

事業項目③: 量子・スパコン連携スケジューラ および量子・HPC連携ソフトウェアの カップリング技術の研究開発

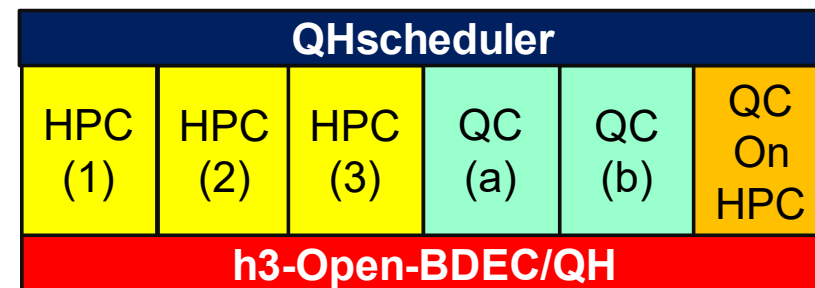
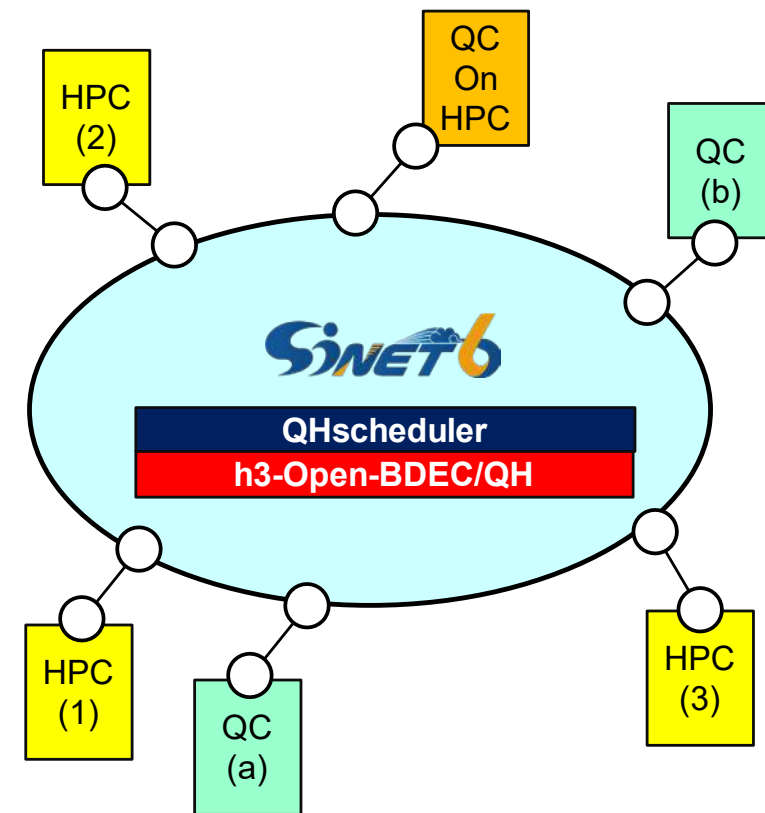


中島 研吾
東京大学情報基盤センター

2025年2月12日

事業項目③: 量子・スパコン連携スケジューラおよび量子・HPC連携ソフトウェアのカップリング技術の研究開発(1/2)

- A) 量子コンピュータ(QC)・スーパーコンピュータ(HPC)連携を効率的に実施するためのジョブスケジューラ(QHScheduler)の研究開発, 整備
- B) 複数の量子コンピュータ・スパコン間でアプリケーションを連携するためのソフトウェア(h3-Open-BDEC/QH)を開発
- 複数のコンポーネントを協調させ, 効率的にデータ転送等を実施するカプラー
 - 異なるアーキテクチャのシステム間の通信・データ転送を実施するライブラリ



事業項目③: 量子・スパコン連携スケジューラおよび量子・HPC連携ソフトウェアのカップリング技術の研究開発(2/2)

C) 量子・スパコン連携アプリケーションのデータ入出力, 通信を模擬するスケルトン型アプリケーションを整備する

- 事業項目④と連携
- 実機(超伝導型, Ion-Trap型)と同じインタフェースを有するシミュレータを整備
- スパコン側のプログラムがC/C++及びFortranの場合に対応

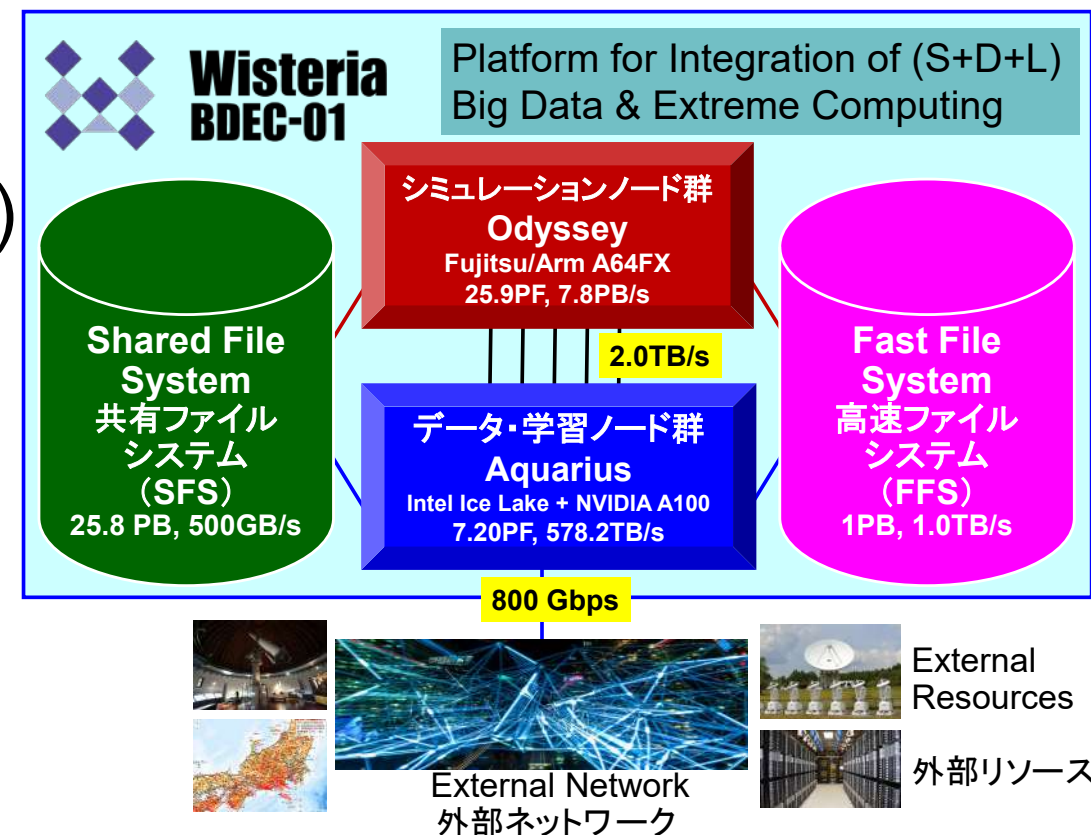
• メンバー

- 中島研吾 全体統括
- 住元真司 B) 連携ソフトウェア(h3-Open-BDEC/QH)
- 胡 曜 A) スケジューラ(QHScheduler), 本プロジェクトで雇用
- 山崎一哉 C) スケルトン型アプリケーション整備

