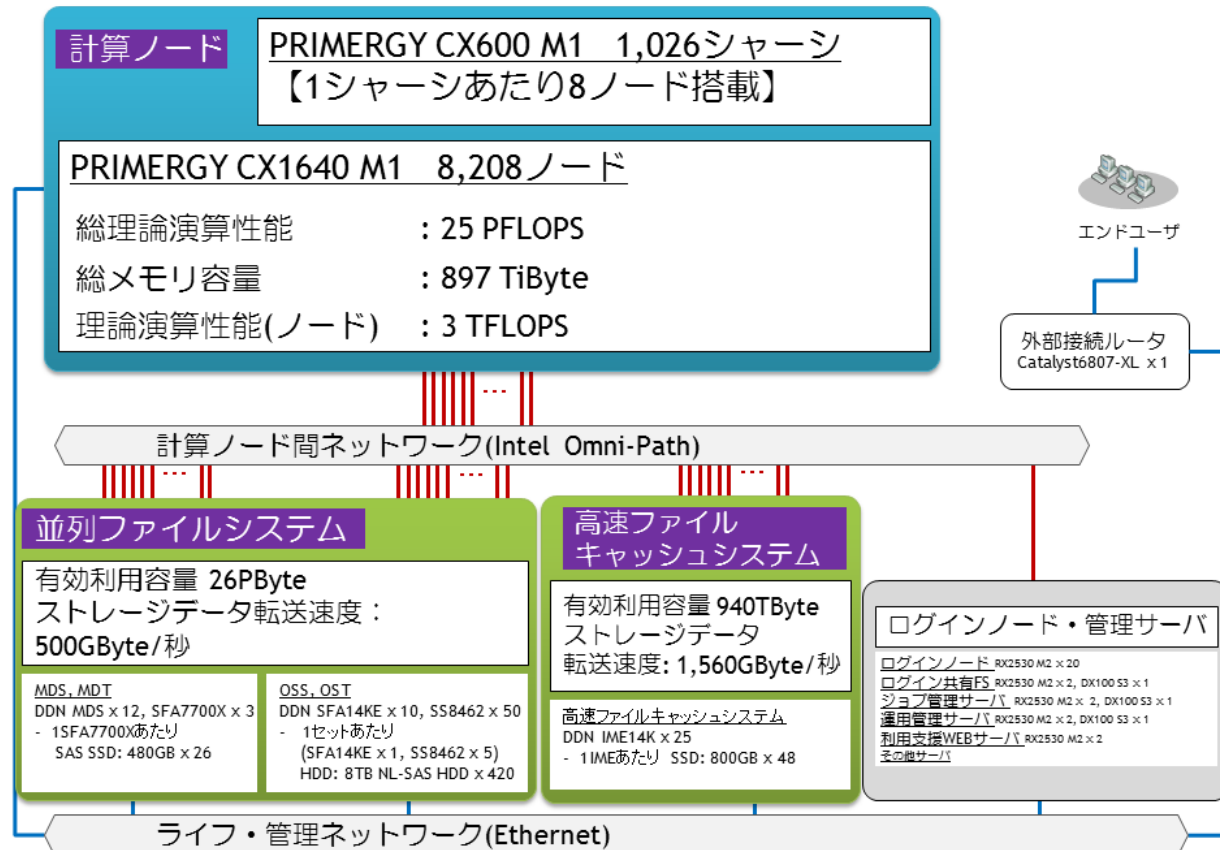
HPCG: 世界第23位
(2021年6月現在)



Fujitsu
PRIMERGY CX1640 M1



Fujitsu PRIMERGY CX600 M1
シャーシ当たりCX1640 M1 × 8搭載



Oakforest-PACS 全景

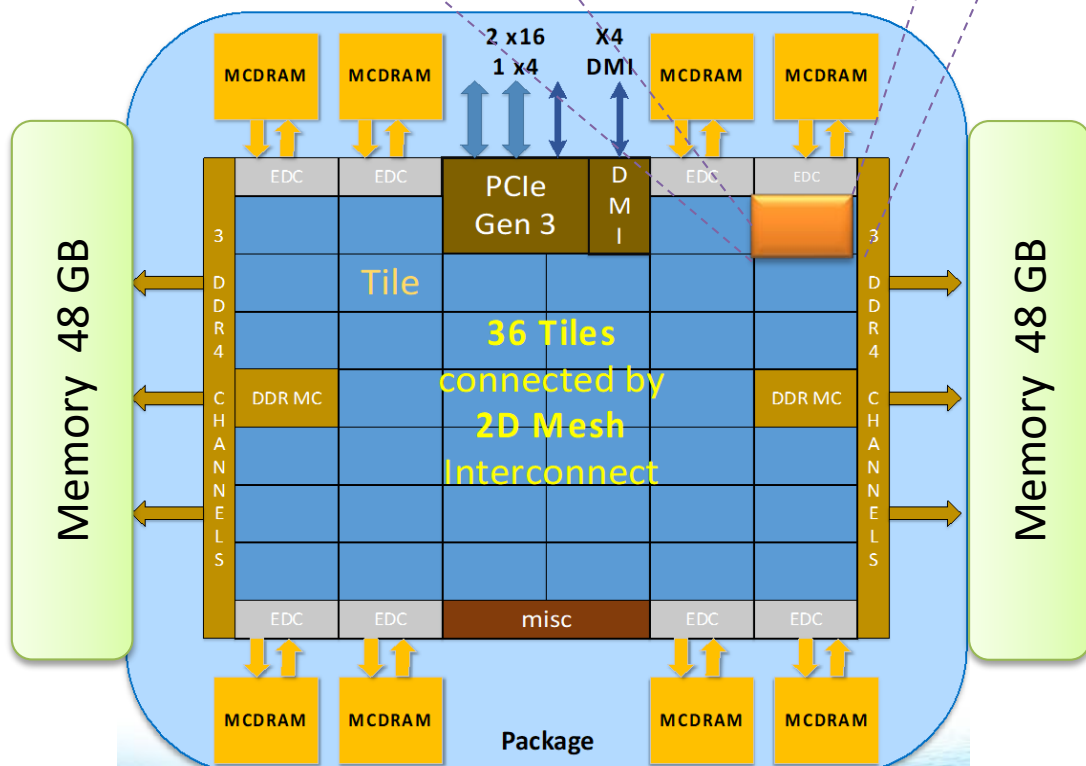
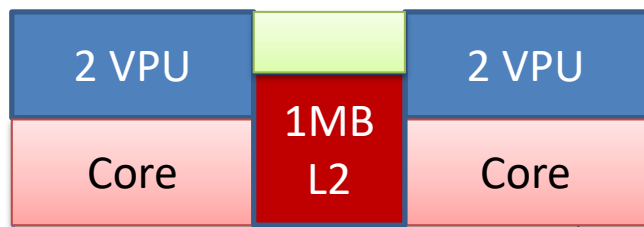


- 全体: **102ラック**
(6列: 計算x4列 + 他x2列)
- 計算ノード: **69ラック**
 - 2ラックで1単位
(1箇所だけ3ラックで1単位)
- インタコネクト (OmniPath)
Director スイッチ: **12ラック**
- ファイルシステム: **16ラック**
 - メタデータ (管理情報): 1ラック
 - 並列ファイルシステム: 10ラック
 - ファイルキャッシュ: 5ラック
- その他: **5ラック**
 - 管理ネットワーク機器: 2ラック
 - ログインノード+プリポスト+Webポータル: 1ラック
 - 管理サーバ: 1ラック
 - 外部接続ルータ: 1ラック
- 消費電力: **3.37 MW**
冷却含め **4.24 MW**
- 水冷 3.0 MW分程度,
残り 0.4 MW分程度は空冷

Oakforest-PACS 計算ノード



HotChips27
KNLスライドより



Intel Xeon Phi 7250 (Knights Landing)

- 1CPU(1ノード) 当たり68コアのメニーコアプロセッサ
- 1.4 GHz
- MCDRAM: オンパッケージ高バンド幅メモリ搭載 16 GB, 490 GB/秒以上(実測)
 - DDR4 96 GB, 85 GB/秒(実測)

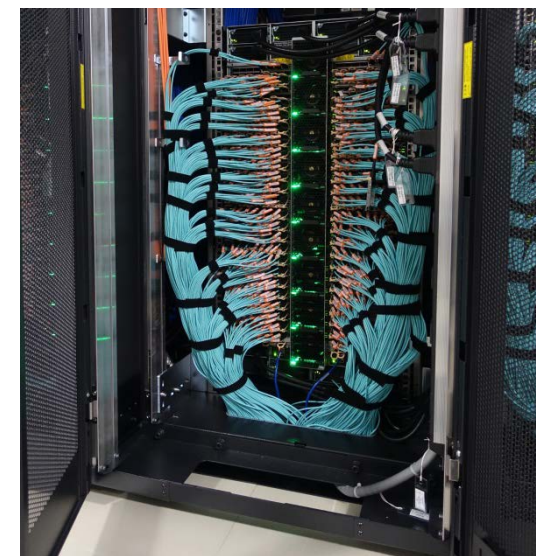
計算ノード・ノード間通信

- 計算ノード

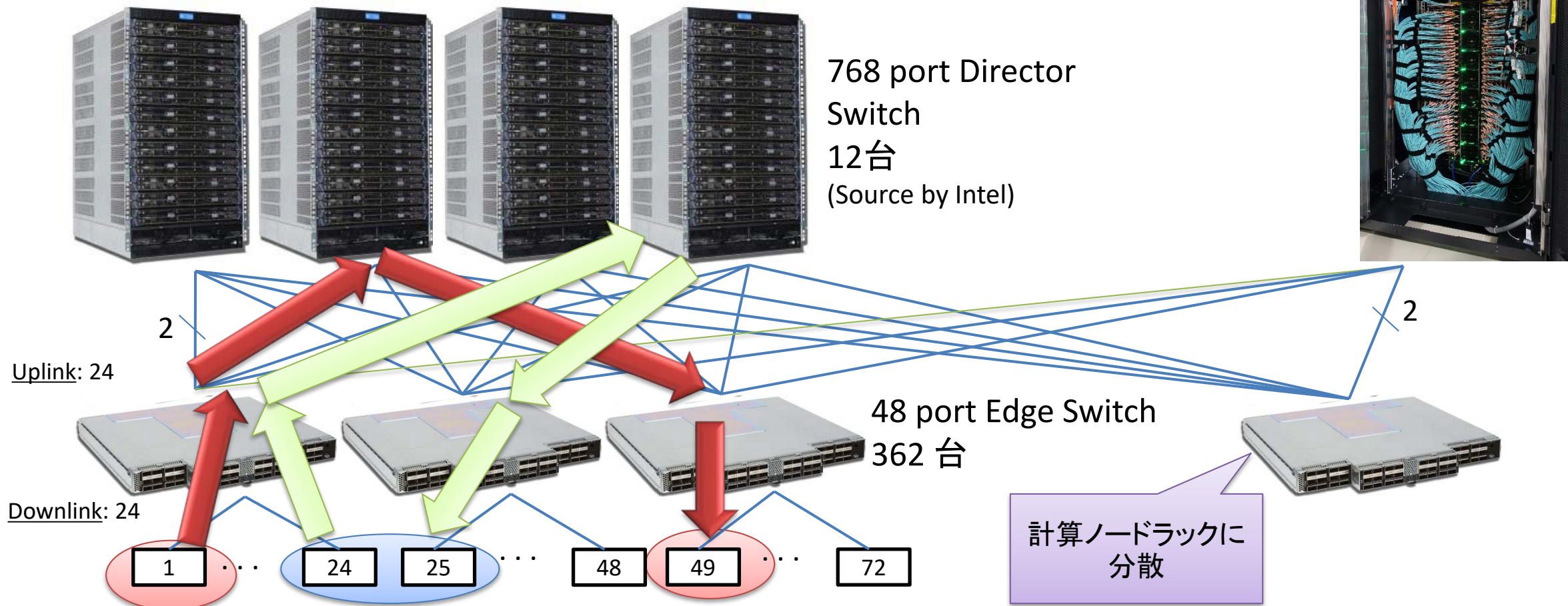
- 1ノード 68コア, 3TFLOPS × 8,208ノード = 25 PFLOPS
- メモリ(MCDRAM(高速, 16GB) + DDR4(低速, 96GB))

- ノード間通信

- フルバイセクションバンド幅を持つFat-Treeネットワーク
- 全系運用時のアプリケーション性能に効果, 多ジョブ運用
- Intel Omni-Path Architecture



Intel Omni-Path Architecture による フルバイセクションバンド幅Fat-tree網



フルバイセクションバンド幅を維持

- フルバイセクションバンド幅: 全ての計算ノードから同時に異なる計算ノードに通信しても性能が低下しない



計算ノードの物理配置を気にせずに
必要な数の計算ノードを確保できる

ファイルシステム (ストレージ)

並列ファイルシステム: Lustreファイルシステム

- 容量: **26.2 PB**
- バンド幅性能: **500 GB/秒**
 - サーバ当たり 50 GB/秒 x 10台
- 全10セット+メタデータ(管理)サーバで構成
 - 1セット当たり、サーバ兼コントローラ+5エンクロージャ
- 合計 4,200個の 8TB ハードディスク, 20%冗長化 (8D2P)

IO500 9位 (2018年11月), 15位 (2019年6月)



高速ファイルキャッシュ(バーストバッファ):
Infinite Memory Engine (IME)

- 容量: **940 TB**
- バンド幅性能: **1.56 TB/秒 = 1,560 GB/秒**
- NVMe接続SSDを使用
 - 合計1,200枚の 800 GB SSD, 冗長化 (Erasure coding)
 - 単体 1.3 GB/s x 48本 x 25台 = 1,560 GB/s

IO-500 **1位** (2017年11月、2018年6月)
4位 (2018年11月), 3位 (2019年6月)



家庭用との比較

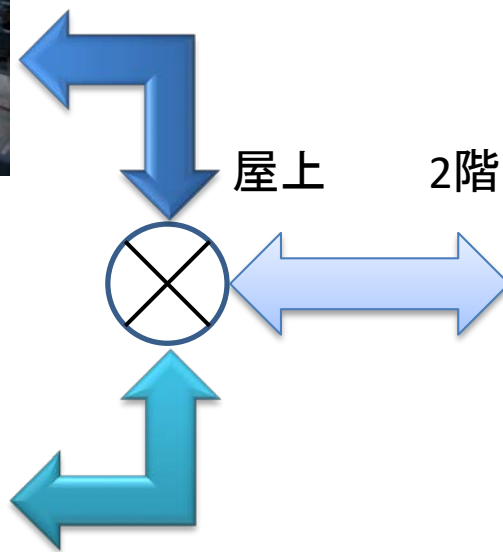
- 2020/10現在、家庭用HDDレコーダーの最大容量: 10TB
- 4K TVに必要なバンド幅: 2.25GB/秒

冷却



チラー
(冷凍機、冷水を作る)と
ポンプ

他の装置は空冷(冷房)



屋上の冷却塔
(気化熱で水を冷やす、
省エネ)

計算ノードのCPUのみ直接水冷、
計算ノードラックはリアドア水冷(ラジエーター)



Oakforest-PACS の仕様

総ピーク演算性能			25 PFLOPS
ノード数			8,208
計算ノード	Product		富士通 PRIMERGY CX600 M1 (2U) + CX1640 M1 x 8node
	プロセッサ		Intel® Xeon Phi™ 7250 (開発コード: Knights Landing) 68コア、1.4 GHz
	メモリ	高バンド幅	16 GB, MCDRAM,実効 490 GB/sec
		低バンド幅	96 GB, DDR4-2400,ピーク 115.2 GB/sec
相互結合網	Product		Intel® Omni-Path Architecture
	リンク速度		100 Gbps
	トポロジ		フルバイセクションバンド幅Fat-tree網

Oakforest-PACS の仕様（続き）

並列ファイルシステム	Type	Lustre File System
	総容量	26.2 PB
	Product	DataDirect Networks SFA14KE
	総バンド幅	500 GB/sec
高速ファイルキャッシュシステム	Type	Burst Buffer, Infinite Memory Engine (by DDN)
	総容量	940 TB (NVMe SSD, パリティを含む)
	Product	DataDirect Networks IME14K
	総バンド幅	1,560 GB/sec
総消費電力		4.2MW(冷却を含む)
総ラック数		102

Oakforest-PACS のソフトウェア

- OS: Red Hat Enterprise Linux (ログインノード)、CentOS および McKernel (計算ノード、切替可能)
 - **McKernel**: 理研AICSで開発中のメニーコア向けOS
 - Linux互換、Linuxに比べ軽量、ユーザプログラムに与える影響なし
 - ポスト京コンピュータにも搭載される予定。
- コパイラ: GCC, Intel Compiler, XcalableMP
 - **XcalableMP**: 理研AICSと筑波大で共同開発中の並列プログラミング言語
 - CやFortranで記述されたコードに指示文を加えることで、性能の高い並列アプリケーションを簡易に開発することができる。
- ライブラリ・アプリケーション: オープンソースソフトウェア
 - **ppOpen-HPC**, OpenFOAM, ABINIT-MP, PHASE system, FrontFlow/blue, LAPACK, ScaLAPACK, PETSc, METIS, SuperLU etc.

運用

- 計算資源は全系を共用（パーティション分けはしない）
 - 全8,208ノード(25PF)を常に全系で運用できるようにしておき、国内最大の計算資源を有効に活用する

利用形態

- 各大学独自の利用コース
- HPCI
 - 全資源の20%を「JCAHPC」として拠出、企業利用可能
- JHPCN(学際大規模情報基盤共同利用共同研究拠点)
 - 全資源の5%程度:企業共同研究、国際共同研究も含む(東大のみ)
- 教育(講義, 講習会)
- 大規模HPCチャレンジ:全ノード占有

Oakforest-PACSに関する情報(1)

- JCAHPC
 - <http://jcahpc.jp>
- Oakforest-PACS がTop500で国内最高性能に認定[プレスリリース]
 - <http://www.jcahpc.jp/pr/pr-20161114.html>
- 運用開始に関する報道等(主なもの)
 - 日本経済新聞
http://www.nikkei.com/article/DGXLASDG01H23_R01C16A2CR0000/
 - 共同通信
<https://this.kiji.is/176995523516057077?c=39546741839462401>
 - ITMedia
<http://www.itmedia.co.jp/news/articles/1612/02/news116.html>
 - マイナビニュース <http://news.mynavi.jp/news/2016/12/02/035/>
 - 他

Oakforest-PACSに関する情報(2)

- 利用全般
 - <https://www.cc.u-tokyo.ac.jp/supercomputer/ofp/service/>
- システム紹介
 - <https://www.cc.u-tokyo.ac.jp/supercomputer/ofp/system.php>
- ジョブクラス
 - <https://www.cc.u-tokyo.ac.jp/supercomputer/ofp/service/job.php>
- 利用申込・利用負担金
 - <https://www.cc.u-tokyo.ac.jp/guide/application/>
- FAQ
 - <https://www.cc.u-tokyo.ac.jp/faq/ofp.php>

3システム: 利用者2,600+, 学外55+%

- Reedbush (HPE, Intel BDW + NVIDIA P100 (Pascal)) (本郷)
 - 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
 - 3.36 PF, 2016年7月～ 2021年11月末(予定)
 - Reedbush-U (CPU only, 2020年6月30日で退役)
 - Reedbush-H (2GPU's/n), Reedbush-L (4GPU's)
 - 東大ITC初GPUクラスタ (2017年3月より), DDN IME (Burst Buffer)
- Oakforest-PACS (OFP) (富士通, Intel Xeon Phi (KNL))
 - JCAHPC (筑波大CCS & 東大ITC), 2022年3月末退役予定
 - 25 PF, TOP500で32位 (日本5位) (2021年6月)
 - Intel Omni-Path アーキテクチャ (OPA), DDN IME (Burst Buffer)
- Oakbridge-CX (OBCX) (富士通, Intel Xeon Platinum 8280)
 - 大規模超並列スーパーコンピュータシステム
 - 6.61 PF, 2019年7月 ～ 2023年6月, TOP500で97位 (2021年6月)
 - 全1,368ノードの内128ノードにSSDを搭載, OPA



Oakbridge-CX (OBCX)

<https://www.cc.u-tokyo.ac.jp/supercomputer/obcx/service/>

- 富士通製
- Intel Xeon Platinum 8280 (Cascade Lake, CLX) (28コア) × 2
 - 合計1,368ノード, Omni-Path Architecture (OPA)
- 共有ファイルシステム (Lustre)
- 128ノードはSSD搭載, 総容量200TBの高速ファイルシステムとして運用可
 - SSD搭載128ノードのうち16ノードは外部ネットワークに直接接続しており (外部接続ノード), 外部リソース (サーバ, ストレージ, センターネットワーク等) との通信可能
- ピーク性能6.61 PFLOPS, TOP500で69位 (2020年11月)
- 更に外部ネットワークに直接接続し, OBCX各計算ノードと連携するGPUサーバ (通称Mini-DP, Wisteria/BDEC-01の「データ・学習ノード」に相当) を利用可能

Oakbridge-CX (OBCX) 世界第97位 2021年6月現在



Fujitsu PRIMERGY CX400 M
シャーシ当たりCX2550 M5 × 4搭載



x1,368 node

体性能

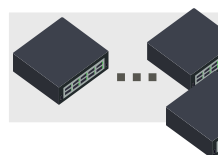
理論演算性能: 6.61PF
記憶容量: 256.5TiB
メモリバンド幅: 385.1TB/s
ノック数: 21ラック
SD搭載: 128ノード

ノード単体

理論演算性能: 4.8384 TF
手記憶容量: 192GiB
メモリバンド幅: 281.6GB/s



計算ノード間ネットワーク (Omni-Path Architecture)
通信性能 100Gbps



x10

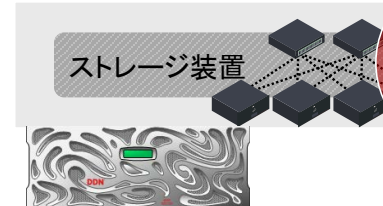
管理サーバ群



x15

FUJITSU Server
PRIMERGY RX2530 M4 x 15
(ジョブ、運用、認証、Web、
セキュリティログ保存)

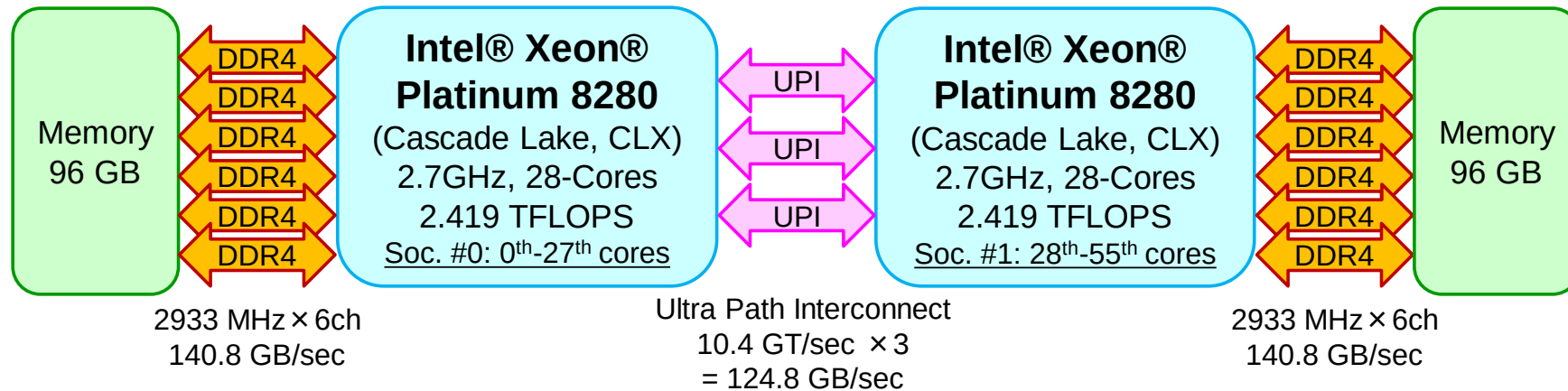
システム



12.4PB

ストレージ装置: DDN ES18KE x2セット
ファイルシステム: DDN ExaScaler
(Lustreベースファイルシステム)

Intel Xeon Platinum 8280 (Cascade Lake-SP, CLX)



Fujitsu PRIMERGY CX2550 M5

- Intel Xeon Platinum 8280 (Cascade Lake, CLX), 富士通
 - 1,368 nodes, 6.61 PF peak, 385.1 TB/sec,
 - **4.2+ PF for HPL**
- Network: Intel Omni-Path, 100 Gbps, Full Bi-Section
- Storage: DDN EXAScaler (Lustre)
 - 12.4 PB, 193.9 GB/sec
- Power Consumption:
 - 950.5 kVA

Oakbridge-CX (OBCX)

計算ノード: ラック当たり最大68ノード, 合計21ラック



全体構成

項目		仕様
総理論演算性能		6.61 PFLOPS
総ノード数		1,368=1,240+112+16
総主記憶容量		256.5 TiB
ネットワークトポロジー		Full-bisection Fat Tree
並列ファイルシステム	システム名	Lustr e ファイルシステム
	サーバ(OSS)	DDN ES18K
	サーバ(OSS)数	8
	ストレージ容量	12.4 PB
	ストレージデータ転送速度	193.9 GB/s

ノードの構成

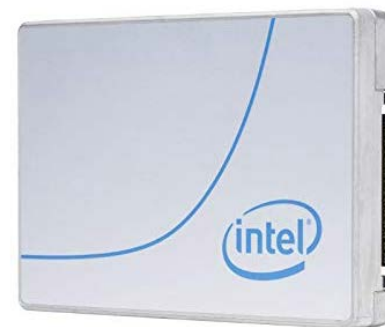
項目		仕様	
製品名		Fujitsu PRIMERGY CX2550 M5	Fujitsu PRIMERGY CX2560 M5
ノード数		1240	112+16
CPU	プロセッサ名	Intel® Xeon® Platinum 8280 (開発コード名：CascadeLake)	
	プロセッサ数(コア数)	2 (28+28)	
	周波数	2.7 GHz	
	理論演算性能	4.8384 TFLOPS	
Memory		192 GiB(DDR4)	
インターコネクト		Intel® Omni-Pathネットワーク (100 Gbps)	
SSD	容量	-	1.6 TB(NVMe)
	読み出し性能		3.20 GB/s
	書き込み性能		1.32 GB/s

ソフトウェア構成

項目	構成
OS	Red Hat Enterprise Linux 7, CentOS 7
コンパイラ	GNUコンパイラ Intelコンパイラ(Fortran77/90/95/2003/2008, C, C++)
メッセージ通信ライブラリ	Intel MPI, Open MPI, Intel Omni-Path Fabric Software
ライブラリ	Intel製ライブラリ(MKL)(BLAS, CBLAS), その他(LAPACK, ScaLAPACK, SuperLU, SuperLU MT, SuperLU DIST, METIS, MT-METIS, ParMETIS, Scotch, PT-Scotch, PETSc, Trillinos, FFTW, GNU Scientific Library, NetCDF, Parallel netCDF, HDF5, Cmake, Anaconda, Xabclib, ppOpen-HPC, ppOpen-AT, MassiveThreads
アプリケーション	Mpijava, OpenFOAM, ABINIT-MP, PHASE, FrontFlow/blue, FrontISTR, REVOCAP-Coupler, REVOCAP-Refiner, OpenMX, xTAPP, AkaiKKR, MODYLAS, ALPS, feram, GROMACS, BLAST, R packages, Bioconductor, BioPerl, BioRuby, BWA, GATK, SAMtools, Quantum ESPRESSO, Xcrypt, Paraview, VisIt, POV-Ray
フリーソフトウェア	Autoconf, automake, bash, bzip2, cvs, emacs, nndutils, gawk, gdb, make, grep, gnuplot, gzip, less, m4, perl, ruby, sed, ubversion, tar, tcsh, tcl, zsh, FUSE, git等
コンテナ仮想化	singularity (dockerイメージ利用可)

SSD搭載ノード

- Regularキューのみで使用可能
 - 最大 112ノード(一時的に数量は変更の場合あり)
 - Intel SSD: DC-P4600 (NVMe IF)
 - 容量: 1.6 TB
 - 性能: Read: 3.2 GB/s, Write: 1.32 GB/s
 - 一時ファイルの置き場所として利用
- ## 利用方法
1. 各計算ノードのSSDを個別に使用
 2. 単一の並列共有ファイルシステムを構成
([BeeGFS on Demand: BeeOND](#))
 - ステージングも可能

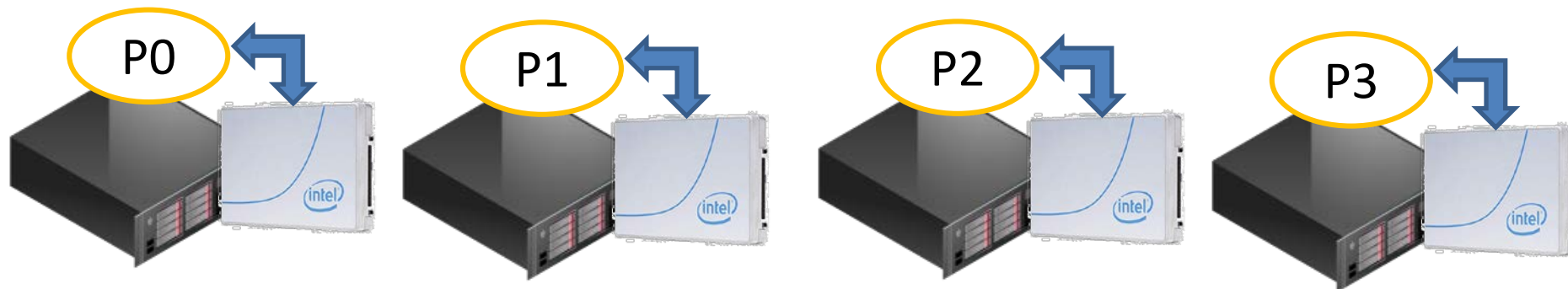


SSDノードの指定(個別に使用)

- 環境変数 PJM_SSDで搭載ノード数を指定
- ジョブ中では **PJM_SSDDIR** 環境変数で割り当てディレクトリ参照

```
$ pjsub -x PJM_SSD=4 run.sh
```

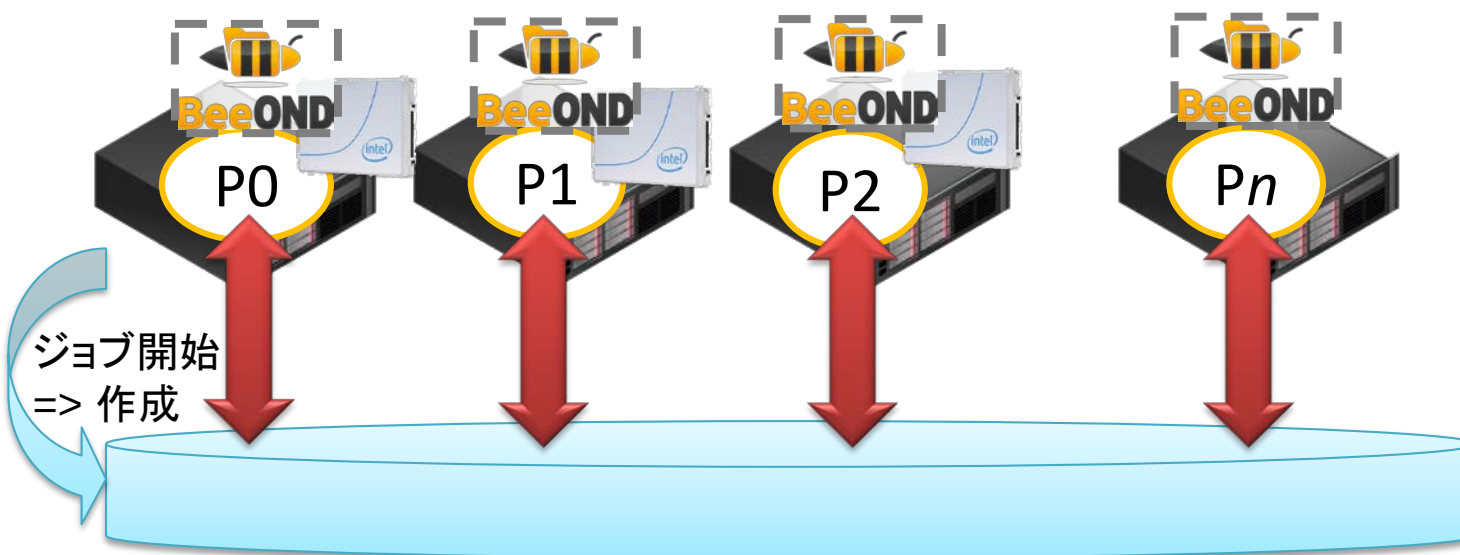
```
export TMPDIR=$PJM_SSDDIR
```



SSDノードの指定 (BeeOND)

- ジョブ内で単一の並列共有ファイルシステムを動的に構成
- 環境変数 PJM_BEEONDでBeeONDの使用を切り替え
 - 環境変数 PJM_SSDでSSD搭載ノード数を指定、SSD無しノードからも利用できる
 - PJM_BEEONDDIR 環境変数でディレクトリ参照

```
$ pjsub -x PJM_SSD=4 -x PJM_BEEOND=1 run.sh
```

