JCAHPC

JCAHPCにおける新システム
Oakforest-PACS
–国内最高性能システムの実現に向けて–

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
塙　敏博
（最先端共同HPC基盤施設: JCAHPC）
（筑波大CCS 客員准教授）

東京大学情報基盤センター
INFORMATION TECHNOLOGY CENTER, THE UNIVERSITY OF TOKYO

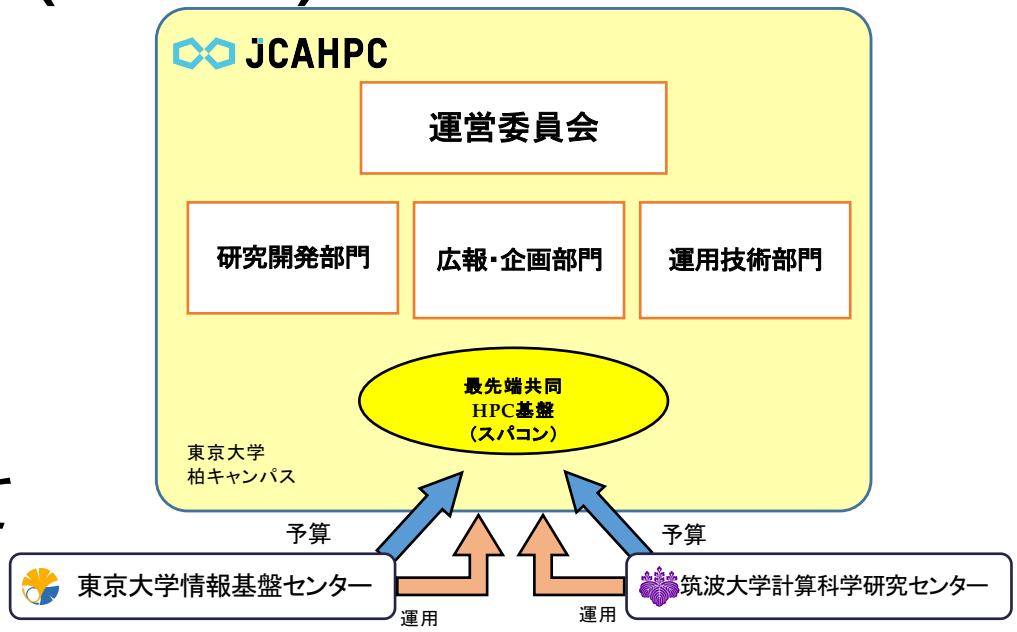
筑波大学
計算科学研究センター
Center for Computational Sciences

最先端共同HPC基盤施設 JCAHPC

- Joint Center for Advanced High Performance Computing (http://jcahpc.jp)
- 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
- 本協定の下、筑波大学計算科学研究センターと東京大学情報基盤センターが JCAHPC を設置
- 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

Oakforest-PACS (OFP) in JCAHPC

- 筑波大学と東京大学の間の密な連携・協力
- 仕様に加え調達プロセスを一本化、単一のシステム
- 2大学が調達と運用に関して責任を持つ
  - 国内初の試み
  - 日本で最大規模のシステムを実現



施設長	中村　宏・東大教授 （東大情報基盤センター長）
副施設長	梅村　雅之・筑波大教授 （筑波大計算科学研究センター長）
運営委員会	メンバー　8名
部門:	
研究開発	**部門長**:中島　研吾・東大教授
運用技術	**部門長**:朴　泰祐・筑波大教授
広報・企画	**部門長**:建部　修見・筑波大教授