Wisteria/BDEC-01@東大

- 2021年5月14日運用開始
 - 東京大学柏Ⅱキャンパス
- 33.1 PF, 8.38 PB/sec., **富士通製**
 - ~4.5 MVA(空調込み), ~360m²
- Hierarchical, Hybrid, Heterogeneous (h3)
- 2種類のノード群**
 - シミュレーションノード群(S, SIM) : Odyssey**
 - 従来のスパコン
 - Fujitsu PRIMEHPC FX1000 (A64FX), 25.9 PF**
 - 7,680ノード(368,640 コア), 20ラック, Tofu-D
 - データ・学習ノード群(D/L, DL) : Aquarius**
 - データ解析, 機械学習
 - Intel Xeon Ice Lake + NVIDIA A100, 7.2 PF**
 - 45ノード(Ice Lake:90基, A100:360基), IB-HDR
 - 外部リソース(ストレージ, サーバー, センサーネットワーク他)に直接接続
 - ファイルシステム: 共有(大容量) + 高速

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



h3-Open-BDEC

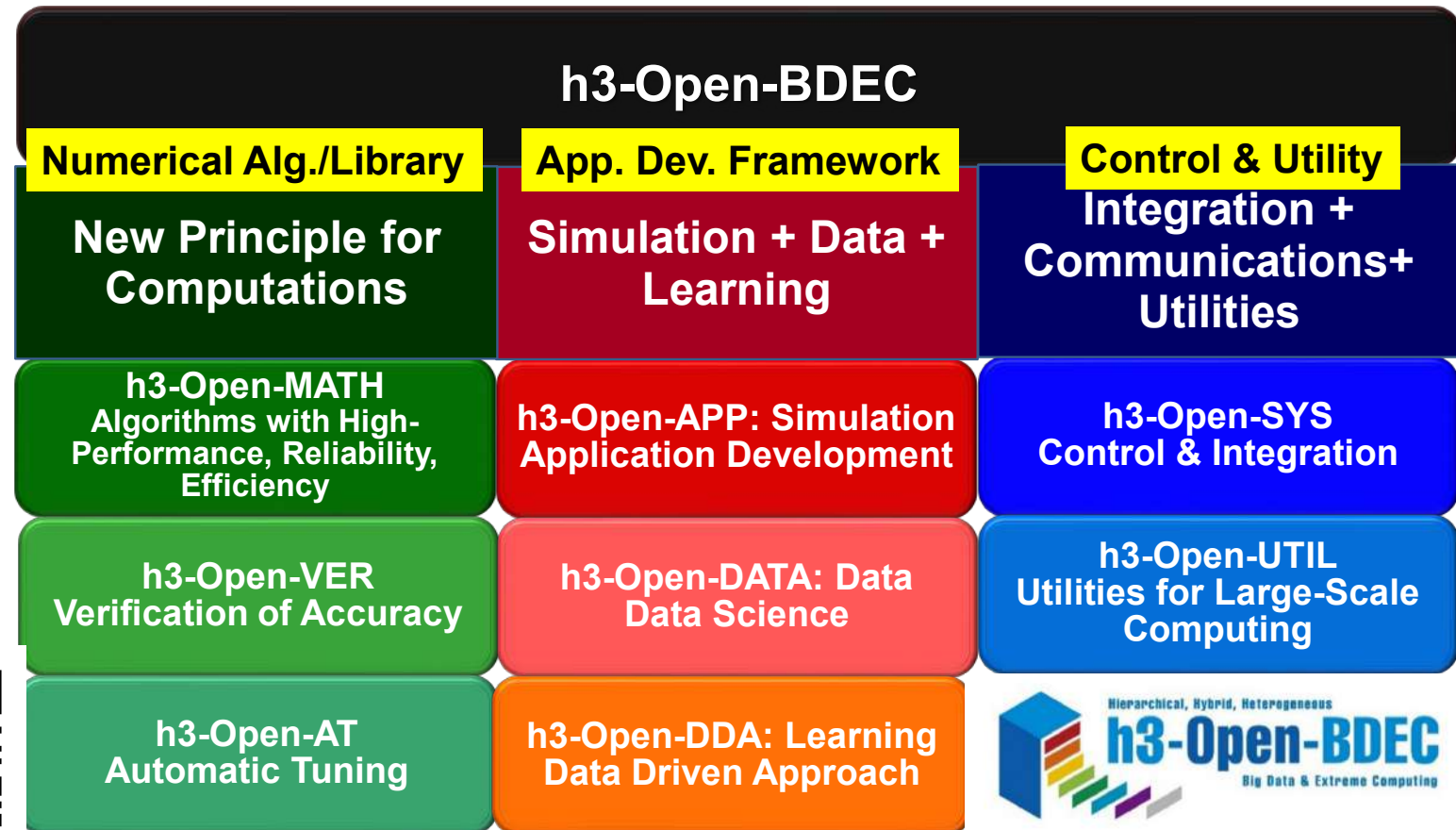
「計算＋データ＋学習」融合を実現する革新的ソフトウェア基盤
科研費基盤研究(S)(2019年度～23年度, 代表: 中島研吾)

