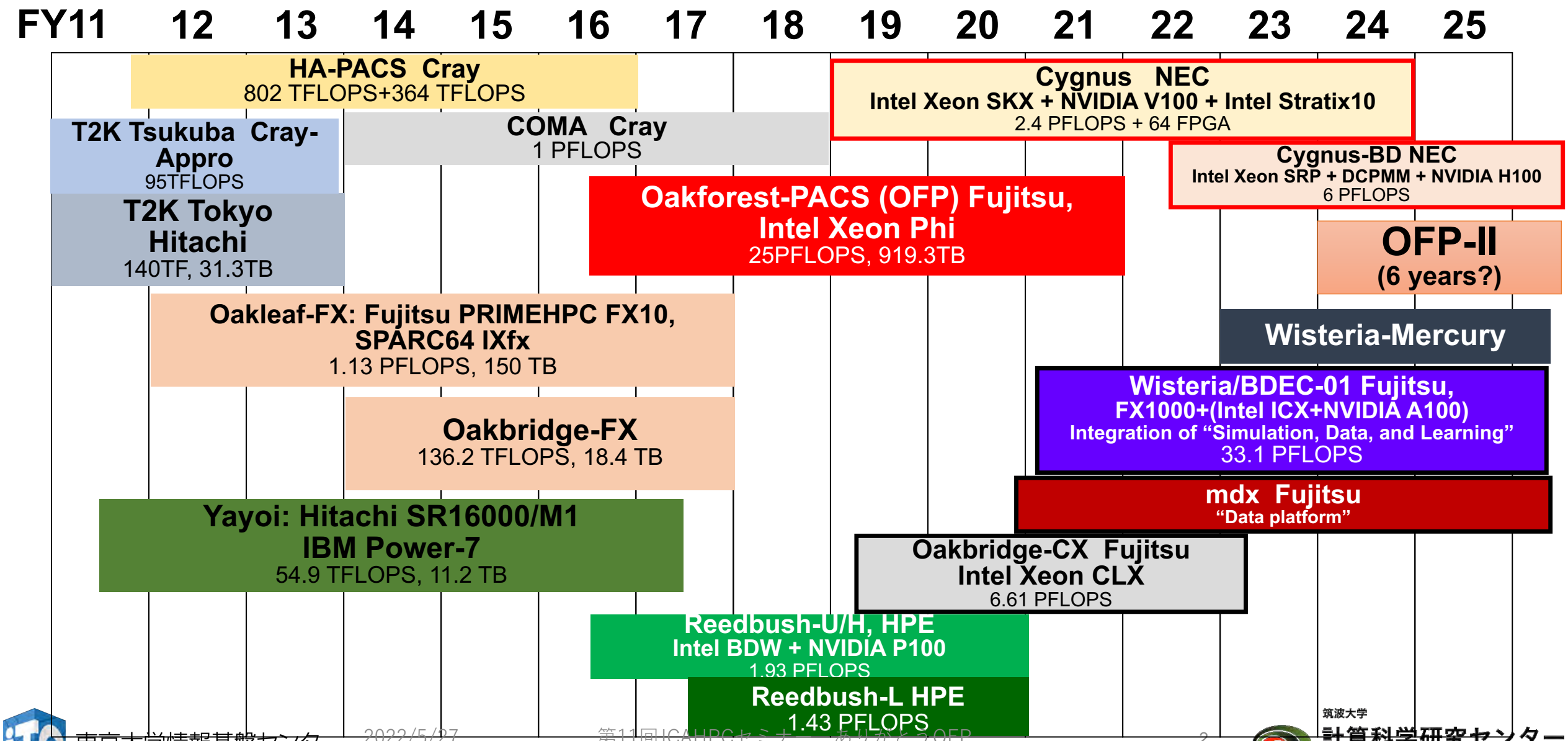
Oakforest-PACS IIに向けて

東京大学 情報基盤センター
最先端共同HPC基盤施設(JCAHPC)
運用支援部門長
(筑波大学計算科学研究センター客員)
塙　敏博

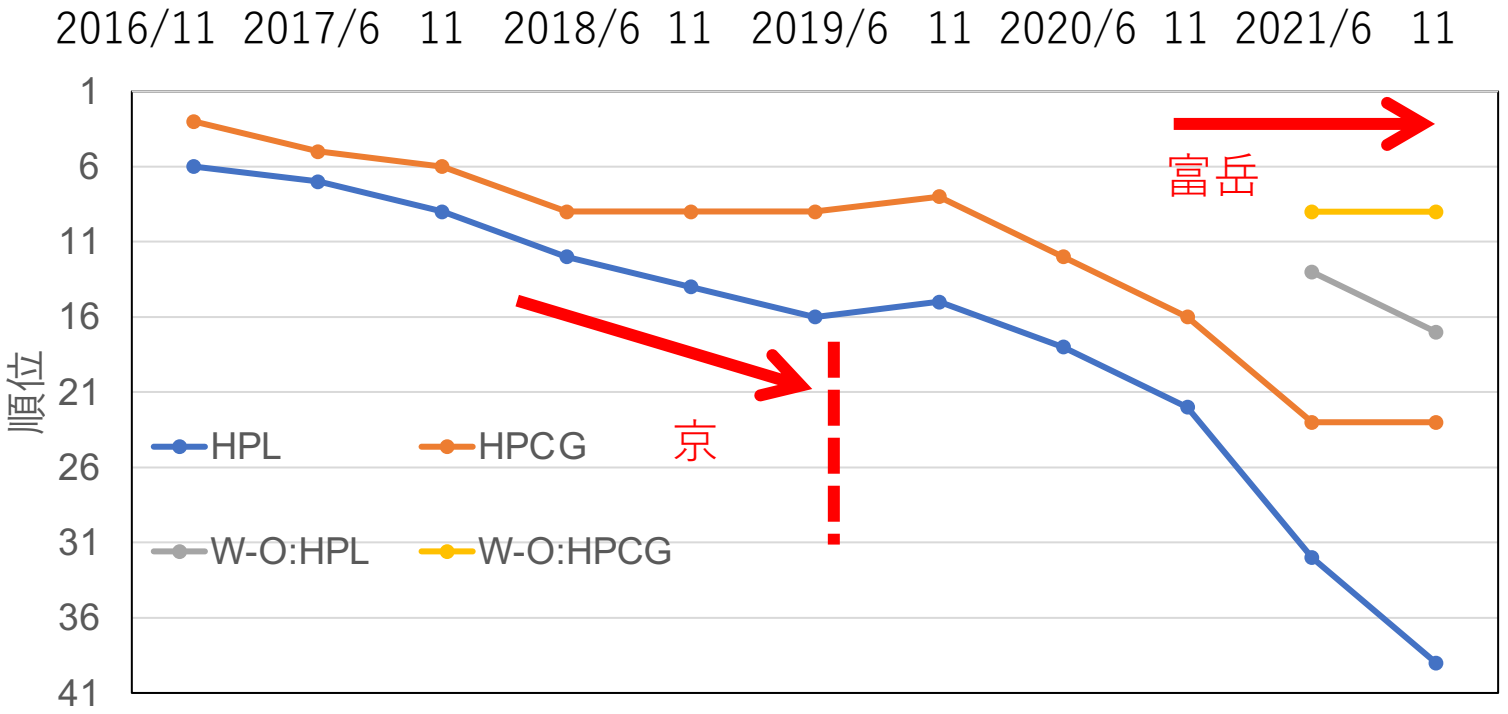