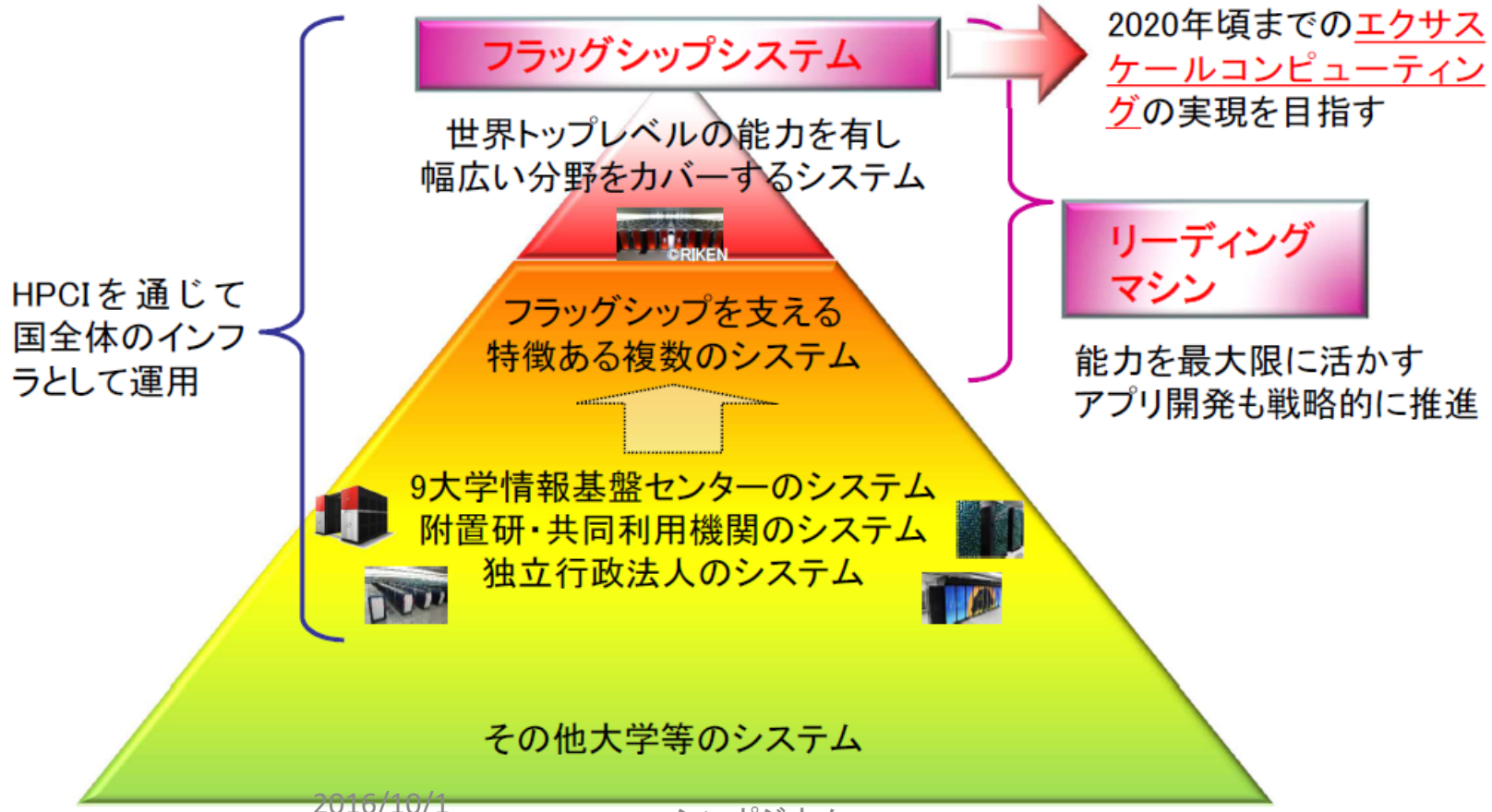
HPCI: High Performance Computing Infrastructure
日本全体におけるスパコンインフラ

