OFPの概要と導入について
～その生い立ちからシステム運用まで～

朴　泰祐
JCAHPC・副施設長／筑波大学計算科学研究センター・センター長

OFP前史～T2K

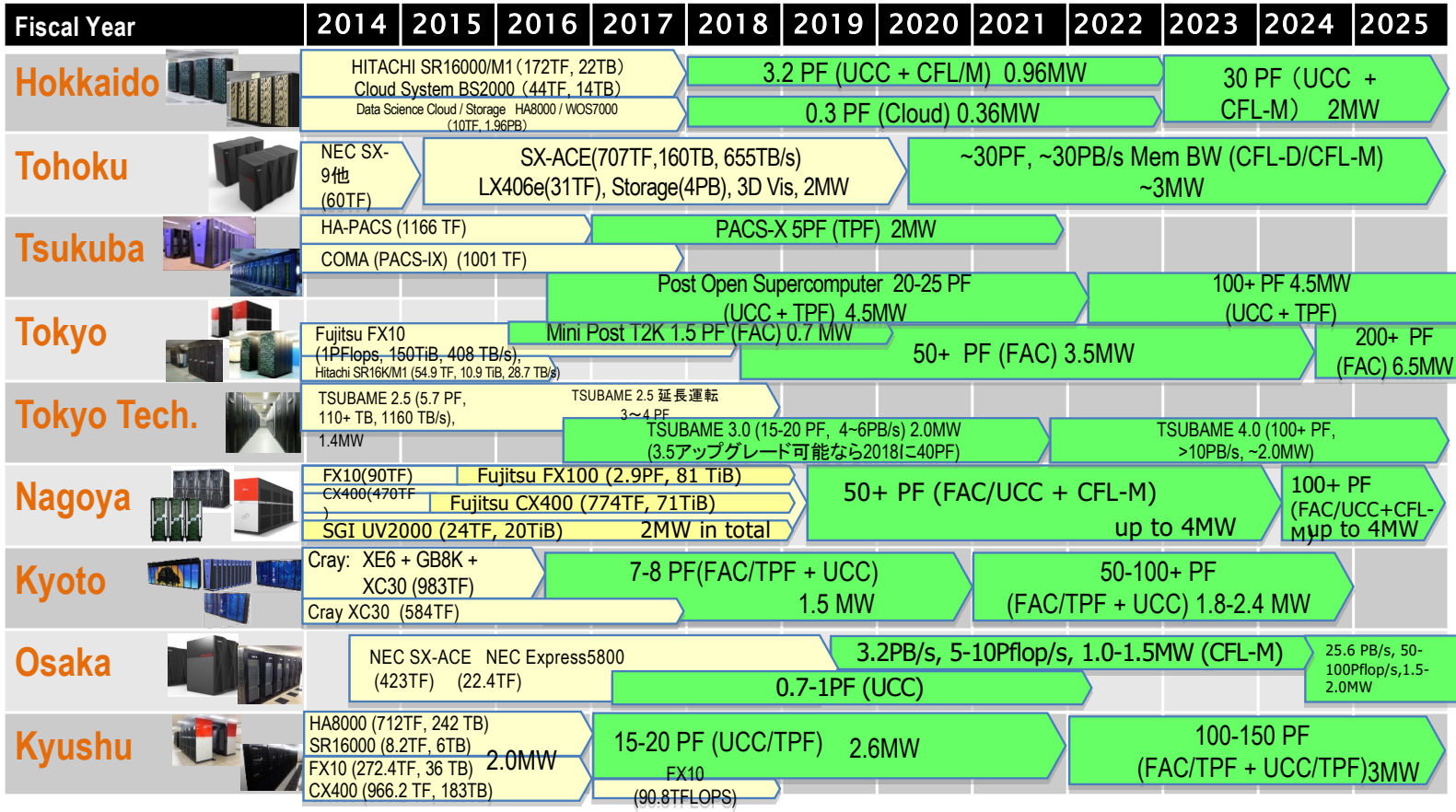
- T2K Open Supercomputer Alliance
  - Tsukuba, Tokyo, Kyoto の3大学による共通基本アーキテクチャに基づくマルチコアCPUクラスタによる基盤センタースーパーコンピュータ調達と連携運用・共同研究の枠組み
  - Vendor Schedule Driven ⇨ Technology Driven によるスーパーコンピュータ調達の走り
    - AMD Opteron quad-core x 4 (16-core/node)
    - InfiniBand or Myrinet x 4
  - 結果：日本のTOP4マシン中、国内で1・2・4位にランクイン（Jun.2008 TOP500）（世界では16・20・34位）
  - これ以降、多くの基盤センターが technology driven 調達を開始