スパコンの変遷@筑波大CCS＆東大ITC



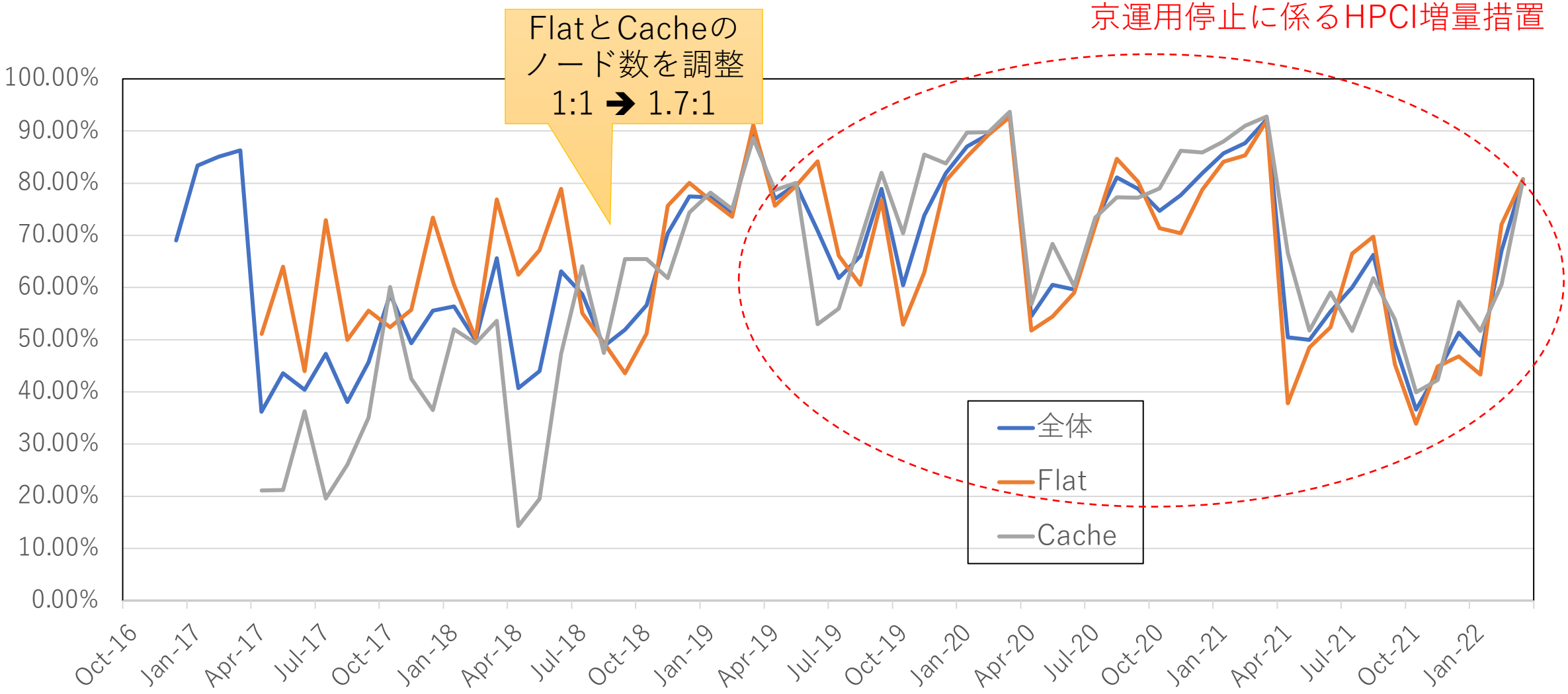
Oakforest-PACSの果たした役割

- OFPは第二階層ではトップクラス
  - 4年半経過でWisteria/BDEC-01に交代
- 京➔富岳移行期の計算資源、大規模ジョブの受け皿として機能
  - 常時2,048ノードジョブ実行可
  - HPCIへの資源提供
    (JCAHPCとして一体で実施)
    - 2017,2018年度 1,600ノード/年
    - 2019年度～ 3,300ノード/年

(参考)2022, 23年度はJCAHPCとしてWisteria-Odysseyを提供
- 2,304ノード/年(2022年度)