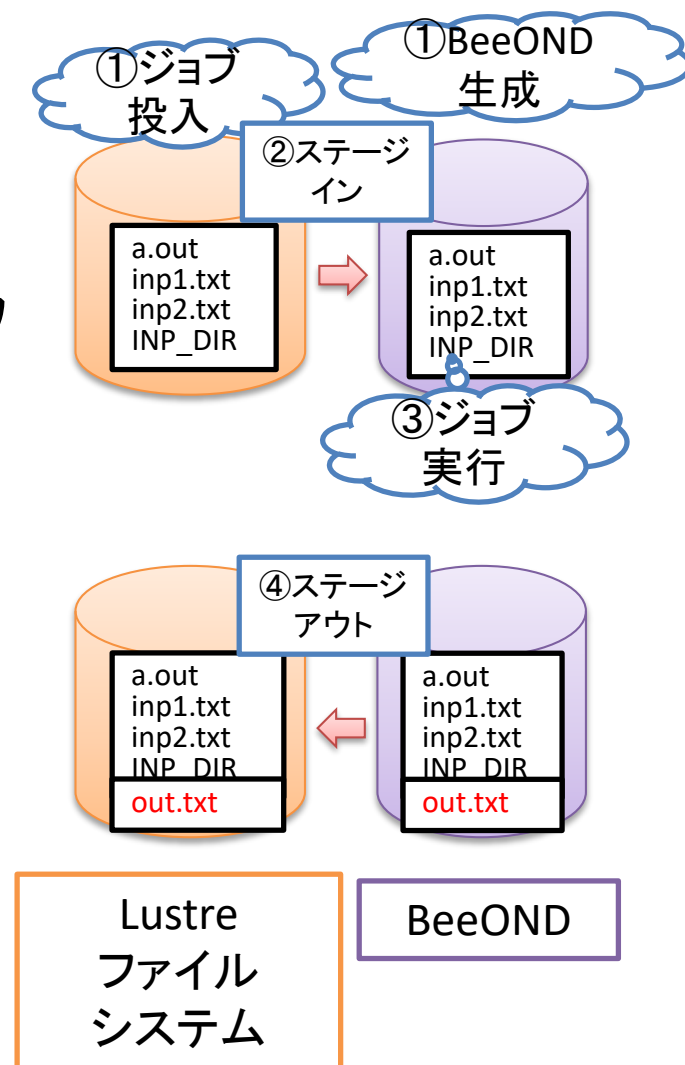


4ノードにSSD搭載
全ノード数は別途
run.shに指定

BeeOND+ファイルステージング

ジョブスケジューラと連携、
高速ファイルキャッシュにステージング

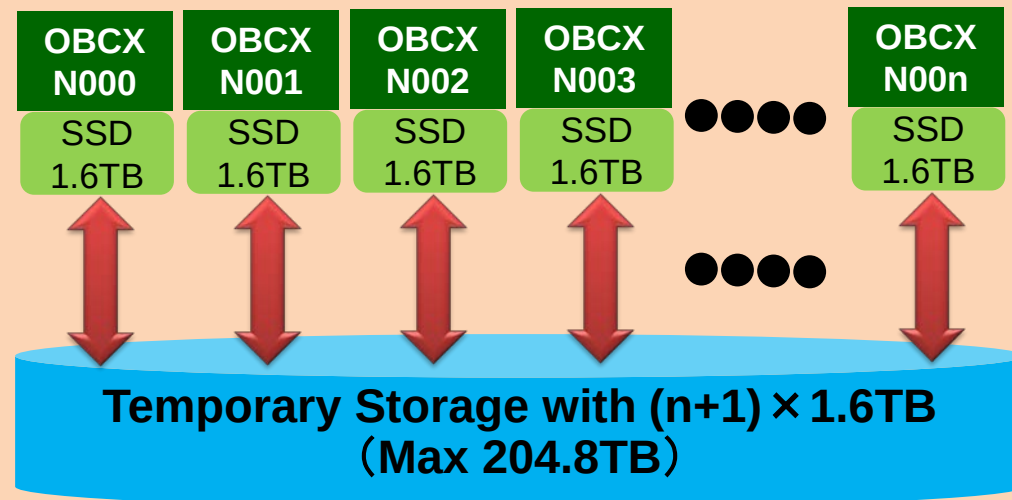
- ステージインリスト、ステージアウトリストをファイルに記述: input_file.txt, output_file.txt
 - ジョブ投入時: 並列ファイルシステム
 - ⇒ **ファイルキャッシュシステム**
 - ジョブ終了時: **ファイルキャッシュシステム**
 - ⇒ 並列ファイルシステム



Oakbridge-CX (OBCX) : BDECに向けた実験システム

- 全1,368ノードのうち128ノードにSSD (Solid State Drive) 搭載
 - Intel SSD + BeeGFS
 - 容量: 1.6 TB/node
 - 読み書き性能: 3.20/1.32 GB/s/node
 - BeeOND (BeeGFS-on-Demand) によって合計 200+TB (128 × 1.6) の高速ファイルシステムとして使用可能
 - データ科学アプリケーション
 - ソフトウェア類も充実
 - ステージング, チェックポイント
 - 128ノードのうち16ノードはSINET経由で外部リソース(サーバー, ストレージ, センサーネットワーク)に直接接続⇒外部接続ノード

BeeGFS on Demand (BeeOND)



Total: 1,368 nodes

**128 nodes
with SSD**

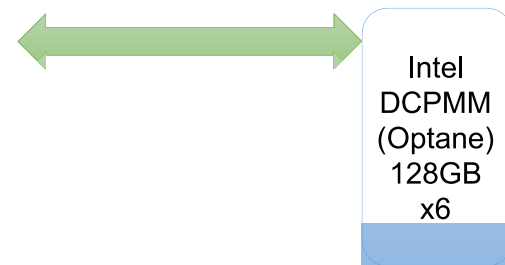
16

OBCXの16ノード(外部接続ノード)
SINET経由で外部計算機資源に直接接続,
BDECにおけるデータ・学習ノード群と同様の
役割



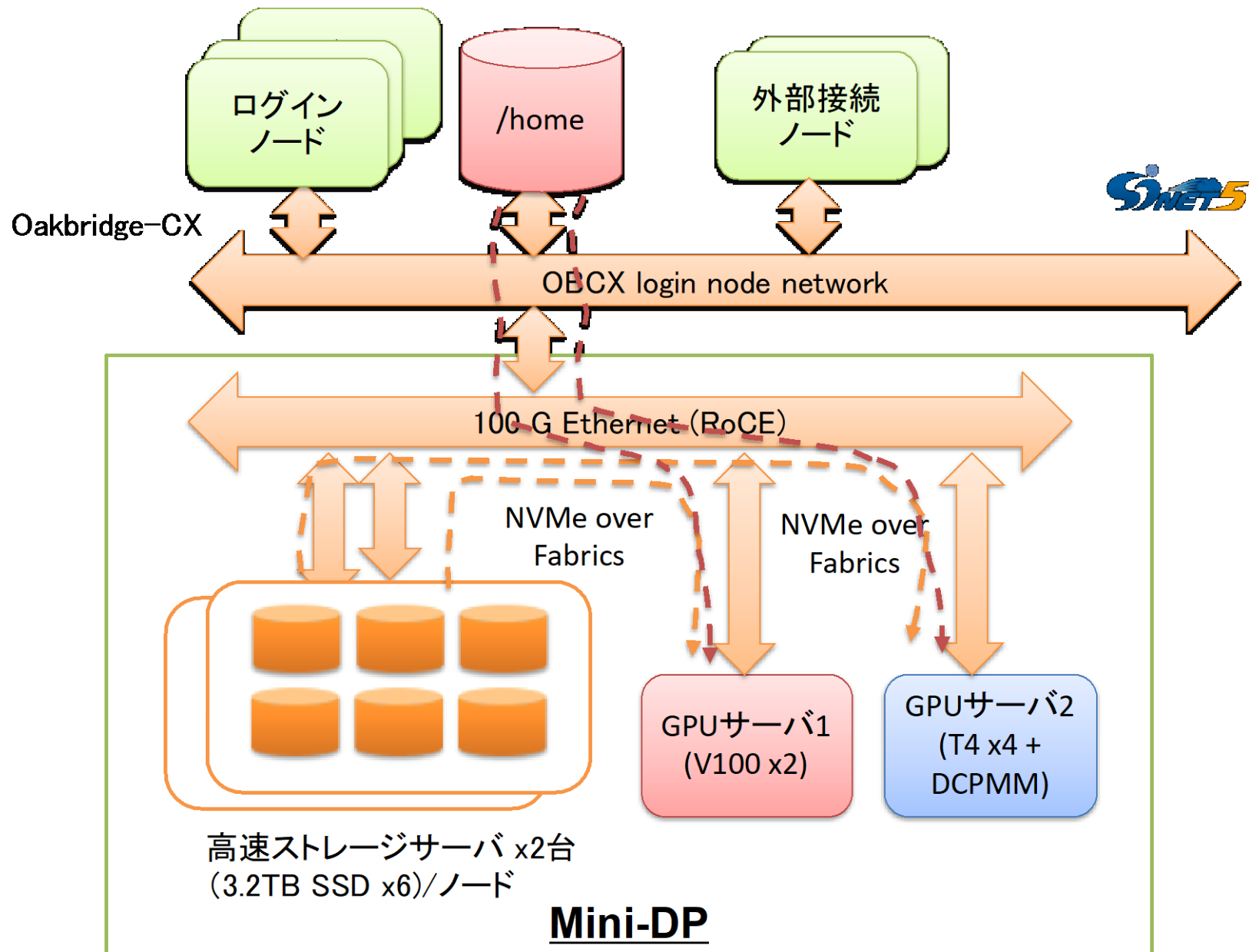
Mini-DP (1/2)

- GPUサーバ1 × 1台
 - Intel Xeon Gold 6230 (Cascade Lake,) (20コア) × 2
 - NVIDIA Tesla V100 32GB × 2
 - 192 GBメモリ, 100 G bit Ethernet
- GPUサーバ2 × 1台
 - Intel Xeon Gold 6230 (CLX) (20コア) × 2
 - NVIDIA Tesla T4 16GB × 4
 - 192 GBメモリ + 1.5 TB不揮発メモリ (Intel DCPMM, Optane DIMM)
 - 100 G bit Ethernet



Mini-DP (2/2)

- 高速ストレージサーバ
× 2台 (3.2TB NVMe
SSD × 6枚搭載)
- GPUサーバ1, 2は
OBCXのログインノード
ファイルシステムをマウ
ント、高速ストレージサー
バのNVMe SSDを
NVMe over Fabricsによ
りブロックデバイスとして
attachして利用可能(構
成変更も応相談)



Mini-DP (Data Platform)

Total: 1,368 nodes

**128 nodes
with SSD**

16

SINET5

**External
Resources**

OBCXの16ノード

SINET経由で外部計算機
資源に直接接続
BDECにおける外部ノード
(EXN)と融合ノード(ITN)
の中間的役割

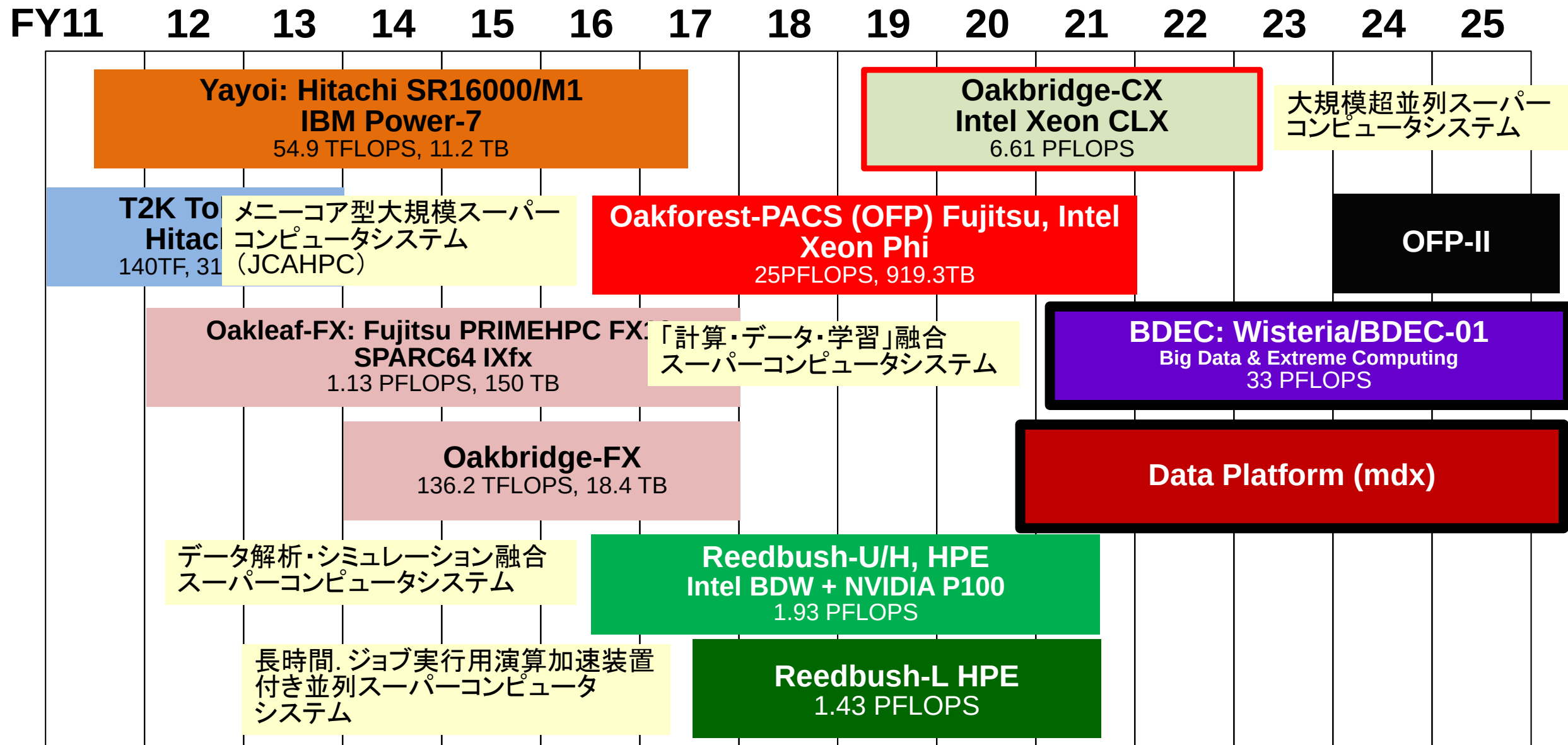
Mini-DP

GPUサーバー with NVMe
SSD, BDECのデータ学習
ノードとmdxの中間的役割

Oakbridge-CXに関する情報

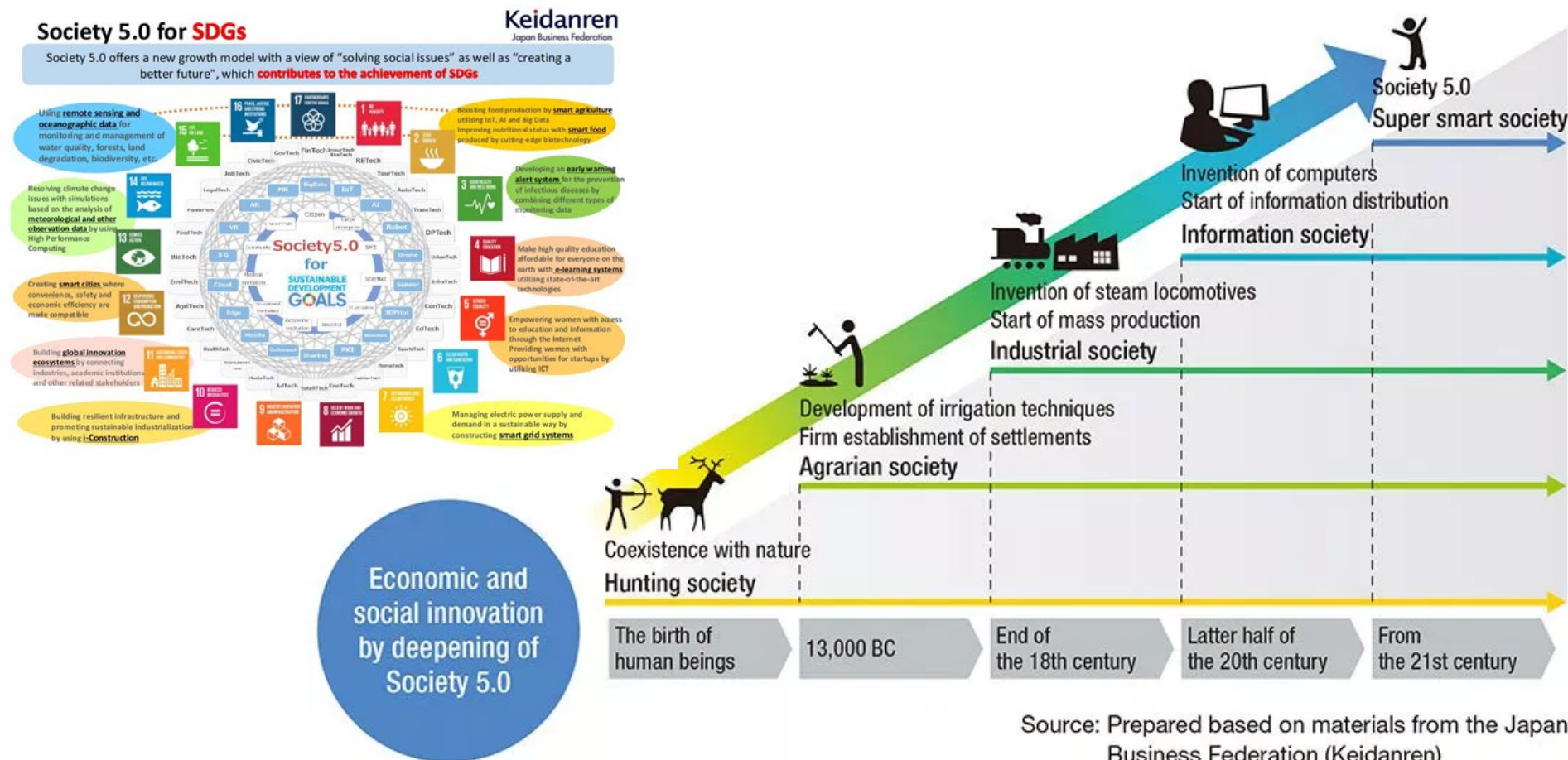
- 全般
 - <https://www.cc.u-tokyo.ac.jp/supercomputer/obcx/service/>
- 利用コース
 - <https://www.cc.u-tokyo.ac.jp/supercomputer/obcx/service/course.php>
- ジョブクラス
 - <https://www.cc.u-tokyo.ac.jp/supercomputer/obcx/service/job.php>
- 利用申込・利用負担金
 - <https://www.cc.u-tokyo.ac.jp/guide/application/>
- FAQ
 - <https://www.cc.u-tokyo.ac.jp/faq/obcx.php>

東大情報基盤センターのスパコン



Society 5.0: 日本が提唱する未来社会のコンセプト

デジタル革新・イノベーション（IoT, AI, ビッグデータ等）により、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システムを構築し，経済発展と社会的課題の解決を両立する，超スマートな人間中心の社会



5.0: Super Smart

4.0: 情報

3.0: 工業

2.0: 農耕

1.0: 狩猟

Source: Prepared based on materials from the Japan Business Federation (Keidanren)

今後のスーパーコンピューティング

- 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

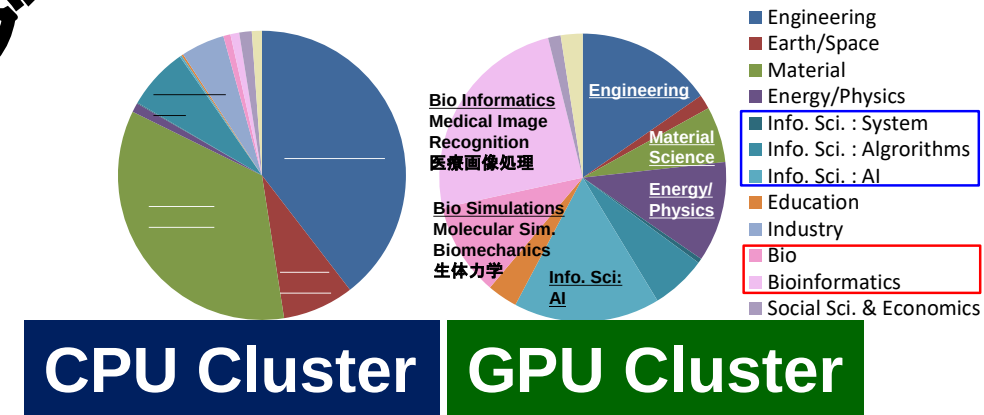
- フィジカル空間とサイバー空間の融合

- 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

- 2つのプラットフォームが2021年春に柏IIキャンパスで始動

- BDEC (Wisteria/BDEC-01): 賢いスパコン
- Data Platform (mdx): Cloud的, よりフレキシブル



BDEC: **S** + **D** + **L**

mdx: s + **D** + **L**



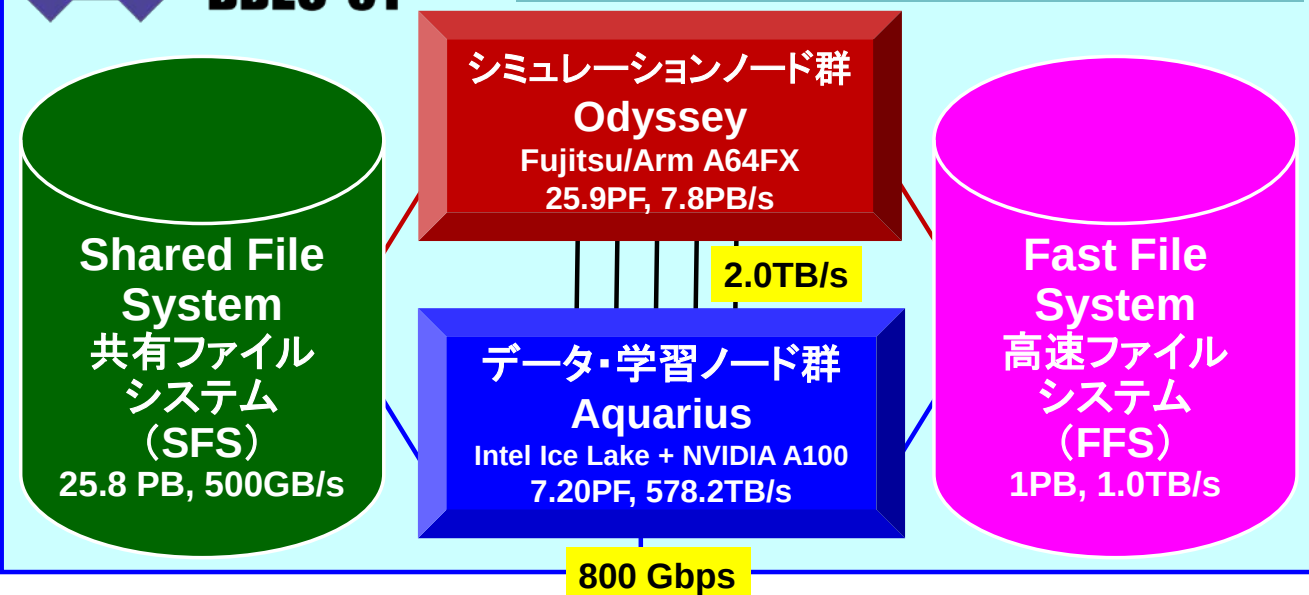
**Wisteria
BDEC-01**

Wisteria/BDEC-01 (S+D+L)融合プラットフォーム



**Wisteria
BDEC-01**

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



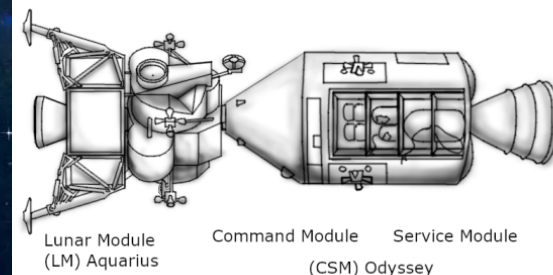
External Network
外部ネットワーク



External
Resources

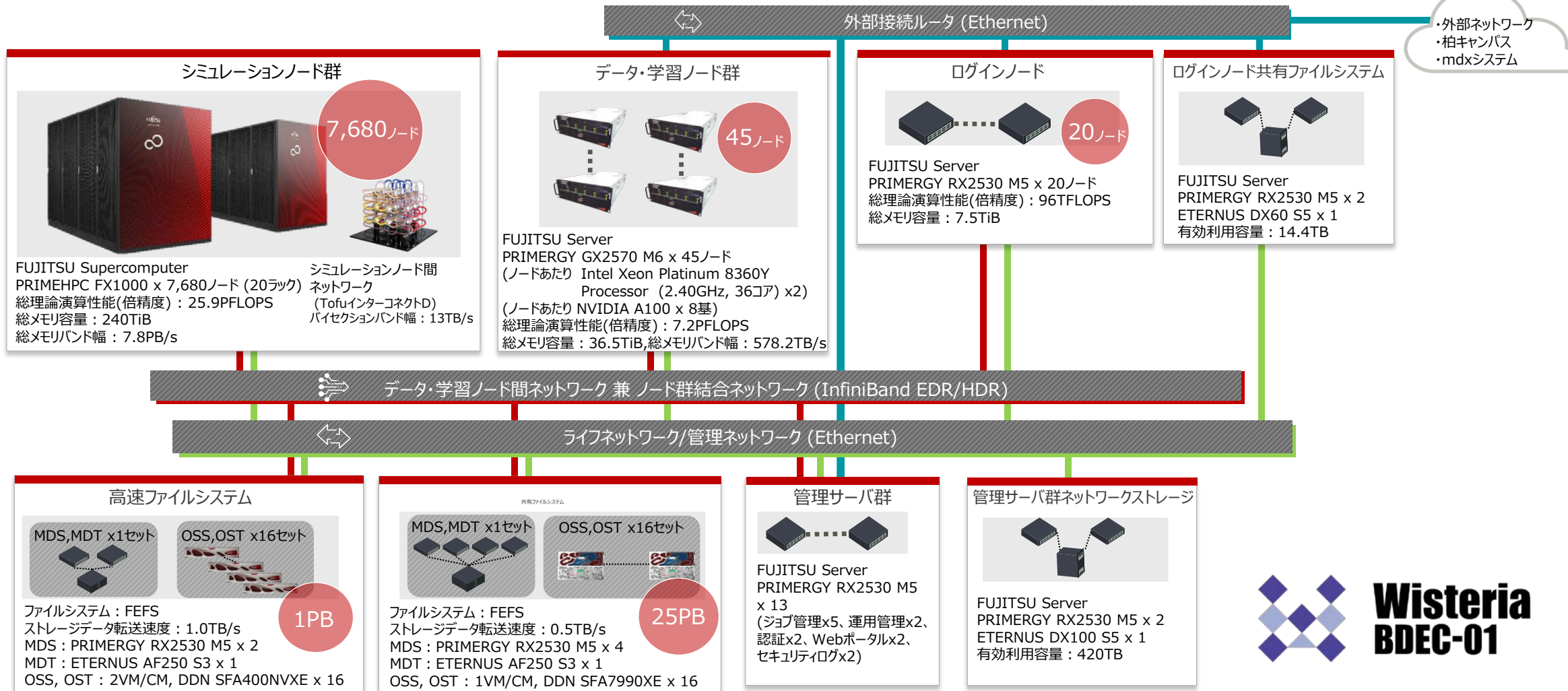
外部リソース

- 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



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)



項目		Wisteria-O (Odyssey)	Wisteria-A (Aquarius)
総演算能力		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
ネットワークポロジ		6次元メッシュトラス	Full-bisection Fat Tree
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	
	スレジ容量	25.8 PB	
	スレジデータ転送速	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	
	スレジ容量	1.0 PB	
	スレジデータ転送速	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コア) システム 2 or 4)	2 (36+36)
	周波数	2.2 GHz	2.4 GHz
	浮動小数演算性能	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
	浮動小数演算性能 (単体)		19.5 TFLOPS
	搭載数		8
	CPU-GPU間接続		PCI Express Gen4 x 16 x 2 (1 x 2あたり片方向32 GB/s)
	GPU間接続		NVLink x 1本 (1本あたり片方向25GB/s)
インターフェース		ToFiber インターフェース D	InfiniBand HDR(200Gbps) x 4

Simulation Nodes Odyssey

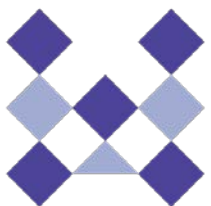
25.9 PF, 7.8 PB/s

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

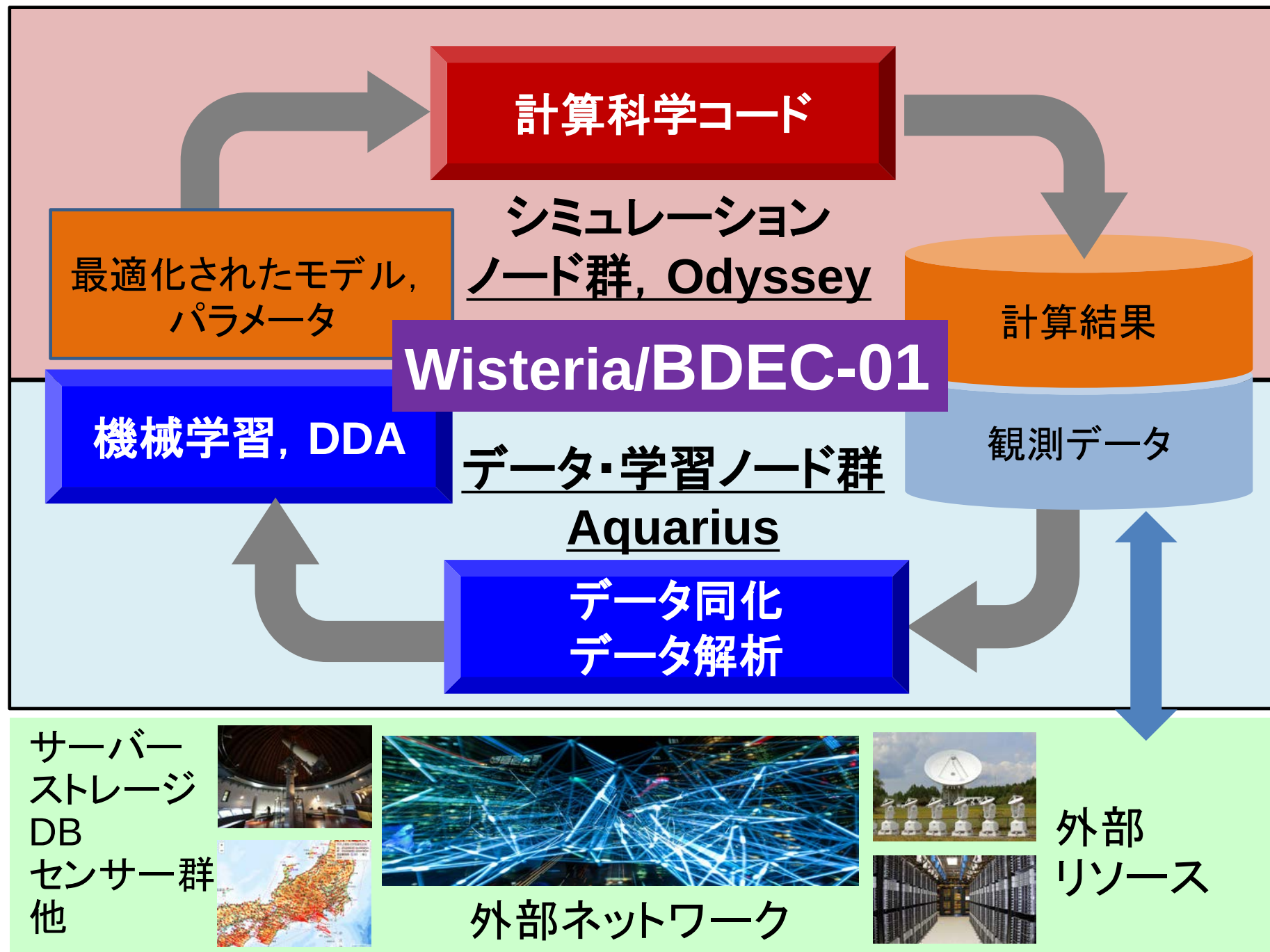
Shared File
System
(SFS)
25.8 PB,
0.50 TB/s

Data/Learning Nodes Aquarius

7.20 PF, 578.2 TB/s



Wisteria BDEC-01



Simulation Nodes Odyssey

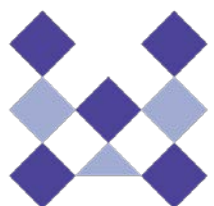
25.9 PF, 7.8 PB/s

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

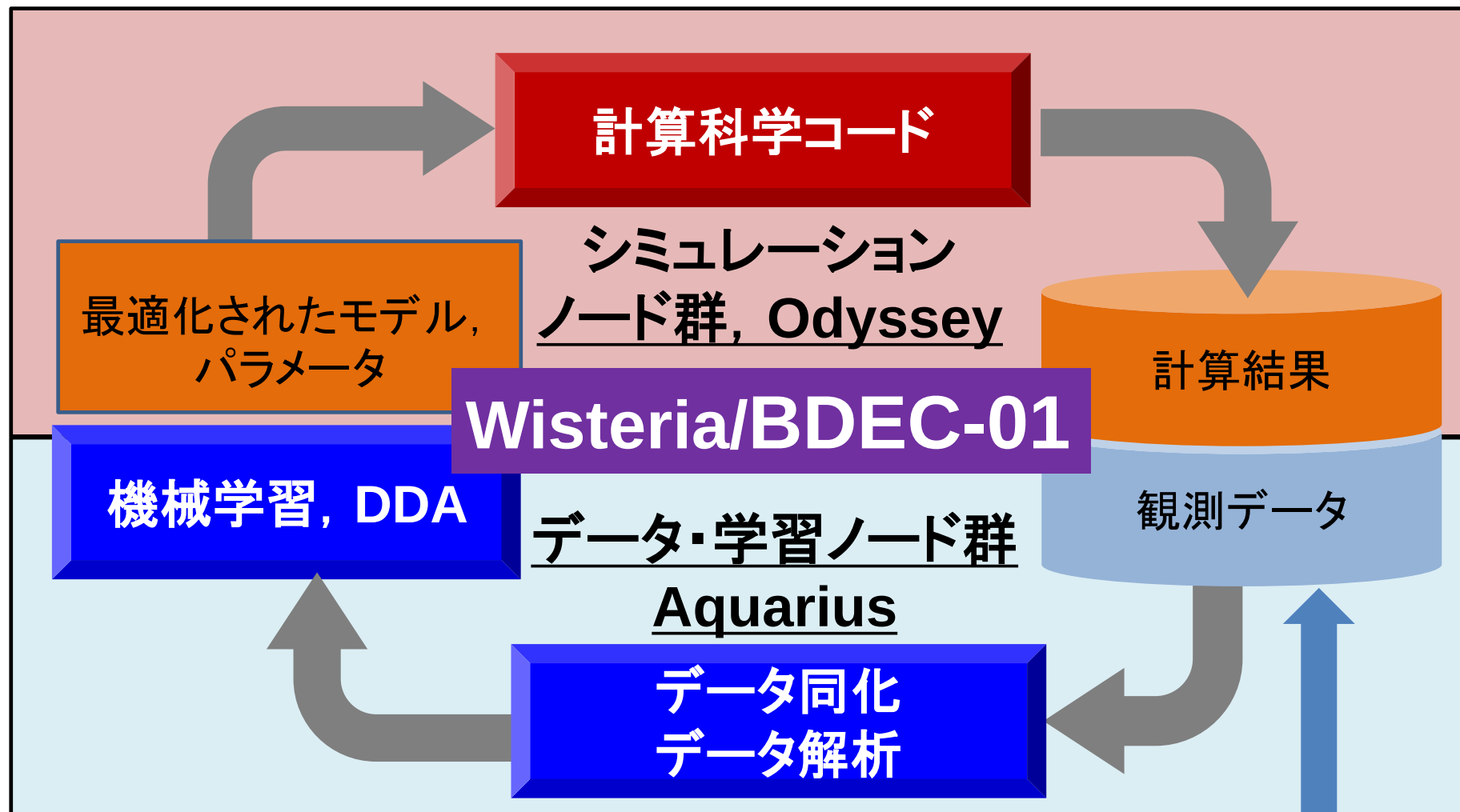
Shared File
System
(SFS)
25.8 PB,
0.50 TB/s

Data/Learning Nodes Aquarius

7.20 PF, 578.2 TB/s



**Wisteria
BDEC-01**



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)

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、P HASE、FrontFlowblue、FrontISTR、REVOCAP-Composer、REVOCAP-Refiner、OpenMX、MODYLAS、GROMACS、BLAST、R packages、bioconductor、BioPerl、BioRuby、BWA、GATK、SAMtools、Quantum ESPRESSO、Xcrys、ROOT、Gart4、LAMMPS、CP2K、NWChem、DeepVariant、Paraview Vst、POV-Ray、TensorFlow、Chainer、PyTorch、Keras、Horovod、MXNet	
		Theano

項目	Wisteria-O (Odyssey)	Wisteria-A (Aquarius)
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	

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

The article on the Wisteria/BDEC-01 appears in *HPCwire*

<https://www.hpcwire.com/2021/02/25/japan-to-debut-integrated-fujitsu-hpc-ai-supercomputer-this-spring/>

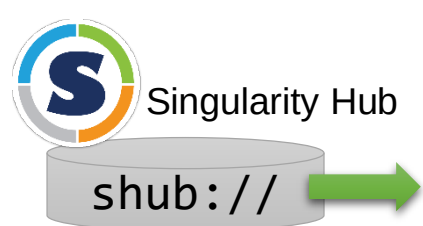
The screenshot shows a web browser displaying the HPCwire article. The article title is "Japan to Debut Integrated Fujitsu HPC/AI Supercomputer This Spring" by Tiffany Trader, dated February 25, 2021. The article text states: "The integrated Fujitsu HPC/AI Supercomputer, Wisteria, is coming to Japan this spring. The University of Tokyo is preparing to deploy a heterogeneous computing system, called 'Wisteria/BDEC-01,' that will tackle simulation and big data 'learning' workloads, in support of Japan's Society 5.0 project, which seeks to achieve economic and social gains through the integration of cyber and physical space. The system comprises two..." The article features a large image of purple wisteria flowers. To the right of the article is a Dell Technologies advertisement with the text "Free HPC from the bounds of physical infrastructure." and an Intel logo. Below the advertisement is a section titled "Leading Solution Providers" with logos for Adaptive, AMD, Ansys, Aspec Systems, ATEMPO, Atos, AWS, Bright Computing, CRAY, ddn, Dell Technologies, and Fujitsu. At the bottom of the article, there are two buttons labeled "Aquarius" and "Odyssey".