T2KからJCAHPCへ

- 次のフェーズの3センターの調達は京大が従来の4年リースのスケジュールを守り非同期に
  ⇨ しかし technology driven による調達は継続
  - Tsukuba, Tokyo の2大学は次のステップとしてさらなる連携を ⇨ JCAHPC
- JCAHP:OFP が目指したもの
  - Technology Driven で次世代の many-core architecture をベースにクラスタを導入すれば K Computer を超えることが可能
  - 物量として、2大学の調達資金では難しい
    ⇨ T2Kより強力な alliance = 予算を合算した共同調達（当時は post-T2K と呼ばれていた）
    ⇨ K Computer を抜いて日本最高性能となるシステムが見えてきた
  - ターゲットの technology = KNL (Knights Landing)
    （同時期、Kyoto も同じ技術に着目）

JCAHPC

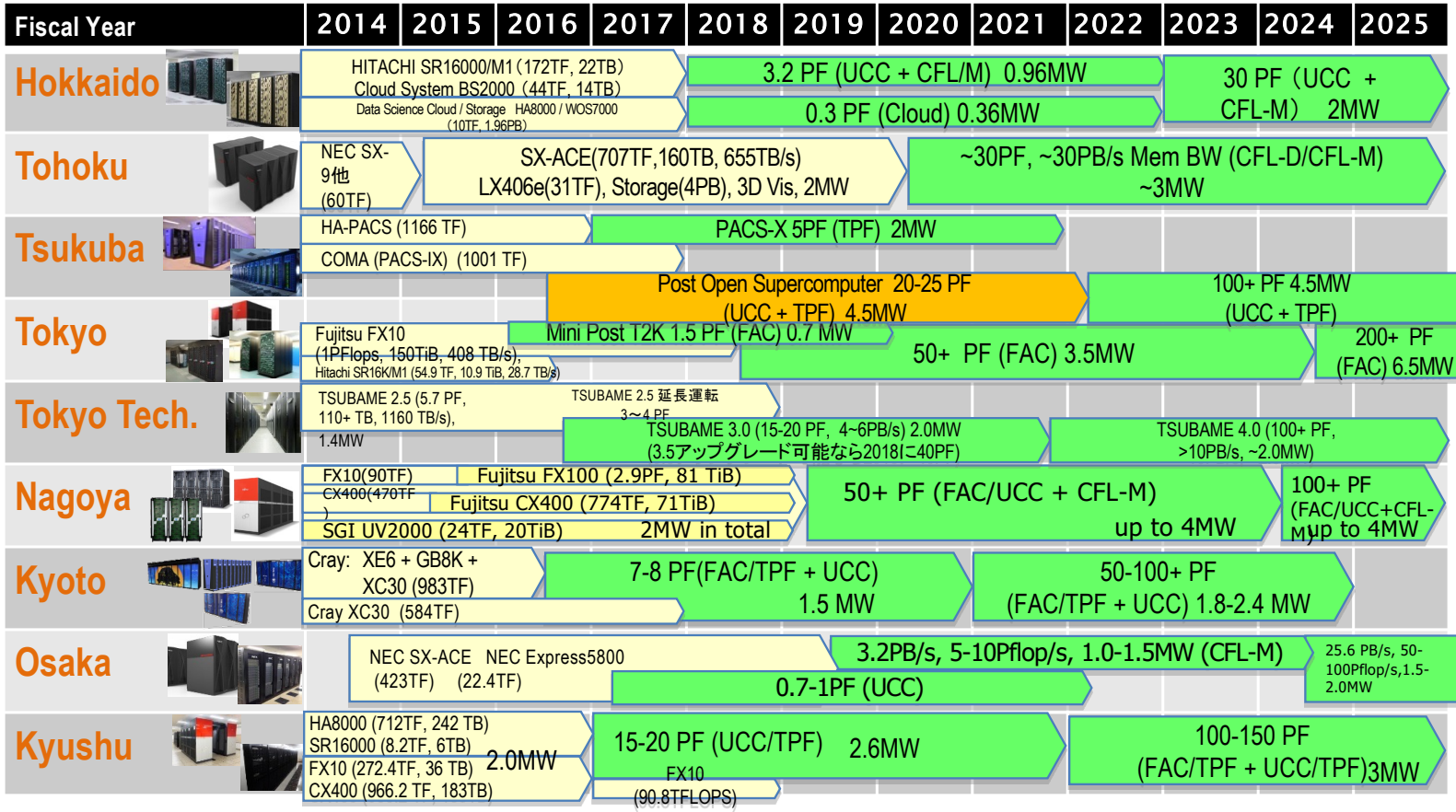
情報基盤センター運用＆整備計画 (2015年9月時点)