今後のHPCI 計画推進の在り方について(H26/3)より

我が国の次期スパコン開発の方向性

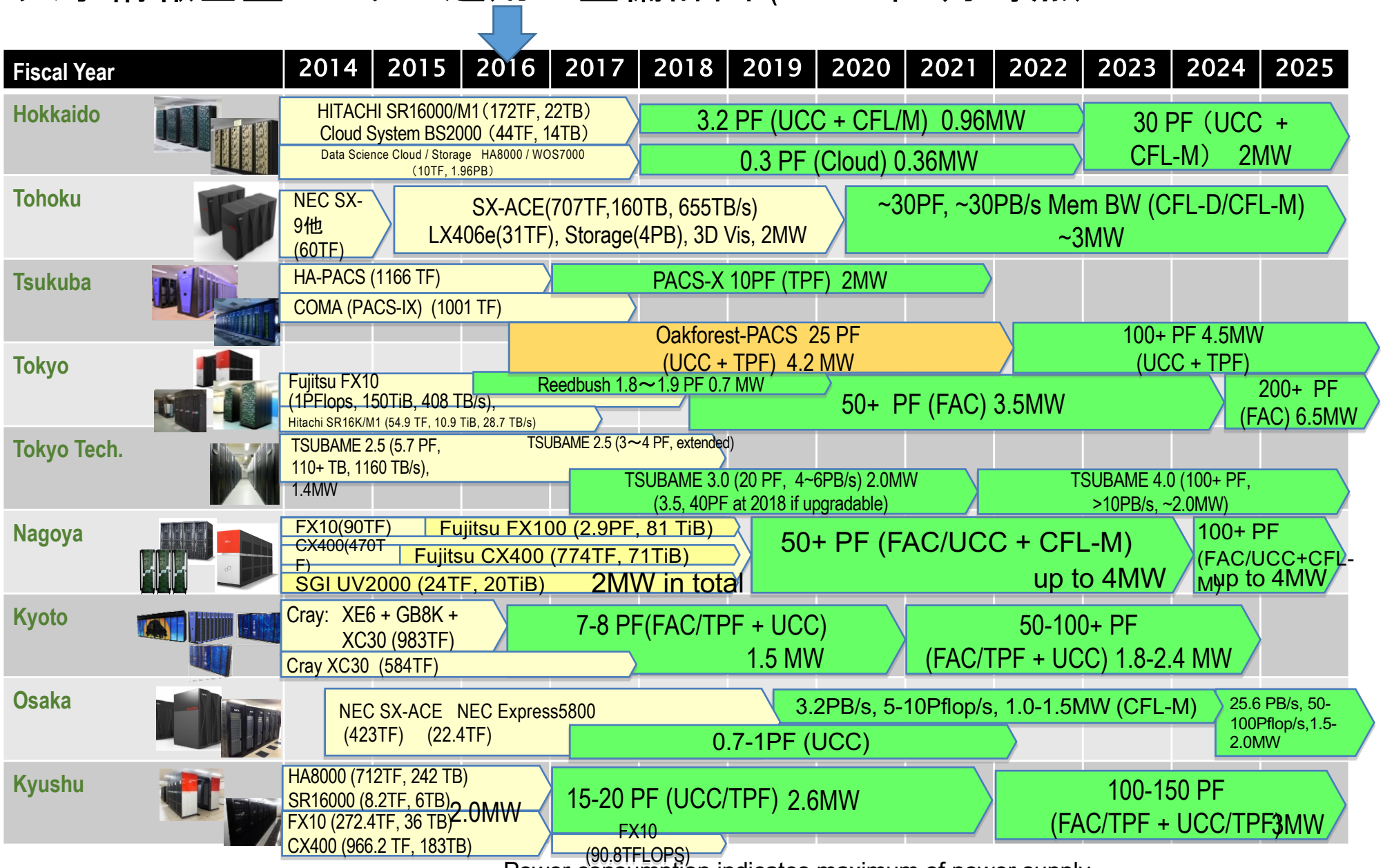
＜我が国の計算科学技術インフラのイメージ＞

http://www.mext.go.jp/b_menu/shingi/chousa/shinkou/028/gaiyou/1348991.htm



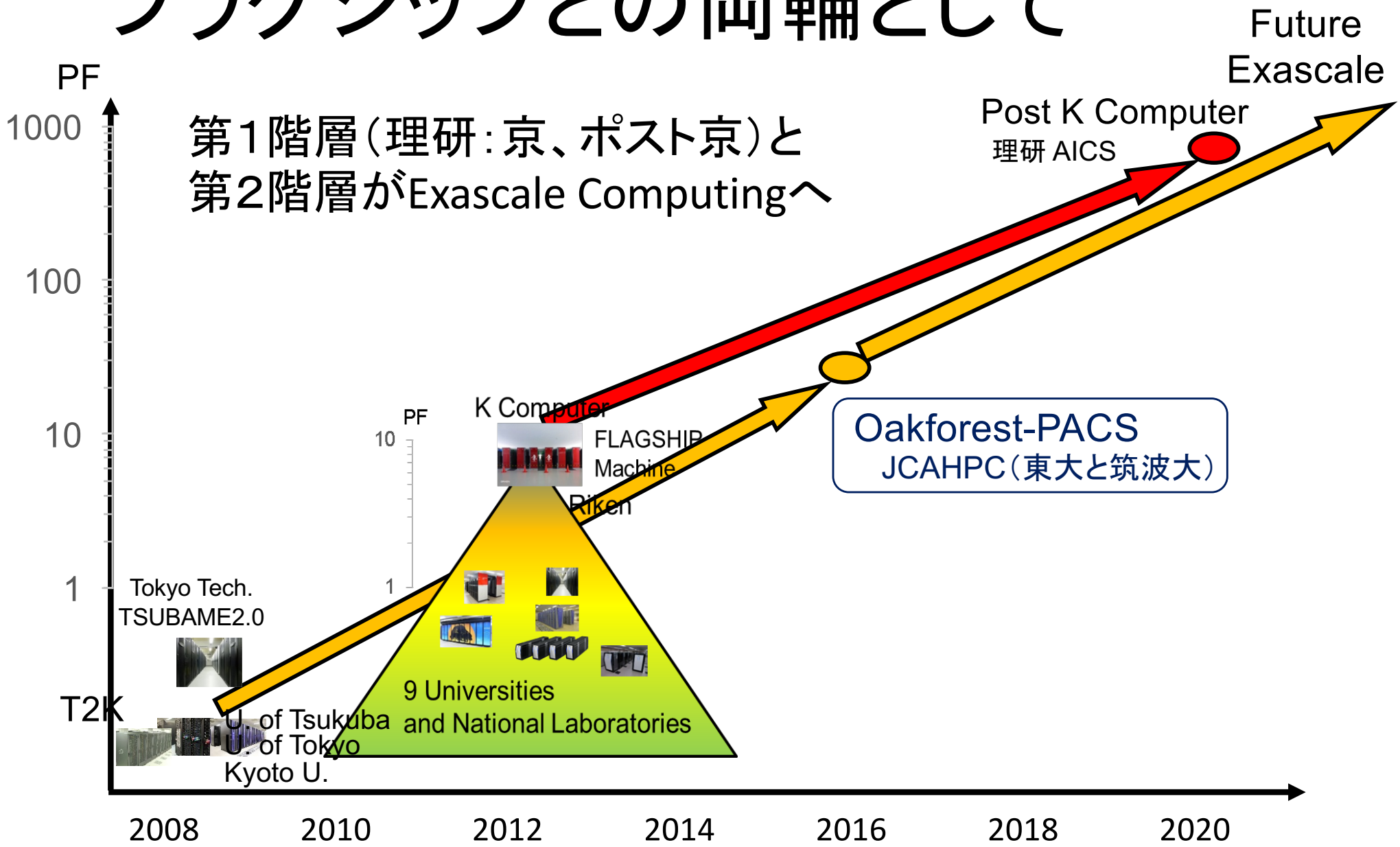
9大学情報基盤センター運用＆整備計画 (2016年5月時点)

JCAHPC



Power consumption indicates maximum of power supply
(includes cooling facility)

フラグシップとの両輪として



JCAHPC：共同調達への道のり

- 2013活動開始
  - 第1期(2013/4-2015/3)：
    施設長：佐藤三久(筑波大学)、副施設長:石川裕(東京大学)
  - 第2期(2015/4-)：
    施設長：中村宏(東京大学)、副施設長:梅村雅之(筑波大学)
- 共同調達・運用へ向けて
  - 2013/7: RFI(request for information)
    共同調達は既定路線ではなかった→1システムとして調達へ
- 複数大学による初めての「1システム」共同調達へ
  - どうして共同調達ができたのか？共同調達は大変・・
    => 目標を共有できる、ことに尽きる