**Wisteria
BDEC-01**

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

- 東大情報基盤センターについて
- サービス概要
- 企業による利用制度
 - 制度概略と利用提案
 - 申込・審査・成果報告
 - スパコン利用への道
 - ①お試しアカウント付き並列プログラミング講習会
 - ②トライアルユース制度
 - ③ 企業利用制度
- スーパーコンピュータシステム概要
 - Reedbush
 - Oakforest-PACS
 - Oakbridge-CX
 - Wisteria/BDEC-01
 - **運用**
- ご質問は

Singularity: コンテナ仮想化



Singularity recipe



コンテナイメージ
ファイル



Production向け (Read only)

```
$ singularity build  
container.img <source>
```

環境構築 (writable image)

```
$ sudo singularity build  
--writable container.img  
<source>
```

環境構築 (sandbox)

```
$ sudo singularity build  
--sandbox container.img  
<source>
```

コンテナイメージ
ファイル



container.img



コンテナ環境内で実行 (shell)

```
$ singularity shell  
container.img  
> python mnist.py
```

コンテナ環境 コマンドで実行(exec)

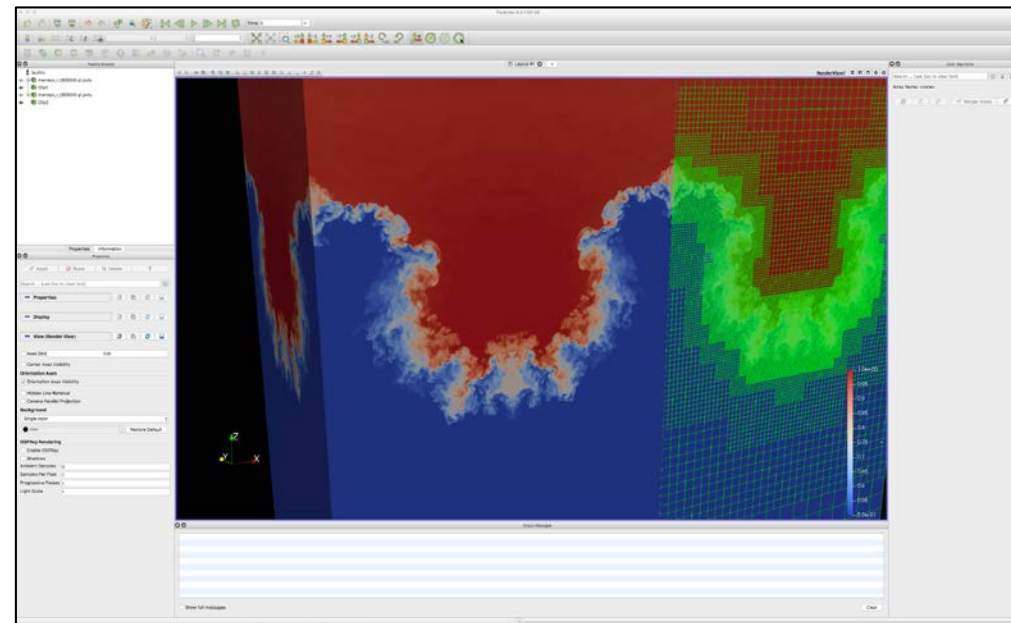
```
$ singularity exec  
container.img python  
mnist.py
```

定義された通りに実行(run)

```
$ singularity run  
container.img  
or  
$ ./container.img
```

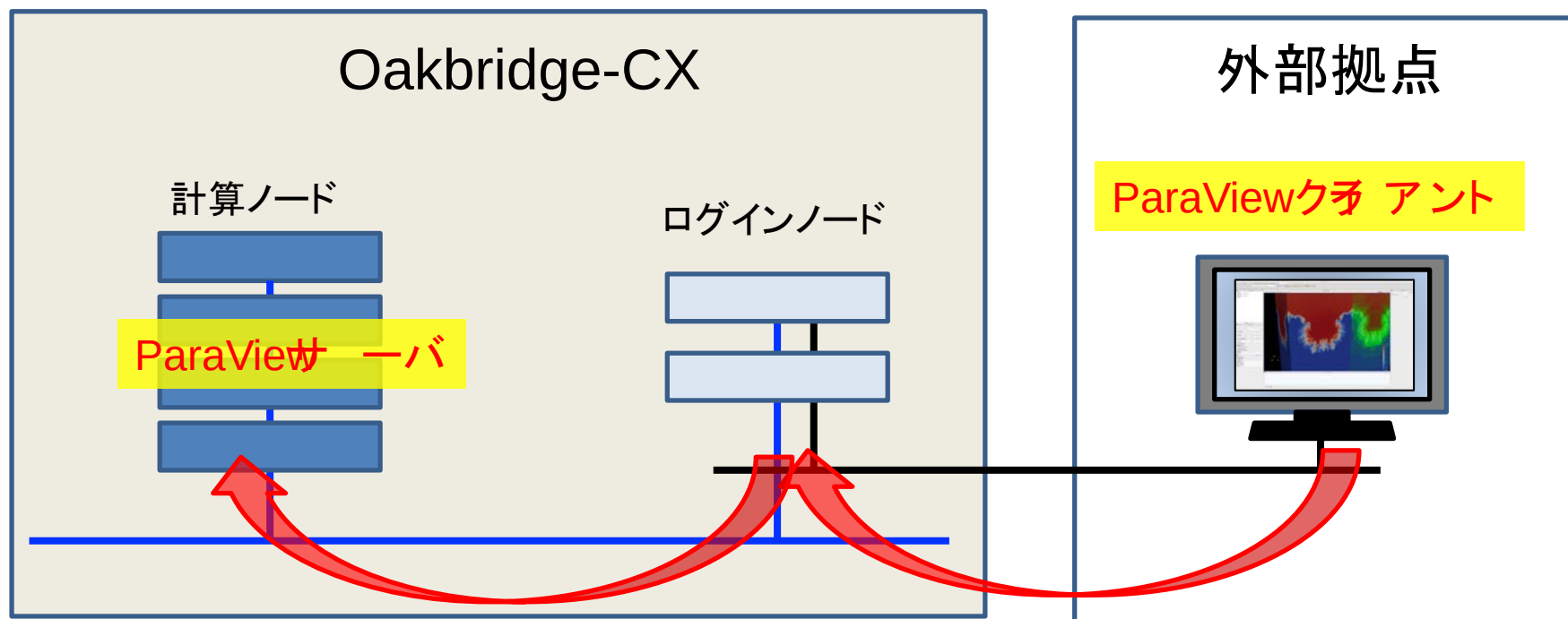
ParaView

- オープンソース、マルチプラットフォームの可視化アプリケーション
- 大容量データの可視化のための分散環境に対応
- クライアント・サーバモデルによるリモート可視化対応
- 様々なデータ形式に対応
- OpenFOAMデータ可視化などに利用される
- <https://www.paraview.org>



リモートからのParaView利用

- クライアント・サーバモデルによるリモート可視化対応
 - 乗客のPCのクライアントからGUI操作によって、Oakbridge-CXのParaViewサーバを制御可能
 - SSHポートフォワーディングを利用



科学技術計算用アプリケーション

■ 商用ソフト

- Altair **HyperWorks** (総合CAEプラットフォーム) <http://altairhyperworks.jp/>

HyperWorks, OpenFOAM,
を用いた講習会を実施

■ スパコン向けオープンソースコードが充実

- OpenFOAM**

(Open source Field Operation And Manipulation)

<http://www.openfoam.com/>

- ppOpen-HPC**

(自動チューニング機構を有するアプリケーションの開発・実行環境)

<http://ppopenhpc.cc.u-tokyo.ac.jp/ppopenhpc/>

- FrontISTR**

(大規模並列有限要素解析アプリケーション・ソースコード)

<http://www.multi.k.u-tokyo.ac.jp/FrontISTR/>

- FrontFlow**

(乱流燃焼解析ソフトウェア)

<http://www.ciss.iis.u-tokyo.ac.jp/dl/>

Altair HyperWorksのライセンス

■国内アカデミックユーザ(大学、短大、大学校、高専等に所属の方)

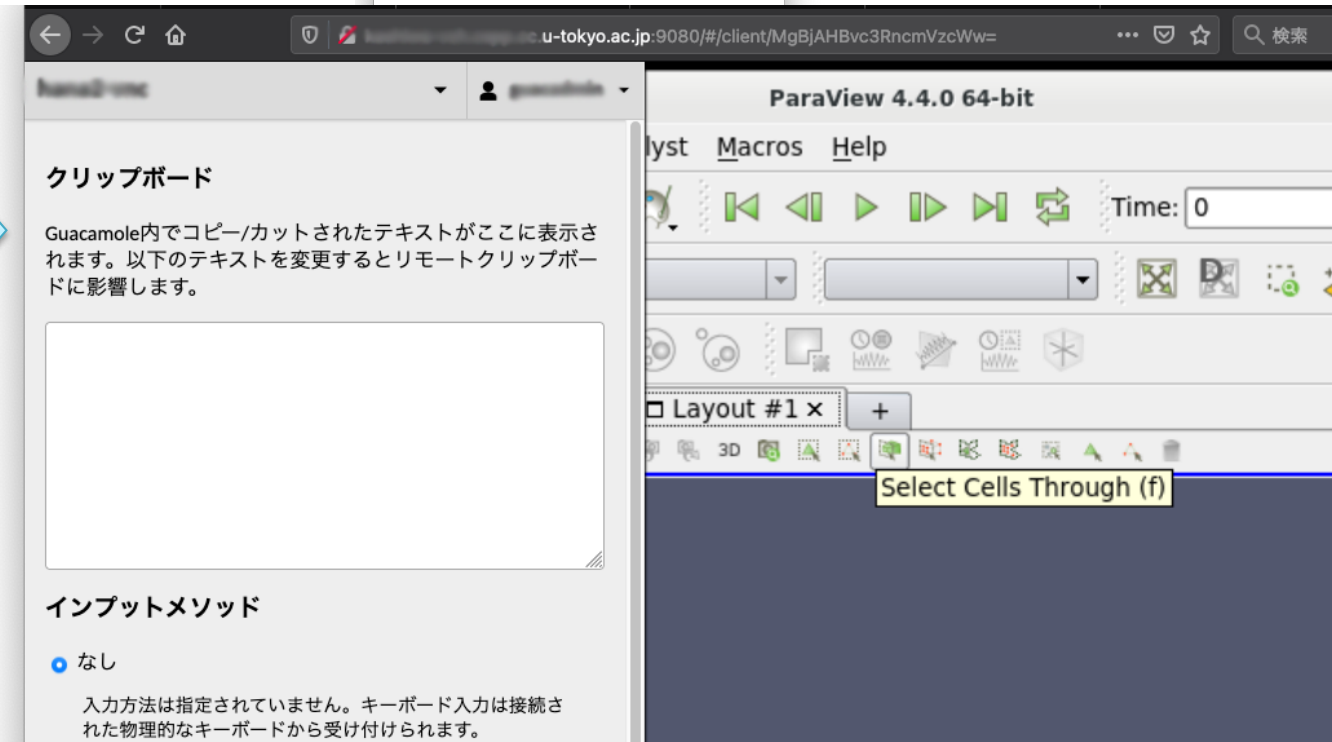
- ・**ライセンス料: 無料**(基盤センター所有のライセンスで利用可能)
- ・HyperMesh, HyperViewなどクライアントPCで動作するソフトも利用可能

■その他のスパコンユーザ(企業や研究機関に所属の方)

- ・基盤センター所有のライセンスは利用不可
- ・**ライセンス個別購入**により利用可能
(ご希望の方はuketsuke@cc.u-tokyo.ac.jpまで問い合わせ下さい)

Apache Guacamoleの利用

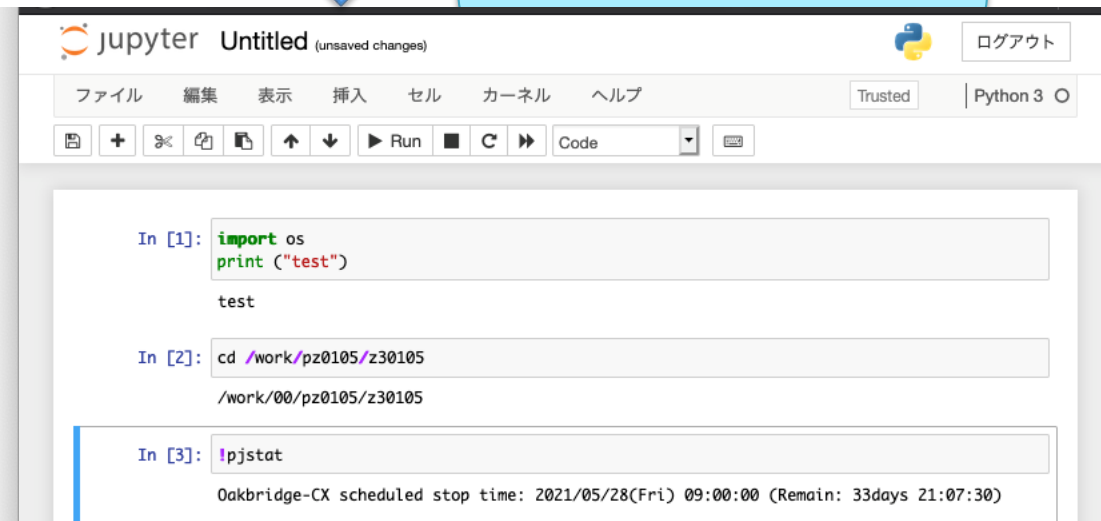
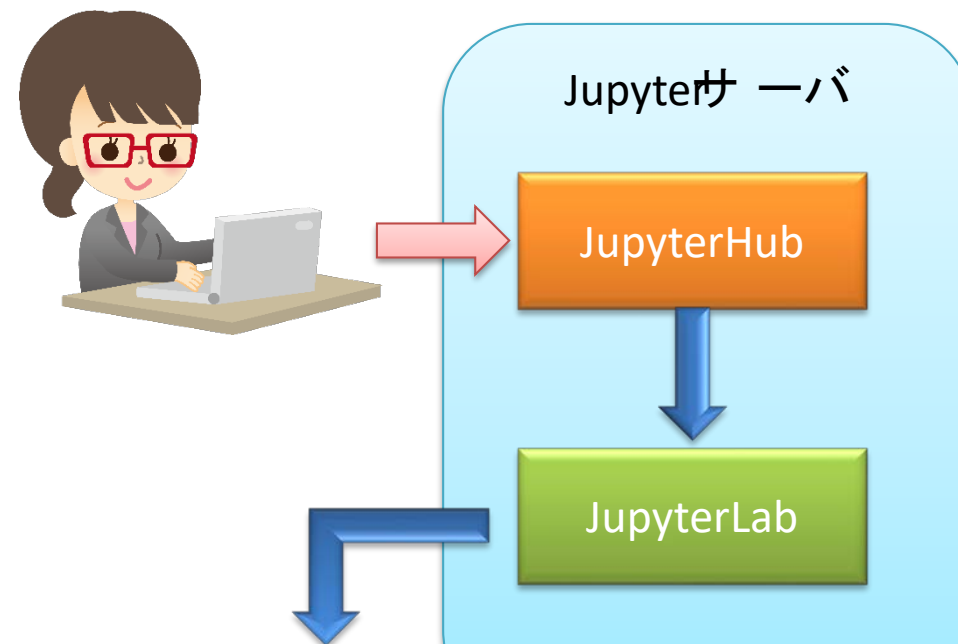
- 一般的なWebブラウザからVNCに接続できる
 - ログインノードでVNCサーバを起動しておく
 - WebブラウザでWisteriaのGuacamoleサーバにアクセス、ブラウザ上でVNC環境を利用



JupyterLabの利用

- WisteriaでJupyterHubを用意
 - 自力でのPython環境構築不要
 - ユーザごとに用意したPython環境も利用できる
- 計算ノードとの連携
 - JupyterLabからバッチジョブを簡便に投入できる(予定)

! psub 行番号



システムの利用方法

- カテゴリー
 - 一般(大学・公共機関), 公募(企業(有料), HPCI等(無料))
 - グループ, パーソナル(Wisteria/BDEC-01除く)
 - 通常, ノード固定(OBCX, Wisteria-Aquarius)
- 通常利用
 - パーソナル
 - 一般(大学・公共機関)のみ
 - グループ
 - 一般(大学・公共機関)
 - 企業(有料・審査あり)
 - その他公募(無料・提案書)
 - 若手・女性支援, HPCI
 - 学際大規模情報基盤 同利典 同研究拠点(JHPCN)

参考情報：利用方法

- ノード固定 (OBCX, Wisteria-A)
 - https://www.cc.u-tokyo.ac.jp/supercomputer/obcx/service/fixed_node.php
 - https://www.cc.u-tokyo.ac.jp/supercomputer/wisteria/service/fixed_node.php
- 若手・女性支援
 - <https://www.cc.u-tokyo.ac.jp/guide/young/>
- HPCI
 - <http://www.hpci-office.jp/>
- 学際大規模情報基盤共同利用・共同研究拠点 (JHPCN)
 - <https://jhpcn-kyoten.itc.u-tokyo.ac.jp/ja/>

計算資源配分のポリシー

- 利用するコース (パーソナル・グループコース),
利用申込したノード数に応じて, 計算ノードの利用可能時間である「トークン」を割当てます。
 - 割り当てられたトークン内であれば (一部のコース, サービスを除き) 利用できるノード数制限はなく, 最大利用可能ノード数まで, バッチジョブの実行を可能。
- 1つのシステムに申し込めば他のシステムへトークンを移し, 複数システムを使用可能な場合もある (トークン移行)
 - カテゴリーとしては「通常・一般 (大学・公共機関)」の場合 (グループ・パーソナル)

トークン(token)

参考: <https://www.cc.u-tokyo.ac.jp/supercomputer/ofp/service/token.php>
<https://www.cc.u-tokyo.ac.jp/supercomputer/obcx/service/token.php>
<https://www.cc.u-tokyo.ac.jp/supercomputer/wisteria/service/token.php>

- トークンは、バッチジョブ実行ごとに消費。(インタラクティブノードを利用したバッチジョブを除く)
 - ノード時間積「経過時間 × ノード数 × 消費係数」により消費
 - 消費係数: Wisteria-Aのみ GPUあたり 3.00, 他システムは 1.00
- トークンを使い果たすとジョブ実行不可: 1日単位モニター
 - トークン不足の場合, ジョブはsubmitできない
 - 計算資源に余裕がある場合にのみ, トークンを追加可能(企業利用: 要審査)
- トークンは, 利用を許可された有効期間内に全量が利用できることを保証するものではありません。
- 利用を許可された期間のみを有効期間としているため, 次年度への繰り越しや返金等は不可。
- 1ヶ月単位の利用負担金は均等割(年額の1/12)になります。

参考：トークン移行

- 対象システム
 - (Reedbush)
 - Oakbridge-CX
 - Oakforest-PACS
 - Wisteria/BDEC-01
- 1つのシステムに申し込めば他システムへトークンを移し、複数システム利用できる(換算表は次頁)
 - 通常・一般(大学・公共機関, グループコース, パーソナル)のみ
 - 公募型(企業, 若手・女性, HPCI, JHPCN等), ノード固定(Oakbridge-CX, Wisteria-A)は不可
 - 「通常かつ一般」の場合「以外」はWisteria-O, Wisteria-A も別々に申し込む必要がある。
 -

トークン移行

参考: https://www.cc.u-tokyo.ac.jp/guide/application/transfer_token.php

1つのシステムに申し込めば他システムへトークンを移し、
複数システム利用可能(企業利用, 公募利用を除く)

現在ご利用のシステム	トークン移 先 のシステム			
	Reedbush	Oakforest-PACS	Oakbridge-CX	Wisteria/BDEC-01
Reedbush	—	1.5	0.75	1.2
Oakforest-PACS	0.6	—	0.5	0.8
Oakbridge-CX	1.3	2.0	—	1.6
Wisteria/BDEC-01	0.8	1.2	0.6	—

- 移行先に追加されるトークン量(ノード時間) = 移行トークン量 × 係数

利用カテゴリーとトークン移行

	トークン移行
パナ ナル(般)(大学・公共機関)	○
グループ 通常・般)(大学・公共機関)	○
グループ (ノ一固 定・般)(大学・公共機関)	×
グループ 通常・公募)(企業・HPC等)	×
グループ (ノ一固 定・公募)(企業・HPC等)	×

Oakforest-PACS:

グループコース負担金(年)

利 益 担 金 年 額、税込)	利用可能 ノ ード 数	割当トーク量 (年 間) 及び消費係数	デイス量
基 セ ヲ 大学公 共 機 関： 50,000円 企業： 60,000円	最大2,048ノード	8,640ノード時間 (1ノード × 360日 相当) 消費係数 : 1.00	1TB グ ル プ 当 り)

Oakforest-PACS: パーソナルコース

大学・公共機関のみ
申込数1口あたり(最大6口)

コース	利用負担金 (年額, 税込) 大学・公共機関等	利用可能 ノード数	トークン	ディスク量
パーソナル コース	50,000円/年	最大 2,048ノード (前頁参照)	8,640 ノード時間 (1ノード×360日相当) 消費係数: 1.00	1TB

バッチジョブキュー: Oakforest-PACS

- パーソナルユーザ, グループユーザ(一般)で共通
- 全てのキューでメモリモードの異なる -cache と -flat 2種類
 - 例: regular-cache, regular-flat
- **interactive_n1/n16はトークンを消費しない**

代表キュー名	キュー名	最大ノド 数	実行制限時間(経過時間)	ノド 当たりメモリ量(GB)
interactive	interactive_n1	1	2 h	82 (-cache)
				96 (-flat)
	interactive_n16	2-16	10 min	82 (-cache)
				96 (-flat)
debug	debug	1-128	30 min	82 (-cache)
				96 (-flat)
regular	small	1-128	48 h	82 (-cache)
	medium	129-512	48 h	96 (-flat)
	large	513-1024	48 h	
	x-large	1024-2048	24 h	
prepost	prepost	1	6 h	222 (Xeon)

Oakbridge-CX: グループコース(ノード固定)

- グループコース(ノード固定)は、研究グループ等で利用するためのコースで、**1ノード単位**で申込みます。
 - 申込み分のノードを占有できる
 - **パーソナルコースグループ(コース通常)**とは別のバッチジョブキューを使います。
 - **グループ(コース通常)**と同じ計算機資源を使用することも可能: その分トークンは減る
 - **トークンがなくなった時点で占有ノードも使えなくなります(資源に余裕があればトークン追加購入可能)**
 - 審査制, 全体の10%程度を上限とする
- **設定のカスタマイズが可能となる**
 - **個別のダインノード・ストレージ(データ隔離等柔軟に対応)**

Oakbridge-CX

グループコース負担金(年)

- 研究グループ等で利用するコース

	利用負担金 (年額,税込)	利用可能 ノード数	割当トークン量(年間 及び消費係数	ディスク量 /work
一般 基本セット	大学 公共機関: 100,00円 企業: 120,00円	最大256ノード	8,640 ノード 時間 (1ノード × 360 相当) 消費係数 : 1.00	4TB (グループ当り)
ノード 固定	大学 公共機関: 150,00円 企業: 180,00円	最大256ノード	8,640 ノード 時間 (1ノード × 360 相当) 消費係数 : 1.00	4TB (グループ当り)

※ 申込ノード数に関わらず、利用可能ノード数（256）までのジョブが実行可能です。

Oakbridge-CX

パーソナルコース負担金

一般(大学・公共機関)のみ

- 研究者が個人単位で使用
- 下記の基本セットを元に最大3口まで申込可能

コース	利用負担金 (年額, 税込)	利用可能 ノード数	トークン	ディスク量 /work
パーソナルコース 申込1口あたり)	100,000円/年	最大256ノード	8,640 ノード時間 (1ノード×360日相当) 消費係数 : 1.00	4TB

※ 申込口数に関わらず、利用可能ノード数（256）までのジョブが実行可能です。

バッチジョブキュー: OBCX

- パーソナルユーザ, グループユーザ(一般)で共通
- **interactive はトークンを消費しない**

代表キュー名	キュー名	最大ノード数	実行制限時間(経過時間)	ノード 当たりメモリ量(GB)
interactive	interactive_n1	1	2 h	168
	interactive_n8	2～ 8	10 min	168
debug	debug	1～ 16	30 min	168
short	short	1～ 8	8 h	168
regular	small	1～ 16	48 h	168
	medium	17～ 64	48 h	
	large	65～ 128	48 h	
	x-large	129～ 256	24 h	

Wisteria/BDEC-01のトークン(token)

- 各サブシステムによってトークンの消費係数が異なります
 - Odyssey (A64FX CPU) 1.0
 - Aquarius (CPU + GPU × 8)
 - GPU当たり 3.0
 - ノード全体 24.0
- 通常・一般(大学・公共機関)利用ではWisteriaに申し込めば, Odyssey, Aquariusのいずれも使うことができます
 - 公募型(企業, HPCI, JHPCN等), ノード固定: 不可
 - 但し若手・女性では可能

Wisteria/BDEC-01 : Aquariusノード固定・GPU専有

■ 申込み

- ・ノード固定: **1ノード** (審査あり)
- ・GPU専有: **1GPU** (審査なし)

■ 利用

- ・申込み分のノードを占有
 - 全体の15%程度が上限 (6ノード)
- ・ **専用キュー: 他のユーザとは別**
- ・占有外の計算機資源の利用も可能: その分トークンは減少
 - トークンがなくなった時点で占有ノードも使えなくなります
(資源に余裕があればトークン追加購入可能)

■ 設定のカスタマイズが可能

- ・個別のログインノード・ストレージ(データ隔離)等柔軟に対応

Wisteria/BDEC-01: 一般利用負担金(年) (大学・公共機関)

- Wisteria以降はパーソナルコース廃止

	利用負担金 (年額、税込)	利用可能 ノード数	割当トークン量(年間) 及び消費係数	ディスク量 /lustre
一般 Wisteria/BDEC-01 基セト	大学公共機関: 60,000円	Odyssey: 2,304 ノード Aquarius: 8ノード	8,640トークン (Odyssey 1ノード × 360日相当) Wisteria-O消費係数: 1.00 ※優先利用ノード群 (全体の15程度)は 消費係数 1.50 Wisteria-A消費係数: 3.00 (GPU対、ノード当対は 24.0)	2TB (グループ当り)

Wisteria : 企業利用負担金(年)

	利用負担金 (年額、税込)	利用可能 ノード数	割当トークン量(年間) 及び消費係数	ディスク量 /lustre
Wisteria-O 基本セット (1ノード当り)	企業: 72,000円	Wisteria-O : 最大2,304ノード	8,640トークン (1ノード×360日相当) Wisteria-O消費係数 : 1.00 ※優先利用メド群 全体の15%程度は 消費係数が1.50	2TB (グループ当り)
Wisteria-A 基本セット (1ノード当り)	企業: 216,000円	Wisteria-A : 最大8ノード	25,920トークン (1GPU×360日相当) Wisteria-A消費係数 : 3.00 (GPU当たり、ノード当りは 24.0)	6TB (グループ当り)

Wisteria: Aquariusノード固定・ GPU専有負担金(年)

	利用負担金 (年額、税込)	利用可能 ノード数	割当トークン量(年間) 及び消費係数	ディスク量 /lustre
GP専有 申込1GP当 たり	大学 公共機関: 270,000円 企業: 324,000円	1 GPU	25,920トークン (1 GPU × 360日 相当)	6TB (グループ 当り)
ノード固定 申込 1ノード当り	大学 公共機関: 2,160,000円 企業: 2,592,000円	1ノード	207,360 ノード時間 (1ノード × 360日 相当)	48TB (グループ 当り)

トークン・ディスク追加

<https://www.cc.u-tokyo.ac.jp/supercomputer/wisteria/service/course.php>

	利用負担金(年 額、税込)	備考
トークン追加	5,00円 (720 トークン) 企 業 : 6,00円)	
ディスク追加	6,48円 (1TB)	1T単 位 申 込 可

注: 企業利用では、追加申請には審査が必要です。

Wisteria/BDEC-01のジョブキュー

<https://www.cc.u-tokyo.ac.jp/supercomputer/wisteria/service/job.php>

- インタラクティブ
- バッチジョブ
- プリポスト
- 現状ではOdyssey, Aquariusは同時利用はできない
- Wisteria-O (Odyssey): シミュレーションノード群
 - XXX-o
 - priority-o: 優先キュー, トークン消費量1.5倍
- Wisteria-A (Aquarius): データ・学習ノード群
 - XXX-a ノード単位
 - share-XXX GPU単位
 - MIG (Multi-Instance GPU) には , GPU内を更に分割可能だが本 々 テムは 採用せず

インタラクティブジョブキュー(トークン消費無し)

<https://www.cc.u-tokyo.ac.jp/supercomputer/wisteria/service/job.php>

- Wisteria-O (Odyssey): シミュレーションノード群

キュー名	ノード数	制限時間(Elapsed)		メモリ容量 (GB)
		試験運用時	正式運用後	
interactive-o				
(interactive-o_n1)	1	30分	30分	28
(interactive-o_n12)	2 ~ 12	10分	10分	28

- Wisteria-A (Aquarius): データ・学習ノード群

キュー名	ノード数	制限時間(Elapsed)		メモリ容量 (GB)
		試験運用時	正式運用後	
interactive-a	1ノード	10分	10分	448
share-interactive	1GPU	10分	10分	56

バッチジョブキュー(Wisteria-Odyssey)

<https://www.cc.u-tokyo.ac.jp/supercomputer/wisteria/service/job.php>

キュー名	ノード数	制限時間(Elapse)		メモリ容量 (GB)
		試験運用	正式運用	
debug-o	1～ 144	30 分	30 分	28
short-o	1～ 72	4時間	8時間	28
regular-o	1～ 144	12時間	48時間	28
(small-o)	145～ 576	//	//	//
(medium-o)	577～ 1152	//	//	//
(large-o)	1153～	6時間	24時間	//
(x-large-o)	2304			
priority-o 優先実行ト ーク消費係数 1.5	1～ 288	12時間	48時間	28

バッチジョブキュー(Wisteria-Aquarius)

<https://www.cc.u-tokyo.ac.jp/supercomputer/wisteria/service/job.php>

キュー名	ノード数 GPU数	制限時間(Elapsed)		メモリ容量 (GB)
		試験運用時	正式運用後	
debug-a	1 ノード	30分	30分	448
short-a	1～ 2 ノード	1 時間	2 時間	448
regular-a	1～ 2 ノード 3～ 4 ノード 5～ 8 ノード	12 時間	48 時間	448
(small-a)		//	//	//
(medium-a)		//	//	//
(large-a)		6 時間	24 時間	//
share-debug	1, 2, 4 GPU	30分	30分	56
share-short	1, 2, 4 GPU	1 時間	2 時間	56
share	1 GPU 2 GPU 4 GPU	12 時間	48 時間	56
(share-1)		//	//	//
(share-2)		//	//	//
(share-4)		6 時間	24 時間	//

プリポストキュー

ログインノードと同じアーキテクチャのノードを使用

<https://www.cc.u-tokyo.ac.jp/supercomputer/wisteria/service/job.php>

キュー名	ノード数	制限時間(Elapsed)		メモリ容量 (GB)
		試験運用時	正式運用後	
prepost 予約無し)	1	3 時間	6 時間	340
prepost1_n1 ～ prepost4_n1	1	1～ 3 時間	1～ 6 時間	340
prepost1_n4	1～ 4	1～ 3 時間	1～ 6 時間	340
prepost1_n8	1～ 8	1～ 3 時間	1～ 6 時間	340

- Wisteria/BDEC-01: 利用例
 - h3-Open-BDEC
 - 予備的計算事例
- 今後の動向等
 - 「共通ストレージ」について

(計算+データ+学習)融合によるエクサスケール時代の革新的シミュレーション手法

- エクサスケール(富岳+クラス)のスパコンによる科学的発見の持続的促進のため、計算科学にデータ科学、機械学習のアイディアを導入した(計算+データ+学習(S+D+L))融合による革新的シミュレーション手法を提案
 - (計算+データ+学習)融合によるエクサスケール時代の革新的シミュレーション手法(科研費基盤S, 代表: 中島研吾(東大情基セ), 2019年度~2023年度)
- 革新的ソフトウェア基盤「h3-Open-BDEC」の開発: 東大BDECシステム(Wisteria/BDEC-01), 「富岳」等を「S+D+L」融合プラットフォームと位置づけ、スパコンの能力を最大限引き出し、最小の計算量・消費電力での計算実行を実現するために、下記2項目を中心に研究
 - 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
 - 階層型データ駆動アプローチ(hDDA: Hierarchical Data Driven Approach)等に基づく革新的機械学習手法
 - Hierarchical, Hybrid, Heterogeneous $\Rightarrow$ h3



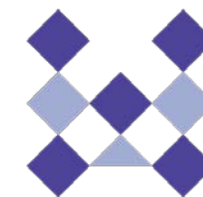
**Wisteria
BDEC-01**



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

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



Wisteria
BDEC-01

h3-Open-BDEC

新しい計算原理
数値アルゴリズム・ライブラリ

シミュレーション＋データ
＋学習 (S+D+L)
アプリ開発フレームワーク

統合＋通信＋
ユーティリティ
制御 & ユーティリティ

h3-Open-MATH
高性能・高信頼性・
混合/変動精度アルゴリズム

h3-Open-APP:
Simulation
計算科学アプリケーション

h3-Open-SYS
制御 & 統合

h3-Open-VER
精度保証

h3-Open-DATA: Data
データ科学

h3-Open-UTIL
大規模計算向け
ユーティリティ群

h3-Open-AT
自動チューニング

h3-Open-DDA:
Learning
データ駆動・機械学習

**Hierarchical,
Hybrid,
Heterogeneous**

**Big Data &
Extreme
Computing**



計算科学シミュレーション

- 非線形: 多数のパラメータスタディ必要

✓ ケース数削減が非常に重要

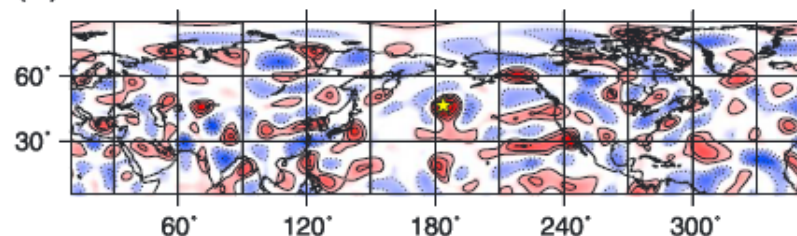
[Miyoshi et al. 2014]