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

Hierarchical,
Hybrid,
Heterogeneous

Big Data &
Extreme
Computing

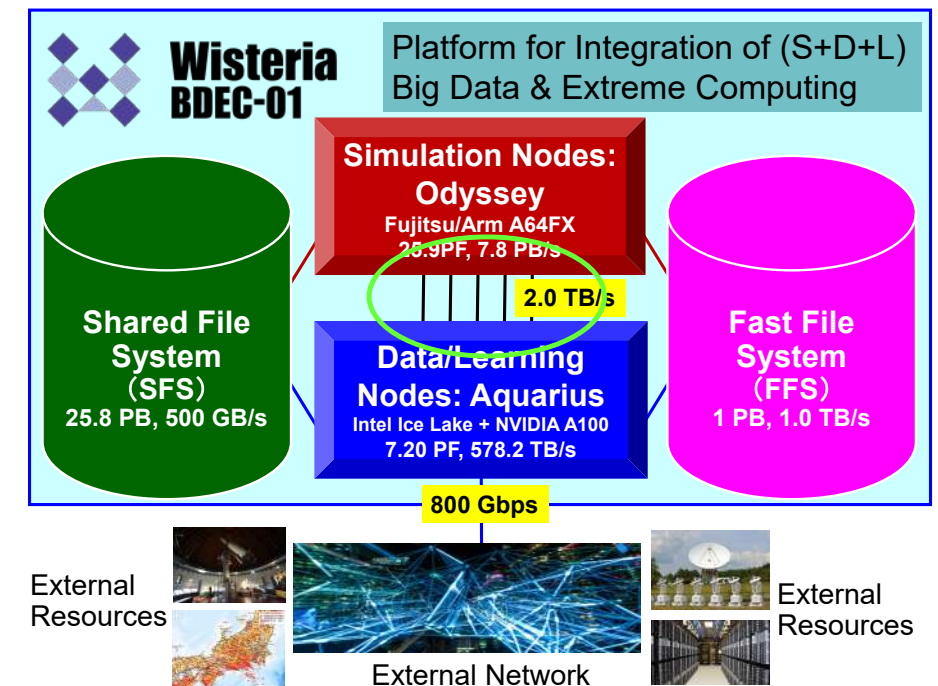
- ① 変動精度演算・精度保証・自動チューニングによる新計算原理に基づく革新的数値解法
- ② 階層型データ駆動アプローチ等に基づく革新的機械学習手法
- ③ ヘテロジニアス環境(e.g. Wisteria/BDEC-01)におけるソフトウェア, ユーティリティ群



h3-Open-SYS/WaitIO

データ受け渡しライブラリ 〔住元他, HPC-181, 2021〕

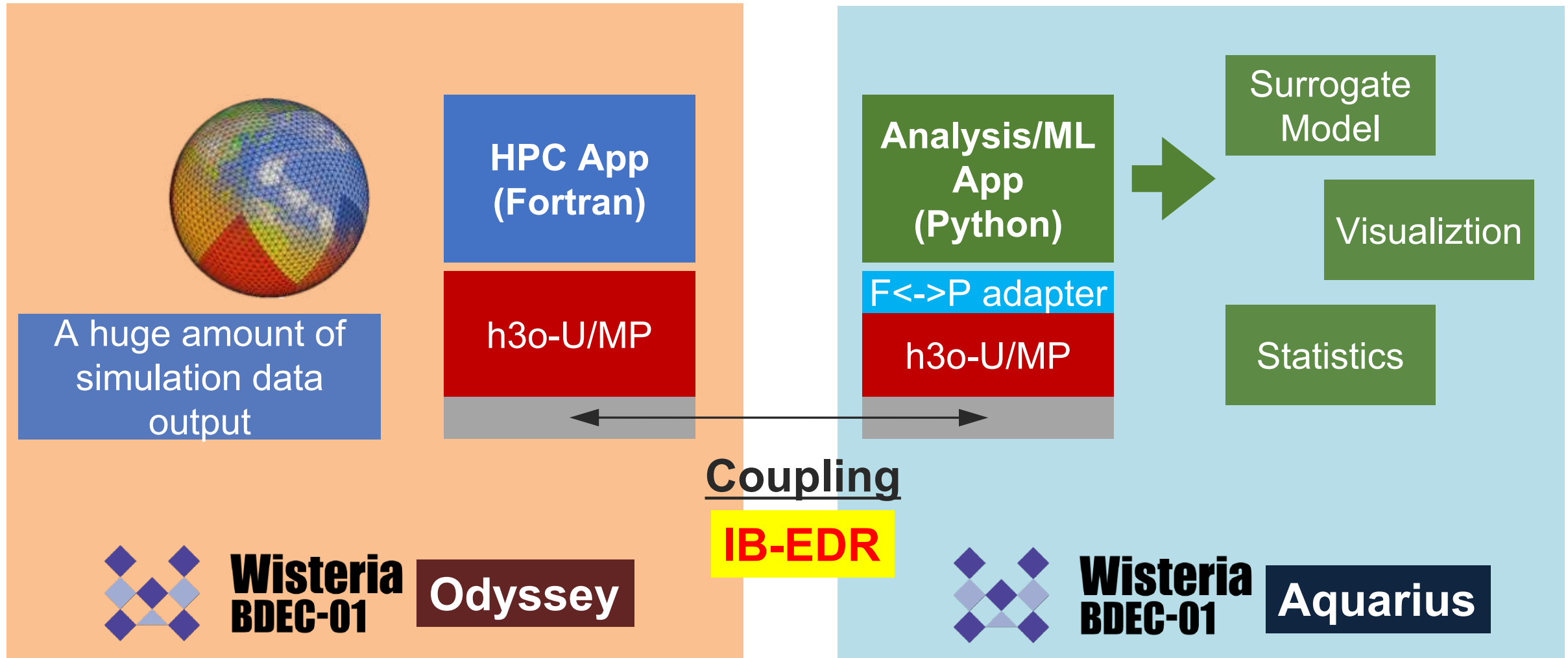
- ヘテロジニアス環境下での異なるコンポーネント間ファイル経由連携ライブラリとして考案
- 機能
 - ✓ Odyssey～Aquarius間連携
 - IB-EDR経由通信 (WaitIO-Socket)
 - ファイル経由 (WaitIO-File)
 - ✓ 外部からのデータ取得 (観測データ等)
 - ✓ 読み込み・書き出しの同期
- API: C/C++, Fortranから呼び出し可能
 - ✓ MPIライクなインタフェースを提供



WaitIO API	Description
waitio_isend	Non-Blocking Send
waitio_irecv	Non-Blocking Receive
waitio_wait	Termination of waitio_isend/irecv
waitio_init	Initialization of WaitIO
waitio_get_nprocs	Process # for each PB (Parallel Block)
waitio_create_group waitio_create_group_wranks	Creating communication groups among PB's
waitio_group_rank	Rank ID in the Group
waitio_group_size	Size of Each Group
waitio_pb_size	Size of the Entire PB
waitio_pb_rank	Rank ID of the Entire PB

h3-Open-UTIL/MP (h3o-U/MP)

計算・機械学習ワークロードを連成



Miyabi (1/2)

2025年1月運用開始



筑波大学
University of Tsukuba



東京大学
THE UNIVERSITY OF TOKYO

• Miyabi-G: CPU+GPU: NVIDIA GH200

– NVIDIA GH200 Grace-Hopper Superchip

- Grace: 72c, 3.456 TF, 120 GB, 512 GB/sec (LPDDR5X)
- H100: 66.9 TF DP-Tensor Core, 96 GB, 4,022 GB/sec (HBM3)
 - CPU-GPU間はキャッシュコヒーレント
 - Grace側からH100のメモリに直接アクセス可能: AMR等CPUが得意な処理
- NVMe SSD for each GPU: 1.9TB, 8.0GB/sec, GPUDirect Storage