2センターのミッション

- 筑波大学計算科学研究センターのミッション:
  - 計算機科学と計算科学の協働:学際的な高性能計算機開発
    → PACSシリーズの開発:CP-PACS@1996 TOP1
  - 先端学際科学共同研究拠点: 最先端の計算科学研究推進
  - これからの計算科学に必要な学際性を持つ人材を育成
- 東京大学情報基盤センターのミッション:
  - 学際大規模情報基盤共同利用・共同研究拠点(8大学の情報基盤センター群からなるネットワーク型)の中核拠点:
    大規模情報基盤を活用し学際研究を発展
  - HPCI資源提供機関:最先端スパコンの共同設計開発及び運用、Capability資源および共用ストレージ資源の提供
  - 人材育成:計算科学の新機軸を創造できる人材の育成

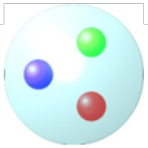
筑波大学計算科学研究センター

JCAHPC

計算科学と計算機科学の協働（コデザイン）

先端計算科学推進室 ＝ **次世代計算機システム開発室**

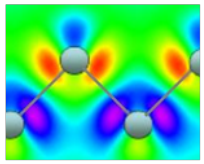
素粒子



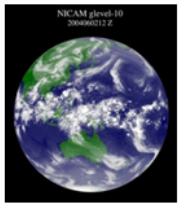
原子核

宇宙

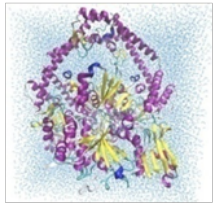
物質



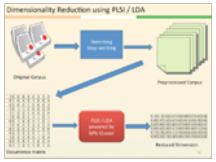
地球環境



生命



データ基盤



PACSシリーズの開発

筑波大学計算科学研究センター

JCAHPC

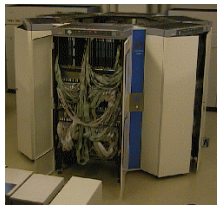
1992年4月　計算物理学研究センター設置(10年計画)。
1996年9月　CP-PACS(2048PU)完成，TOP500で世界第1位。
2004年4月　改組拡充し，計算科学研究センターを設置。
2007年4月　融合型宇宙シミュレータFIRST完成。
2008年6月　T2K-Tsukubaオープンスーパーコンピュータ運用開始。
2010年4月　共同利用・共同研究拠点「先端学際計算科学共同研究拠点」認定。
2013年3月　東京大学との協定に基づき「最先端共同HPC基盤施設」を設置。

- 科学者と計算機工学者の協力による，application –drivenな超高速計算機の開発・製作＝学際計算科学
  世界的に見てもユニーク
- 高い計算パワーの集中による計算科学の最重点課題・最先端課題の研究

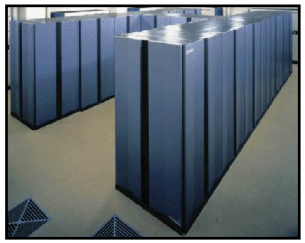
1978
1st PACS-9

1980
2nd PAXS-32

1989
5th QCDPAX



1996
6th CP-PACS
(1996年世界最高速を達成)



2006
7th PACS-CS

2012
8th HA-PACS

2014
9th COMA

Year	System	Performance
1978	PACS-9 (PACS I)	7 KFLOPS
1980	PACS-32 (PACS II)	500 KFLOPS
1983	PAX-128 (PACS III)	4 MFLOPS
1984	PAX-32J (PACS IV)	3 MFLOPS
1989	QCDPAX (PACS V)	14 GFLOPS
1996	CP-PACS (PACS VI)	614 GFLOPS
2006	PACS-CS (PACS VII)	14.3 TFLOPS
2012	HA-PACS (PACS VIII)	1.166 PFLOPS
2014	COMA (PACS IX)	1.001 PFLOPS