電力は最大供給量（空調システム含む）

JCAHPC

情報基盤センター運用＆整備計画 (2015年9月時点)



電力は最大供給量（空調システム含む）

JCAHPC

産みの苦しみ

- technology driven であるということは、その technology が市場にちゃんと出てきてくれなければいけないということ
- technology を待ちながら...
  - 資料招請（CFI）までは両大学は独立に行ったがその後は仕様を統一し、いよいよ共同調達を行う体制に
  - その後、諸般の事情により2年間を経て意見招請（CFR、仕様書原案）を実行
    ⇨ JCAHPCとして両大学共同調達
  - 2016年にシステム導入できる目処が立ち、入札（CFP）へ
    ⇨ 富士通による落札
  - 2016年10月にシステム導入を開始、12月からフルシステムでの試験運用を開始
    ⇨ TOP500 Nov. 2016 に登録
- ネーミング
  - 時間切れ？
    １年ぐらい考えたが，結局，東大が考えていた柏（Oakforest）と筑波大伝統のPACSを合わせた
  - 長い名前は失敗？ ⇨ 新聞は縦書き，かつカタカナ

Oakforest-PACS (OFP)

- 2016年12月1日稼働開始
- 8,208 Intel Xeon/Phi (KNL)，ピーク性能25PFLOPS
  - 富士通が構築
- **TOP 500初出　6位（国内1位），HPCG 3位（国内2位）**
- **最先端共同HPC 基盤施設(JCAHPC: Joint Center for Advanced High Performance Computing)**
  - 東京大学情報基盤センター
  - 筑波大学計算科学研究センター
    - 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

TOP500 list on Nov. 2016 (#48)

Machine	Architecture	Country	Rmax (TFLOPS)	Rpeak (TFLOPS)	MFLOPS/W
TaihuLight, NSCW	MPP (Sunway, SW26010)	China	93,014.6	125,435.9	6051.3
Tianhe-2 (MilkyWay-2), NSCG	Cluster (NUDT, CPU + KNC)	China	33,862.7	54,902.4	1901.5
Titan, ORNL	MPP (Cray, XK7: CPU + GPU)	United States	17,590.0	27,112.5	2142.8
Sequoia, LLNL	MPP (IBM, BlueGene/Q)	United States	17,173.2		2176.6
Cori, NERSC-LBNL	MPP (Cray, XC40: KNL)	United States	14,014.7		???
Oakforest-PACS, JCAHPC	Cluster (Fujitsu, KNL)	Japan	13,554.6	25,004.9	4985.1
K Computer, RIKEN AICS	MPP (Fujitsu)	Japan	10,510.0	11,280.4	830.2
Piz Daint, CSCS	MPP (Cray, XC50: CPU + GPU)	Switzerland	9,779.0	15,988.0	7453.5
Mira, ANL	MPP (IBM, BlueGene/Q)	United States	8,586.6	10,066.3	2176.6
Trinity, NNSA/ LABNL/SNL	MPP (Cray, XC40: KNL)	United States	8,100.9	11,078.9	1913.7

Cori (50.2%)を上回る
54.2%のHPL効率

Oakforest-PACS (OFP)

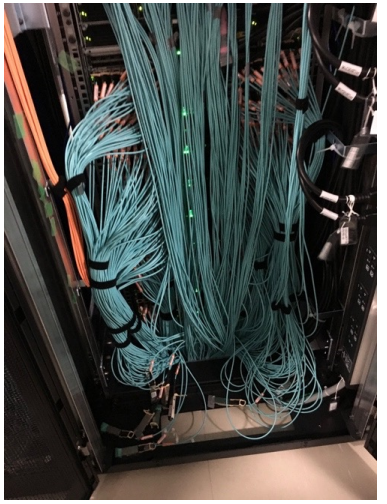
- ピーク性能25 PFLOPS
- 8208 KNL CPUs
- OmniPathによるFBB Fat-Tree
- HPL 13.55 PFLOPS
  2016/11:
  国内第1位
  世界第6位
- HPCG 0.385PFLOPS
  （2.8% of HPL)
  2016/11: 世界第3位
- Green500
  2016/11: 世界第6位
- IO500 (best throughput)
  2017/11: 世界第1位

計算ノードとシャーシ

Computation node (Fujitsu PRIMERGY CX1640 M1)
with single chip Intel Xeon Phi (Knights Landing, 3+TFLOPS)
and Intel Omni-Path Architecture card (100Gbps)

Chassis with 8 nodes, 2U size
(Fujitsu PRIMERGY CX600 M1)