- 天気予報: データ同化

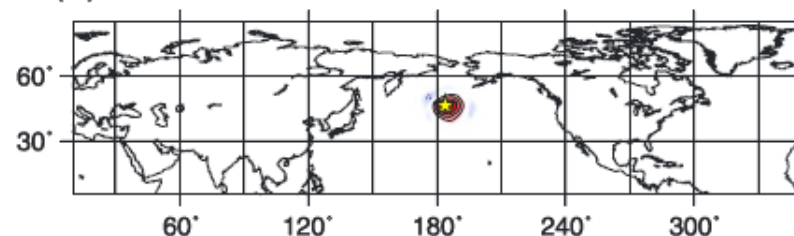
✓ 中期予報(半月程度): 50-100程度のアサンブル実行, 精度良い解のためには1,000程度必要(時間, 計算資源の制限)

✓ 機械学習等でパラメータを正確に予測できれば, 50-100(あるいはそれ以下)で十分かも知れない

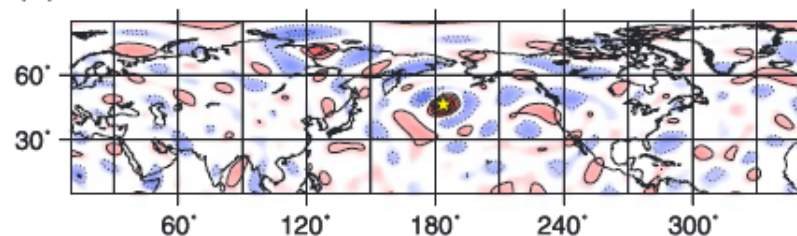
(a) 20 members w/o localization



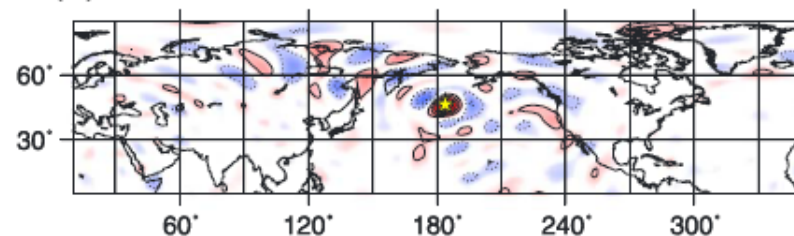
(b) 20 members w/ 700-km localization



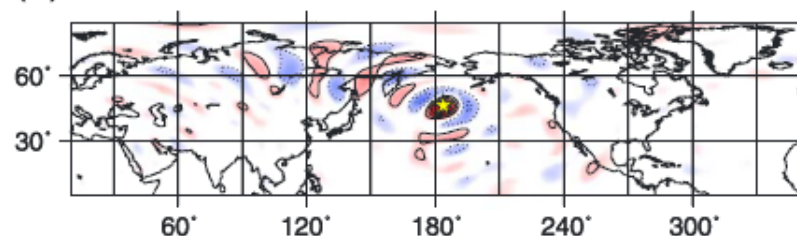
(c) 80 members w/o localization



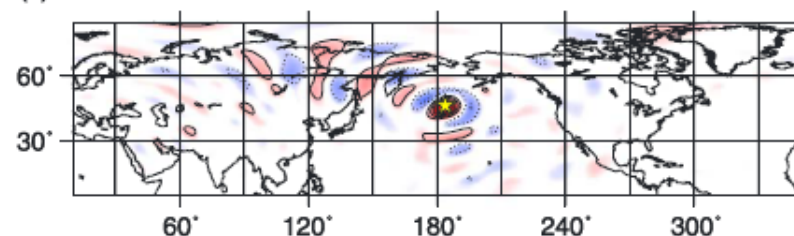
(d) 320 members w/o localization



(e) 1280 members w/o localization

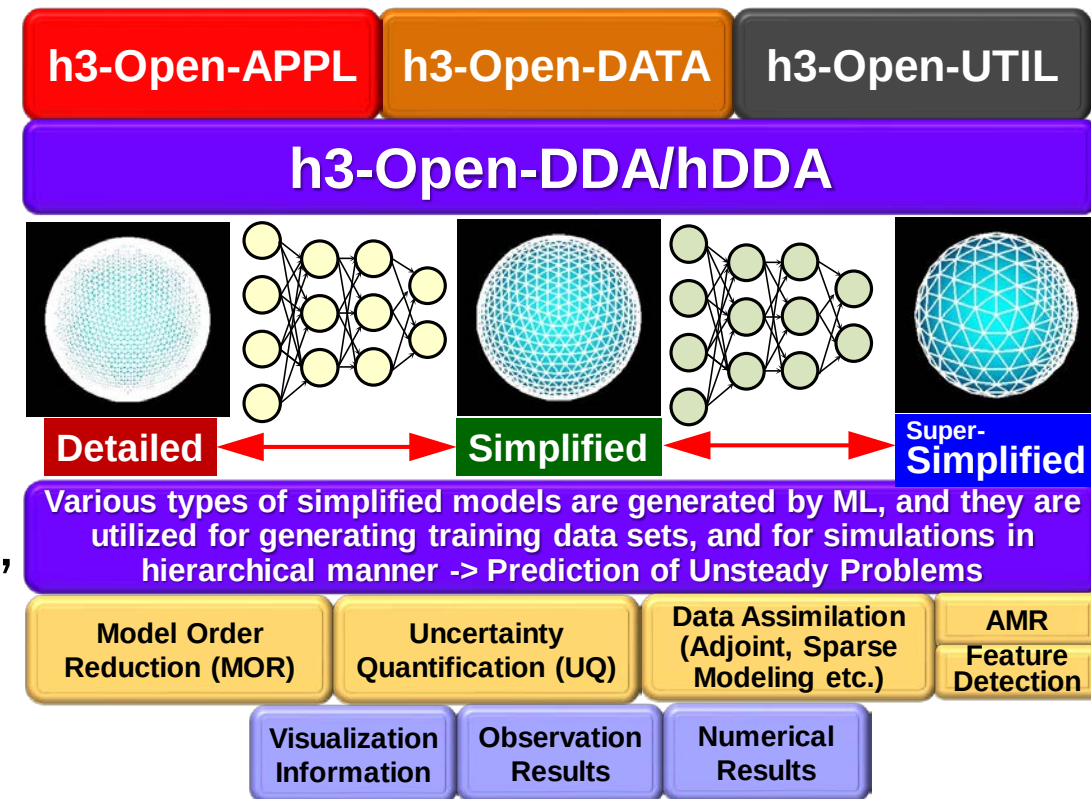
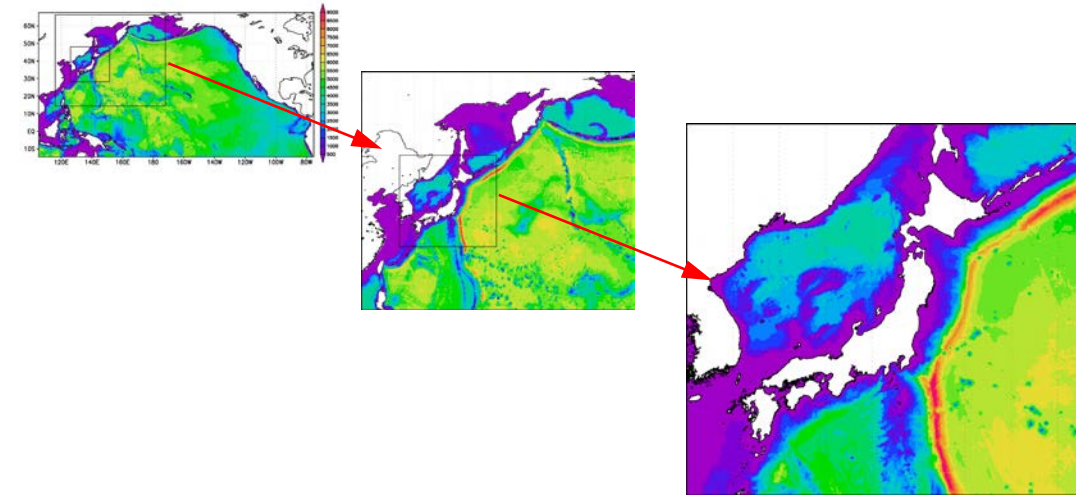


(f) 10240 members w/o localization

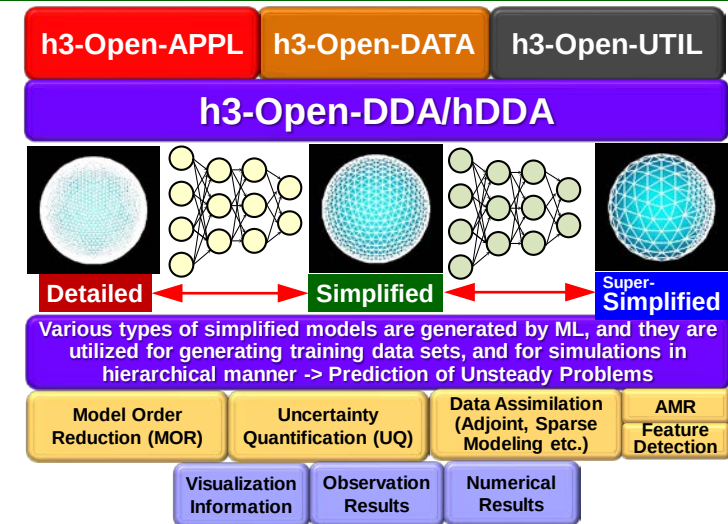
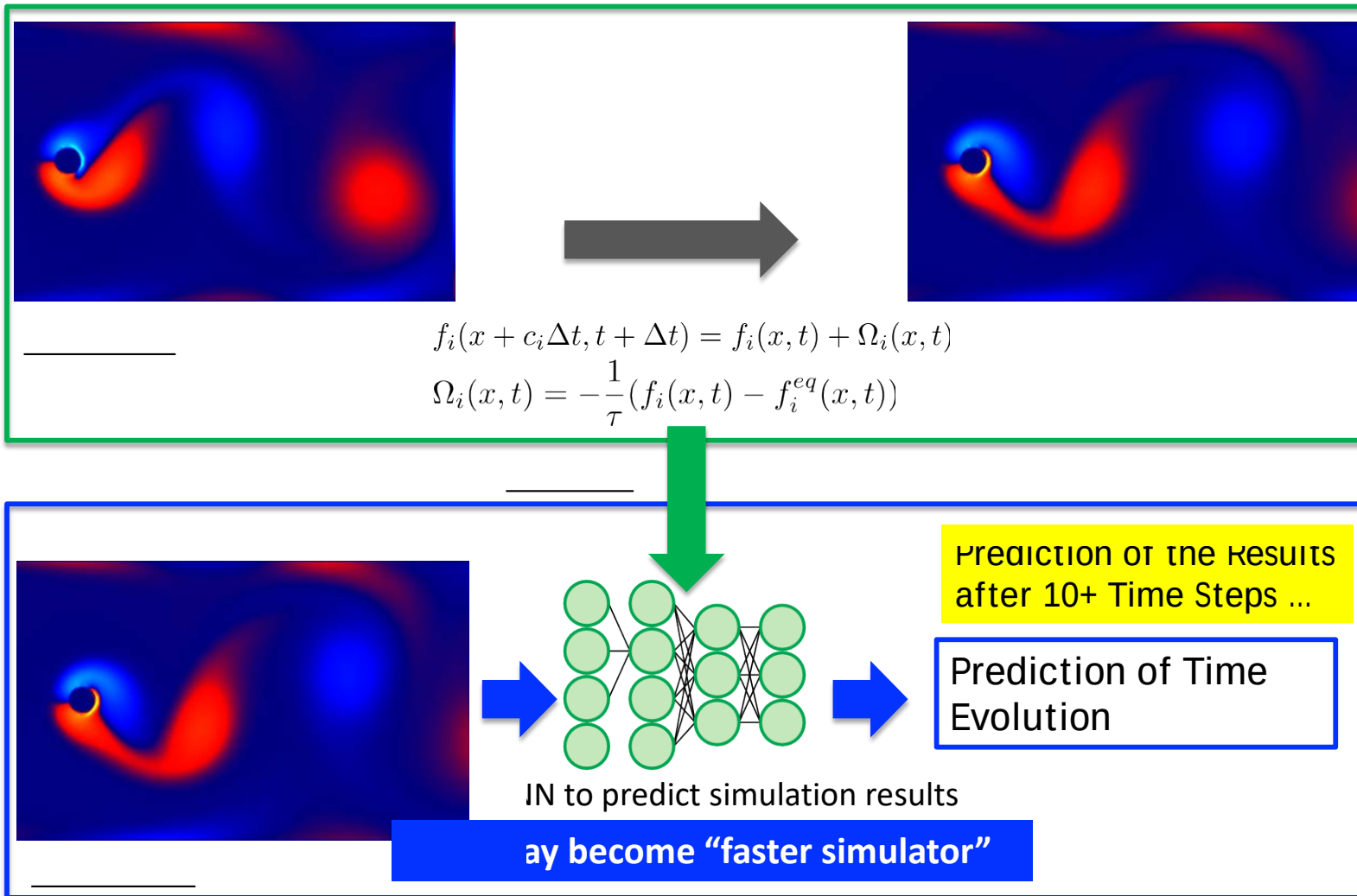


階層型データ駆動アプローチ： hDDA

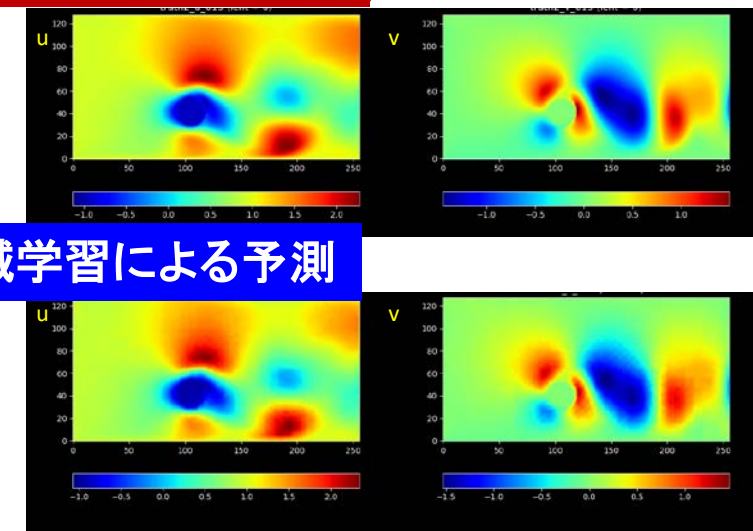
- シミュレーションに機械学習を適用して異なるパラメータでの解を予測するデータ駆動アプローチ (DDA, Data Driven Approach) では, 計算を繰り返して教師データを生成する必要がある。
- 階層型DDA (*hDDA*) は, 特徴検知, MOR (Model Order Reduction), UQ (Uncertainty Quantification), スパースモデリング, 適応格子等の諸機能を駆使して, 計算量(メッシュ数, 粒子数)を削減した簡易モデルを, 機械学習により自動生成, 教師データ生成用モデルとして利用する



機械学習による非定常数値流体力学シミュレーション高速化 (計算+データ+学習(S+D+L))融合の実例



シミュレーション結果



機械学習による予測

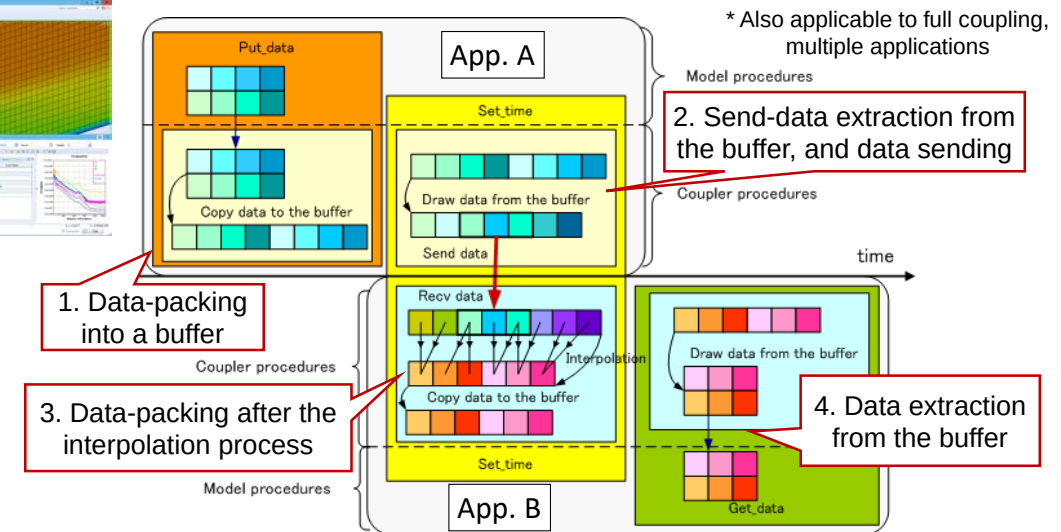
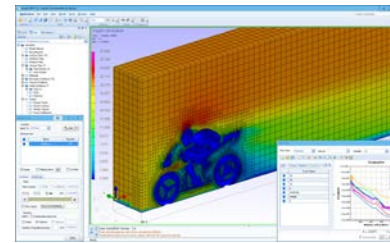
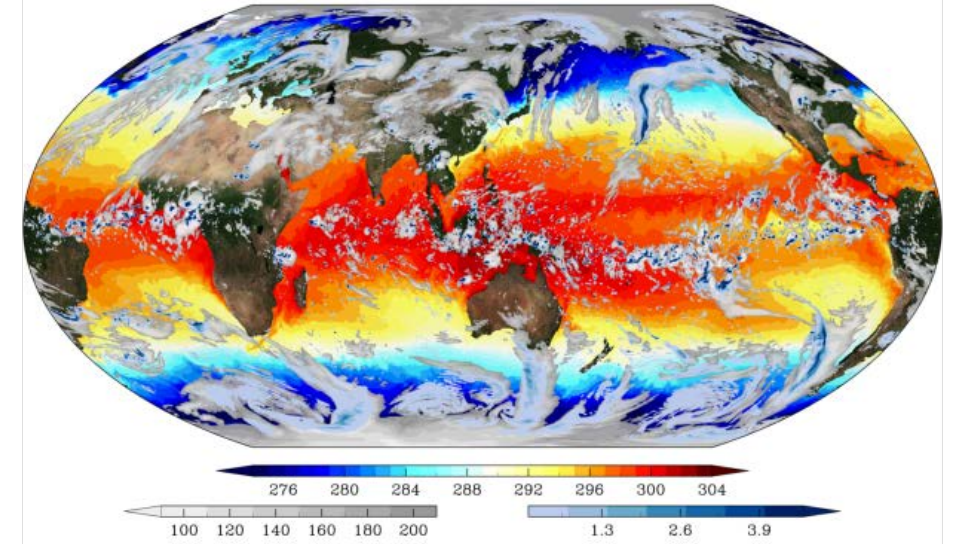
期待される成果と意義



- 計算科学の専門家のみで(S+D+L)融合を容易に実現
 - 機械学習の専門家のサポートを必要としない
- ソースコード, マニュアル類も含めて一般に公開, 様々なエクサスケールシステムでの普及を目指す
 - ポスト富岳も含めたポストムーア時代への展開
- h3-Open-BDEC利用による(S+D+L)融合シミュレーションにより従来手法と同等の正確さを保ちつつ, 大幅な計算量・消費電力削減を目指す(10分の1が目標)。
- シミュレーション高度化: パラメータスタディのケース数を削減できる
- リアルタイム災害シミュレーション等への適用

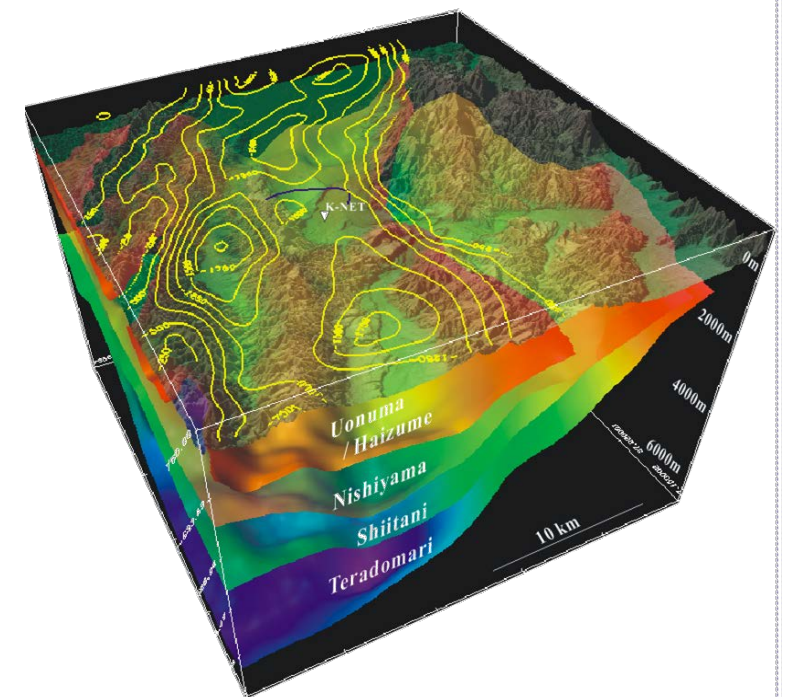
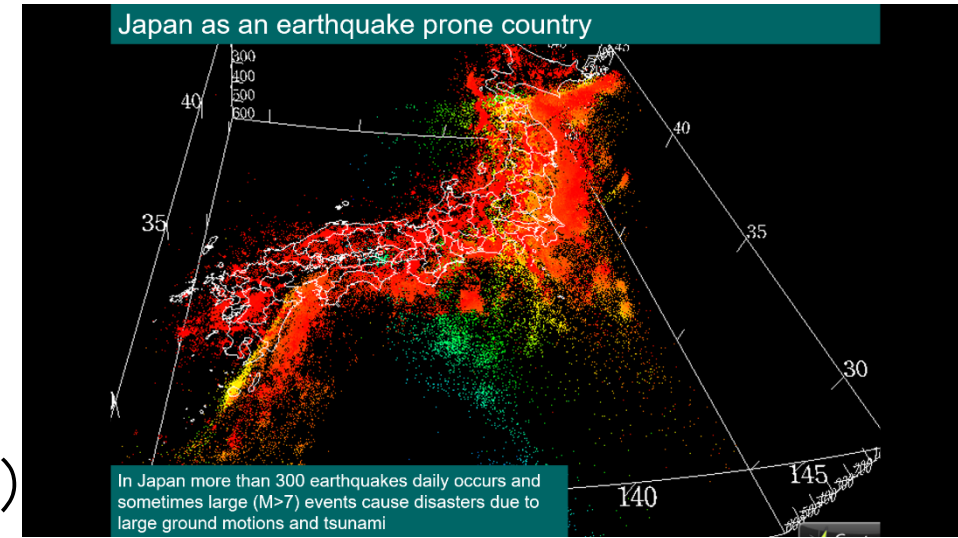
Wisteria/BDEC-01上における h3-Open-BDECを使用した(S+D+L)融合

- シミュレーションとデータ同化の融合
 - 典型的・伝統的な(S+D+L)融合
- 気候・気象のための大気海洋連成シミュレーション
 - 東大大気海洋研, 理研, 国立環境研他
- **リアルタイム同化＋三次元強震動シミュレーション**
 - **東大地震研**
- リアルタイム災害シミュレーション
 - 洪水, 津波
- 既存シミュレーションコードの(S+D+L)融合による高度化
 - OpenFOAM



地震シミュレーション： 不確実性(Uncertainty)と 隣り合わせ

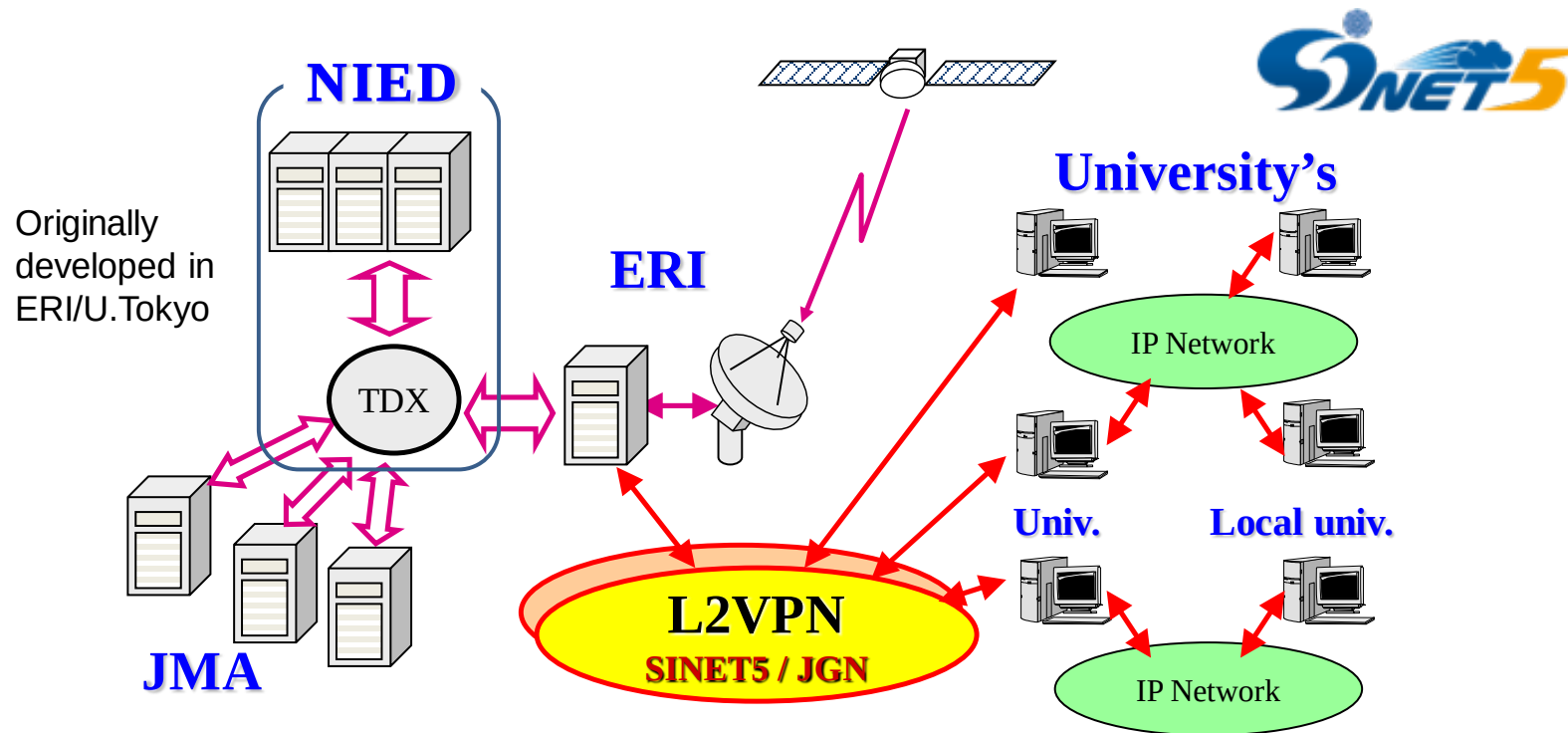
- 地震シミュレーション(強震動シミュレーション)
 - 応力蓄積過程⇒動的破壊⇒地震波動伝播(強震動)
- 地下構造
 - 不均質, 不確定
- シミュレーション・観測融合が不可欠
- 伝統的なシミュレーション
 - いわゆるフォワードモデリング
 - 「メカニズムの理解」の域を出ない
- データ同化・リアルタイム観測と融合した手法の開発が必要
 - シミュレーション: 予測 + 観測・データ同化: 補正



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

全国地震観測データ流通ネットワーク「JDXnet」

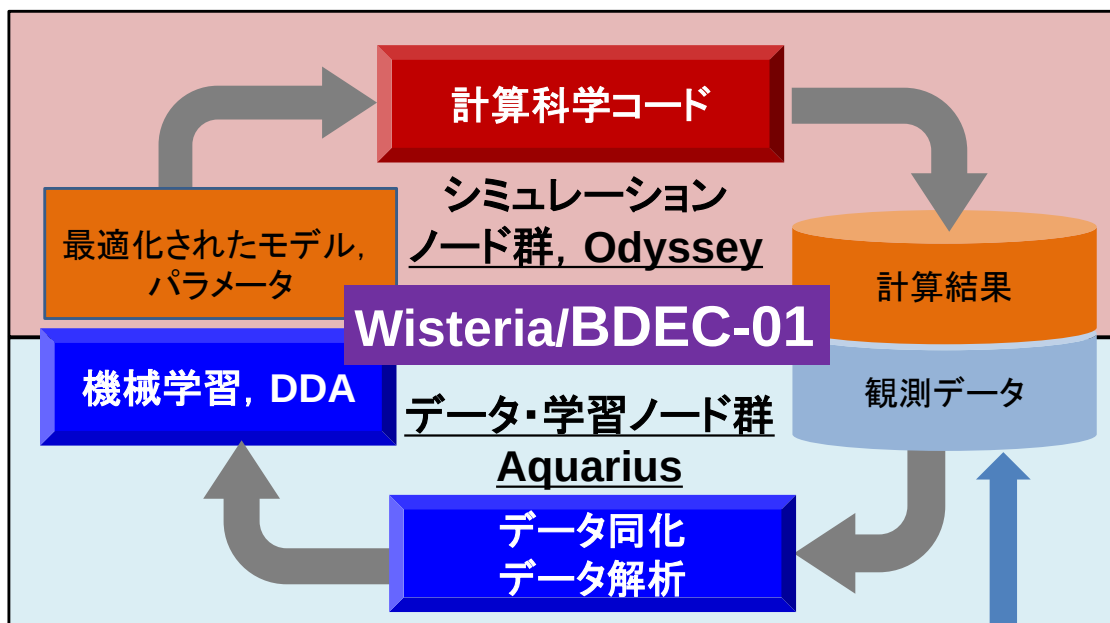
- 国内地震観測点の観測データ(約2,000点, 100Hz, 3方向) をSINET経由でリアルタイムに取得可能
 - 気象庁, 東大地震研, 防災科技研, 各大学
 - 1日のデータ量: 100GB級



〔資料提供: 鶴岡弘准教授(東大・地震研)〕

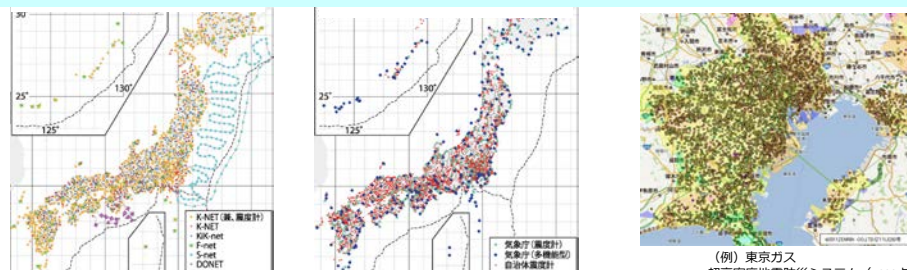
三次元地震シミュレーション＋リアルタイムデータ同化/観測

JDXnetの観測データを利用したリアルタイムデータ同化/観測

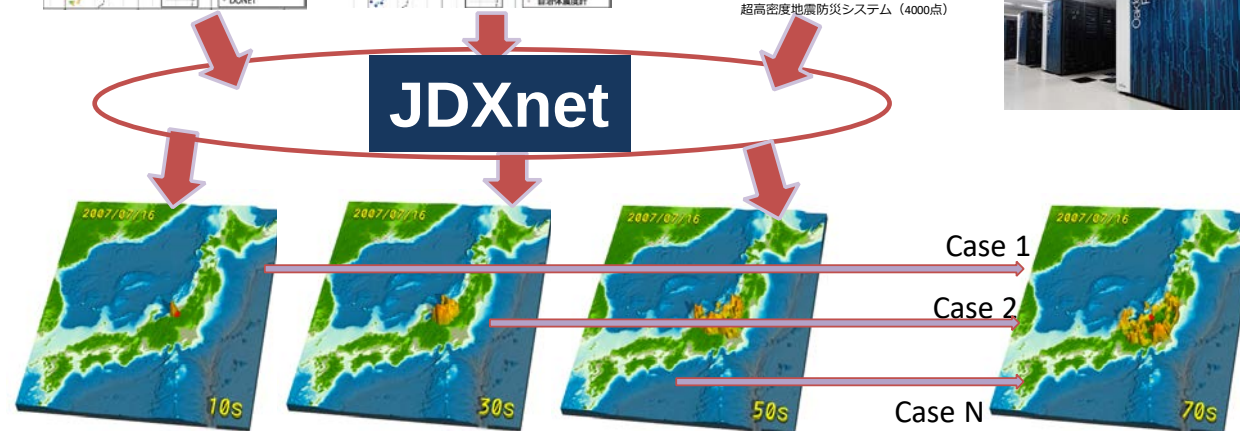
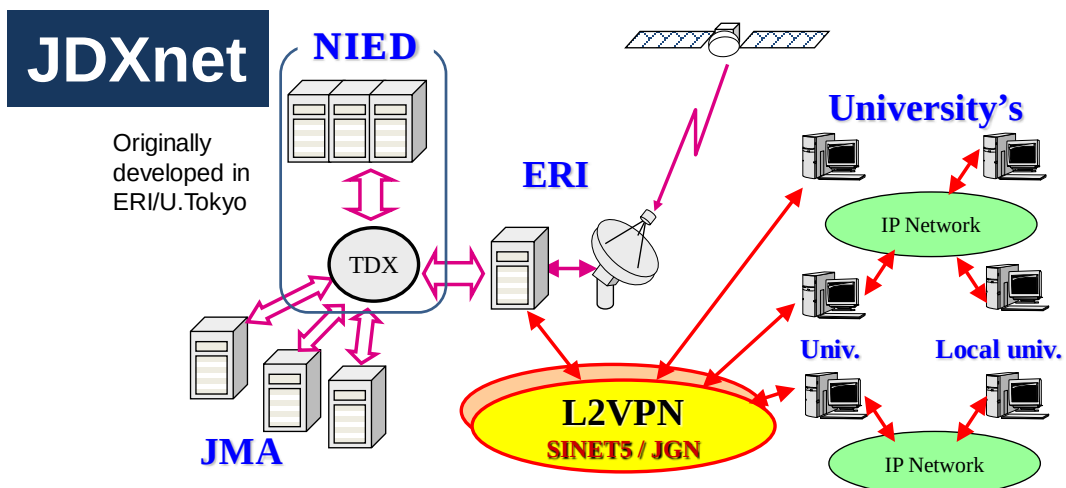