2007
FIRST
(Hybrid Simulator)

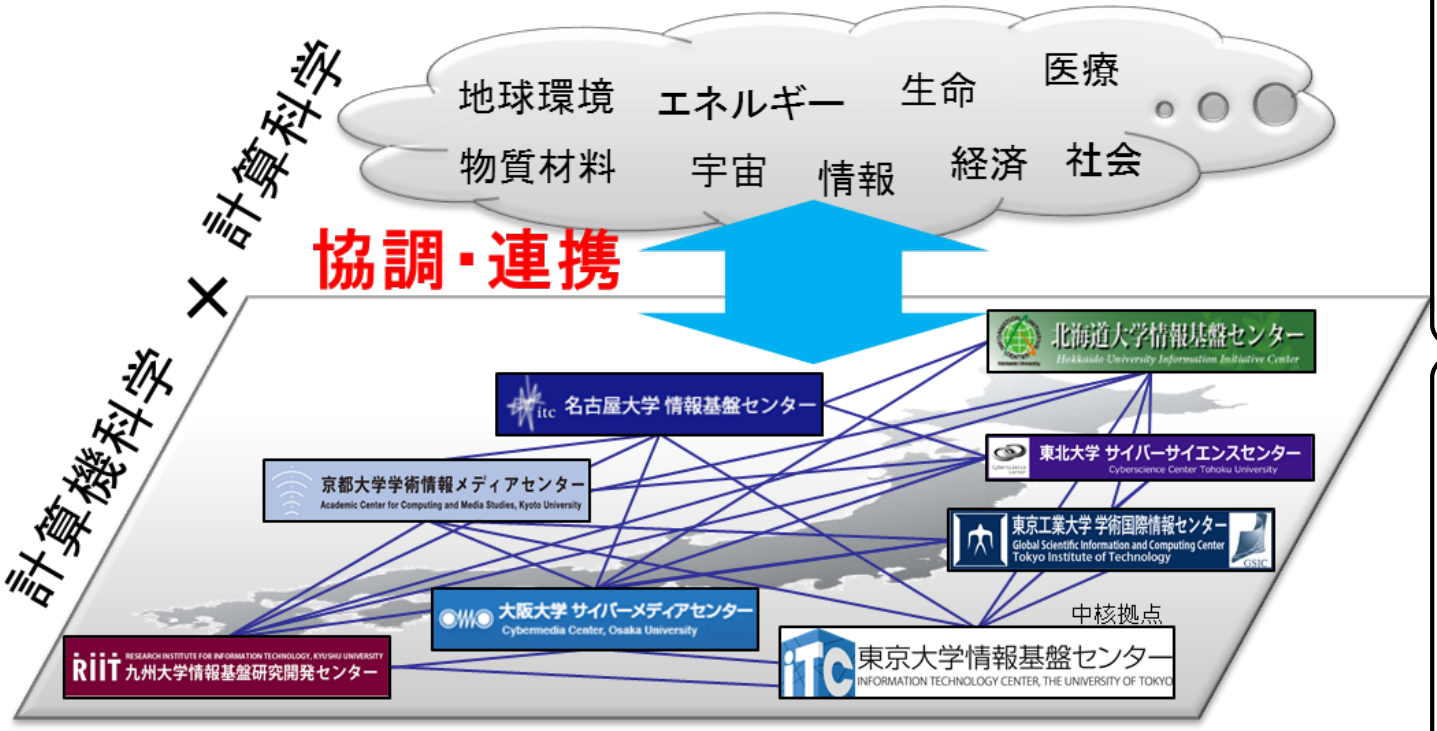
36TFLOPS
Host 3TFOPS
Accelerator 33TFLOPS

東京大学情報基盤センター

Research Center for Extreme Scale Computing and Data

- 学際大規模情報基盤共同利用・共同研究拠点の中核拠点

学際研究：計算科学・工学の問題解決に向け計算機科学と協調・連携



大規模情報基盤：８大学のスーパーコンピュータ群と利用技術

- 解決や解明が困難と考えられていた課題の解決へ
- 学術基盤としての大規模情報基盤の活用による研究コミュニティへの貢献
- 多様で大規模な計算資源
- 公募型の学際的共同研究（萌芽段階を含む）を遂行

東京大学情報基盤センター

JCAHPC

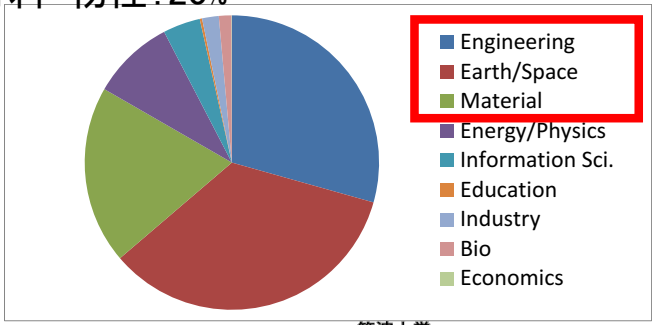
- HPCI資源提供機関として
  - 機関連携による最先端スパコンの共同設計開発及び運用、Capability資源および共用ストレージ資源の提供
  - Data Intensive Applicationに対応したシステムの整備
- 人材育成機関として
  - 計算科学の新機軸を創造できる人材の育成
  - 学内各部局，利用者，共同利用・共同研究拠点との連携
- 合計約2,000人のユーザー（学外が半分）
  - 大学（研究，教育），研究機関，企業

Oakleaf-fx
(Fujitsu PRIMEHPC FX10)

Total Peak performance:	1.13 PFLOPS
Total number of nodes:	4800
Total memory:	150 TB
Peak performance / node:	236.5 GFLOPS
Main memory per node:	32 GB
Disk capacity:	1.1 PB + 2.1 PB

SPARC64 Ixfx 1.84GHz

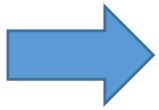
利用の多い分野（2015年度）
- 工学（流体・構造・電磁気等）：30%
- 地球宇宙科学（大気海洋・地震等）：35%
- 材料・物性：20%



→ 大規模シミュレーション，特に連成解析
- 全球規模大気海洋カップリング
  - ppOpen-HPC, ppOpen-MATH/MP
- 地震シミュレーション
  - 地震発生＋破壊伝播＋強震動
  - 地盤強震動＋都市・建造物振動
- 流体・構造シミュレーション