水冷パイプ、リアパネル冷却、OPA、ファイルサーバ



JCAHPC

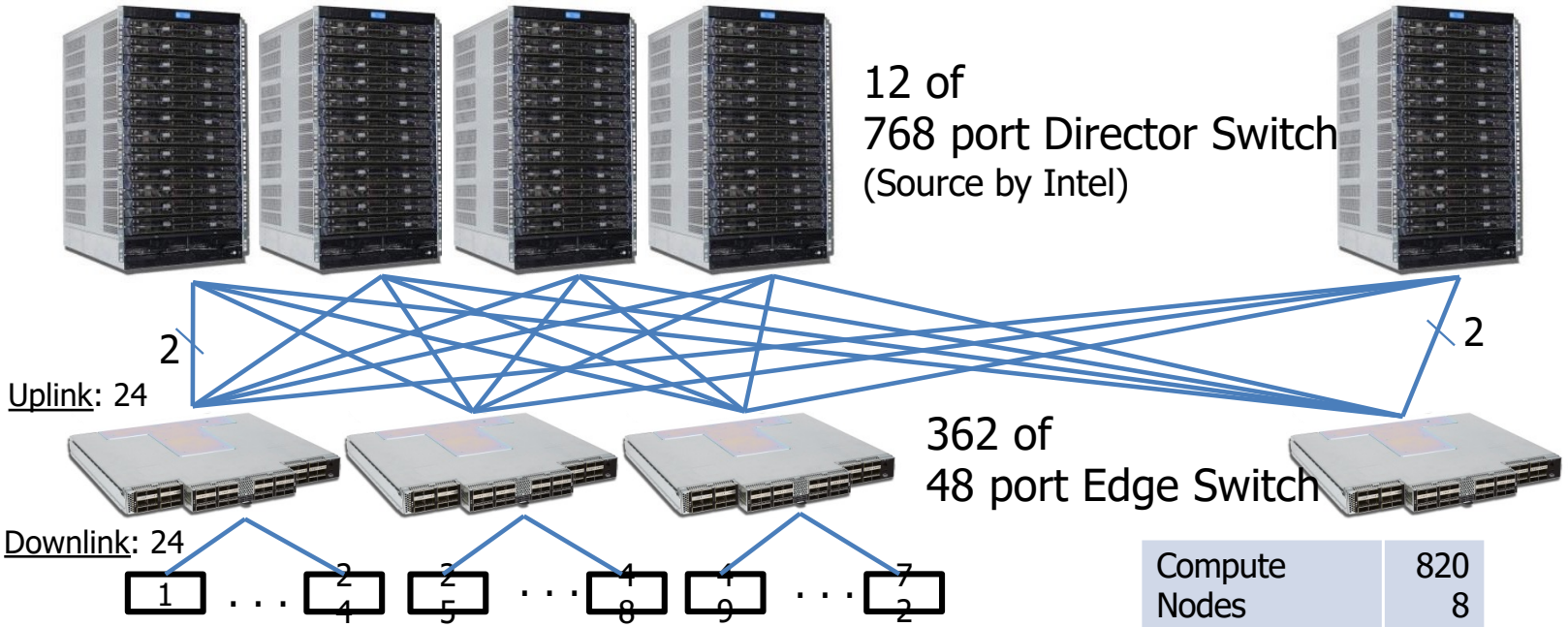
Oakforest-PACSのシステム仕様

Total peak performance			25 PFLOPS
Total number of compute nodes			8,208
Compute node	Product		Fujitsu Next-generation PRIMERGY server for HPC (PRIMERGY CX1640 M1)
	Processor		Intel® Xeon Phi™ (Code name: Knights Landing), 68 cores
	Memory	MCDRAM	16 GB, > 400 GB/sec (effective rate)
		DDR4	96 GB, 115.2 GB/sec (DDR4-2400 x 6ch, peak rate)
Inter-connect	Product		Intel® Omni-Path Architecture
	Link speed		100 Gbps
	Topology		Fat-tree with (completely) full-bisection bandwidth
Login node	Product		Fujitsu PRIMERGY RX2530 M2 server
	# of servers		20
	Processor		Intel Xeon E5-2690v4 (2.6 GHz 14 core x 2 socket)
	Memory		256 GB, 153 GB/sec (DDR4-2400 x 4ch x 2 socket)