– 合計 (CPU+GPU)

- 1,120 nodes, 78.8 PF, 5.07 PB/sec, IB-NDR 200

• Miyabi-C: CPU Only: Intel Xeon Max 9480 (SPR)

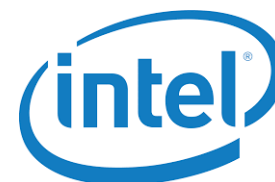
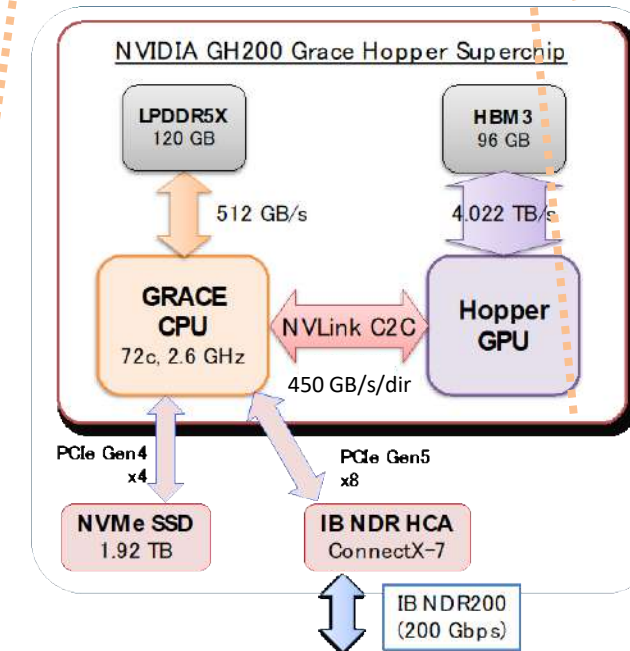
- Intel Xeon Max 9480 (1.9 GHz, 56c) x 2
 - 6.8 TF, 128 GiB, 3,200 GB/sec (HBM2e only)

– 合計

- 190 nodes, 1.3 PF, IB-NDR 200
- 372 TB/sec for STREAM Triad (Peak: 608 TB/sec)



NVIDIA



Miyabi (2/2)

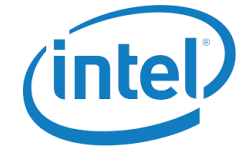
2025年1月運用開始



筑波大学
University of Tsukuba



東京大学
THE UNIVERSITY OF TOKYO



NVIDIA®



- ファイルシステム: DDN EXA Scalar, Lustre FS
 - 11.3 PB (NVMe SSD) 1.0TB/sec, “Ipomoea-01” (26 PB) も利用可能
- 全計算ノードはフルバイセクションバンド幅で接続
 - $(400\text{Gbps}/8) \times (32 \times 20 + 16 \times 1) = 32.8 \text{ TB/sec}$
- 2025年1月運用開始、Miyabi-G/C間の通信は h3-Open-SYS/WaitIO により実現

IB-NDR(400Gbps)		
IB-NDR200(200)		IB-HDR(200)
Miyabi-G NVIDIA GH200 1,120 78.2 PF, 5.07 PB/sec	Miyabi-C Intel Xeon Max (HBM2e) 2 x 190 1.3 PF, 608 TB/sec	File System DDN EXA Scaler 11.3 PB, 1.0TB/sec

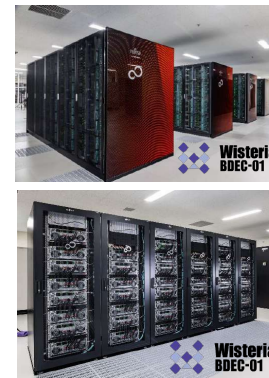
Ipomoea-01
Common Shared Storage
26 PB



	Site	Computer/Year Vendor	Cores	$R_{\max}$ (PFLOPS)	R_{peak} (PFLOPS)	GFLOPS/W	Power (kW)
1	<u>El Capitan, 2024, USA</u> DOE/NNSA/LLNL	HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS	11,039,616	1,742.00 (=1.742 EF)	2,746.38 63.4 %	58.99	29,581
2	<u>Frontier, 2021, USA</u> DOE/SC/Oak Ridge National Laboratory	HPE Cray EX235a, AMD Optimized 3 rd Gen. EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11	9,066,176	1.353.00	2,055.72 65.8 %	54.98	24,607
3	<u>Aurora, 2023, USA</u> DOE/SC/Argonne National Laboratory	HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel	9,264,128	1,012.00	1,980.01 51.1 %	26.15	38,698
4	<u>Eagle, 2023, USA</u> Microsoft	Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR	2,073,600	561.20	846.84 66.3 %		
5	<u>HPC 6, 2024, Italy</u> Eni S.p.A.	HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9	3,143,520	477.90	606.97 66.3 %	56.48	8,461
6	<u>Fugaku, 2020, Japan</u> R-CCS, RIKEN	Fujitsu PRIMEHPC FX1000, Fujitsu A64FX 48C 2.2GHz, Tofu-D	7,630,848	442.01	537.21 82.3 %	14.78	29,899
7	<u>Alps, 2024, Switzerland</u> Swiss Natl. SC Centre (CSCS)	HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11	2,121,600	434.90	574.84 75.7 %	61.05	7,124
8	<u>LUMI, 2023, Finland</u> EuroHPC/CSC	HPE Cray EX235a, AMD Optimized 3 rd Gen. EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11	2,752,704	379.70	531.51 71.4 %	53.43	7,107
9	<u>Leonard, 2023, Italy</u> EuroHPC/Cineca	BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64GB, Quad-rail NVIDIA HDR100	1,824,768	241.20	306.31 78.7 %	32.19	7,494
10	<u>Tuolumne, 2024, USA</u> DOE/NNSA/LLNL	HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS	1,161,216	208.10	288.88 72.0 %	61.45	3,387
13	<u>Venado, 2024, USA</u> DOE/NNSA/LANL	HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11	481,440	98.51	130.44 75.5 %	59.29	1,662
16	<u>CHIE-3, 2024, Japan</u> SoftBank, Corp.	NVIDIA DGX H100, Xeon Platinum 8480C 56C 2GHz, NVIDIA H100, Infiniband NDR400, Ubuntu 22.04.4 LTS	163.200	91.94	138.32 66.5 %		
17	<u>CHIE-2, 2024, Japan</u> SoftBank, Corp.	NVIDIA DGX H100, Xeon Platinum 8480C 56C 2GHz, NVIDIA H100, Infiniband NDR400, Ubuntu 22.04.4 LTS	163.200	89.78	138.32 64.9 %		
18	<u>JETI, 2024, Germany</u> EuroHPC/FZJ	BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, RedHat Linux, Modular OS	391,680	83.14	94.00 88.4 %	63.43	1,311
22	<u>CEA-HE, 2024, France</u> CEA	BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail BXI v2, EVIDEN	389,232	64.32	103.48 62.2 %	52.17	1,233
28	<u>Miyabi-G, 2024, Japan</u> JCAHPC	Fujitsu, Supermicro ARS 111GL DNHR LCC, Grace Hopper Superchip 72C 3GHz, Infiniband NDR200, Rocky Linux	80,640	46.80	72.80 64.3 %	47.59	983
36	<u>TSUBAME 4.0, 2024, Japan</u> Institute of Science Tokyo	HPE Cray XD665, AMD EPYC 9654 96C 2.4GHz, NVIDIA H100 SXM5 94 GB, Infiniband NDR200	172,800	39.62	61.60 64.3 %	48.55	816
58	<u>Wisteria/BDEC-01 (Odyssey), 2021, Japan</u> U.Tokyo	Fujitsu PRIMEHPC FX1000, A64FX 48C 2.2GHz, Tofu D	368,640	22.12	25.95 85.2 %	15.07	1,468