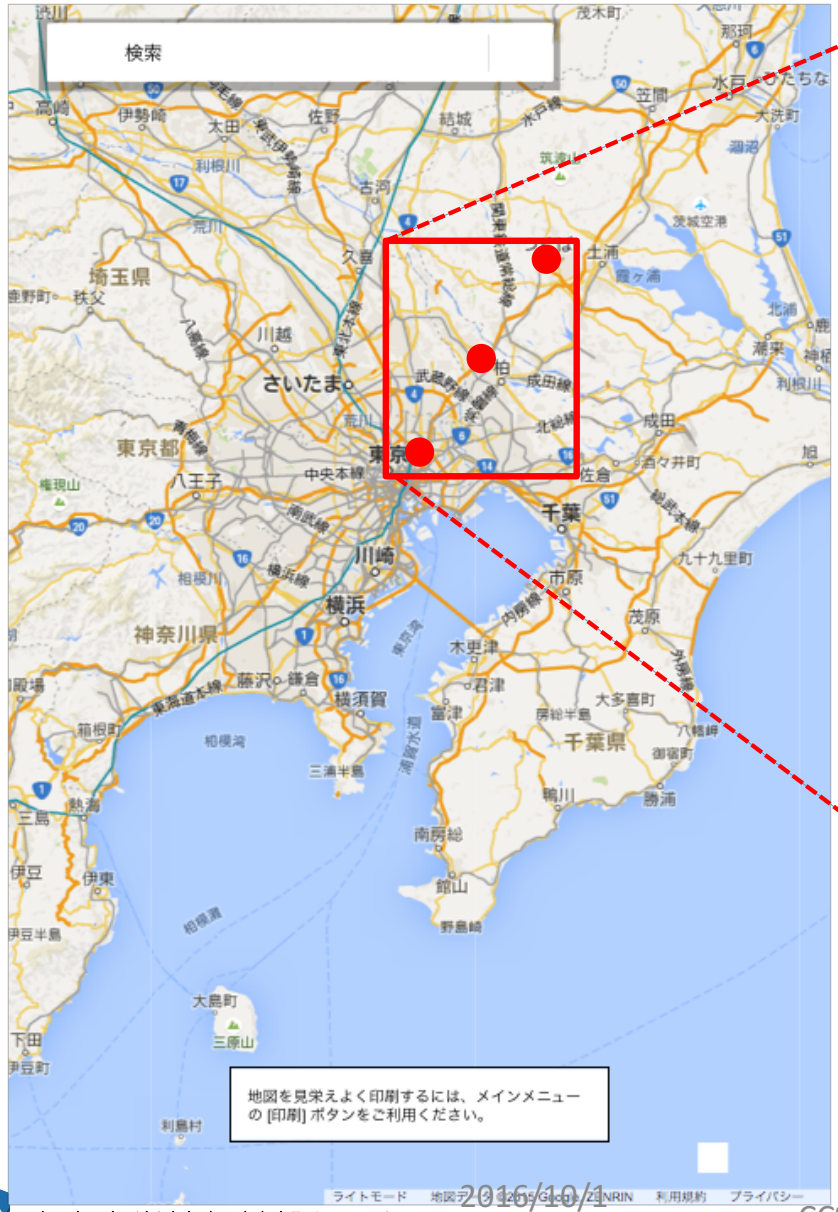
JCAHPC共同調達のポリシー

- T2Kの精神に基づき、オープンな最先端技術を導入
  - T2K：2008年に始まったTsukuba, Tokyo, Kyoto の3大学でのオープンスパコンアライアンス、3機関の研究者が仕様策定に貢献、システムへの要求事項を共通化
- システムの基本仕様
  - 超並列PCクラスタ
  - HPC用の最先端プロセッサ、アクセラレータは不採用
    → 広範囲なユーザとアプリケーションのため
    → ピーク性能追求より、これまでのコードの継承を優先
  - 使いやすい高効率相互結合網
  - 大規模共用ファイルシステム
- 2大学による共同調達
  - 国立大のスパコンシステムとして最大級
  - PCクラスタとしても国内最大級
- スケールメリットを活かす
  - 超大規模な単一ジョブ実行も可能とする