外部リソース
Server, Storage, DB, Sensors他

Observation Network for Earthquake: $O(10^5)$ Points



[c/o Furumura]



Real-Time Data/Simulation Assimilation
Real-Time Update of Underground Model

〔資料提供: 古村孝志教授
(東大・地震研)〕

強震動シミュレーション＋データ同化・観測融合に向けた取り組み(古村孝志教授(東大地震研))

Seism3D/OpenSWPC-DAF (Data-Assimilation-Based Forecast)

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

- Furumura, T., Maeda, T., & Oba, A., Early forecast of long - period ground motions via data assimilation of observed ground motions and wave propagation simulations, Geophysical Research Letters, 46, 138–147, 2019
- Oba, A., Furumura, T., & Maeda, T., Data assimilation - based early forecasting of long - period ground motions for large earthquakes along the Nankai Trough, Journal of Geophysical Research: Solid Earth, 125, 2020

Optimal Interpolation Techniqueによるリアルタイムデータ同化 104

[c/o Oba & Furumura]

- Data Assimilation of Wave Propagation by “Optimal Interpolation Technique”

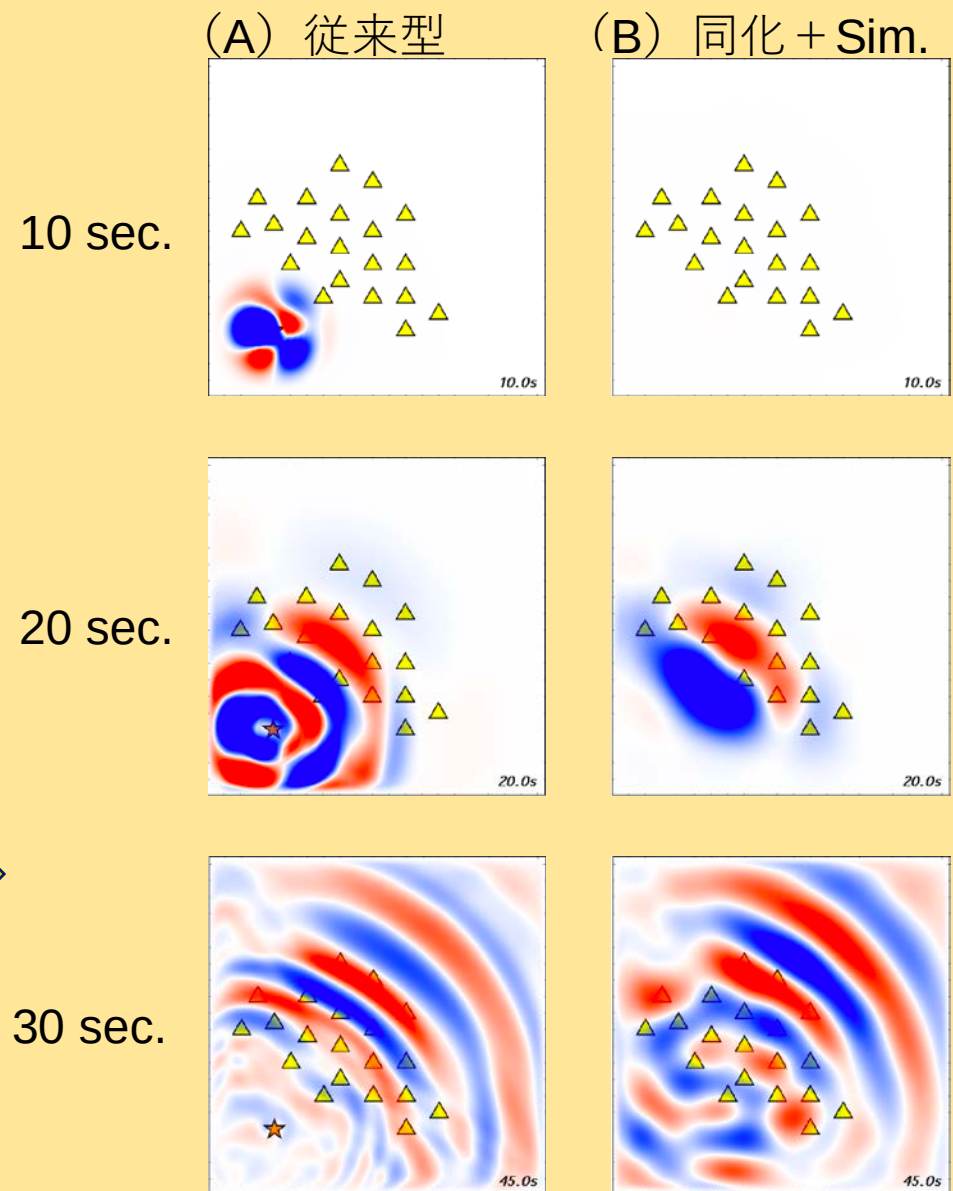
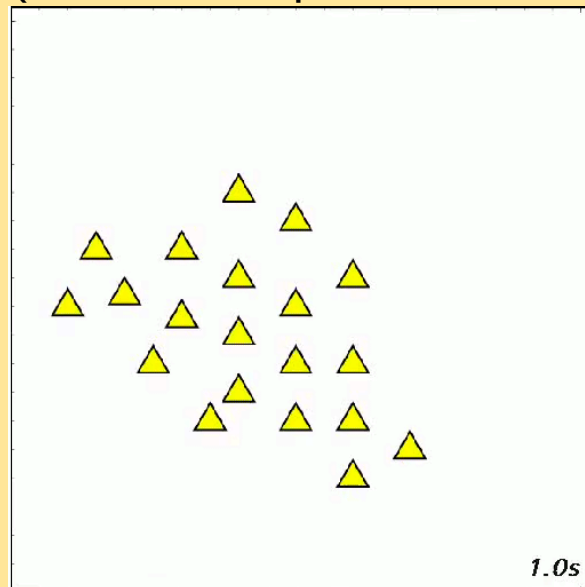
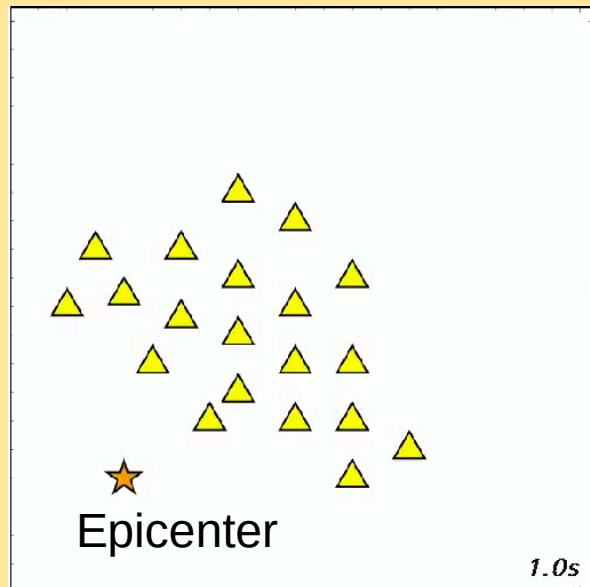
$$\begin{array}{l} \text{Assim.} \quad \text{Comp.} \quad \text{Residual} \quad \text{Comp.} \\ \mathbf{x}_n^a = \mathbf{x}_n^f + \mathbf{W}(\mathbf{y}_n - \mathbf{H}\mathbf{x}_n^f) \\ \text{Comp.} \quad \text{Assim.} \\ \mathbf{x}_{n+1}^f = \mathbf{F}\mathbf{x}_n^a \end{array}$$

n : Time Step
 $\mathbf{W}$: Weighting Matrix
 $\mathbf{F}$: Wave Propagation simulation

(A) 従来型Sim.

(B) 同化 + Sim.

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



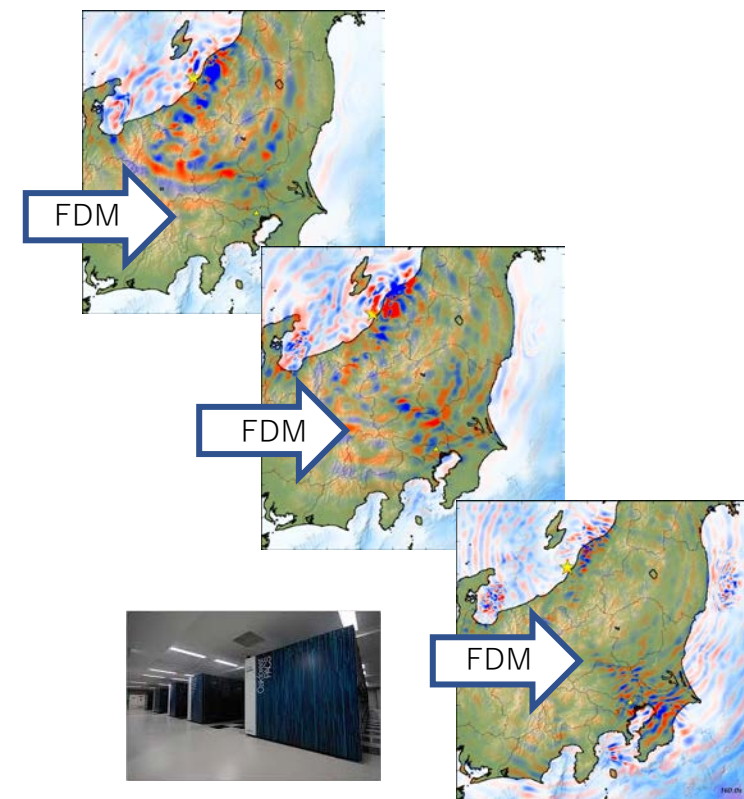
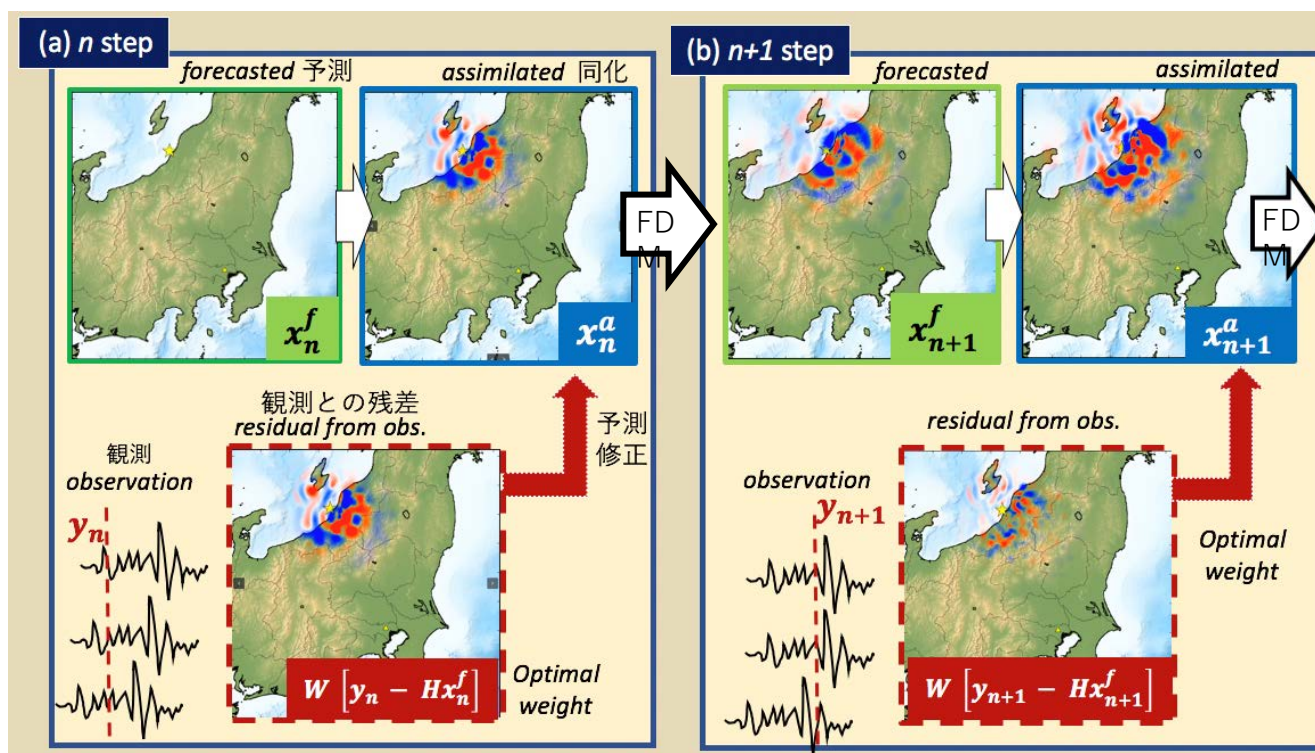
最初は「A:同化+シミュレーション」⇒「B:シミュレーション予測」へ移行

$$\begin{aligned}
 & \text{Assim.} \quad \text{Comp.} \quad \text{Obs.} \quad \text{Residual} \quad \text{Comp.} \\
 & x_n^a = x_n^f + W(y_n - Hx_n^f) \\
 & \text{Comp.} \quad \text{Assim.} \\
 & x_{n+1}^f = Fx_n^a \quad F: \text{Wave Propagation simulation}
 \end{aligned}$$

n : Time Step
 W : Weighting Matrix

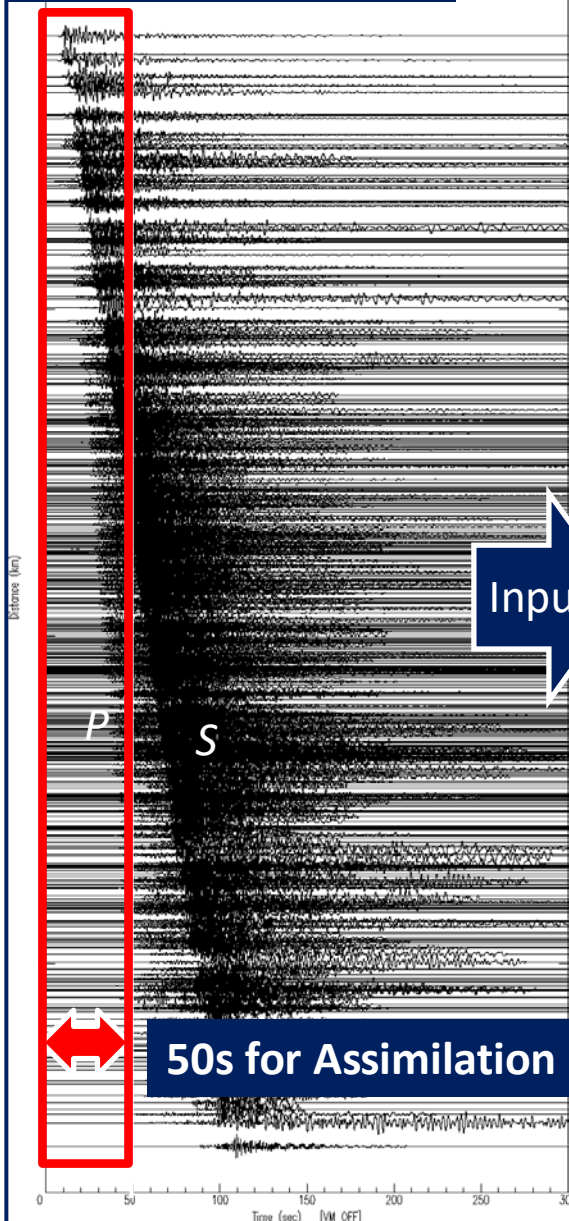
(A) Assimilation+Simulation

(B) Pure Simulation/Forecast

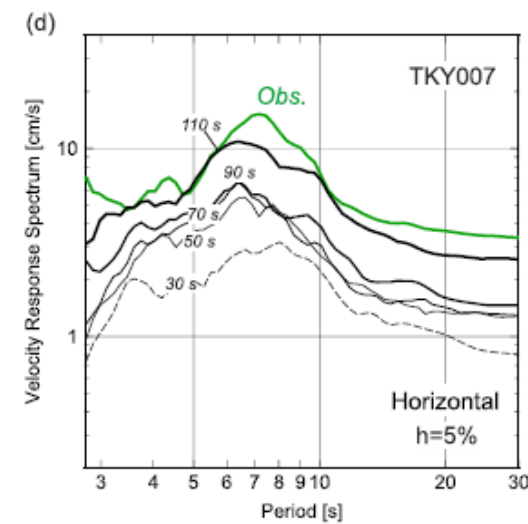
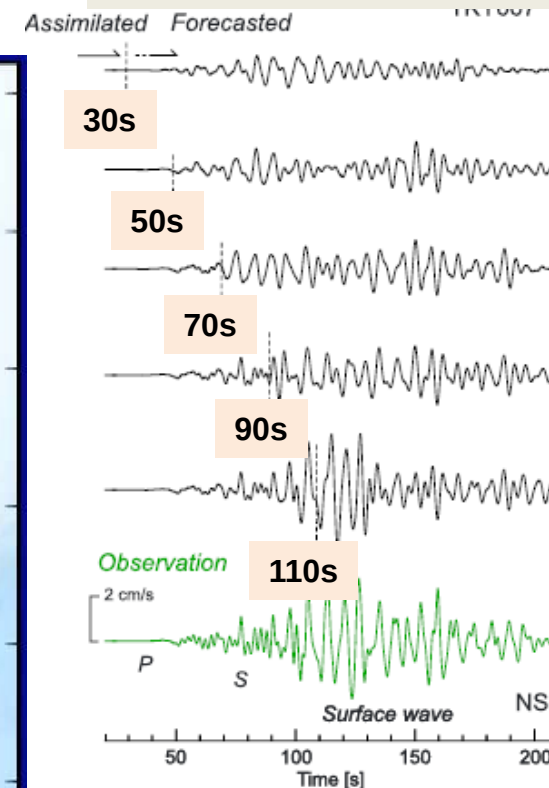


リアルタイム観測・データ同化+シミュレーション

482 K-NET, KiK-net Observation

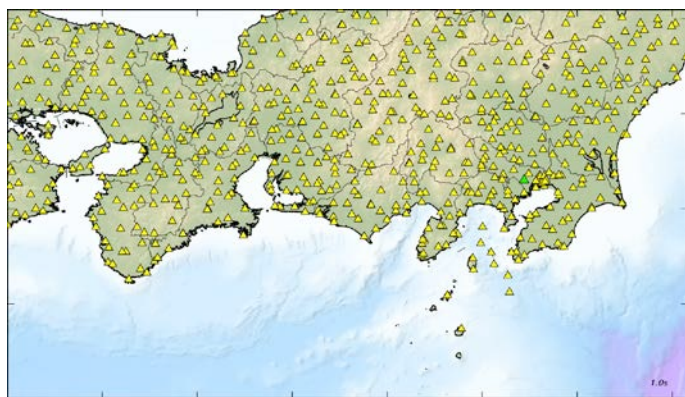
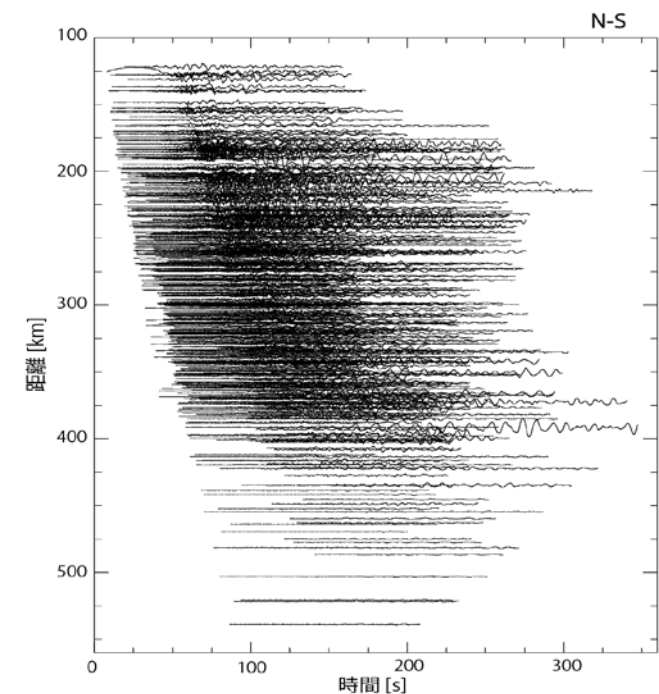


新宿における計測 ▲

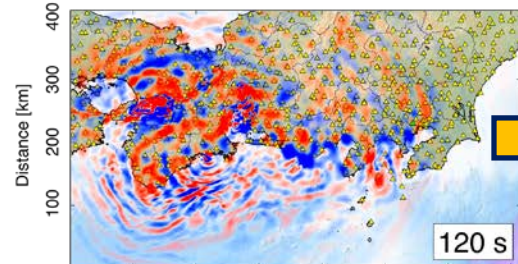
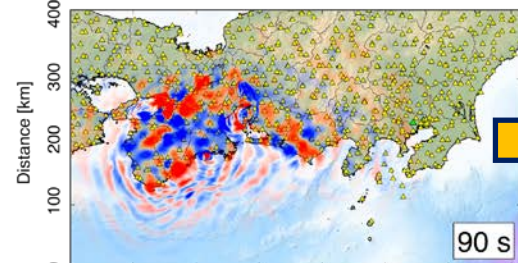
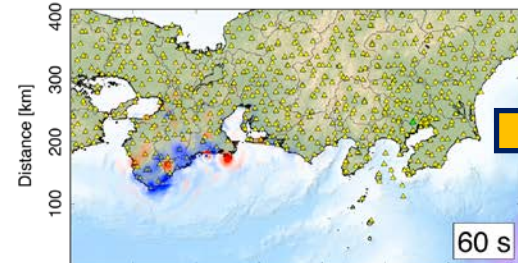


紀伊半島南東沖地震 (2004年, Mw 7.4) [c/o Oba & Furumura]

○ Observation (K-NET, KiK-net 446 pts)

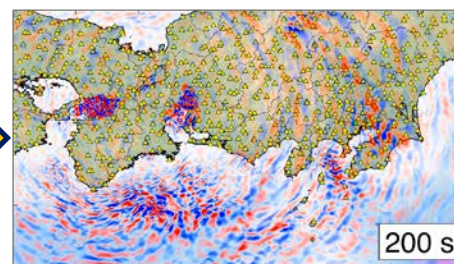
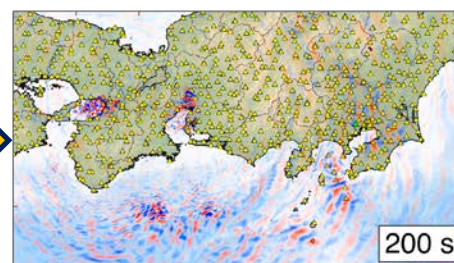


(a) Sim. + Assim.



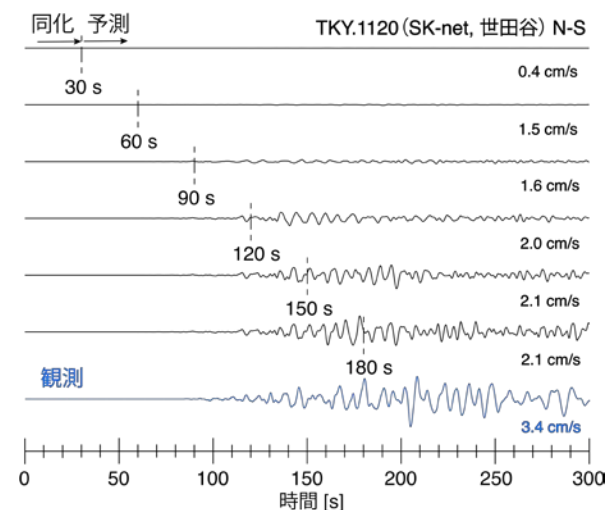
Distance [km]

(b) Pure Simulation

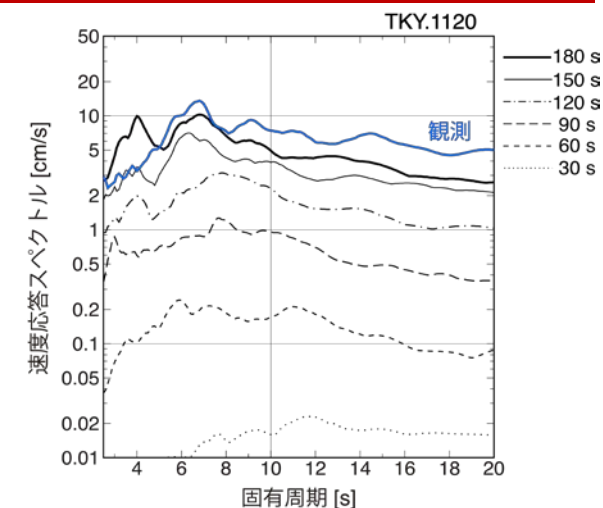


Distance [km]

Long Wave Propagation in Tokyo



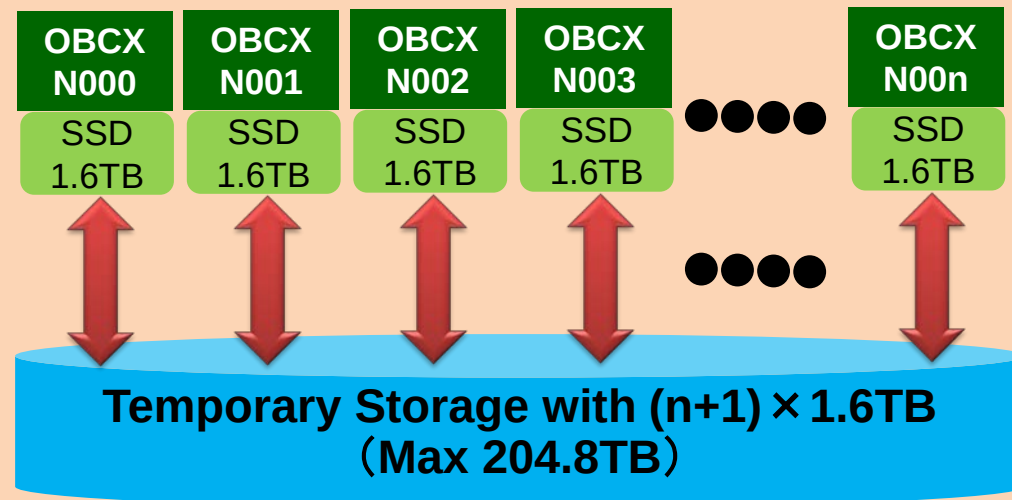
Response Spectrum



Oakbridge-CX (OBCX) : BDECに向けた実験システム

- 全1,368ノードのうち128ノードにSSD (Solid State Drive) 搭載
 - Intel SSD + BeeGFS
 - 容量: 1.6 TB/node
 - 読み書き性能: 3.20/1.32 GB/s/node
 - BeeOND (BeeGFS-on-Demand) によって合計 200+TB (128 × 1.6) の高速ファイルシステムとして使用可能
 - データ科学アプリケーション
 - ソフトウェア類も充実
 - ステージング, チェックポイント
 - 128ノードのうち16ノードはSINET経由で外部リソース(サーバー, ストレージ, センサーネットワーク)に直接接続⇒外部接続ノード

BeeGFS on Demand (BeeOND)



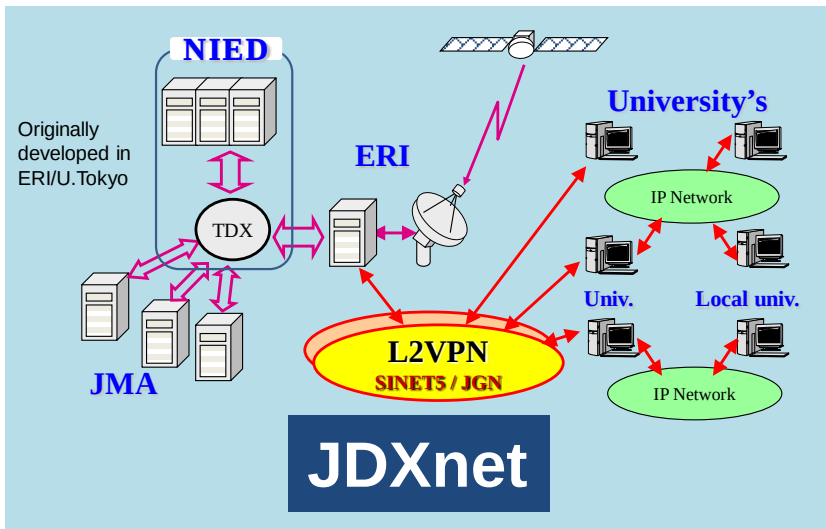
Total: 1,368 nodes

128 nodes
with SSD

16

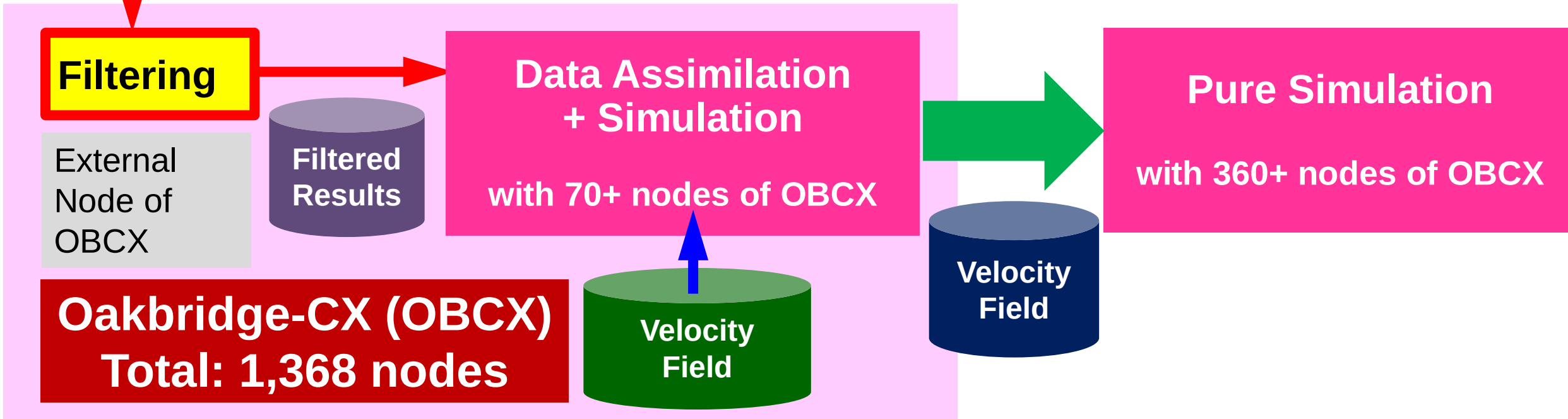
OBCXの16ノード(外部接続ノード)
SINET経由で外部計算機資源に直接接続,
BDECにおけるデータ・学習ノード群と同様の
役割





OBCXによる実験システム

外部接続ノードで観測結果受信⇒フィルタリング
 古村教授の試算によれば60秒先の結果を6秒
 間で予測するのに, OBCX 360ノード必要



The article on my presentation@CSE21 appears in *SIAM News*



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

The screenshot shows a web browser displaying the SIAM News website. The article title is "Supercomputer Simulations of Earthquakes in Real Time" by Jillian Kunze. The article text discusses the integration of simulation, data, and learning in computational science and engineering, mentioning a presentation at the 2021 SIAM Conference on Computational Science and Engineering. A diagram illustrates the Wisteria/BDEC-01 system architecture, showing the flow of data between simulation nodes, file systems, and data/learning nodes. The diagram includes components like Simulation Nodes Odyssey, Fast File System (FFS), Shared File System (SFS), Machine Learning, DDA, Data/Learning Nodes Aquarius, and Data Assimilation Data Analysis. The article also mentions the University of Tokyo's supercomputing center and the upcoming operations of the Wisteria/BDEC-01 system in May 2021.