全体計画と展望(1/2)

- 事業項目(①, ②)とは密に連携する
- 2025年度末まで
 - 量子(QC)・スパコン(HPC)連携の達成
 - まずは, 1-QC, 1-HPCの連携
 - 可能な限り複数HPCからQCを利用可能なところまで持っていく
 - 2025年度は事業項目①②で作成中のスケジューラを引き継ぐ
 - QC, HPCともに「優先キュー」的な特別なリソースグループを想定
 - 2023年度・2024年度はWisteria/BDEC-01上でプロタイプ・SW整備
 - Odyssey=HPC, Aquarius=QCとして容易に疑似環境を構築可能



	2023年度	2024年度	2025年度
量子・スパコン連携スケジューラ QHScheduler	設計・プロトタイプ開発・実装 試作・設計	スパコン上での予備検討・設計 実装・動作試験	量子コンピュータ実機・スパコン連携環境での動作試験を行い、試験運用 実機での動作試験・統合テスト
量子・スパコン連携アプリ開発基盤 h3-Open-BDEC/QH	スパコン上での予備検討・設計 試作・設計	プロトタイプ開発・検証, スケジューラと連携した動作試験 実装・動作試験	SINET環境上でのソフトウェア開発, スケジューラと連携した動作試験 動作試験

項目A): スケジューラ

• 2024年度実施内容

– (A-1) 2025年度末向けスケジューラ(HPC: 優先利用グループ前提)(事業項目①②)

- 現状は「富岳」向け

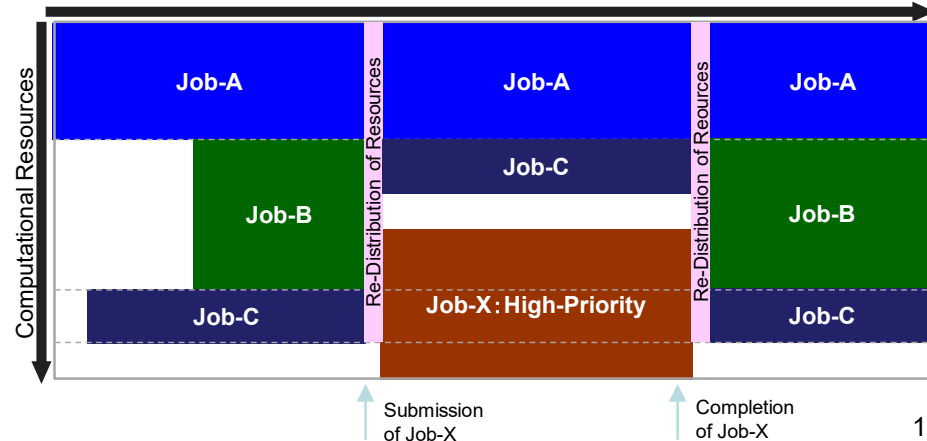
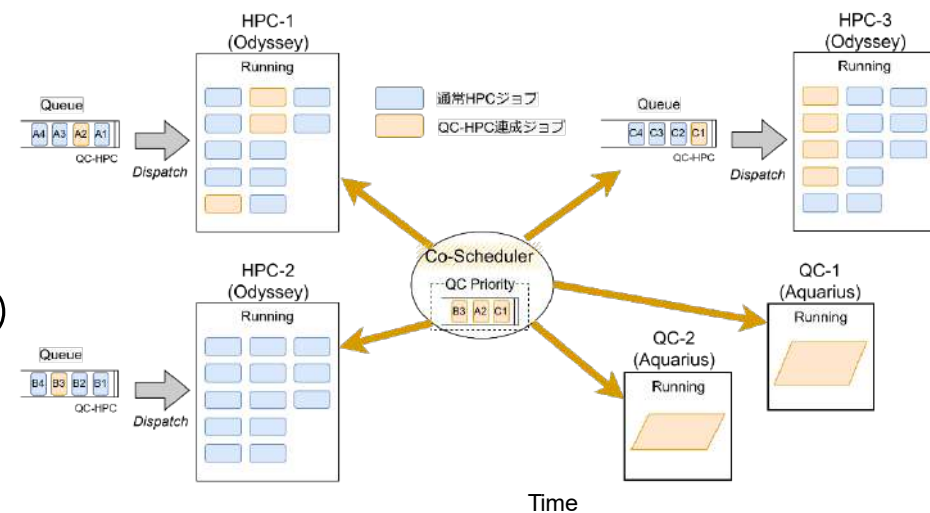
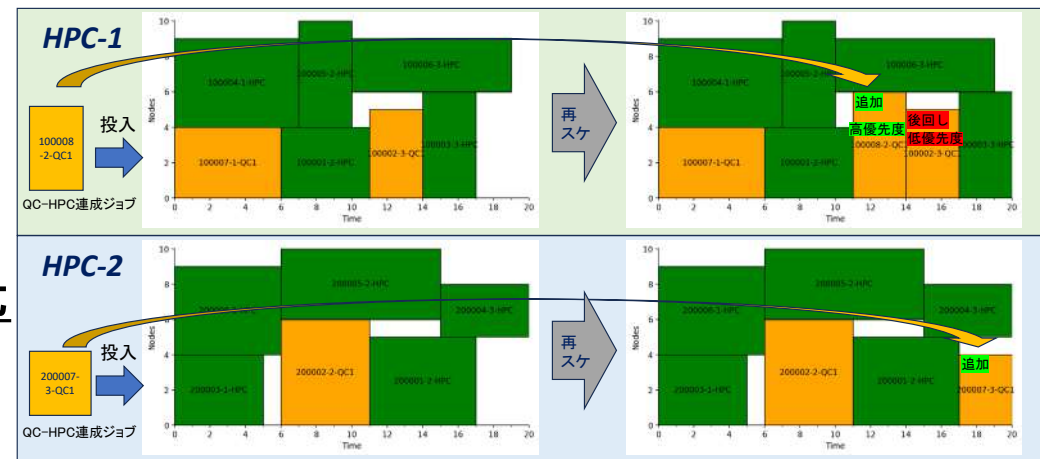
– (A-2) 2026年度以降向け簡易シミュレータ