Oakforest-PACS利用率推移



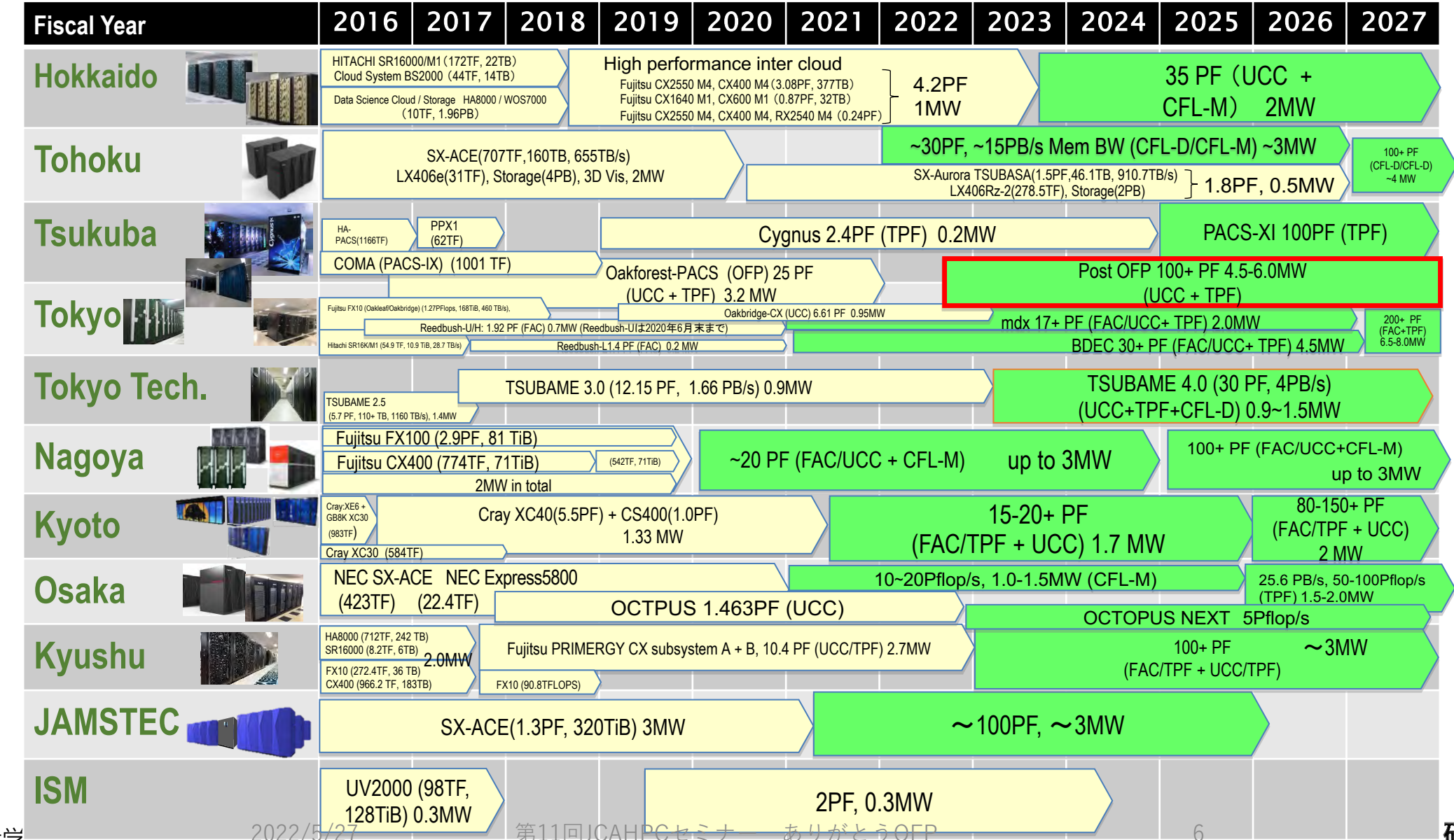
Oakforest-PACS II 計画

- 2019年11月頃： JCAHPCとして2代目も継続して設計・運用することを確認
- 2021年2月：システム名を “Oakforest-PACS II”とすることに決定

- OFPは第2階層システムのトップマシンであった
  - 京=>富岳の移行期に、常時2,000ノードまでのジョブが実行可能という貴重な存在
- OFP2もそうありたい

- OFPの継承：これまでと同様、大規模アプリケーションのユーザを支える
- 新しい使い方: AI for HPCなど
- アプリケーションとのコデザイン

HPCI第2階層システムの開発・整備・運用計画 (2020年11月現在)