Oakforest-PACSのI/O仕様

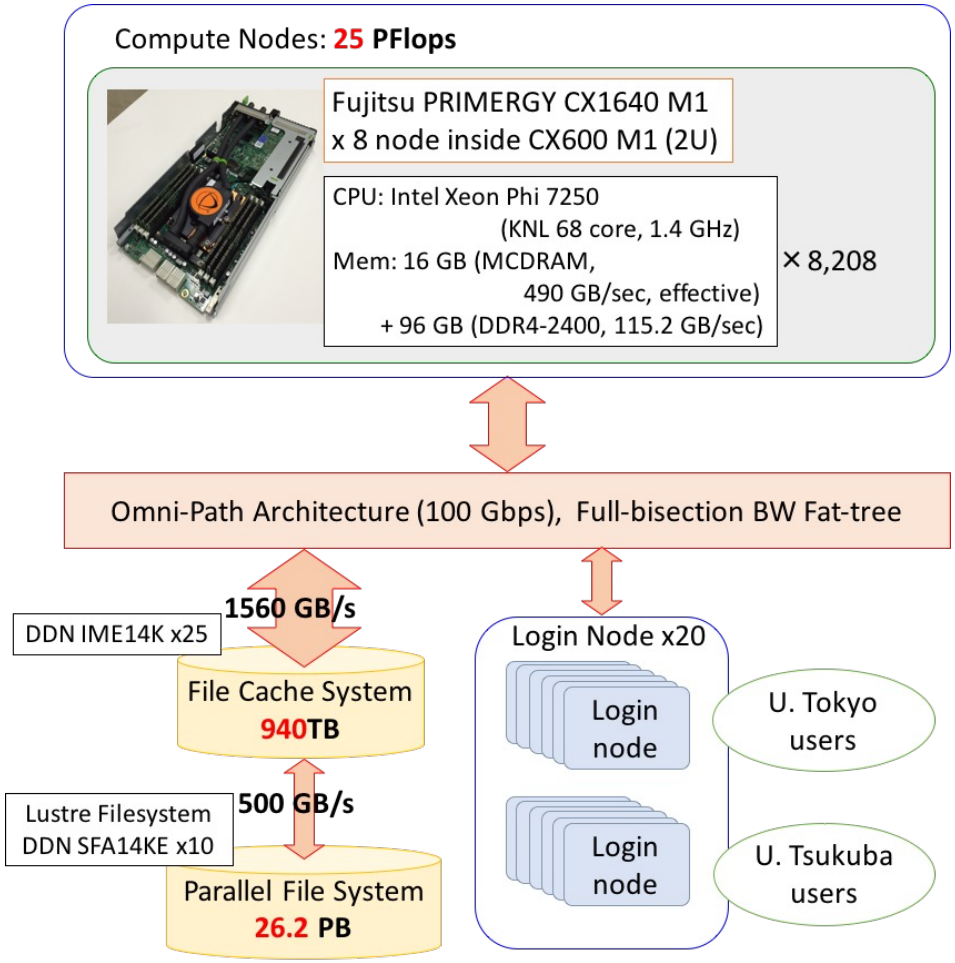
Parallel File System	Type		Lustre File System
	Total Capacity		26.2 PB
	Meta data	Product	DataDirect Networks MDS server + SFA7700X
		# of MDS	4 servers x 3 set
		MDT	7.7 TB (SAS SSD) x 3 set
	Object storage	Product	DataDirect Networks ES14K
		# of OSS (Nodes)	10 (20)
		Aggregate BW	500 GB/sec
Fast File Cache System	Type		Burst Buffer, Infinite Memory Engine (by DDN)
	Total capacity		940 TB (NVMe SSD, including parity data by erasure coding)
	Product		DataDirect Networks IME14K
	# of servers (Nodes)		25 (50)
	Aggregate BW		1,560 GB/sec

Intel® Omni-Path Architectureによる フルバイセクションバンド幅の相互結合網



Compute Nodes	8208
Login Nodes	20
Parallel FS	64
IME	200
Mgmt, etc.	8
Total	**8600**

利用者から見たシステム



Oakforest-PACSの特徴

- 最先端のHPC向けメニーコアCPUの利用
  - 極めて高い並列性（ノード内，ノード間）を持つHPCアプリケーションが主要ターゲット（3TFLOPS）
  - 高度なチューニングにより性能を大幅に向上可能
  - 初期プログラム移植の容易さ（OpenMP+MPI）
- 最先端の高性能相互結合網をfull-bisectionバンド幅で装備
  - 100Gbpsのリンク速度
  - ジョブスケジューラによるノード配置の自由度が高い
  - ノード位置にかかわらず全てのファイルを高速にアクセス可能
  - 超並列アプリケーションを容易に実行可能
- 高性能並列ファイルシステムとバーストバッファ
  - Lustreによる高並列・高性能アクセス（500GB/s）
  - ファイルキャッシュ（バーストバッファ）による1TB/s越えの超高速アクセス