- 複数QC, 複数HPCの連携可能
 - 複数QC=1-リアルQC+シミュレータ
- フレキシブルなリソースマネジメント手法開発(HPC側)
 - チェックポイントリスタート必要
- 情報処理学会HPC研究会(2025年3月)発表予定

– (A-3) Wisteria/BDEC-01上のシミュレータ環境

• 2025年度

- (A-1) 引き継ぎ⇒Miyabi, Wisteria/BDEC-01, 複数HPCのための簡易Co-Scheduler整備(mdx等)
- (A-2) (A-3): 2026年度以降向け研究開発を継続



項目B): 連携ソフトウェア

- QC-HPC間の通信とデータ転送をオンラインおよびリアルタイムで効率的に実装および統合

- 2024年度実施内容

- (B-1) WaitIO/Router

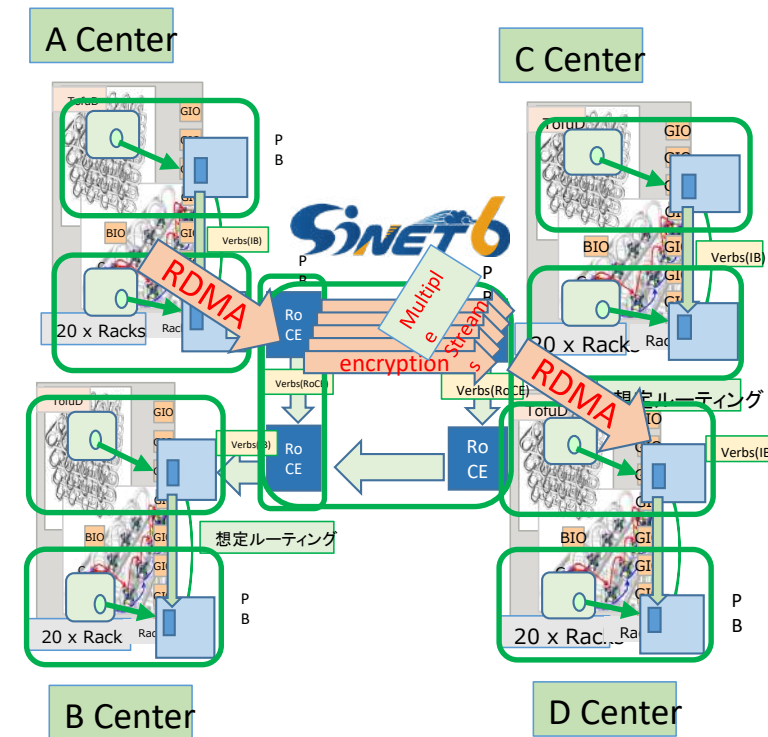
- Sinet6結合したHPC/QCセンター群を高速・安全に結合
 - Wisteria/BDEC-01を対象にセンター内向けプロトタイプ整備, センター間機能予備設計
 - HPC Asia 2025 WSで発表予定(2025年2月)

- (B-2) QC-HPCカップリング

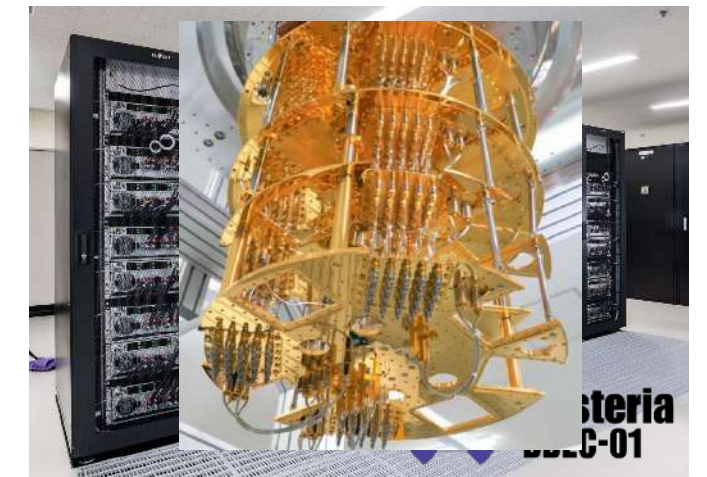
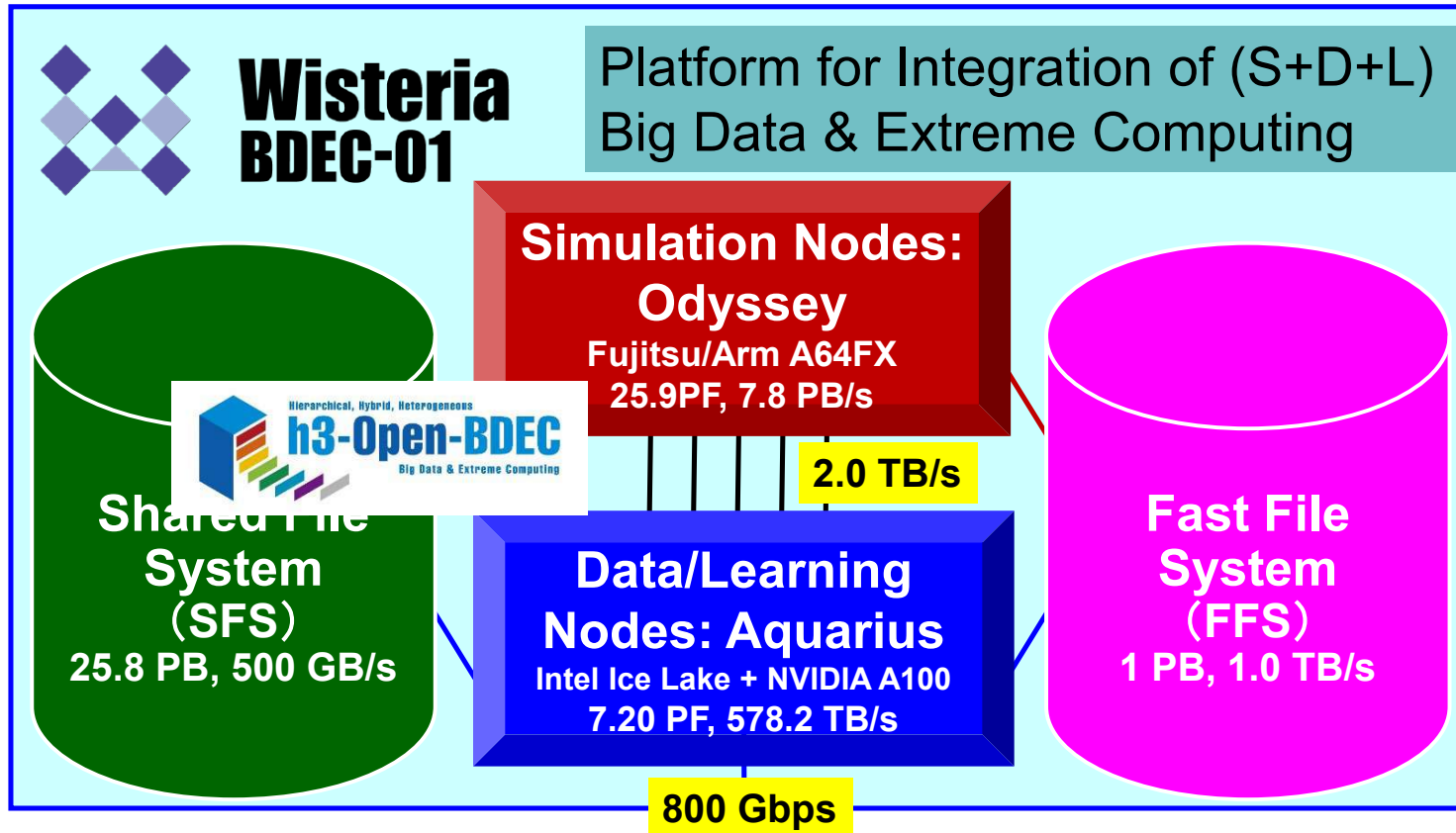
- WaitIO4Py(Python向けインタフェース), 簡易カプラーによる疑似QC-HPCハイブリッド連携環境をWisteria/BDEC-01上に構築, チュートリアル実施(次頁)
 - 情報処理学会量子ソフトウェア研究会(2025年3月)発表予定