Power is the maximum consumption including cooling

HPCI第2階層システム分類

- Flagship-Aligned Commercial Machine (FAC)：フラグシップシステムと同様のマシン
  - フラグシップシステムユーザの多くを抱えるセンターやフラグシップシステムと同様のシステムを整備することによりユーザニーズに合致するだけでなくよりフラグシップシステムへの橋渡しができると判断するセンターが、フラグシップシステム同様のシステムを整備していく。しかし、スパコン調達では、要求性能および要求機能を仕様とし製品固有機能をMUSTとすることはないために、フラグシップシステムと同系列のシステムが入るとは限らない。
- Complimentary Function Leading Machine (CFL-M, CFL-D):フラグシップシステムがカバーできない応用領域を支援するマシン
  - センターが抱えるユーザの応用領域をフラグシップシステムで実行しても必ずしも効率よく実行できるとは限らない。そのようなセンターはユーザニーズに沿ったマシンを設置していく。スパコンメーカの開発動向から従来のスパコン調達で設置する場合（CFL-M)と、ユーザニーズに沿った何らかの開発を含めた調達が考えられる（CFL-D)は、CFL-Dに関しては、その必要性を考慮の上、競争的資金による開発が行われることが望まれる。なお、フラグシップシステムがカバーしない応用領域については、フラグシップシステム開発元が情報開示しないと議論できない。
- Upscale Commodity Cluster Machine (UCC): コモディティクラスタからの大規模並列処理を支援するマシン
  - フラグシップシステムを含むスパコンが研究室レベルにまで下方展開できない限り、研究室レベルではコモディティクラスタが利用され続ける。センターは、そのようなユーザがより大規模並列処理へと向かうような大規模コモディティクラスタを整備していく。
- Technology Path-Forward Machine (TPF): 将来のHPC基盤に向けた先端マシン
  - 既存アプリケーションを動かしたいというレベルのユーザニーズではなく、ユーザ応用分野が要求する計算手法や計算資源量を勘案しながら、市場には投入されていない先端マシンを設計試作し、調達手続きを経てマシンを整備していく。ユーザと共にそのような先端マシン上のアプリケーションを開発していくことになる。さらにこのようなシステムを通じて次の世代のフラグシップシステムへとつながっていくだろう。

OFP(+OBCX)ユーザの継続性：コア数

- メニーコア、汎用CPU
  - OFP: 68コア (Intel Xeon Phi KNL)
  - OBCX: 56コア (Intel Xeon Cascade Lake x2ソケット)
- 想定：ノード当たり数十コア～100コア超
  - 48コア: 富士通A64FX
  - ～78コア: Intel Xeon IceLake x2ソケット
  - ～128コア: AMD EPYC Rome/Milan x2ソケット
  - …

OFPユーザの継続性：メモリBW

- 高バンド幅メモリ
  - OFP: 実効 500 GB/s弱 (MCDRAM, 16GB)
  - DDR4に対して4~5倍のバンド幅，1/6の容量
  - Flatモード & Cacheモード
- 想定：高いメモリバンド幅を維持
  - DDR5+スロット数増加
  - LPDDR5X
  - HBM2, HBM2e, HBM3
  - ノード当たり TB/秒 クラスが望ましい

新しい応用に向けて

- 「計算＋データ＋学習」融合によるシミュレーションの高度化：
  より緊密な連携を可能に
  - データ同化＋シミュレーション
  - 機械学習によるシミュレーションパラメータ推定＋アンサンブル計算
    ➔ AI for HPC
  - Wisteria/BDEC-01 or Cygnus-EC+BD での取り組みを高度化
- 脱炭素化、および省電力化への要請

GPUの導入
（ただし全体ではなく一部ノード）

GPUの導入

- 数千人の利用者のGPUクラスタへの完全移行は困難
  - Wisteria/BDEC-01導入時も，「シミュレーションノード群(Odyssey)」も含めGPUクラスタとする案もあった
    ➡アプリ移行への準備に見通しがつかず（主として人手確保）、断念

- 一部ノードにGPU導入を検討中
  - GPUの多様化：複数ベンダがHPC向けGPUを提供
  - プログラミング環境も多様化：OpenMP+MPIの利用者にもハードルが下がると期待
  - HPCIでのGPU需要を支える側面も考慮

OFP-IIの設計ポイント

- CPUノード数とGPUノード数のバランス
  - CPUノード：1,000ノード規模のジョブを常時流せる，OFPの規模感を維持したい
    - OFPのノード数： 8,208ノード，Flatモード/Cacheモードに分けて運用
  - GPUノード：多いに越したことはない，が
    - きっと電力より予算がボトルネックになる，，
  - 参考：NERSC Perlmutter
    - CPUノード: 3,072ノード，AMD EPYC Milan 2ソケット
    - GPUノード: 1,536ノード，Milan 1ソケット+ NVIDIA A100 x4基
- GPUとCPU間の接続
  - キャッシュコヒーレントなら細粒度でのオフロードが可能になり、GPUへ移行し易く
  - 一方で、CPUとGPUが不可分になり、選択肢が狭まる