運用状況

- 2016/12~2017/3 試験運用（無償）
  - システム安定稼働チェック
  - 機能・性能の確認
  - （特別）大規模HPCチャレンジ：GBP
- 2017/4~ 公開運用
  - HPCI, 各大学の個別運用プログラム
  - 実運用だがユーザによってはここでチューニング開始
  - 試験運用で実績を積んだグループはジャンプスタート（素粒子、光物性等）
  - 月末のメンテナンス前に大規模HPCチャレンジとして約24時間の全系占有利用
- 稼働状況
  - 実稼働率（メンテナンス、停電等を除く）：実質ほぼ100%
  - 利用率：50～80%程度（時期による）

運用：KNLのメモリモード

- メモリモード
  - Cache:Flat = 50:50 (4096+4096 nodes)
  - 特定時期の需要に応じて月単位で変更
- 動的メモリモード変更
  - オンデマンドでCache:Flatの比率を緩やかに動的変更することを計画していた
  - ノードのリブートが予想より時間がかかること，ネットワークの安定動作のため見送り
- 大規模ジョブ（最大2048ノード）が通常利用可能
  - メモリモードの関係もあり、リソース確保が難しい
  - キューイングの自由度に制限（fair shareを実行していない）
  - 利用率の抑制原因なのでは？⇒ さらに解析・改善を続けた

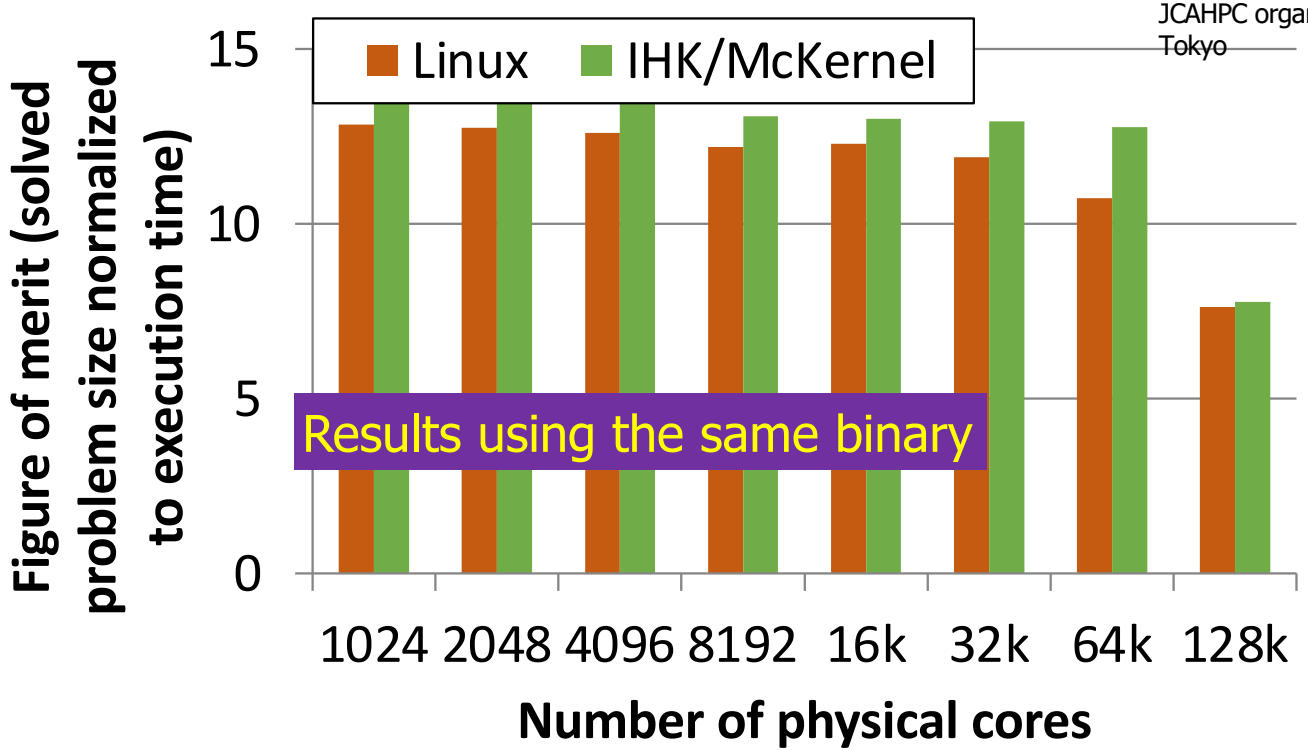
Oakforest-PACS のソフトウェア

- 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。
- コンパイラ：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.