Simulation Nodes Odyssey
25.9 PF, 7.8 PB/s

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

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

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

Simulation Codes
Simulation Nodes Odyssey

Optimized Models & Parameters

Machine Learning, DDA

Wisteria/BDEC-01

Data/Learning Nodes, Aquarius

Data Assimilation Data Analysis

Results
Observation Data

Server, Storage

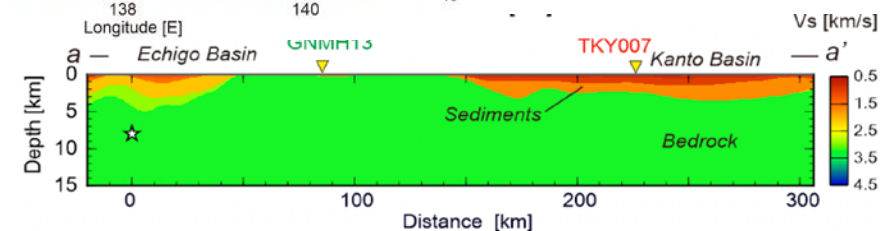
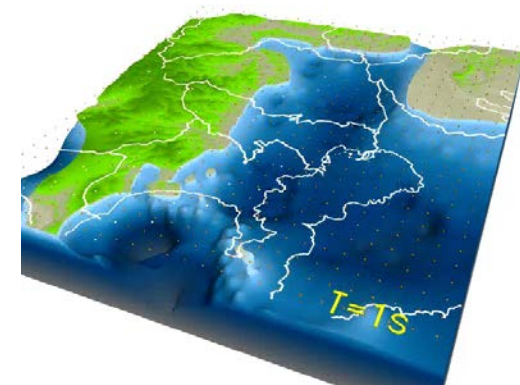
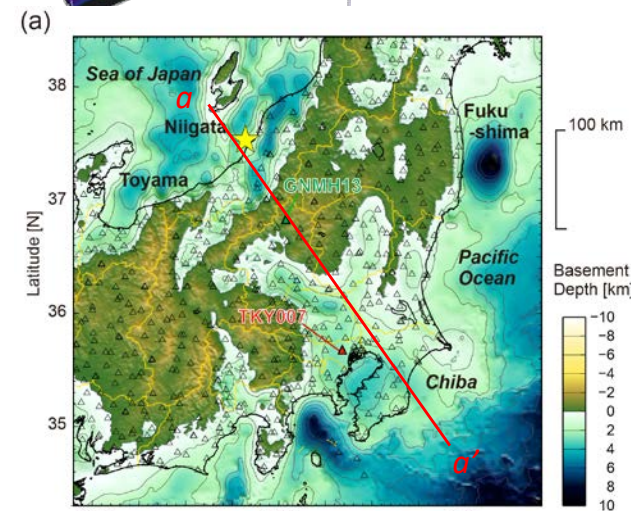
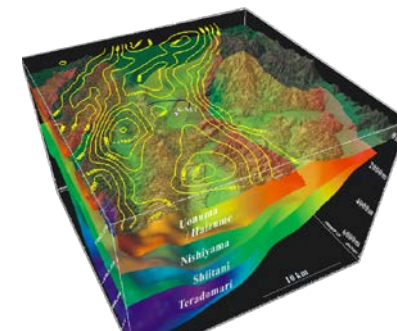
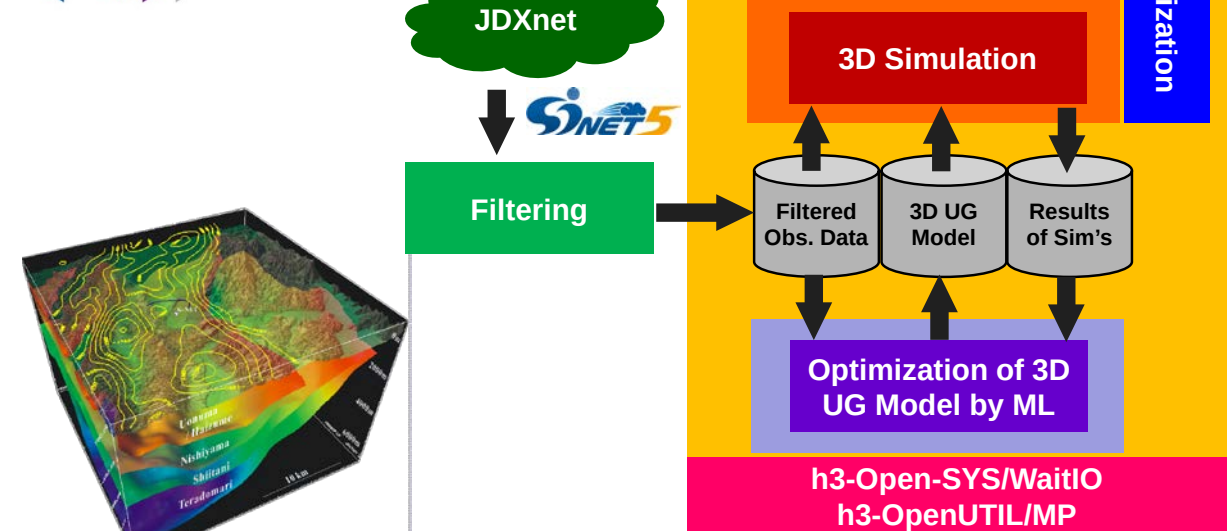
Most Recent

Happening Now
SIAM Unwrapped - March 2021

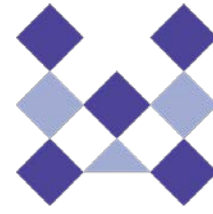
Get Involved
Machine Learning Accelerates

リアルタイムデータ同化＋ 3D強震動Sim. 融合(1/2)

- 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
 - 60秒先の予測を10秒以内に計算
 - 的確な避難計画の策定
- 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

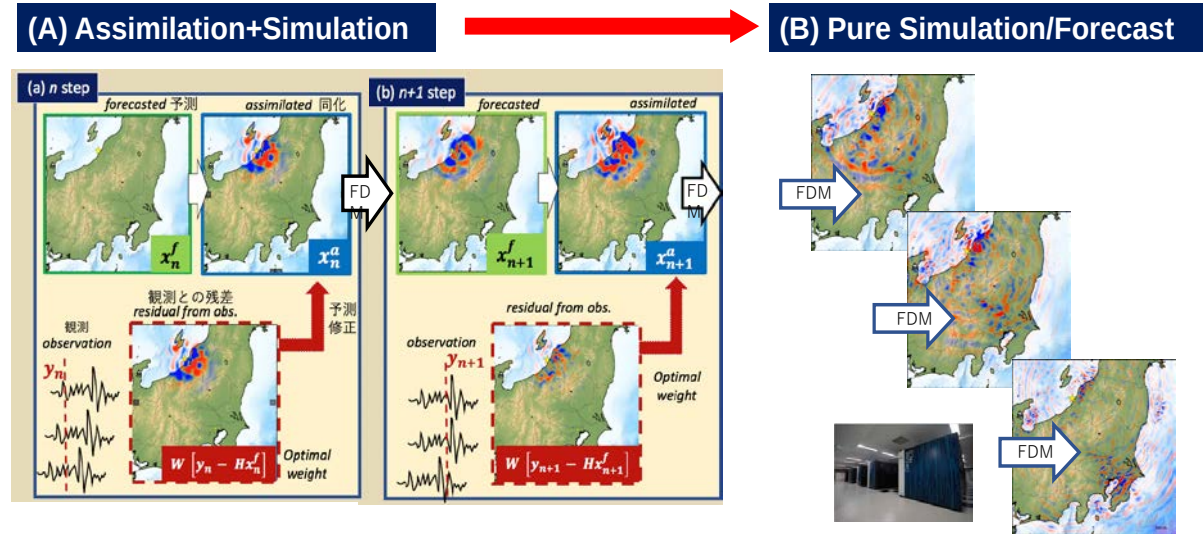


リアルタイムデータ同化＋ 3D強震動Sim 融合 (2/2)

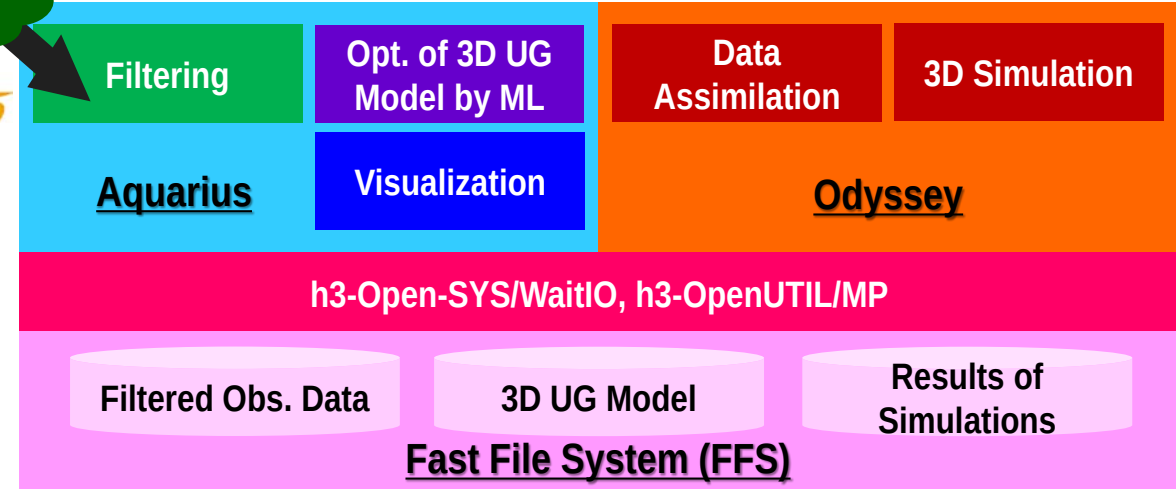


**Wisteria
BDEC-01**

- Wisteria/BDEC-01の利用
- データ同化＋シミュレーション
 - Optimal Interpolation Technique: 高速
- 三次元地下構造モデル高度化
 - リアルタイム性はそれほど重要ではない
 - より高度なデータ同化手法(e.g.四次変分法)を適用
- Odyssey
 - データ同化, シミュレーション
- Aquarius
 - フィルタリング, 機械学習, 可視化



JDXnet



Wisteria/BDEC-01

における計算



**Wisteria
BDEC-01**

- Wisteria/BDEC-01

- Aquarius (IceLake + A100)
 - データ・学習ノード群, 外部通信機能
- Odyssey (A64FX)
 - シミュレーションノード群

- 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
 - IB-EDR経由のライブラリを開発中
- データ受け渡しライブラリ(h3-Open-SYS/WaitIO)及び高機能力プラー(h3-Open-UTIL/MP)活用
- 高速ファイルシステム(FFS)経由の連携も可能

h3-Open-BDEC

新しい計算原理
数値アルゴリズム・ライブラリ

シミュレーション+データ
+学習(S+D+L)
アプリ開発フレームワーク

統合+通信+
ユーティリティ
制御 & ユーティリティ

h3-Open-MATH
高性能・高信頼性・
混合/変動精度アルゴリズム

h3-Open-APP:
Simulation
計算科学アプリケーション

h3-Open-SYS
制御 & 統合

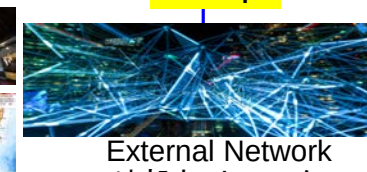
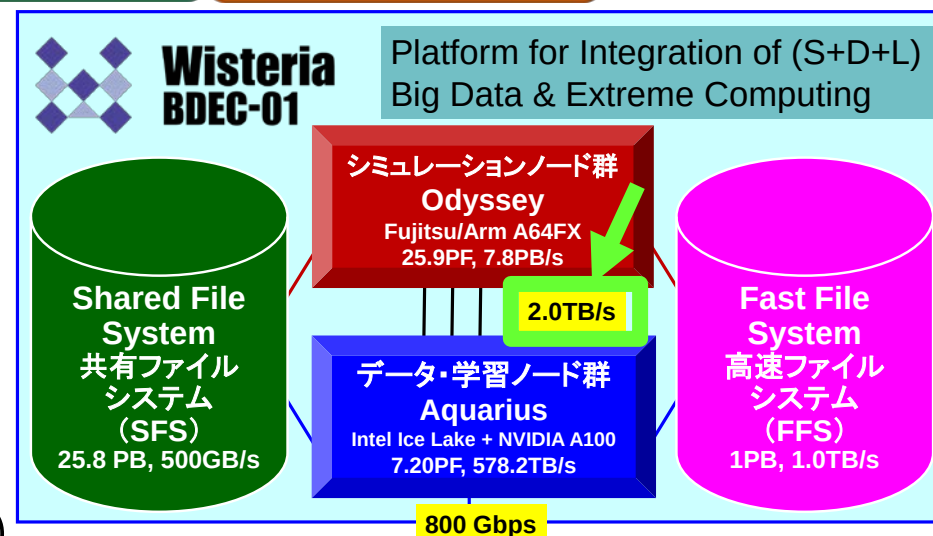
h3-Open-VER
精度保証

h3-Open-DATA: Data
データ科学

h3-Open-UTIL
大規模計算向け
ユーティリティ群

h3-Open-AT
自動チューニング

h3-Open-DDA:
Learning
データ駆動・機械学習



External
Resources

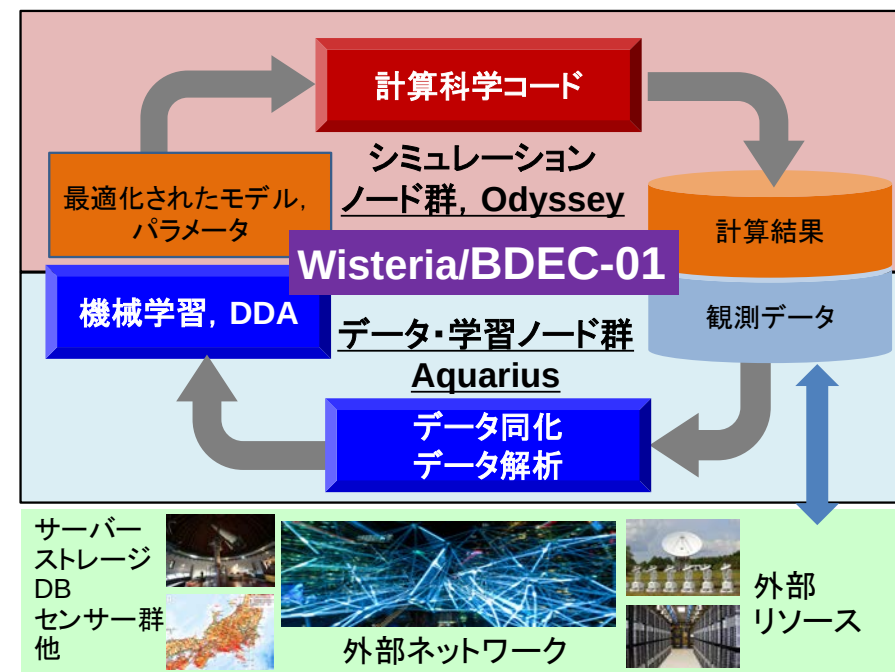
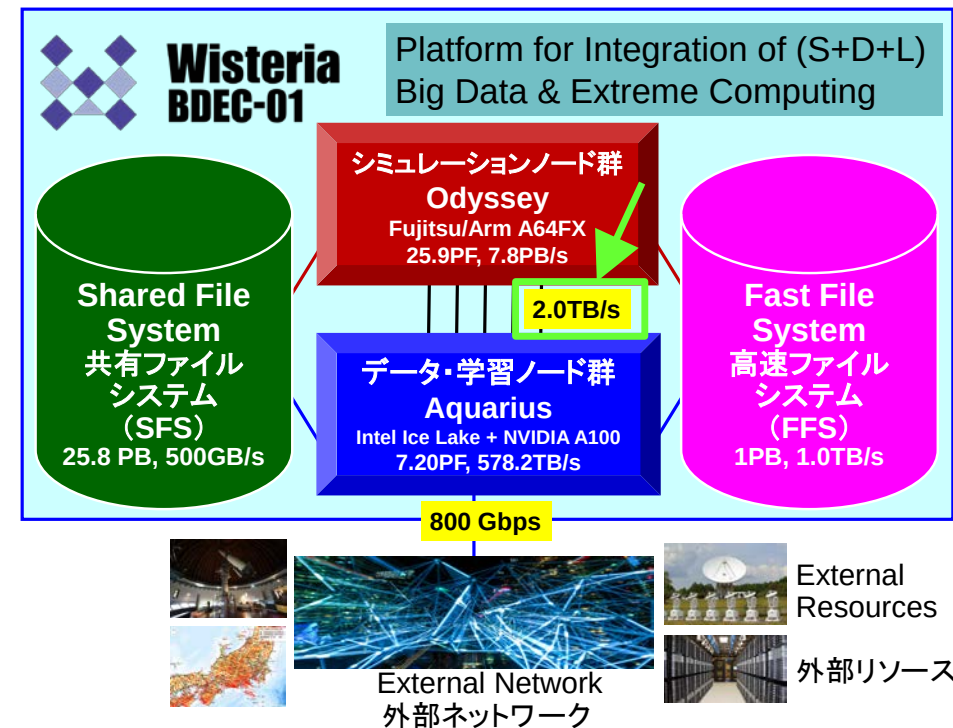
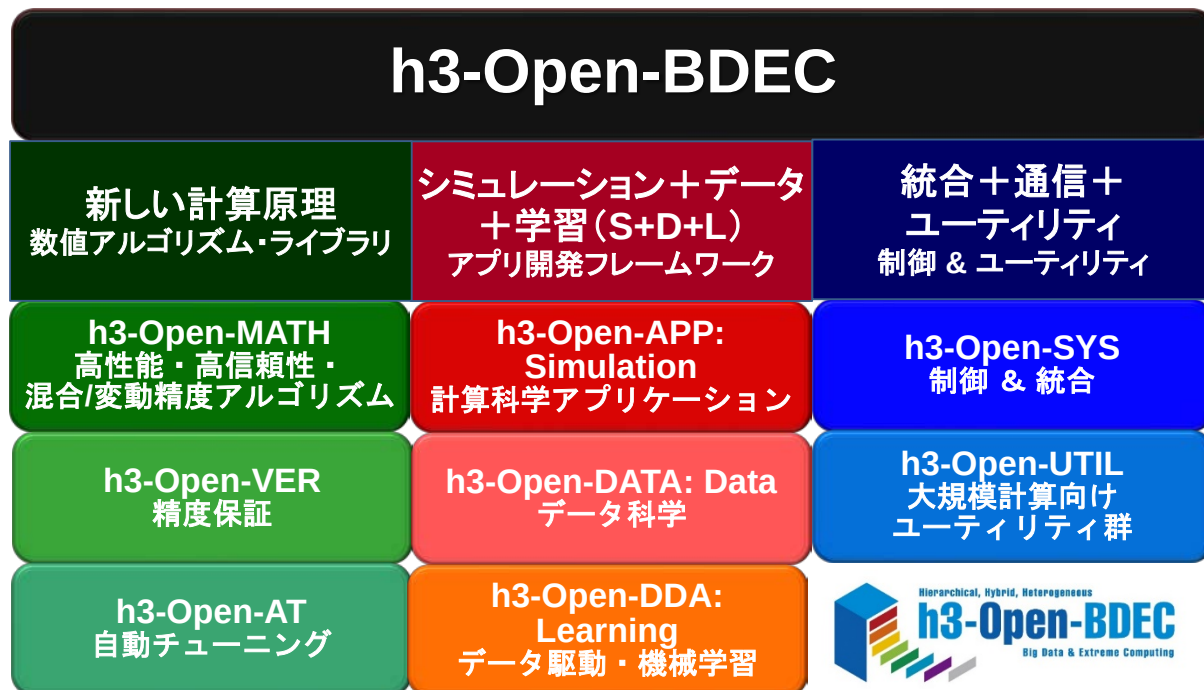
外部リソース

External Network
外部ネットワーク

h3-Open-SYS/WaitIO

データ受け渡しライブラリ[松葉, 2020]

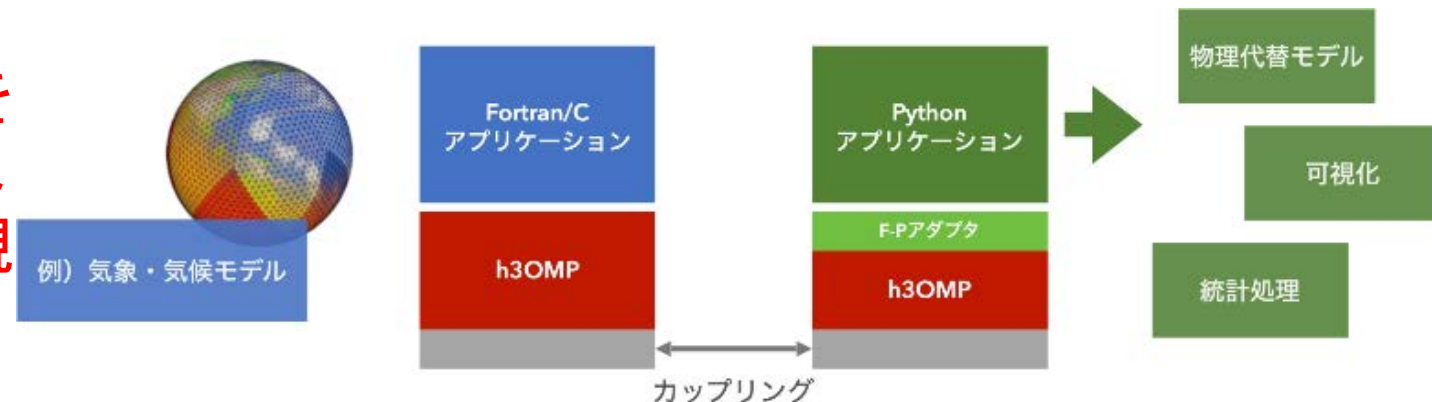
- Odyssey～Aquarius間のIB-EDR経由通信
- 読み込み・書き出しの同期
- API: C/C++, Fortranから呼び出し可能
 - ✓ MPIライクなインタフェースを提供
- カプラーとの連携



「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

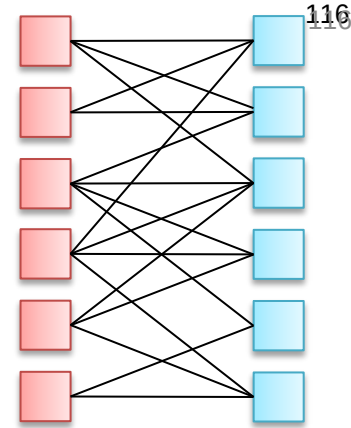
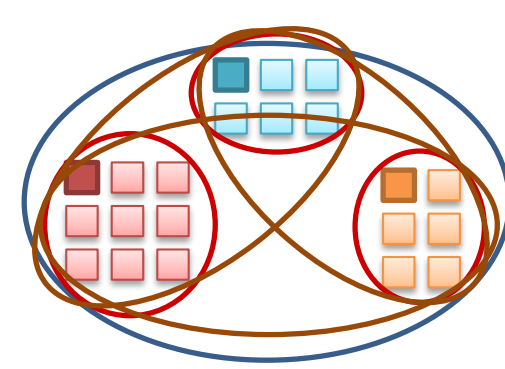


- 異なる物理モデル連成のアンサンブル実行を支援・統合するための機能
 - MPI通信、時刻同期、格子系間マッピング等の管理機能の他、従来のカプラーには無い、複数の弱連成結合シミュレーションのアンサンブル実行、片側のモデルのみをアンサンブル実行する多対1の弱連成結合が可能
 - スパコン上で、全地球大気海洋連成シミュレーションによって動作検証済み
- Fortran/Cコード(物理モデル)とPythonコードの弱連成を実現する機能
 - FortranやCで記述されたプログラム同士の連成計算に限って開発を行ってきたカプラーを、Pythonによって記述されたAI・機械学習、可視化処理系のワークロ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



Fortran/CアプリとPythonアプリの連成計算の模式図

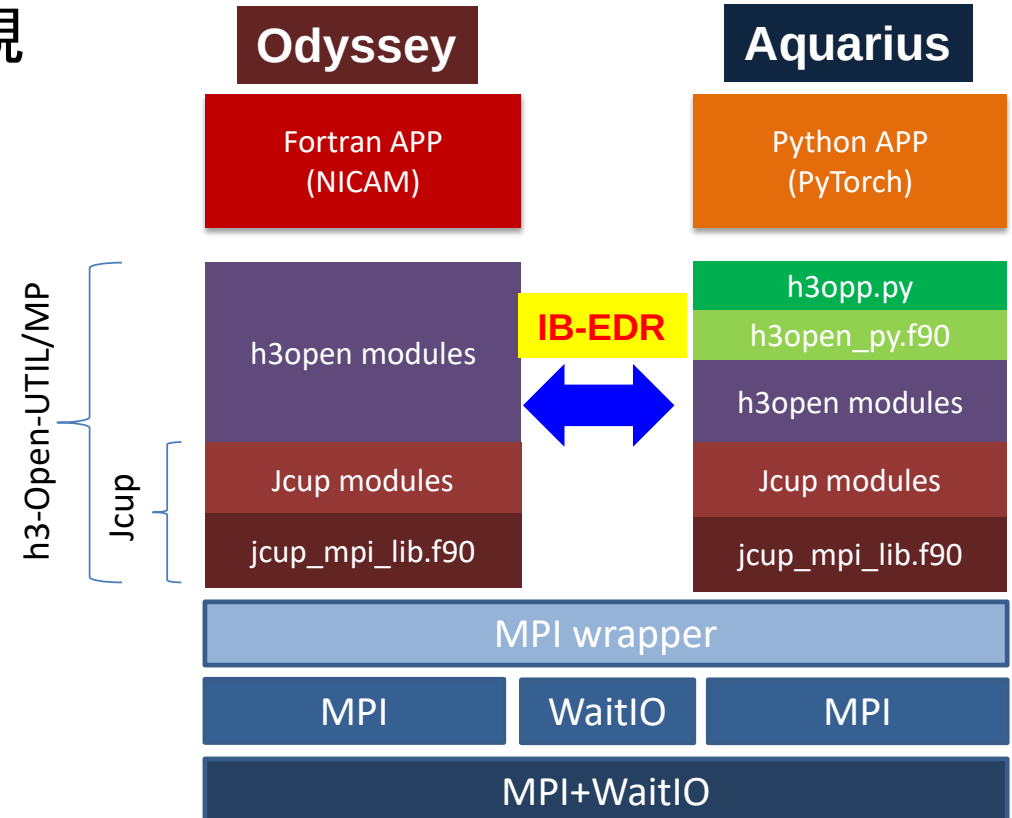
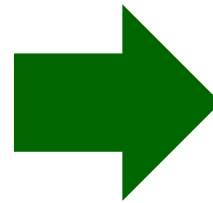
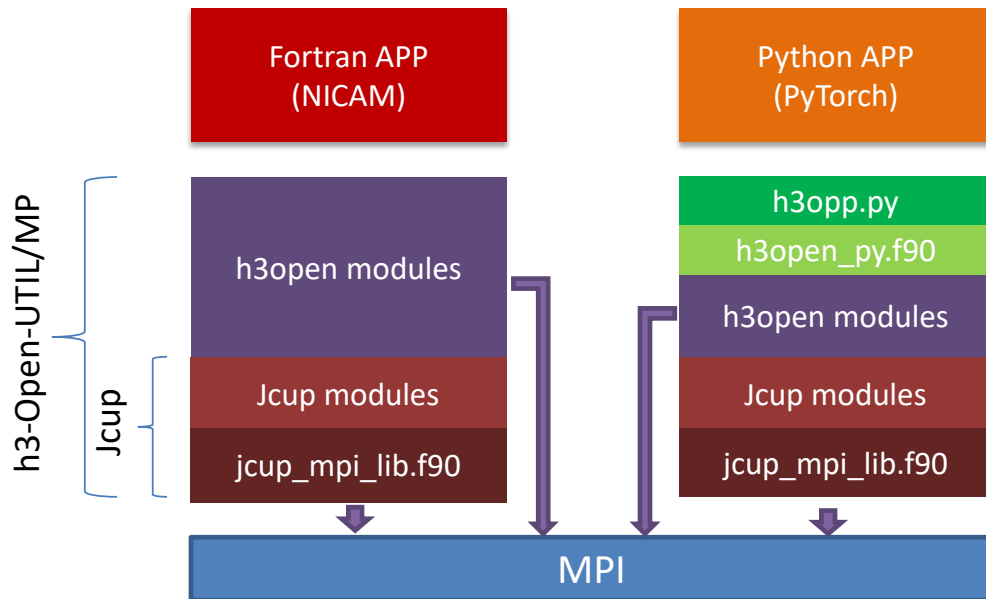
h3-Open-UTIL/MP・ h3-Open-SYS/WaitIO連携



116

- h3-Open-UTIL/MP

- (現状)MPIによるコンポーネント間通信:1対1, 集団通信
- Odyssey-Aquarius間はMPIによる通信は不可⇒
h3-Open-SYS/WaitIOによりO-A間通信実現

**Wisteria
BDEC-01**

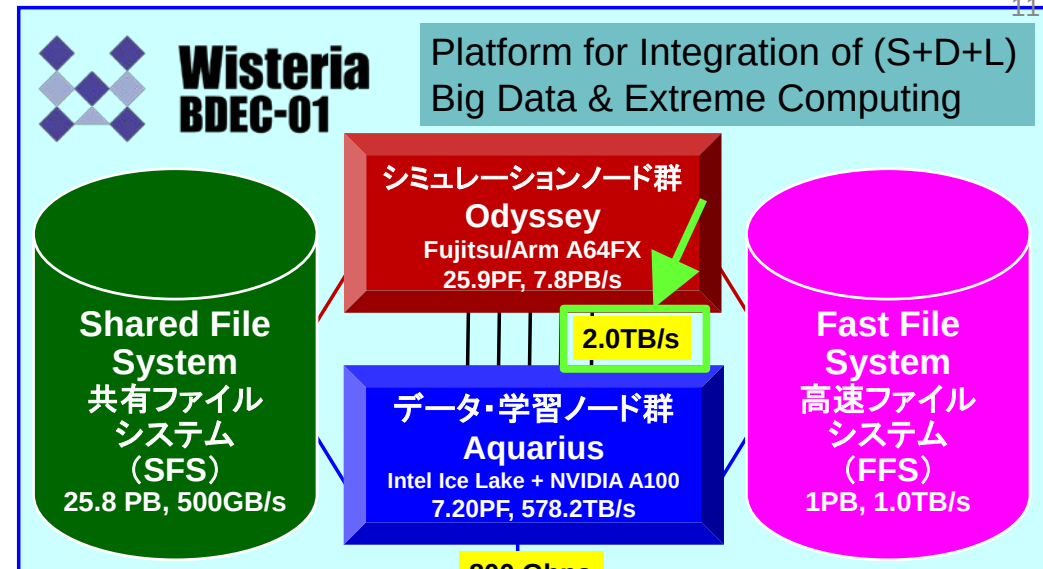
現状:MPI通信可能な環境を前提

整備・公開のスケジュール

- h3-Open-SYS/WaitIO
 - 2021年10月 (Odyssey+Aquarius, 直接通信)
 - 2022年度 (O+A, FFS経由)
- h3-Open-UTIL/MP (HPC+Python)
 - 2021年10月 (Oのみ)
- h3-Open-UTIL/MP+h3-Open-SYS/WaitIO
 - 2022年1月～4月 (O+A, 直接通信)
 - 2022年度 (O+A, FFS経由)
- 協力者求む！
 - ユーザーアカウント+トークン付与
 - AI for HPCへの応募も可



Wisteria
BDEC-01

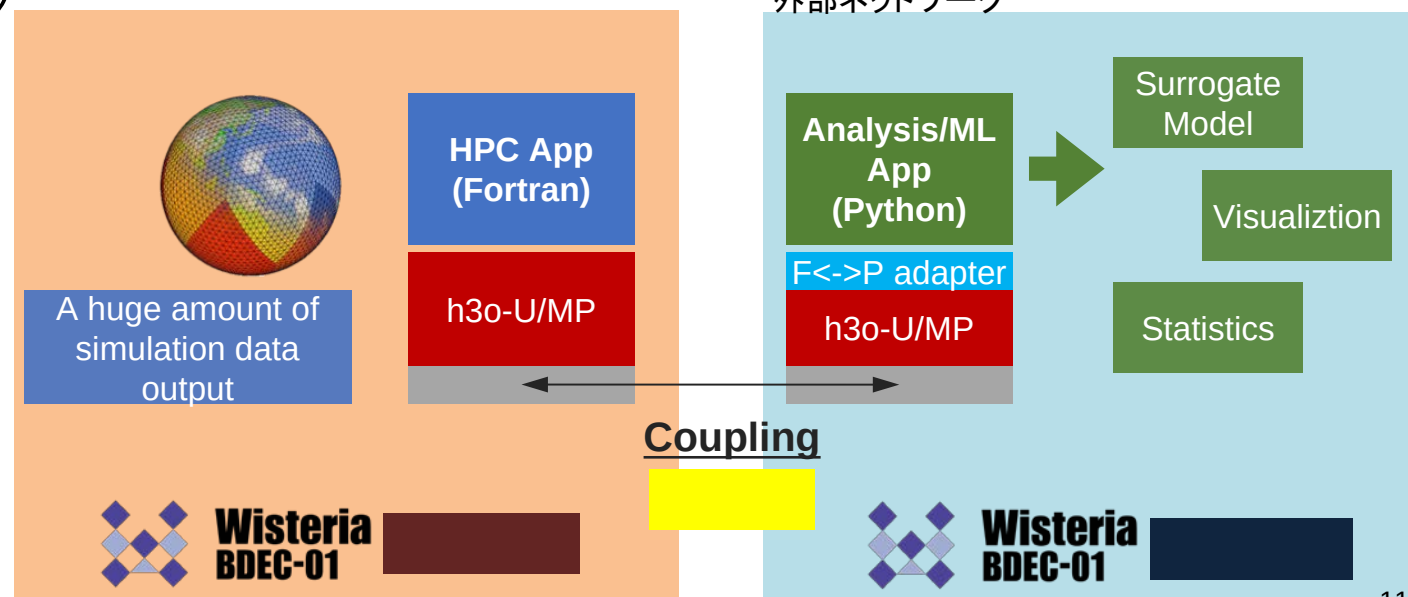


External Network
外部ネットワーク



External Resources

外部リソース



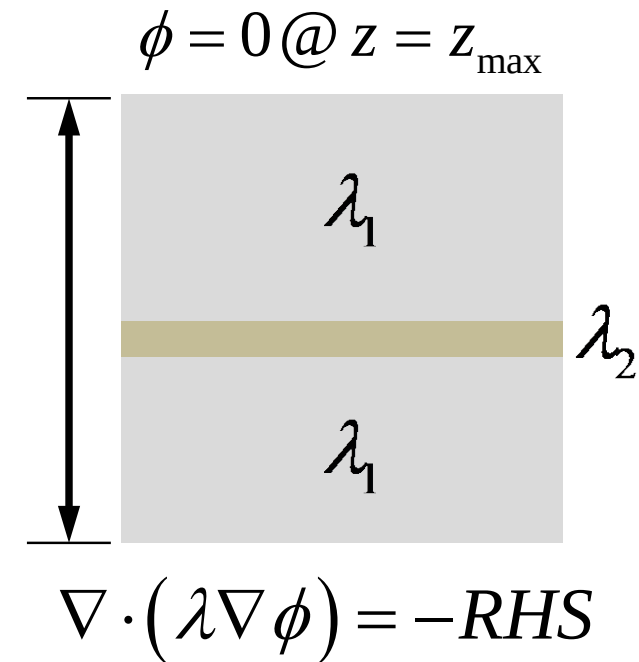
- Wisteria/BDEC-01: 利用例
 - h3-Open-BDEC
 - 予備的計算事例
- 今後の動向等
 - 「共通ストレージ」について

予備的計算事例

- Odyssey
 - FX700による評価 (A64FX搭載)
 - 疎行列演算, 半精度実数 (FP16) [中島他 HPC175, 2020]
 - ステンシル計算 [星野他 HPC178, 2021]
- Aquarius
 - 宇宙物理シミュレーション [三木, 2021]
 - P100, V100との比較

FX700(A64FX搭載)による混合精度演算

- 三次元不均質場熱伝導方程式, FVM, ICCG法
- 倍精度(FP64), 単精度(FP32), 半精度(FP16)
- 反復法の前処理とそれ以外(SpMV, DAXPY, 内積)
- 「それ以外」(SpMV, DAXPY, 内積)
 - 倍精度・単精度
- 前処理
 - 倍精度・単精度・半精度
- 半精度前処理
 - ベクトルは単精度にしないと破綻した
 - 係数行列のみ半精度となっている
- λ_1/λ_2 を変化させる, CRS, 128^3



GFLOPS(ピーク性能)当たり利用負担(円):電気代 GFLOPS/W(Green 500)

***1: Fugaku, *2 NVIDIA DGX SuperPOD, TOP/Green 500@Nov. 2020**

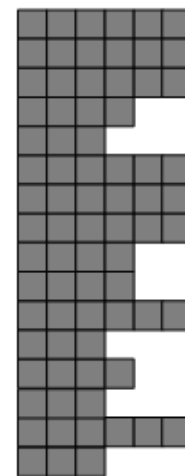
System	JPY/GFLOPS Small is Good	GFLOPS/W Large is Good
Oakleaf-FX/Oakbridge-FX (Fujitsu) (Fujitsu SPARC64 IXfx)	125	0.866
Reedbush-U (HPE) (Intel Xeon Broadwell (BDW))	61.9	2.310
Reedbush-H (HPE) (Intel BDW+NVIDIA P100x2/node)	15.9	8.575
Reedbush-L (HPE) (Intel BDW+NVIDIA P100x4/node)	13.4	10.167
Oakforest-PACS (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	14.66^{*1}
Wisteria-Aquarius(Intel Xeon Ice Lake + NVIDIA A100x8)	9.00	26.20^{*2}

システム名	Oakforet-PACS	Oakbridge-CX	Oakleaf-7 (FX700)
略称	OFP	OBCX	OL7
CPU 名 称	Intel Xeon Phi 7250 (Knights Landing, KNL)	Intel Xeon Platinum 8280 (Cascade Lake, CLX)	Fujitsu A64FX 1.8GHz
コア数 / ソケット	68	28	48
ソケット数 / ノード	1	2	1
理論演算能力 (GFLOPS / ノード)	3,046	4,838	2,765
主記憶容量 GB / ノード	MCDRAM: 16 DDR4: 96	192	32
メモ性能 (GB/sec / ノード) STREAM Triad	MCDRAM: 490 DDR4: 84.5	202	809
コンパイラ	Intel Parallel Studio 2019		Fujitsu FCC 4.0.0