インターコネクト、ストレージ

インターコネクト

- OFP: OmniPath Architecture 100 Gbps

✓OFP-IIでは400~800 Gbpsクラスが利用可能

➢トポロジ：フルバイセクションを維持するか？ 数千ノード程度なら問題ない

ストレージ

- OFP: Burst Bufferによるファイルキャッシュ + Lustre

✓ 階層ストレージを検討

- 第1階層：不揮発性メモリ、ストレージクラスメモリ
- 第2階層：All flash並列ファイルシステム
- 第3階層：センターワイド共有ファイルシステム ➔ Ipomoea-01をすでに導入

設置条件、スケジュール

- スケジュール
  - 2024年4月稼働開始
- 設置場所：柏キャンパス 第2総合研究棟 1F+2F
  - 1F: Oakbridge-CX　(~2023年6月?)
  - 2F: Oakforest-PACS (~2022年3月)
- 利用可能な電力(冷却除く)
  - 1F: 1.6 MVA
  - 2F: 3.3 MVA + 0.4 MVA (UPS)
- 電力性能が 50+ GFLOPS/W 達成できれば
  単純計算で ≒ 250 PFLOPS ➔ 目標(私見): 200+ PFLOPS

対象アプリケーションとのコデザイン

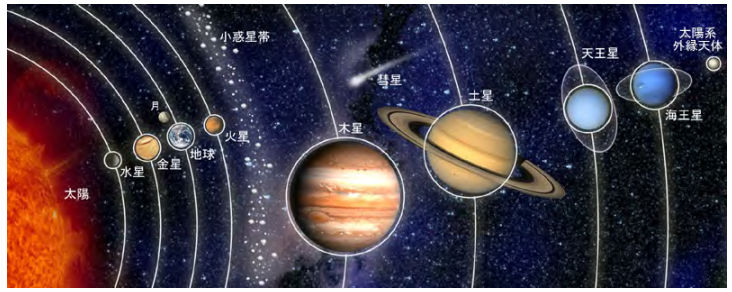
JHPCN（学際大規模情報基盤共同利用・共同研究拠点）課題

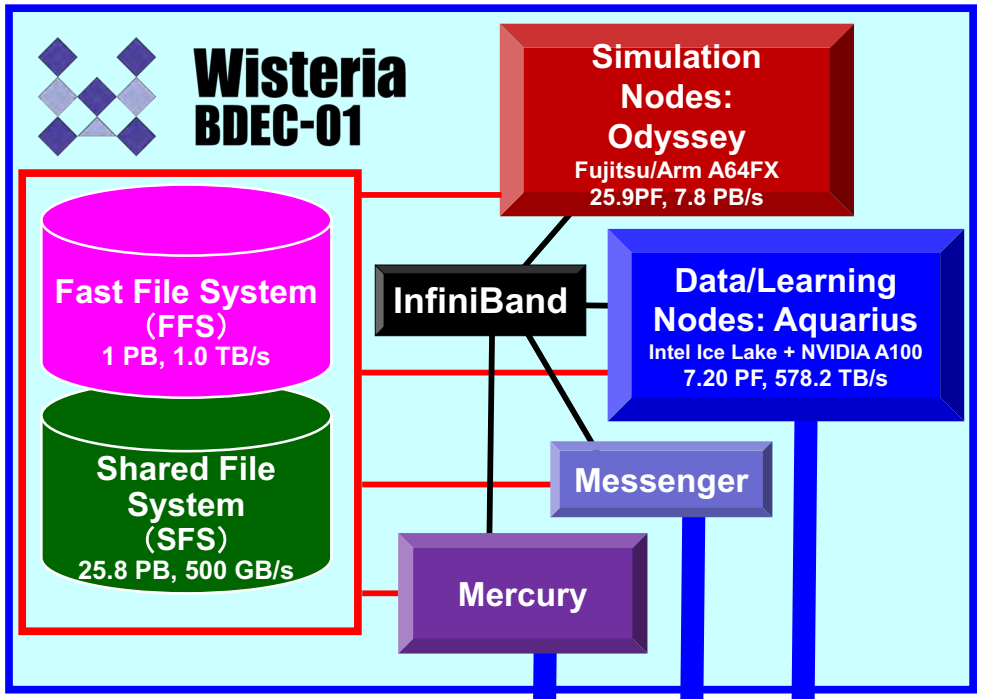
- 次世代演算加速装置とそのファイルIOに関する研究
  （代表：塙@東大，副代表：建部@筑波大）
  - GPU上データの直接ファイル IO、あるいは計算とファイル IO のオーバラップの両者を容易に取り扱い可能にする手法を確立
  - 実アプリケーションにおいて GPU-ファイルIO間の効率的な処理を実現
    - 宇宙物理、City LES、機械学習
- 大規模アプリケーションの高性能な実用的アクセラレータ対応手法
  （代表：下川辺@東大，額田@筑波大）
  - 既存のスパコン向けアプリをGPUスパコンに移植
  - 指示文などをベースに最小限の修正で、可搬性に考慮しつつ高い性能を目指す