McKernel評価：GeoFEM (University of Tokyo)

- ICCG with Additive Schwartz Domain Decomposition - weak scaling
- Up to 18% improvement

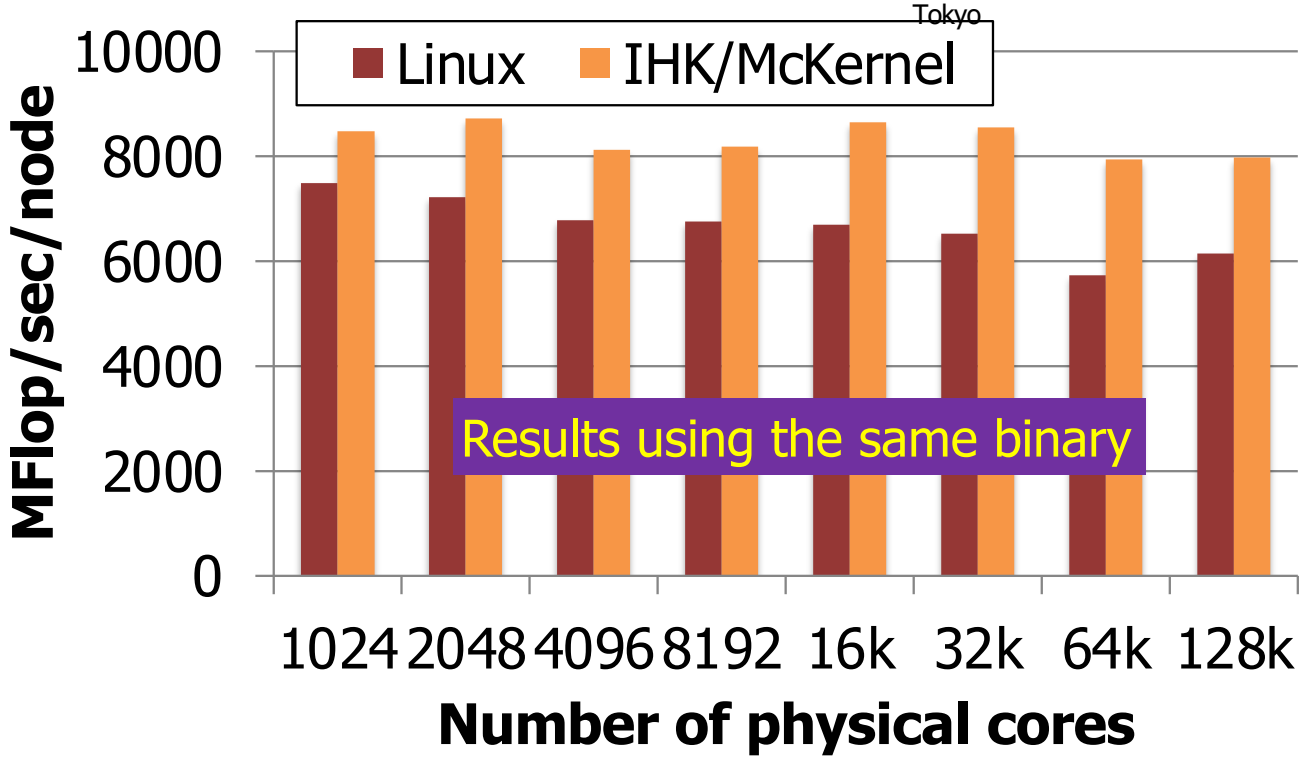
Acknowledgement: Kengo Nakajima, University of Tokyo, for providing GeoFEM. This result is on Oakforest-PACS supercomputer, 25 PF in peak, at JCAHPC organized by U. of Tsukuba and U. of Tokyo



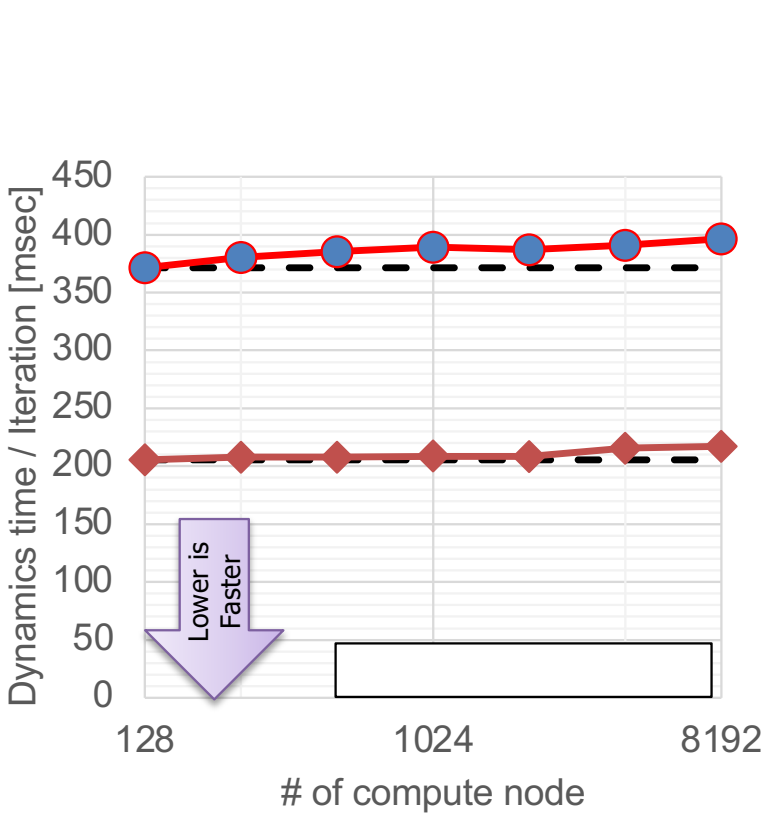
McKernel評価：CCS-QCD (University of Tsukuba)