Oakforest-PACS

設置予定場所：東京大学柏キャンパス

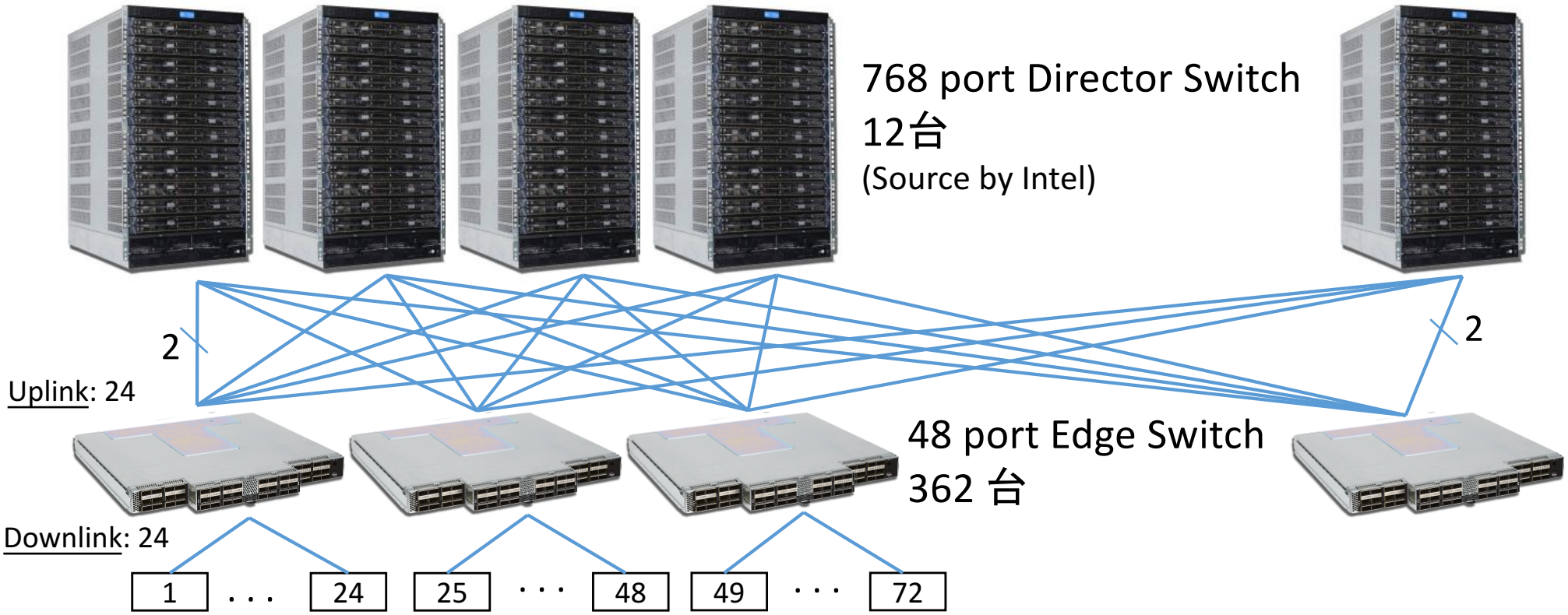


東大＠柏(Oak)
筑波大：PACS

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

Intel® Omni-Path Architecture を用いたフルバイセクションバンド幅Fat-tree網



コストはかかるがフルバイセクションバンド幅を維持
- システム全系使用時にも高い並列性能を実現
- 柔軟な運用：ジョブに対する計算ノード割り当ての自由度が高い

Oakforest-PACS の仕様（続き）

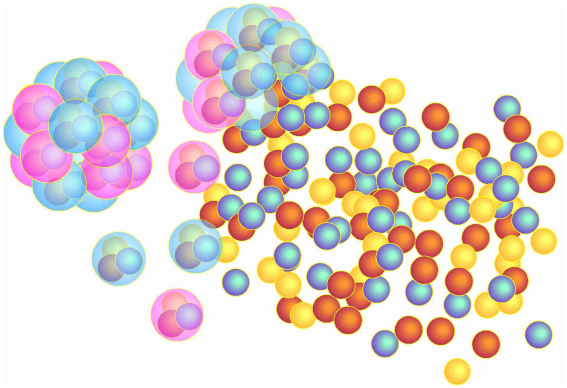
並列ファイルシステム	Type	Lustre File System
	総容量	26.2 PB
	Product	DataDirect Networks SFA14KE
	総バンド幅	500 GB/sec
ファイルキャッシュシステム	Type	Burst Buffer, Infinite Memory Engine (by DDN)
	総容量	940 TB (NVMe SSD, パリティを含む)
	Product	DataDirect Networks IME14K
	総バンド幅	1,560 GB/sec
総消費電力		4.2MW（冷却を含む）
総ラック数		102

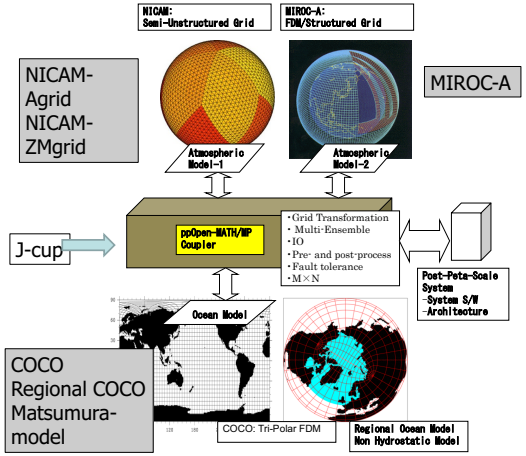
Oakforest-PACS のソフトウェア

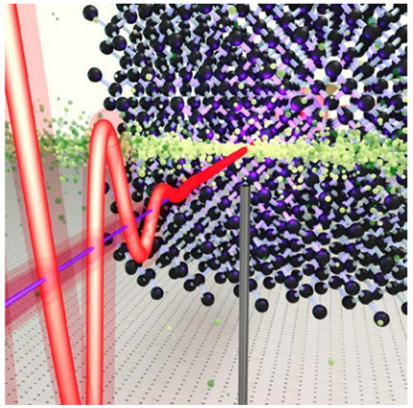
- OS: Red Hat Enterprise Linux (ログインノード)、CentOS および McKernel (計算ノード、切替可能)
  - McKernel: 理研AICSで開発中のメニーコア向けOS
    - Linuxに比べ軽量、ユーザプログラムに与える影響なし
    - ポスト京コンピュータにも搭載される予定。
- コンパイラ：GCC, Intel Compiler, XcalableMP
  - XcalableMP: 理研AICSと筑波大で共同開発中の並列プログラミング言語
    - CやFortranで記述されたコードに指示文を加えることで、性能の高い並列アプリケーションを簡易に開発することができる。
- 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など

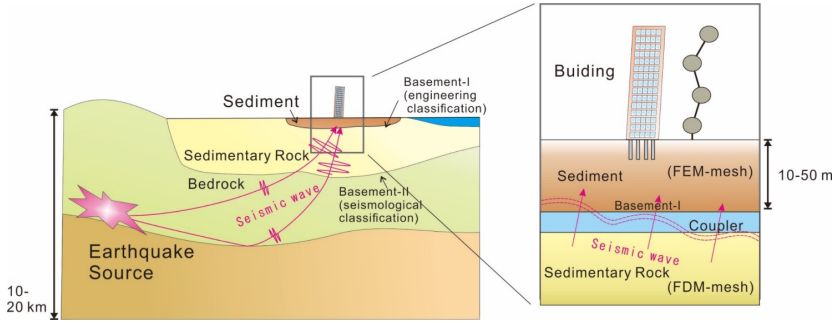
OFPにおけるアプリケーション例

- ARTED: 電子ダイナミクスの第一原理計算
- Lattice QCD: 量子色力学計算
- NICAM & COCO: 大気＋海洋連成シミュレーション
- GAMERA/GOJIRA: 地震シミュレーション
- Seism3D: 地震波伝搬シミュレーション