OFP-IIの設計にフィードバック
OFP-II調達時のベンチマーク作成

OFP-IIプロトタイプとしてのWisteria-Mercury

- 当初、Wisteria/BDEC-01 Aquarius 「データ・学習」ノードを拡充するためのCPUのみによるサブシステムとして計画
- OFP-II向けアプリ準備，GPU対応のためのテストベッド
- GPU検討のため各社にベンチマーク依頼
  - 7種，計算科学系



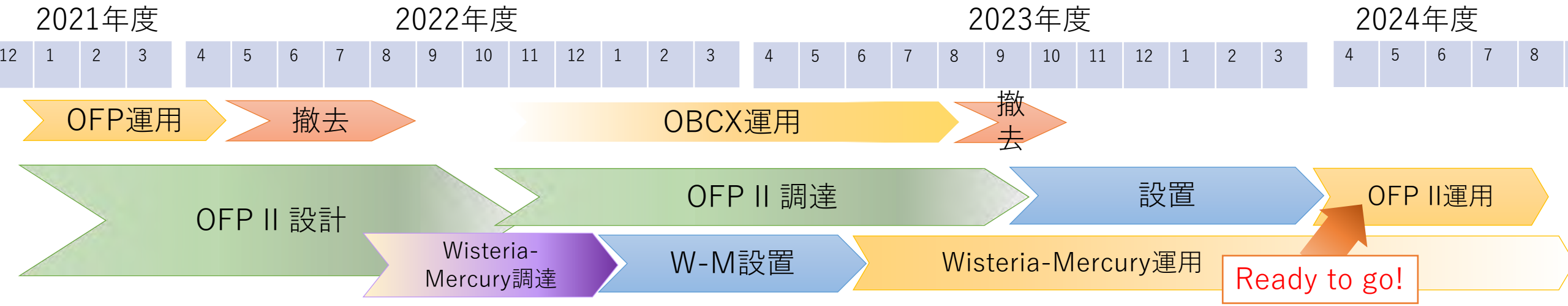


A	CPU版のコードに対してホストCPUでの性能を評価	A-1	提供コードをそのまま実行（As-Is）
		A-2	最適化したコードを使用
B	CPU版のコードをGPU化	B-1	OpenACC, OpenMP, Standard Language等の手法でGPU化
		B-2	当該GPUで最大限の性能が出せるようチューニング
C	GPU化済みのコードを最適化	C-1	提供コードまたは既存コードを実行
		C-2	当該GPUで最大限の性能が出せるようチューニング

おわりに

Oakforest-PACS II (OFP-II)構想

- 2024年4月稼働を目指して設計中
- 200+ PFLOPS，汎用メニーコアCPU + 一部GPU
- HPCI第二階層をリードするUCC+TPFマシンとして設計・開発
  - アプリケーションとのコデザインを進める
- Wisteria-Mercuryを用いて既存アプリケーションのGPU化を推進



（注）矢印は正確な期間を表すものではありません