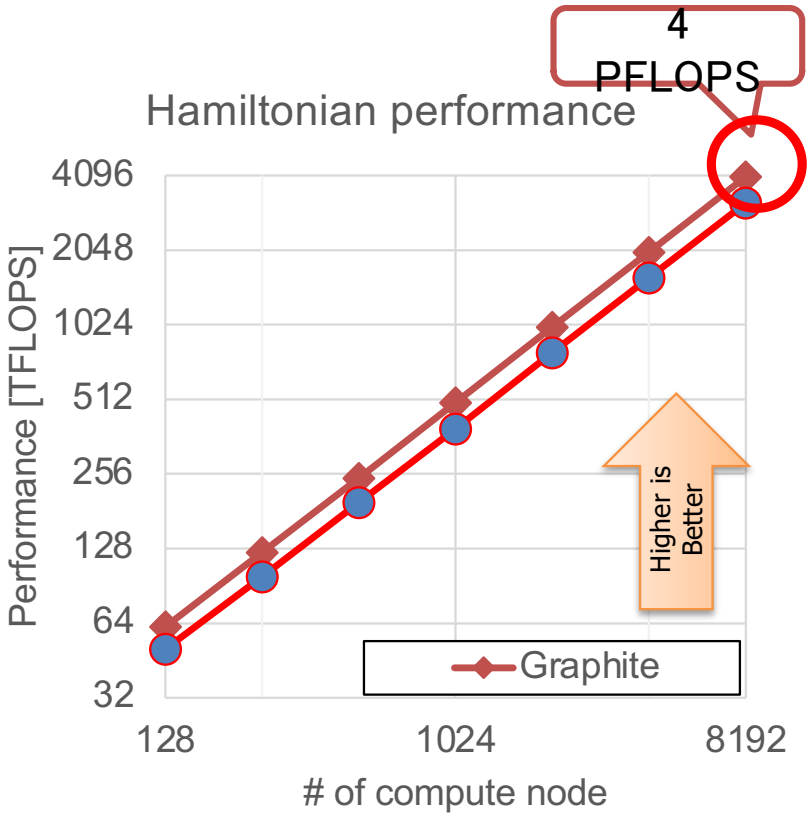
- Lattice quantum chromodynamics code - weak scaling
- Up to 38% improvement

Acknowledgement: Ken'ichi Ishikawa, Hiroshima University, providing CCS-QCD. This result is on Oakforest-PACS supercomputer, 25 PF in peak, at JCAHPC organized by U. of Tsukuba and U. of Tokyo



SALMON (ex-ARTED) Weak scaling on OFP full system





まとめ

- OFPは実運用に供されているスーパーコンピュータとして2016/11に国内最高性能を達成し、HPCIをはじめとする筑波大・東大の様々な利用プログラムに活用された
- 同時期にNERSC Cori，京大 Camphor，さらに少し遅れてKISTI Nurion，北大Polaireなど，KNLに基づく technology driven system が続いた
- McKernel, XcalableMPといったシステムソフトウェア開発はOFPの性能改善、プログラミング環境改善だけでなくポスト「京」（＝「富岳」）の開発にもつながった
- 大口・大規模ユーザは性能チューニングをよく行い、数々のノウハウはその後の many-core Flgaship である富岳につながった
- 2018/08に K Computer が運用停止してからの1年半ぐらいは HPCI 第２階層の代表的マシンとしてHPCIを支えた
- 複数大学の共同調達による国内最高システムの構築という実績は文部科学省にも大きく着目された
- システムは結局人が作る，テクノロジがそれを支える
- JCAHPCの継続 ⇨ OFP2へ