- 2025年度

- (B-1) 機能拡張(RDMA-ソケットルータ, ソケットマルチストリーム化, セキュリティ強化機能)⇒理研(富岳, IBM, Quantinuum)-東大(Wisteria/BDEC-01, Miyabi)結合
 - (B-2) 項目A), C)と連携, QC-HPCハイブリッド連携環境への統合



Odyssey-AquariusにQC-HPCハイブリッドシミュレーション環境を構築



External Resources



External Network



External Resources

『CUDA-Q+Wisteria/BDEC-01+h3-Open-BDEC』で『量子・HPCハイブリッド』を体験してみよう

- CUDA-Q
 - NVIDIAが開発している量子古典ハイブリッド計算のためのオープンソースプラットフォーム
- 疑似QC-HPCハイブリッド環境でVQE (Variational Quantum Eigensolver) を体験
 - Aquarius (GPU) with CUDA-Q: QC
 - Odyssey (CPU): HPC
 - Pythonで記述した簡単なプログラム
 - h3-Open-BDEC
 - WatIOのPython版によるQC-HPC連携
- 2024年12月18日(水)にチュートリアル実施
 - <https://www.cc.u-tokyo.ac.jp/events/lectures/238/>
- 世界で初めての「本当の」QC-HPCがつながったハイブリッド環境(QCは本物ではないが)
- 次回はMiyabi利用(6月頃, QC-HPC HB環境構築)



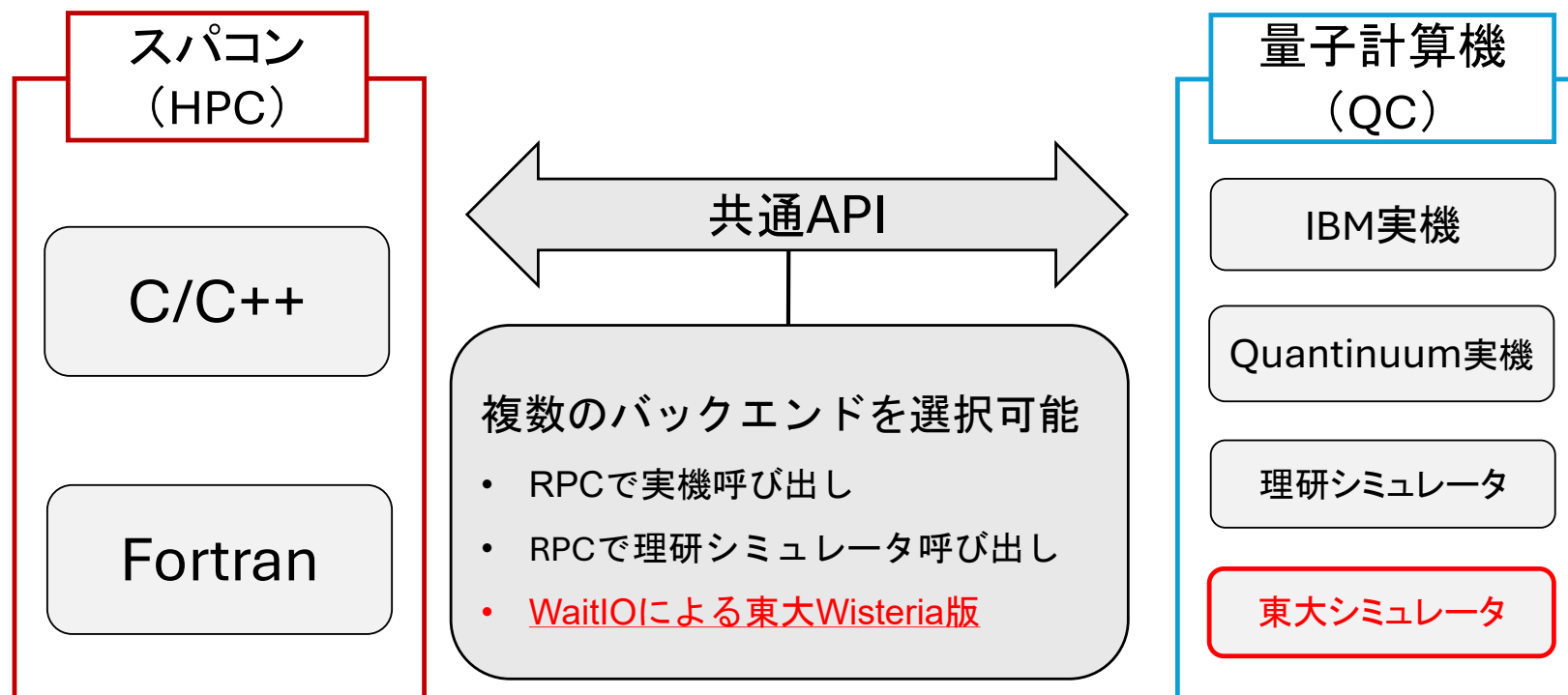
項目C): スケルトン型アプリ・インタフェース

- 2024年度実施内容

- Odyssey上のC/FortranプログラムからAquarius上のPython量子パッケージQiskitを呼ぶためのラッパーを, WaitIOを使って2024年度に実装(項目B)と連携)