計算ノードの写真

2Uサイズのシャーシ
（富士通 PRIMERGY CX600 M1）に
8計算ノードを搭載

計算ノード (富士通 PRIMERGY CX1640 M1)
Intel Xeon Phi 1ソケット、Intel Omni-Path Architecture card (HFI)搭載

システム運用

- 通常運用
  - 各大学単位にハードウェアの分割はしない
  - 「CPU時間」を2大学で按分、柔軟な運用を可能に
  - HPCIに対してはJCAHPCとして一括で資源提供
  - 各大学の利用プログラムは「CPU時間」に基づく按分で決定
- 特別運用
  - 限られた時間だけ、全系を1システムとして、超大規模な単一ジョブの実行(ex. Gordon Bell Challenge)
- 省電力運用
  - 状況に応じて、総電力にキャッピングをかける省電力運用（夏季節電など ）

計算ノードの動作モード

- Xeon Phi (Knights Landing)には複数の動作モード

候補: Flat/Cache + Quadrant/SNC-4

- メモリモード：3種類
  - Flat:　MCDRAMとDDR4 が独立したアドレス
  - Cache: MCDRAMはDDR4メモリのキャッシュとして動作
  - Hybrid
- クラスタリングモード: 5種類
  - (All-to-all: アドレス情報が全体に分散… 非推奨)
  - Quadrant, Hemisphere: 内部でアドレス情報が4(または2)に分割（ユーザからは見えない）
  - SNC-4, SNC-2: NUMAドメインが明示的に4 (or 2)に分割

モードの変更には再起動が必要
=> ジョブスケジューラによりプロビジョニング
（現時点では、各モードのジョブキューを用意）

運用予定

- 2016/10/1：一部のシステム稼働（400ノード）
- 2016/12/1：全系システム稼働
- 2017/4：オープンな資源提供（HPCI資源を含む）

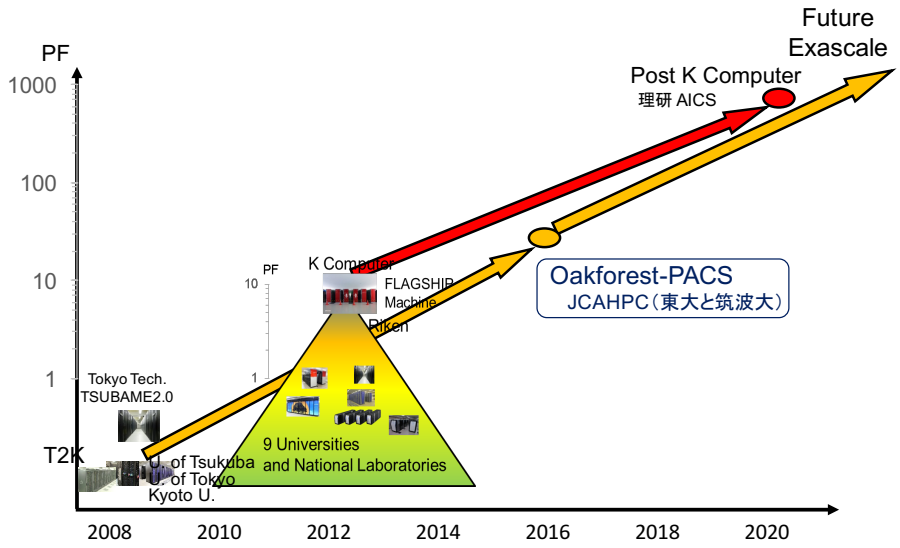
Oakforest-PACS (OFP)
（2016年9月29日現在）

ビジュアル公開前につき

写真非表示にてWeb掲載

Oakforest-PACSの果たす役割

- OFPは国内第1位のシステムになる（はず）
  - OFP: 25 PF, 京コンピュータ: 11.2 PF
  - 京コンピュータ以後、OFPがポスト京システムに向けたブリッジシステムとして期待されている
- ポスト京コンピュータはOFPと同様メニーコアシステムになる予定
- OFPはポスト京コンピュータに向けて重要な役割
  - 大規模アプリケーションの開発基盤
  - McKernelやXMPなどのメニーコアシステムに向けたシステムソフトウェアの開発基盤



おわりに

- JCAHPC（最先端共同HPC基盤施設）
  - 筑波大学計算科学研究センターと東京大学情報基盤センターが設置
  - 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に連携
- Oakforest-PACS：ピーク性能 25 PFLOPS
  - Intel Xeon Phi (Knights Landing) と Omni-Path Architecture
  - CPU時間を2大学で按分することで柔軟な運用を可能
  - 全系を1システムとして超大規模単一ジョブの実行も可能に
  - 全系システムの稼働は2016/12を予定
  - HPCI資源を含めオープンは資源提供は2017/4を予定
- JCAHPC：最先端HPC研究に寄与する計算資源の提供を目指し、コミュニティに貢献していく予定