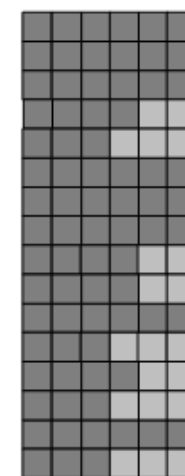
ICCG法計算時間比による性能比 (Small is Good, 倍精度)

OFP・CRSで無次元化, $\lambda_1/\lambda_2 = 1$

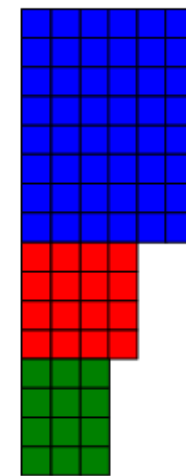
[中島他 SWoPP 2020]



CRS

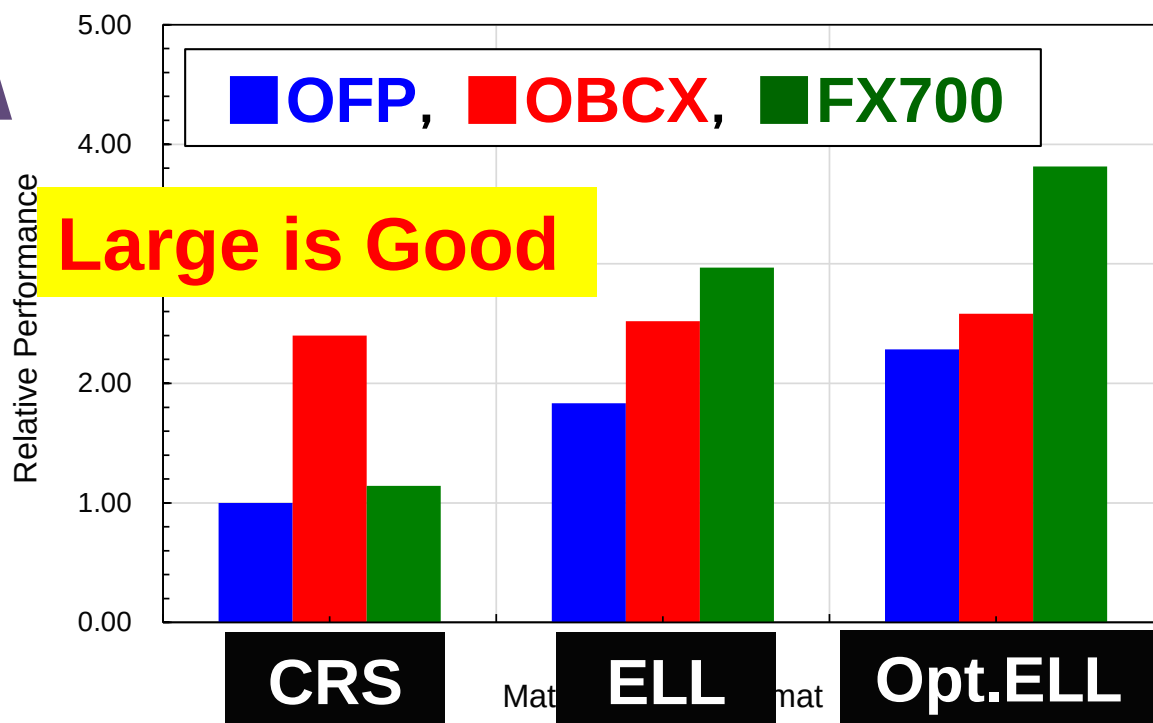


ELL

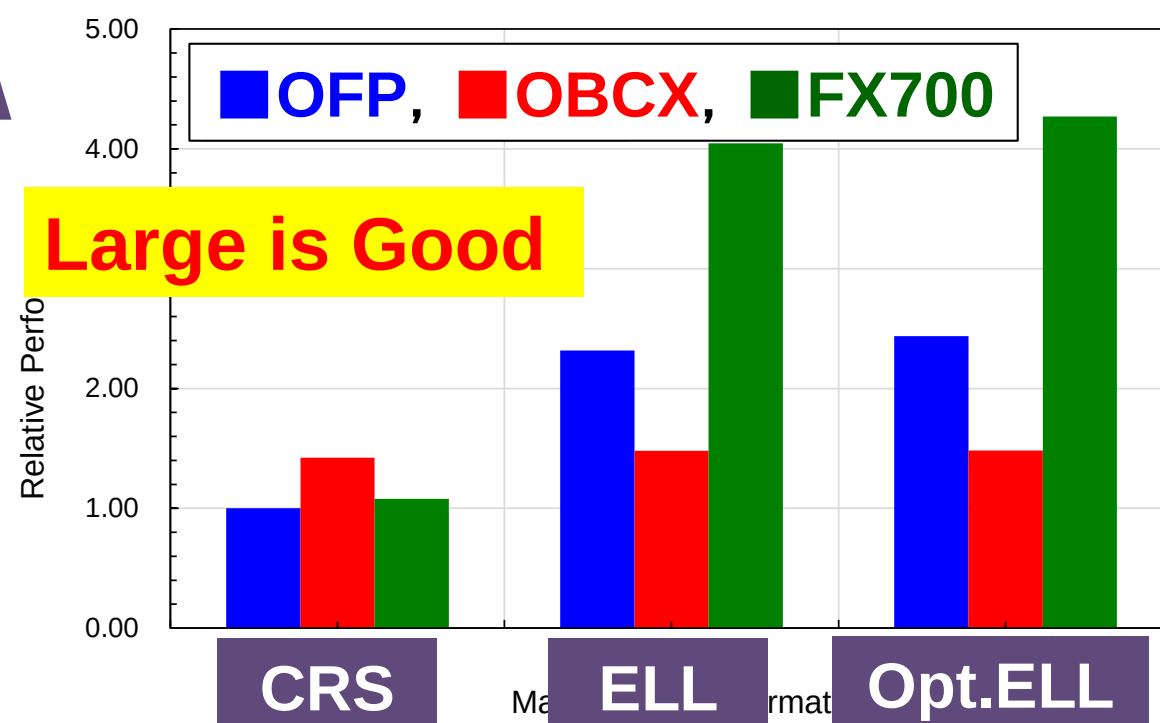


Sliced ELL

Medium: 128^3



Large: 256^3



P3D ICCG法ソルバー混合精度演算・FX700

	反復法主要部	前処理	前処理部分 ベクトル
D-D	FP64	FP64	FP64
D-S	FP64	FP32	FP32
D-H	FP64	FP16	FP32
S-S	FP32	FP32	FP32
S-H	FP32	FP16	FP32

P3D ICCG法ソルバー混合 精度演算・FX700 前進代入部(CRS)実装

[KN et al. SWoPP 2020]

FP64

FP32

FP16

```
!$omp parallel do private(ip, i)
do ip= 1, PEsmptTOT
do i= SMPindex((ip-1)*NCOLORtot)+1, SMPindex(ip*NCOLORtot)
  Ws(I, Z) = W(I, R)
enddo
enddo
```

D-S

```
!$omp parallel private(ic, ip, ip1, i, WVALs, k)
do ic= 1, NCOLORtot
!$omp do
do ip= 1, PEsmptTOT
ip1= (ip-1)*NCOLORtot + ic
do i= SMPindex(ip1-1)+1, SMPindex(ip1)
  WVALs= Ws(i, Z)
do k= indexL(i-1)+1, indexL(i)
  WVALs= WVALs - ALs(k) * Ws(itemL(k), Z)
enddo
  Ws(i, Z)= WVALs * Ws(i, DD)
enddo
enddo
enddo
!$omp end parallel
```

(Backward Substitution)

```
!$omp parallel do private(ip, i)
do ip= 1, PEsmptTOT
do i= SMPindex((ip-1)*NCOLORtot)+1, SMPindex(ip*NCOLORtot)
  W(I, Z) = Ws(I, Z)
enddo
enddo
```

```
!$omp parallel do private(ip, i)
do ip= 1, PEsmptTOT
do i= SMPindex((ip-1)*NCOLORtot)+1, SMPindex(ip*NCOLORtot)
  Ws(i, Z) = Ws(i, R)
enddo
enddo
```

S-H

```
!$omp parallel private(ic, ip, ip1, i, WVALs, k)
do ic= 1, NCOLORtot
!$omp do
do ip= 1, PEsmptTOT
ip1= (ip-1)*NCOLORtot + ic
do i= SMPindex(ip1-1)+1, SMPindex(ip1)
  WVALs= Ws(i, Z)
do k= indexL(i-1)+1, indexL(i)
  WVALs= WVALs - ALh(k) * Ws(itemL(k), Z)
enddo
  Ws(i, Z)= WVALs * Wh(i, DD)
enddo
enddo
enddo
!$omp end parallel
```

```
!$omp parallel do private(ip, i)
do ip= 1, PEsmptTOT
do i= SMPindex((ip-1)*NCOLORtot)+1, SMPindex(ip*NCOLORtot)
  Ws(i, Z) = W(i, R)
enddo
enddo
```

D-H

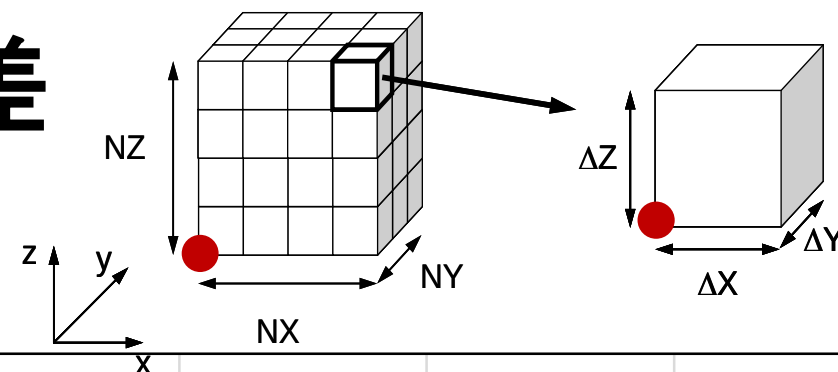
```
!$omp parallel private(ic, ip, ip1, i, WVALs, k)
do ic= 1, NCOLORtot
!$omp do
do ip= 1, PEsmptTOT
ip1= (ip-1)*NCOLORtot + ic
do i= SMPindex(ip1-1)+1, SMPindex(ip1)
  WVALs= Ws(i, Z)
do k= indexL(i-1)+1, indexL(i)
  WVALs= WVALs - ALh(k) * Ws(itemL(k), Z)
enddo
  Ws(i, Z)= WVALs * Wh(i, DD)
enddo
enddo
enddo
!$omp end parallel
```

(Backward Substitution)

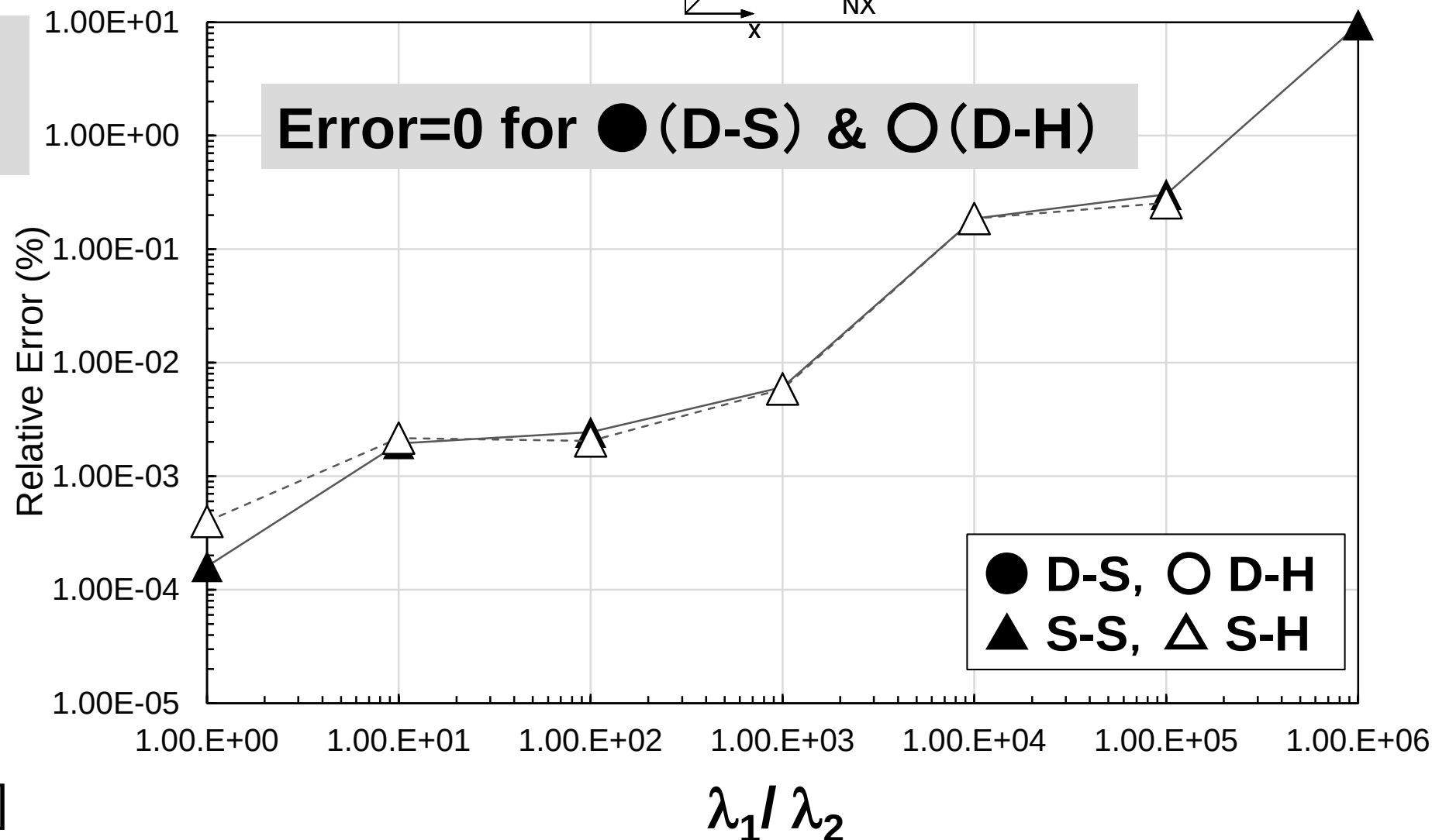
```
!$omp parallel do private(ip, i)
do ip= 1, PEsmptTOT
do i= SMPindex((ip-1)*NCOLORtot)+1, SMPindex(ip*NCOLORtot)
  W(I, Z) = Ws(I, Z)
enddo
enddo
```

混合精度演算計算結果：相対誤差

D-H, S-Hは $\lambda_1 / \lambda_2 = 10^6$ で収束せず



●点におけるD-D
との相対誤差(%)



混合精度演算計算結果：反復回数

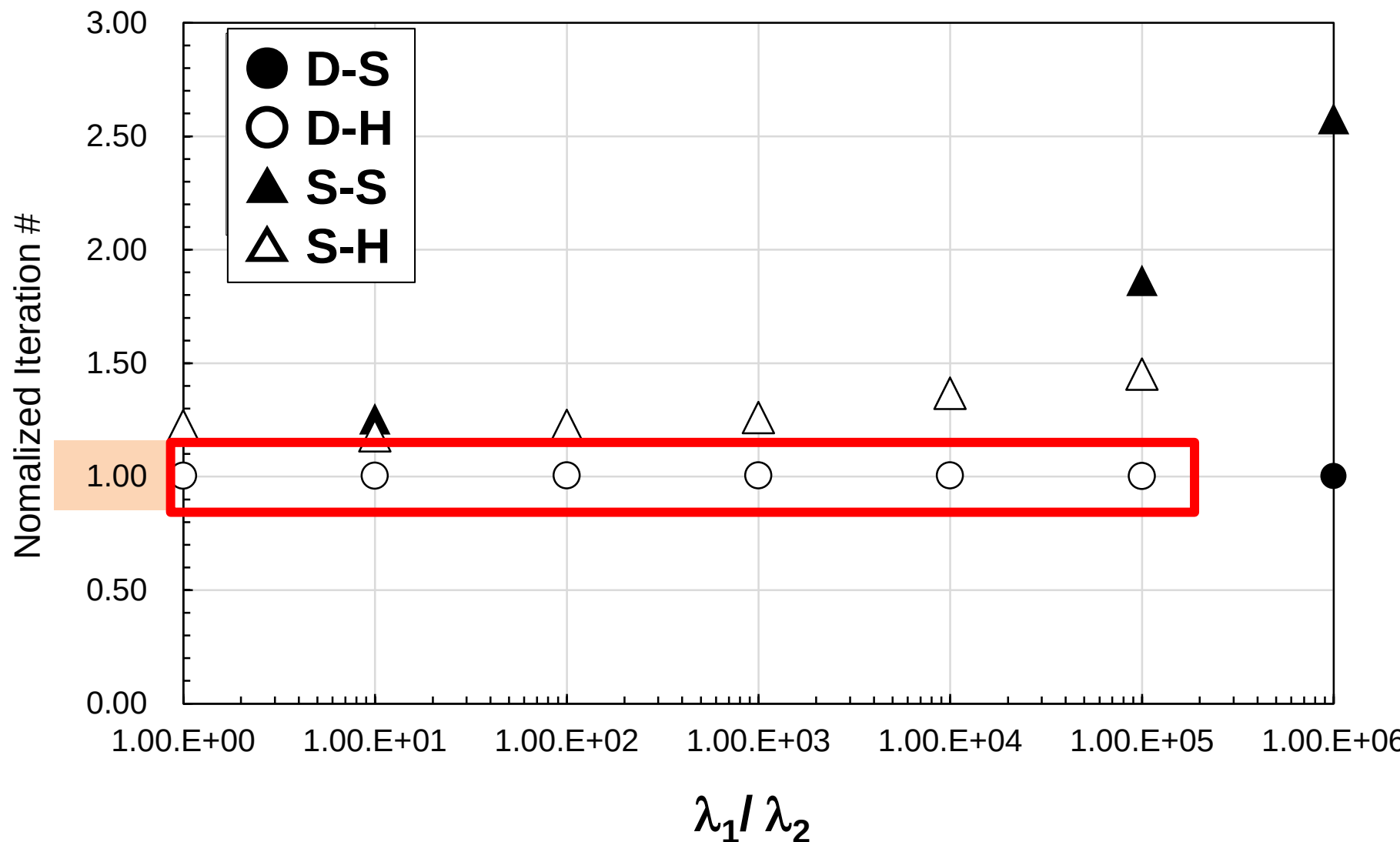
[KN et al. SWoPP 2020]

D-H, S-Hは $\lambda_1/\lambda_2 = 10^6$ で収束せず

D-D 反復回数で
無次元化

● ~ ○ ~ D-D, ▲
~ △

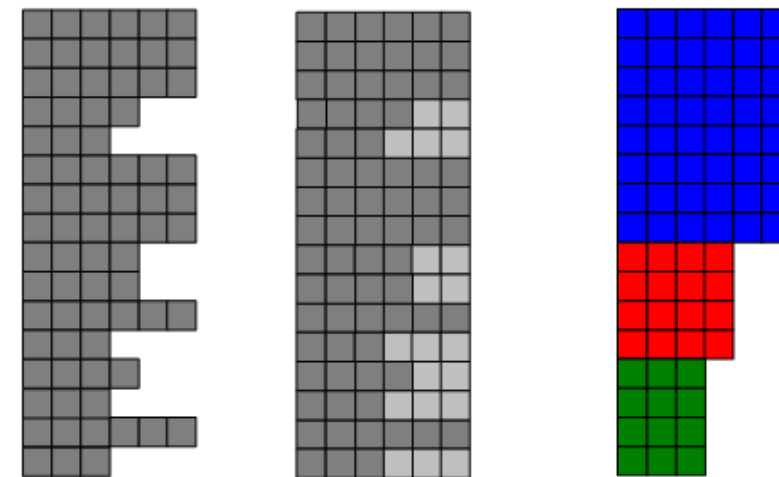
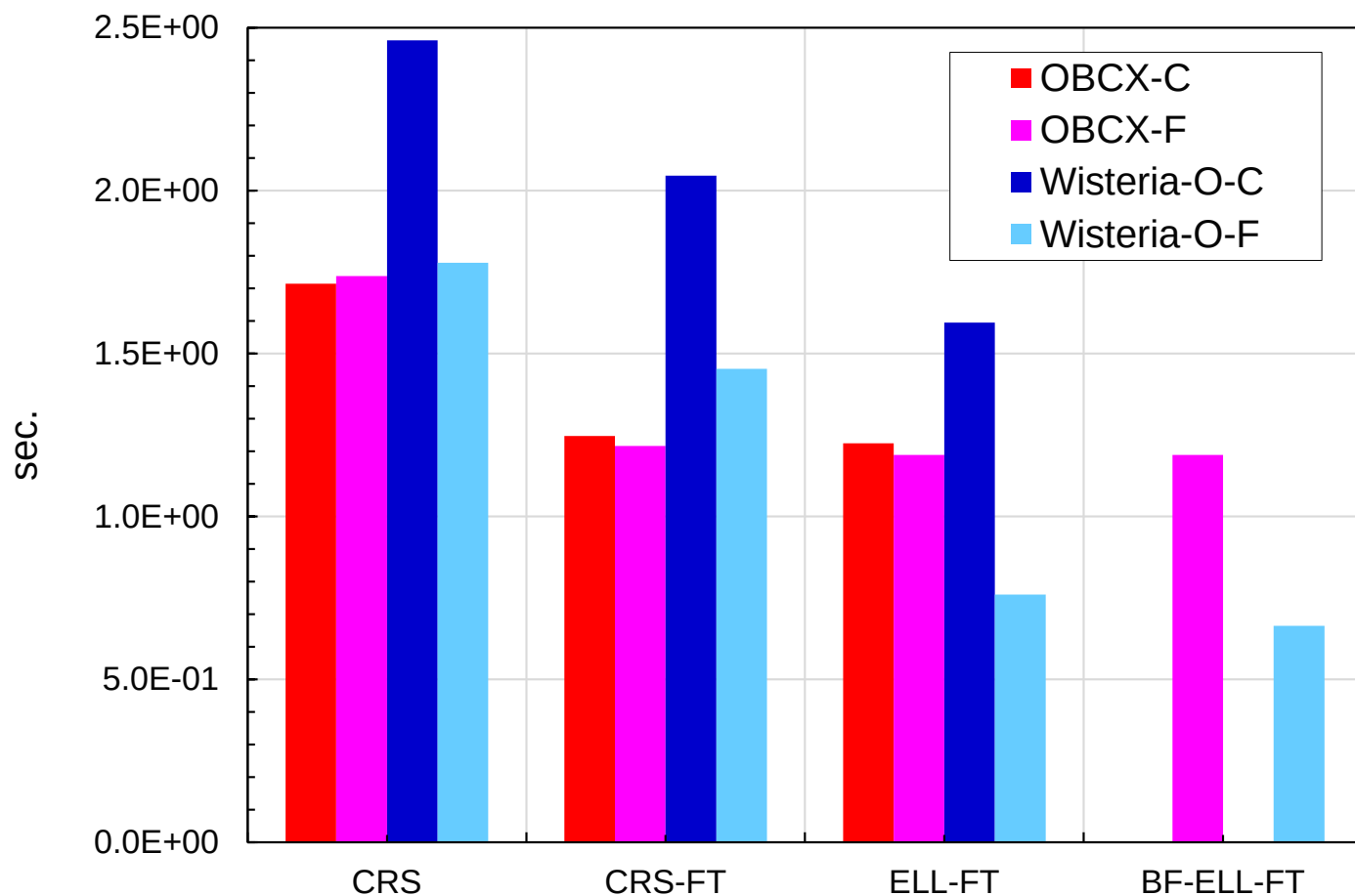
(D-S, D-H) の結果
はD-Dと一致 (if $\lambda_1/\lambda_2 \leq 10^5$)



7点ステンスルPoisson方程式

PCG法, $N=128^3$

1ノード48コア, 4CMGまたぎ



CRS

ELL

Sliced ELL

- CRS
- CRS-FT: First Touch
- ELL-FT
- BF-ELL-FT: Barrier Free
- OBCXではC ■, Fortran ■ と同様の傾向
- Wisteria-OではC ■ が極遅
– -Kfast,openmp

FX700 vs. CLX vs. KNL

(7点ステンシル)

- 拡散方程式のカーネル
 - 自身と前後左右上下の7点を参照して値を更新
- ディリクレ境界条件
 - 境界では自身の値を参照
- 倍精度
- サイズ: 256*384*384
 - A64FXが48コアのため、 $48*8 = 384$
- Byte/Flop = 16/13
 - 各ステップで読み込んだデータは全てキャッシュに乗ると仮定して性能評価
- 7点ステンシルへよく知られた一連の最適化を適用
 - 基本的には直前までの実装で速かったものに最適化を加算

→ 7点ステンシル
のプログラム

```

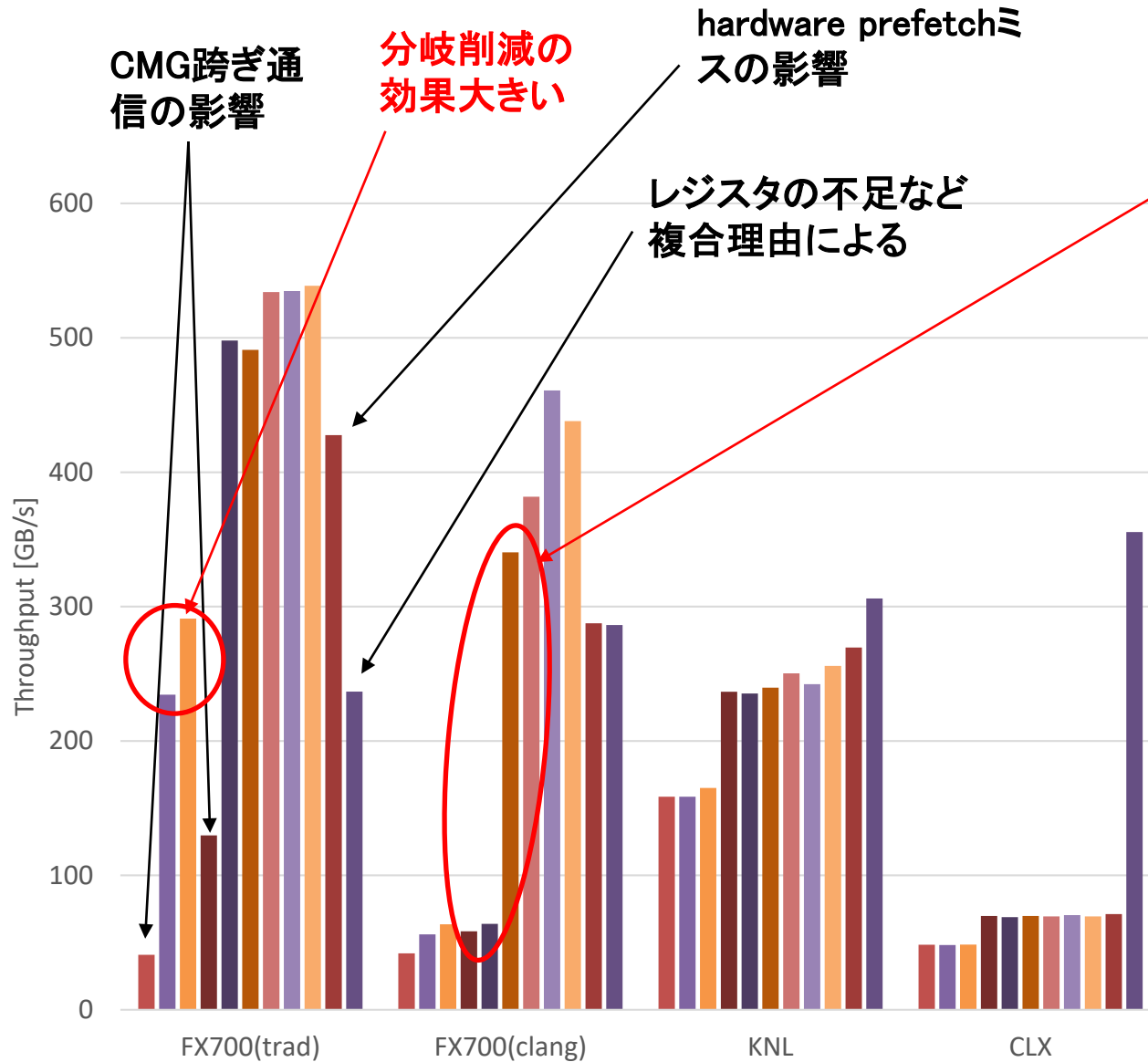
1  do {
2  #pragma omp parallel for
3    for (int z = 0; z < nz; z++) {
4      for (int y = 0; y < ny; y++) {
5        for (int x = 0; x < nx; x++) {
6          int c = x + y * nx + z * nx * ny;
7          int w = (x == 0) ? c : c - 1;
8          int e = (x == nx-1) ? c : c + 1;
9          int n = (y == 0) ? c : c - nx;
10         int s = (y == ny-1) ? c : c + nx;
11         int b = (z == 0) ? c : c - nx * ny;
12         int t = (z == nz-1) ? c : c + nx * ny;
13         f2_t[c] = cc * f1_t[c]
14           + cw * f1_t[w] + ce * f1_t[e]
15           + cs * f1_t[s] + cn * f1_t[n]
16           + cb * f1_t[b] + ct * f1_t[t];
17       }
18     }
19   }
20   double *tmp = f1_t;
21   f1_t = f2_t;
22   f2_t = tmp;
23   time += dt;
24 } while (time + 0.5*dt < 0.1);

```

↓ 計算環境

略 称	FX	KNL	CLX
名称	Fujitsu A64FX (1.8GHz)	Intel Xeon Phi 7250 (Knights Landing)	Intel Xeon Platinum 8280 (CascadeLake)
コア数	48	68	28
理論演算 性能	2,765 GFLOPS	3,046 GFLOPS	2,419 GFLOPS
主記憶容 量	32 GB	MCDRAM: 16 GB DDR4: 96 GB	96 GB
STREAM Tri性能	809 GB/sec	MCDRAM: 490 GB/sec DDR4: 80.1 GB/sec	101 GB/sec

FX700 vs. CLX vs. KNL (7点ステンシル)



LLVMモードでは手動によるベクトル化が重要

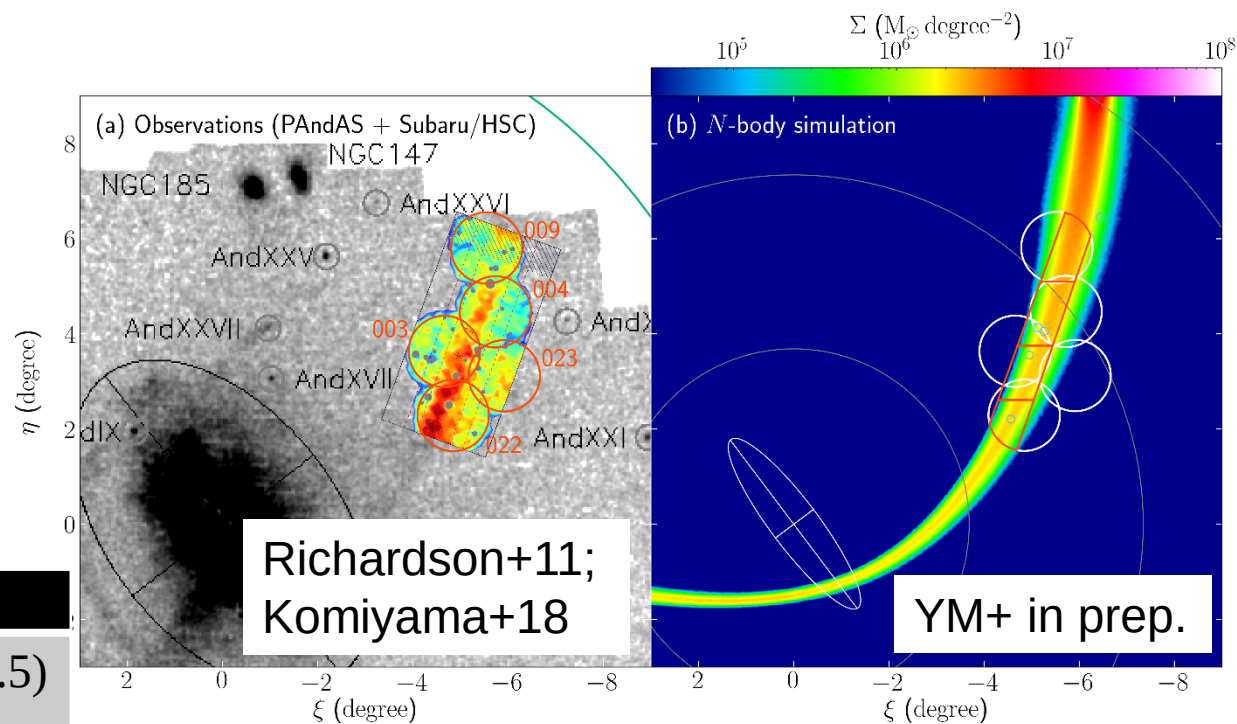
	最適化内容
BASE	ベースライン実装
1stT	ファーストタッチ
PEEL	loop peeling (branch hoisting)により最内ループでの分岐の削減
Ydim	Yループに #pragma omp for nowaitを適用
Y-Zdim	ZループをCM分割、YループをCM内スレッド分割
Intrin	Intrinsicsを用いた実装
w/oFMA	Fused Multiply Addを敢えて使わず、MUとADDのみを用い、パイプライン処理を効率化
UNR	ループアンローリングによりパイプライン処理の効率化
REG	レジスタブロッキング
TILE	タイリングによるキャッシュ利用の効率化
TB	テンポラルブロッキング

〔星野他 HPC178, 2021〕

A100上での性能評価：重カツリーコード(1/2)