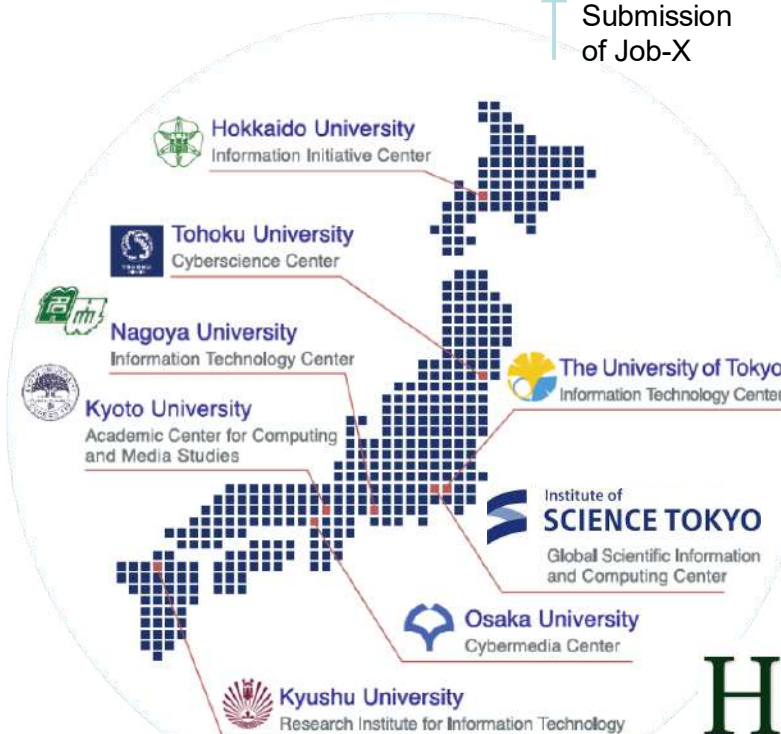
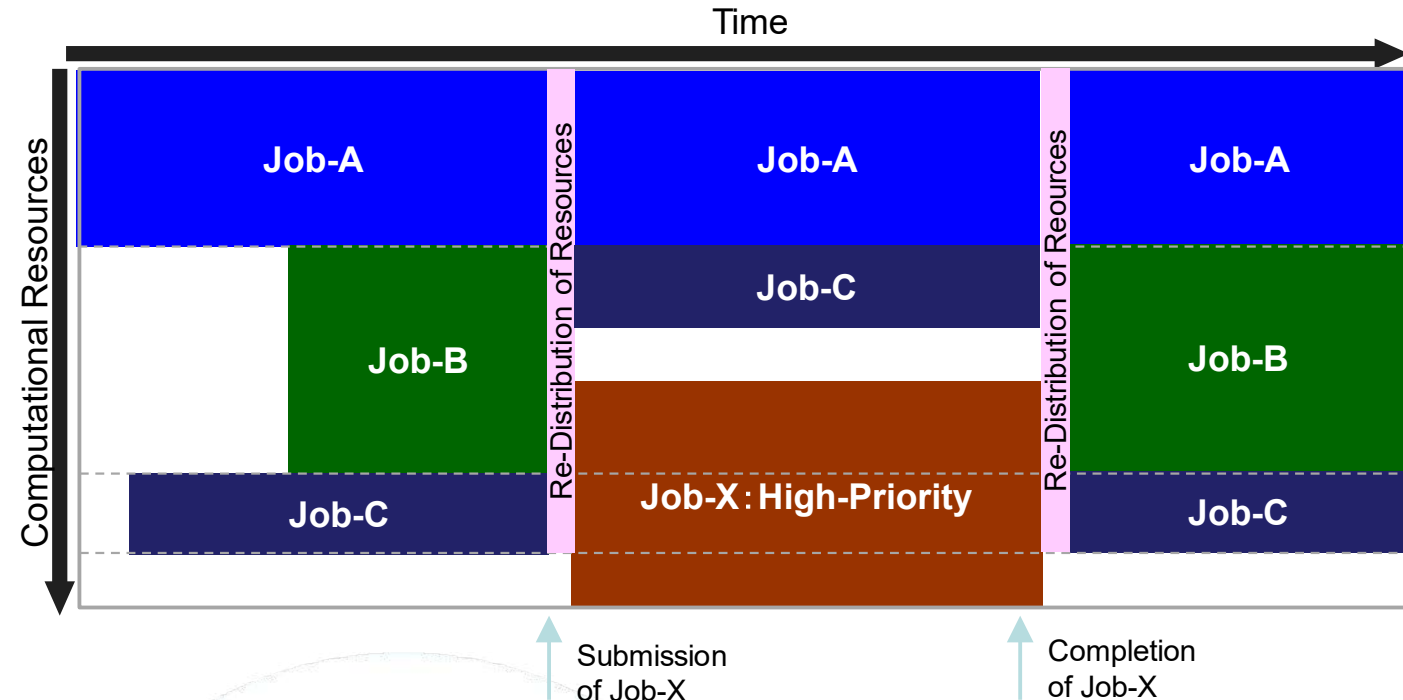
- 2025年度

- プロジェクト共通のAPIのバックエンドの一種として共通APIに統合する
 - プロジェクト共通のAPI: 量子コンピュータ実機も呼び出し可能, 事業項目①②にて開発中



全体計画と展望(2/2)

- 2026年度以降
 - 複数QC, 複数HPCの連携可能
 - 複数QC=1-リアルQC+シミュレータ
 - フレキシブルなリソースマネジメント手法開発(HPC側)
- その他
 - HPCI, JHPCN(学際大規模情報基盤共同利用共同研究拠点)等, 全国のスパコンの共同利用の枠組みで「量子コンピュータも使える」ことをアピール
 - 産業利用枠もある



HPCI High Performance Computing Infrastructure