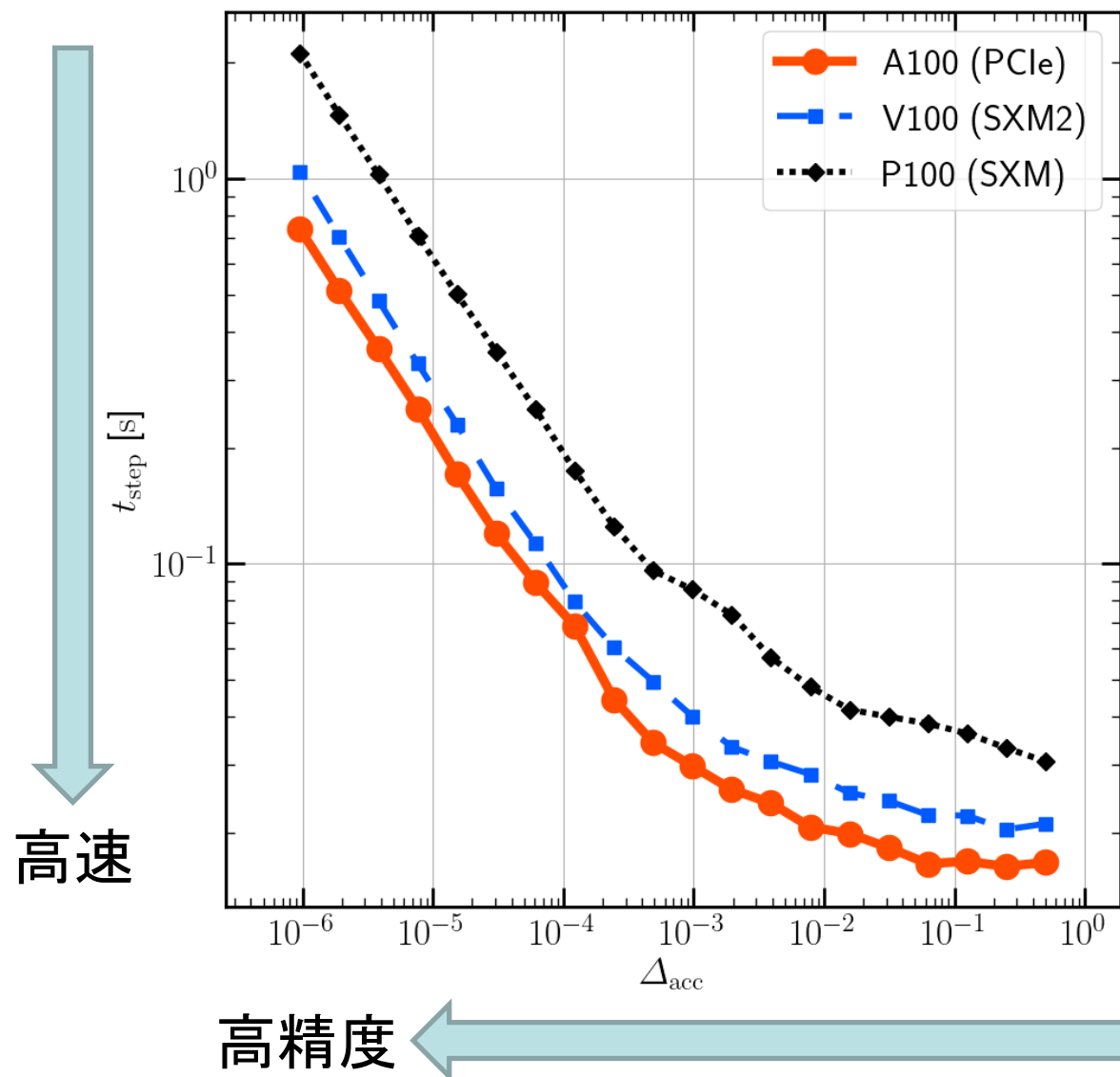
- GPU向けに最適化された重カツリーコードGOTHIC(実装はCUDA)
 - Miki & Umemura (2017), New Astronomy, 52, 65
 - Miki (2019), International Conference on Parallel Processing (ICPP 2019)
- ほとんどの浮動小数点演算は単精度
- 幅優先探索を採用
- 動的な最適化を実装
- 各世代のGPU向けの詳細な調整
 - A100ではL2キャッシュ制御を追加
 - テンソルコア(TC)は使用していない

GPU名 称	P100	V100	A100
理論演算性能 (TFLOPS)	DP: 5.3 SP: 10.6	DP: 7.8 SP: 15.7	DP: 9.7 (+TCで 19.5) SP: 19.5
主記憶容量 (GB)	16	32	40
メモリ性能 (GB/sec)	732	900	1,555

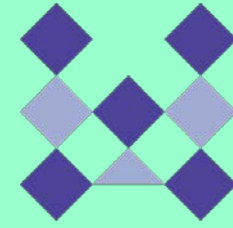


A100上での性能評価：重カツリーコード(2/2)

- A100(PCIe版)上での測定結果
 - M31 model, $N = 8M$
 - 4096ステップの計算時間を測定
- 理論ピーク性能比よりも高速！
 - P100(SXM:Reedbush-H/L搭載)
よりも2.6倍高速
(理論ピーク性能比は1.8倍)
 - V100(SXM2)よりも1.3倍高速
(理論ピーク性能比は1.2倍)
 - SXM4版(Wisteria-A搭載)はPCIe
版よりも1割程度高速と期待される



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
 - 論理コア数が多いが、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

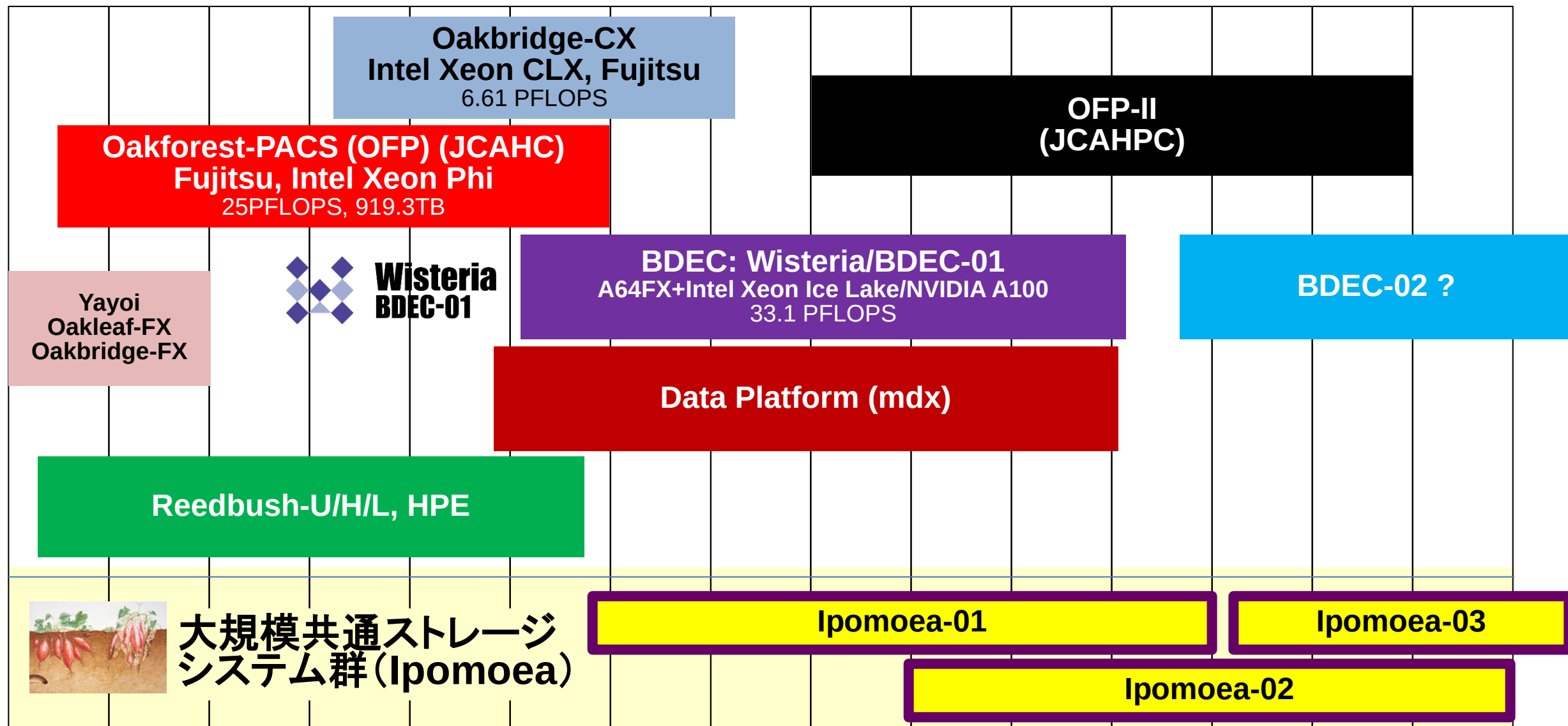
- Wisteria/BDEC-01 : 利用例
 - h3-Open-BDEC
 - 予備的計算事例
- 今後の動向等
 - 「共通ストレージ」について

今後へ向けた様々な技術的課題

- アプリケーションの多様化(計算・データ・学習)
 - 利用方法の多様化
 - それに対応した環境整備(ハードウェア, ソフトウェア(オープンソース含む))
- アーキテクチャ多様化
 - 演算加速装置不可避
 - プログラミング環境多様化
 - それなりの準備期間・体制が必要(2,600人以上のユーザー)
 - 一方で汎用CPUのみのシステムは必要
- センターによる利用者サポートの重要性は更に増加
 - センターとしての継続的な研究開発が必要
 - センター間協力の重要性
 - 様々なツール, 教育体制等の共有
- 増大するデータ量
 - 共通データストレージシステム構想(次節)

東大情報基盤センターのスパコン

FY16 17 18 19 20 21 22 23 24 25 26 27 28 29 30



今後のスケジュール

- Reedbush-H/L : 2021年11月末退役
- Oakforest-PACS (OFP) : 2022年3月末退役
- Oakbridge-CX (OBCX) : 2023年6月末退役 (延長可能性あり)
- Wisteria/BDEC-01 (2021年5月-2027年4月末 (予定))

- Ipomoea-01 (大規模共通ストレージ) (2022年1月～2028年3月 (予定))
- Wisteria-Mercury (Aquariusサポート) (2023年4月～ (予定))
- OFP-II (OFP後継機) (2024年4月～ (予定))

大規模共通ストレージシステム「Ipomoea」(1/2)

- 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
 - 「計算・データ・学習」融合分野では、大量の観測データ、パラメータスタディの結果ファイルなどを処理する必要がある
- 当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
 - 近年は当センターのシステム数も増加、利用者も目的や手法に応じて複数のシステムを同時に利用する事例が増加
 - システムがリプレイスされる場合には大量のデータをバックアップする必要があった
 - 個別のワークロードのデータ量も増加
- 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、当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

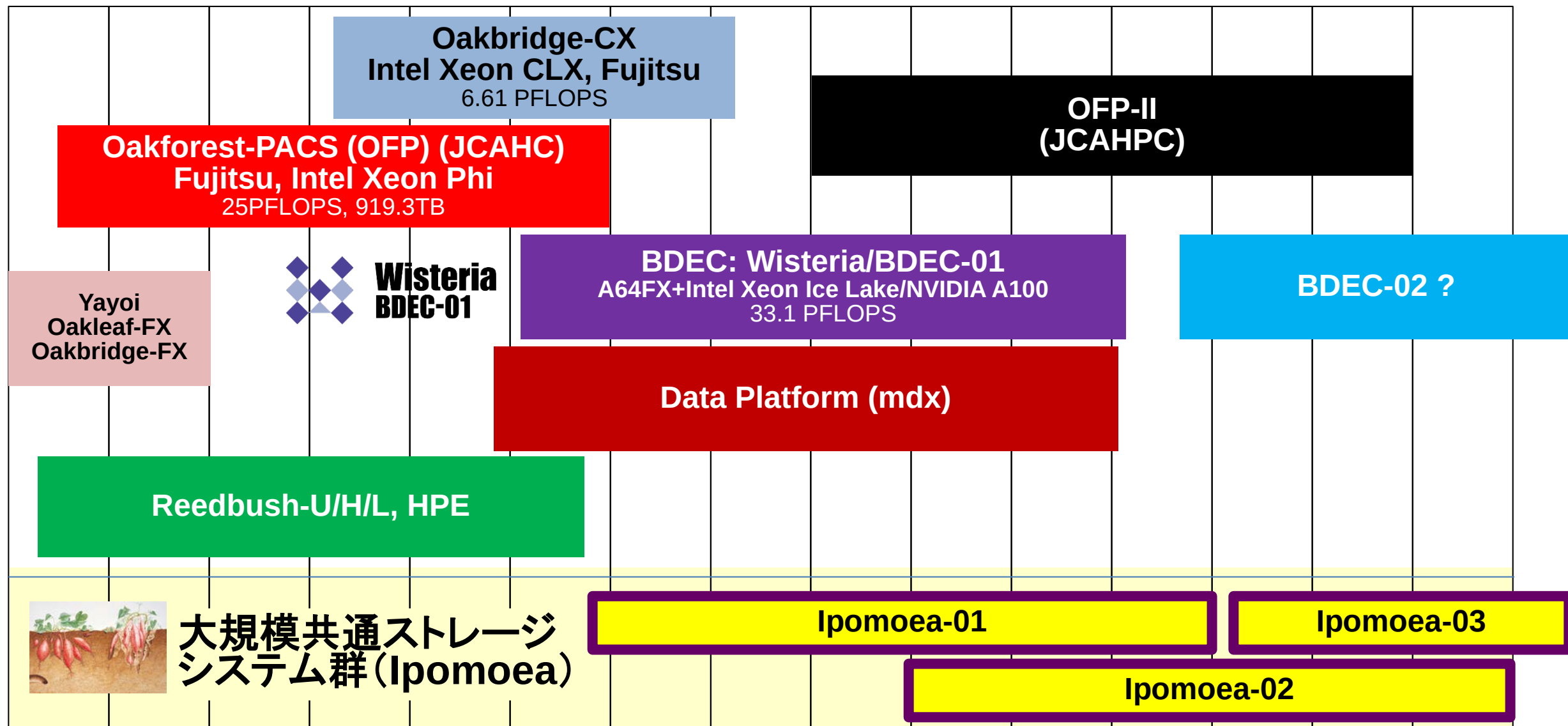
大規模共通ストレージシステム「Ipomoea」(2/2)

- 筑波大学と共同で最先端共同HPC 基盤施設(JCAHPC)によって運用しているOakforest-PACS(OFP)は, ピーク性能25PFLOPS を有するナショナルフラッグシップ級のシステム、2022 年3 月末に運用終了予定
- 2021年春, mdx , Wisteria/BDEC-01 が運用を開始
- 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)」導入を決定
 - 「大規模共通ストレージシステム(Ipomoea)」は約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



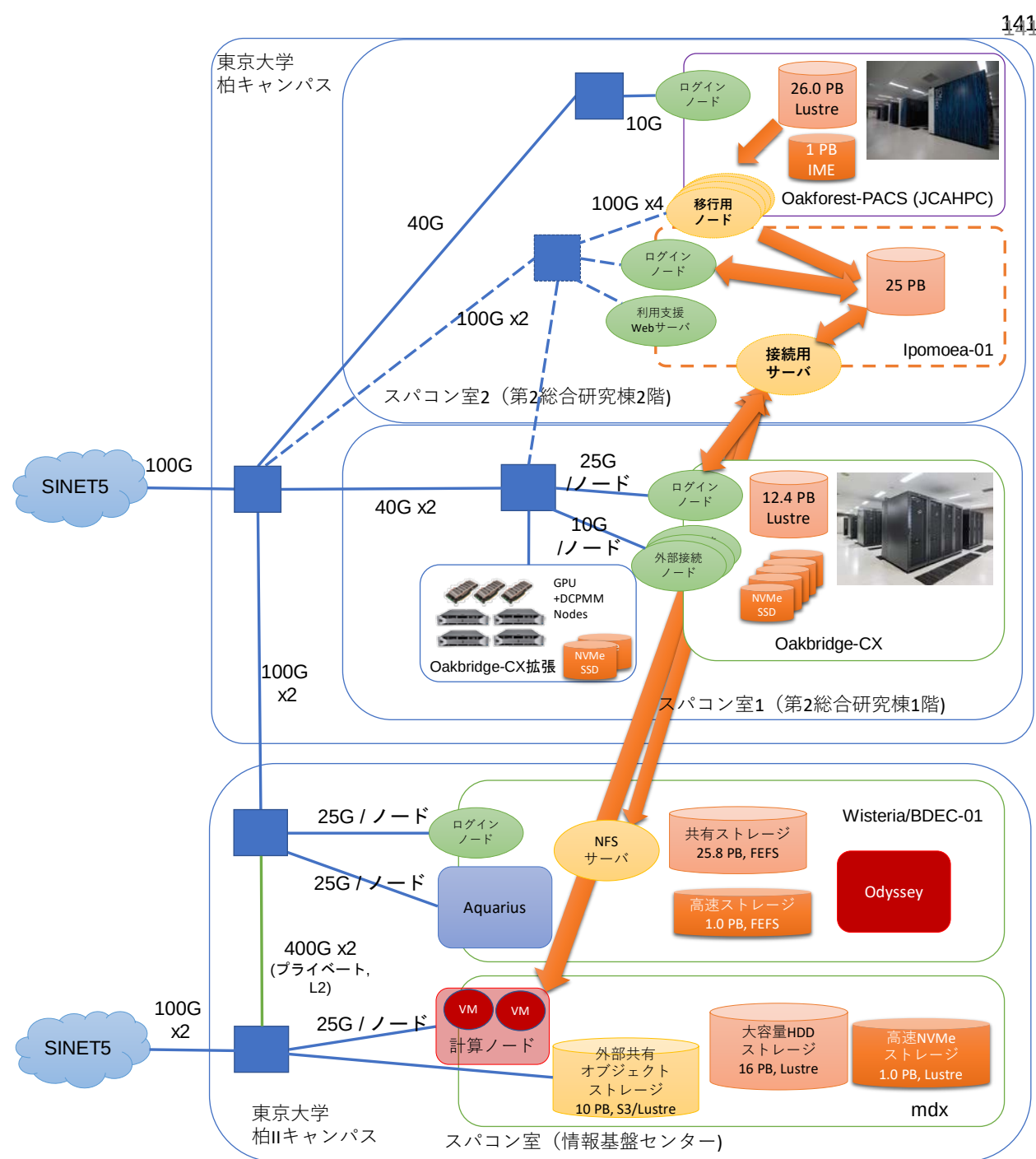
東大情報基盤センターのスパコン

FY16 17 18 19 20 21 22 23 24 25 26 27 28 29 30



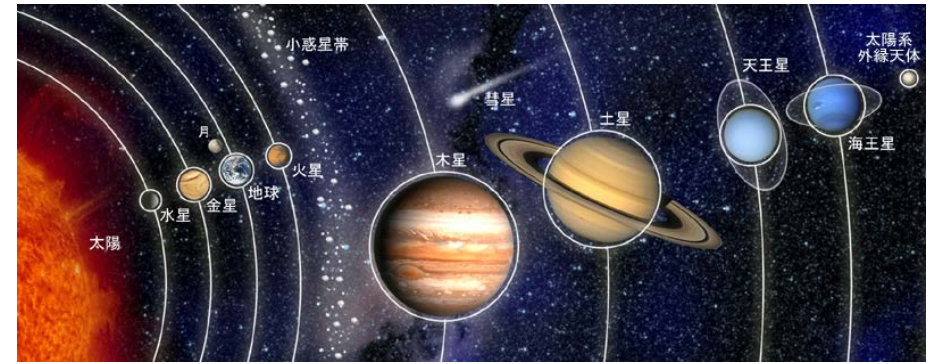
Ipomoea-01

- 2022年1月運用開始予定, 25+PF
- 2022 年6 月末までにOFPのLustre 領域の必要ファイルについて移行完了
- 既存システム(2021年3月末)
 - OFP 50億ファイル 11 PB
 - OBCX 8.5億ファイル 2.6 PB
 - OPA

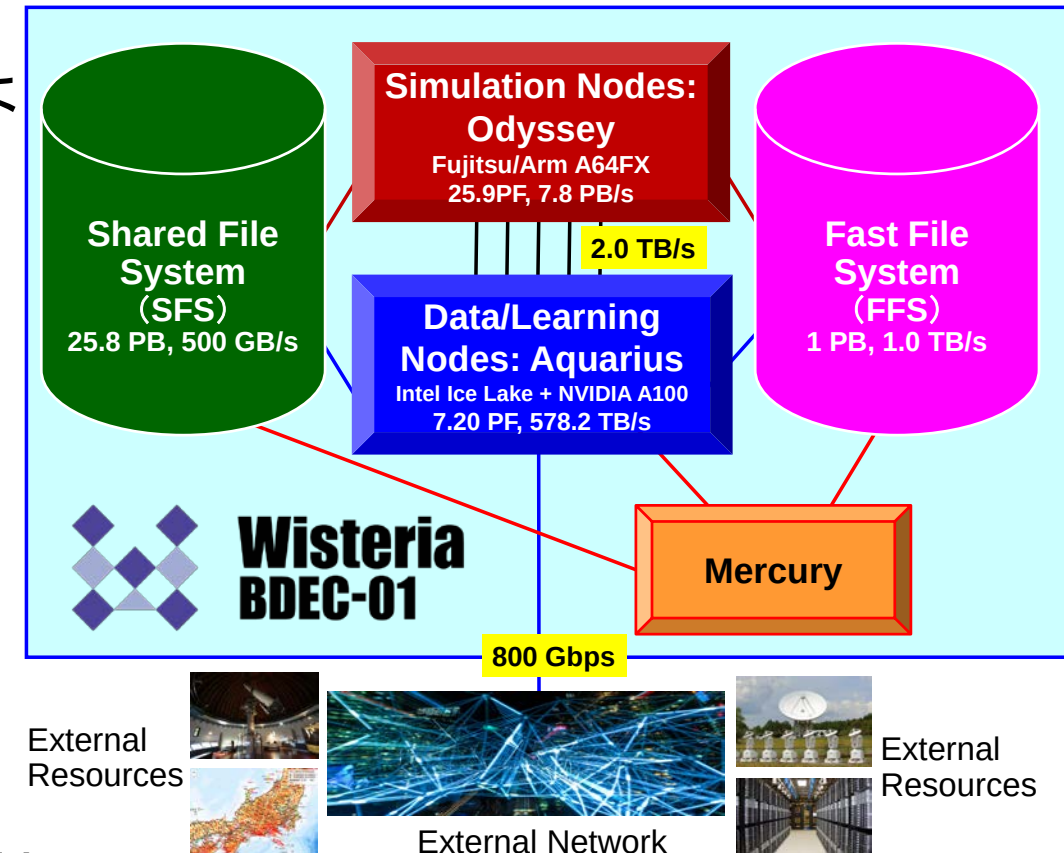


Wisteria-Mercury

翼の生えた使い (Winged Messenger)



- Wisteria/BDEC-01-Aquarius
 - CPUのみ使用⇒GPU8枚分負担金:割高
 - グラフ演算, 整数演算, 推論等Odysseyを使うよりも高速な場合ある
- Mercury: 汎用CPU, 100+ノード
 - Aquariusのサポート, 連携
 - Aquarius~Mercuryの単独MPIジョブが走ると猶良い
 - HPC向けにも使用
 - SFS, FFSへのアクセス
- 2023年4月運用開始に向けて導入
 - 様々なアーキテクチャ: Intel, AMD, Arm
 - 部分的にDPU, 特殊用途加速装置等導入も検討



- 東大情報基盤センターについて
- サービス概要
- 企業による利用制度
 - 制度概略と利用提案
 - 申込・審査・成果報告
 - スパコン利用への道
 - ①お試しアカウント付き並列プログラミング講習会
 - ②トライアルユース制度
 - ③ 企業利用制度
- スーパーコンピュータシステム概要
 - Reedbush
 - Oakforest-PACS
 - Oakbridge-CX
 - Wisteria/BDEC-01
 - 運用
- **ご質問は**

ご質問・連絡先

- 東京大学情報システム部情報戦略課研究支援チーム
 - [uketsuke\(at\)cc.u-tokyo.ac.jp](mailto:uketsuke(at)cc.u-tokyo.ac.jp)

backup

利用コース(一般利用)

<https://www.cc.u-tokyo.ac.jp/supercomputer/wisteria/service/course.php>

グループ	利用負担金 (年額、税込)	利用可能 ノード・GPU数	割当トークン量(年間) 及び消費係数	ディスク量
申込セクタ当り	60,000円 (企業: 72,000円)	Wisteria-O : 最大2,304ノード Wisteria-A : 最大64GPU (8ノード)	8,640トークン (Wisteria-O 1ノード、 2時間 × 360日 相当) <u>(6.9円 / NH)</u> Wisteria-O消費係数 : 1.00 (1ノード 当たり) ※優先利用ノード群 全体の15%程度は 消費係数が1.50となります Wisteria-A消費係数 : 3.00 (1GPU 当たり)	•グループ (/work) 2TB •利用者 (/home) 50GB

利用コース(ノード固定: Aquarius) (1ノード・8GPU)

審査あり

<https://www.cc.u-tokyo.ac.jp/supercomputer/wisteria/service/course.php>

	利用負担金 (年額、税込)	利用可能 ノード・GPU数	割当トークン量(年間) 及び消費係数	ディスク量
ノード固定 Wisteria-A 申込セクタ当り	2,160,000円 ※セクタのみ 申込可能 (企業: 2,592,000円)	Wisteria-O : 最大230ノード Wisteria-A : 最大64GPU (8ノード)	207,360トークン (8GPU, 2時間 × 360日 相当) Wisteria-O消費係数 : 1.00 (1ノード当たり) ※優先利用ノード群 (全体の15%程度)は 消費係数が1.5となります。 Wisteria-A消費係数 : 3.00 (1GPU当たり)	•グループ (/work) 48TB •利用者 (/home) 50GB

- Wisteria-O (Odyssey), Wisteria-A (Aquarius) 複数ノードを使用することもできるが、その分トークンが消費されるため、ノード固定で利用できるトークン数は減る

利用コース(GPU専有: Aquarius) 審査無し

<https://www.cc.u-tokyo.ac.jp/supercomputer/wisteria/service/course.php>

	利用負担金 (年 額、税込)	利用可能 ノード数	割当 トークン量(年間 及び消費係数	ディスク量
GP専 有 Wisteria-A 申込セ ャ 当り	270,00円 ※申込単位は 1,2,セ ャ のみ (1,2,4GPU (企業: 324,00円)	Wisteria-O : 最大230ノード Wisteria-A : 最大64GPU (ノード)	25,920トークン (1GPU 2時間 × 36日 相当) Wisteria-O消費係数 : 1.00 (ノード 当たり) ※優先利用ノード群 (全体の15程 度)は 消費係数が1.5セ になります。 Wisteria-A消費係数 : 3.00 (1GPU 当 たり)	•グループ (/work) 6TB •利用者 (/home) 50GB

- Wisteria-O (Odyssey), Wisteria-A (Aquarius) 複数ノードを使用することもできるが、その分トークンが消費されるため、GPU専有で利用できるトークン数は減る

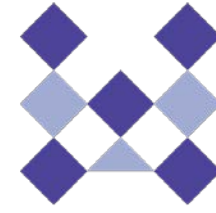
参考情報: 利用形態(1/3)

- 一般利用コース
 - <https://www.cc.u-tokyo.ac.jp/supercomputer/wisteria/service/course.php>
- トライアルユース(8月以降, 大学他・企業で別制度)
 - <https://www.cc.u-tokyo.ac.jp/guide/trial/>
- お試しスパコン利用(無料体験)(8月以降, 1ヶ月無料)
 - https://www.cc.u-tokyo.ac.jp/guide/trial/free_trial.php
- 教育利用(8月以降, 無料)
 - <https://www.cc.u-tokyo.ac.jp/guide/education/>
 - 大学・高専の講義, 集中講義等(学内外問わず)
- 講習会(随時)
 - <https://www.cc.u-tokyo.ac.jp/events/lectures/>
- 企業利用(2021年度後期申込: 8月中旬締切)
 - <https://www.cc.u-tokyo.ac.jp/guide/company/>
 - 審査あり, 大学・研究機関の2割増しの負担金

利用形態(2/3):公募型課題

- HPCI(本年度分受付終了, 一部募集中課題あり, 来年度:10月下旬締切)
 - https://www.hpci-office.jp/pages/using_hpci
- JHPCN(本年度分受付終了, 来年度:1月上旬締切)
 - <https://jhpcn-kyoten.itc.u-tokyo.ac.jp/ja/>
 - 学際大規模情報基盤共同利用・共同研究拠点(全国8国立大学センター)
- 若手・女性(2021年度後期申込:8月下旬締切)
 - <https://www.cc.u-tokyo.ac.jp/guide/young/>
 - 若手研究者(2021年4月1日時点で40歳以下, 学生含む), 女性研究者または学生
- 萌芽課題: AI for HPC(公募中(締切6月30日))
 - 「AI for HPC: Society 5.0実現へ向けた人工知能・データ科学による計算科学の高度化(試行)」: 翌年度JHPCNへの応募を目標, Wisteria/BDEC-01がターゲット
 - <https://www.cc.u-tokyo.ac.jp/guide/exploratory/AIforHPC/>
 - 公募説明会(5月20日(木)13:00-15:00, オンライン): 事前申込
 - <https://www.cc.u-tokyo.ac.jp/guide/exploratory/AIforHPC/application.php#meeting>

利用形態(3/3): 公募型課題(続き)



Wisteria
BDEC-01

- 大規模HPCチャレンジ(8月以降, 月1回(月末)実施)
 - <https://www.cc.u-tokyo.ac.jp/guide/hpc/>
 - 全系を24時間占有利用, 年3回募集(3~4ヶ月分をまとめて募集)
 - 当面はOdyssey, Aquariusは別々, 8時間(09:00-17:00)
 - ニュースレターへの寄稿, セミナー等での発表が義務

一般利用コース

<https://www.cc.u-tokyo.ac.jp/supercomputer/wisteria/service/course.php>

- グループコース(1人またはそれ以上から構成されるグループ)

- Wisteria/BDEC-01から「パーソナルコース」は廃止
- 代表者は大学・公共機関所属者



- トークン(ノード時間)を購入

- Odyssey, Aquariusを利用できる
- O/Aでそれぞれ消費係数が異なる

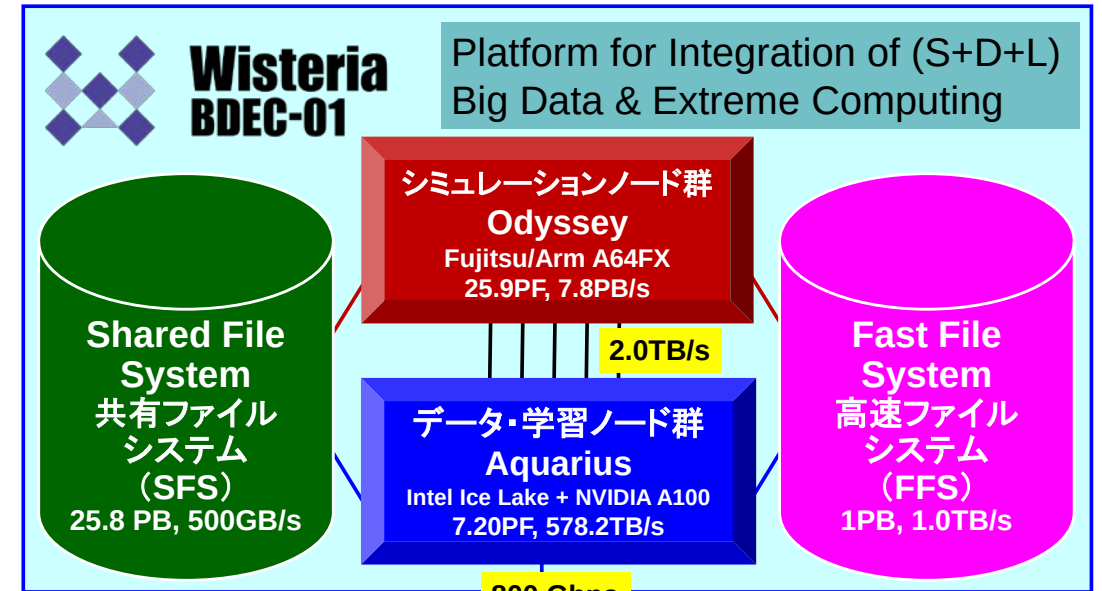
- 一般利用

- ノード固定

- Aquariusの1ノード(8GPU)を占有して利用
- 審査有り(ヒアリング)

- GPU占有

- Aquariusの1・2・4GPUを占有して利用
- 審査無し



External Resources

External Network
外部ネットワーク

外部リソース

公募型課題(再掲)

- HPCI(本年度分受付終了, 一部募集中課題あり, 来年度:10月下旬締切)
 - https://www.hpci-office.jp/pages/using_hpci
- JHPCN(本年度分受付終了, 来年度:1月上旬締切)
 - <https://jhpcn-kyoten.itc.u-tokyo.ac.jp/ja/>
 - 学際大規模情報基盤共同利用・共同研究拠点(全国8国立大学センター)
- 若手・女性(2021年度後期申込:8月下旬締切)
 - <https://www.cc.u-tokyo.ac.jp/guide/young/>
 - 若手研究者(2021年4月1日時点で40歳以下, 学生含む), 女性研究者または学生
- 萌芽課題:HPC for AI(公募中(締切6月30日))
 - 「AI for HPC:Society 5.0実現へ向けた人工知能・データ科学による計算科学の高度化(試行)」:翌年度JHPCNへの応募を目標, Wisteria/BDEC-01がターゲット
 - <https://www.cc.u-tokyo.ac.jp/guide/exploratory/AIforHPC/>
 - 公募説明会(5月20日(木)13:00-15:00, オンライン):事前申込
 - <https://www.cc.u-tokyo.ac.jp/guide/exploratory/AIforHPC/application.php#meeting>

若手・女性利用

<https://www.cc.u-tokyo.ac.jp/guide/young/>

- 対象
 - 若手研究者(2021年4月1日現在において40歳以下), 女性研究者, 学生
- 一般枠(年2回募集)
 - 個人研究
 - 4月開始:1年, 10月開始:半年
- インターン制度(6月頃募集)
 - 学部学生・大学院生を対象, 主に夏期
 - 1人または複数人
- 一般枠課題のうち, 特に優れた課題は, 本センターよりJHPCN萌芽型共同研究課題として推薦する予定
 - 7月のシンポジウムでポスター発表できる

AI for HPC: Society 5.0実現へ向けた人工知能・データ科学による計算科学の高度化(試行)

<https://www.cc.u-tokyo.ac.jp/guide/exploratory/AIforHPC/>

- (計算+データ+学習(S+D+L))融合実現, データ科学, 機械学習, 人工知能による計算科学の高度化を目指す提案を募集
- 原則として, 計算科学シミュレーション(自作またはオープンソース)を, データ科学, 人工知能, 機械学習等によって高度化, 効率化することを目的とする
 - 大規模データ同化と人工知能を融合するような研究も受け付ける。
 - プログラム本体のチューニング, アルゴリズム高度化などは対象外ですが, 自動チューニングによって最適アルゴリズムを選択するような提案は歓迎いたします。
- 応募者グループ・センター教員の共同研究として実施
 - 代表者: 居住者(大学・研究機関・企業), メンバー: 非居住者参加も可能
 - 2022年度JHPCN共同研究課題応募を目指す
 - <https://jhpcn-kyoten.itc.u-tokyo.ac.jp/ja/>
- 公募説明会(5月20日(木)13:00-15:00, オンライン): 事前申込
 - <https://www.cc.u-tokyo.ac.jp/guide/exploratory/AIforHPC/application.php#meeting>

- 東大情報基盤センターの概要
- Wisteria/BDEC-01
 - 導入の背景
 - システム概要
- Wisteria/BDEC-01: 利用プログラム
 - 一般, 企業, 萌芽課題, JHPCN, HPCI, 大規模HPCチャレンジ
 - 負担金, リソースグループ等
 - 講習会, GPUミニキャンプ, JCAHPCセミナー
- **Wisteria/BDEC-01: 利用例**
 - **h3-Open-BDEC**
 - 予備的計算事例
- 今後の動向等
 - 「共通ストレージ」について

- 東大情報基盤センターの概要
- Wisteria/BDEC-01
 - 導入の背景
 - システム概要
- Wisteria/BDEC-01: 利用プログラム
 - 一般, 企業, 萌芽課題, JHPCN, HPCI, 大規模HPCチャレンジ
 - 負担金, リソースグループ等
 - 講習会, GPUミニキャンプ, JCAHPCセミナー
- **Wisteria/BDEC-01: 利用例**
 - **h3-Open-BDEC**
 - **予備的計算事例**
- 今後の動向等
 - 「共通ストレージ」について