JCAHPC

メニーコア型大規模スーパーコンピュータシステム Oakforest-PACSの現状と動向

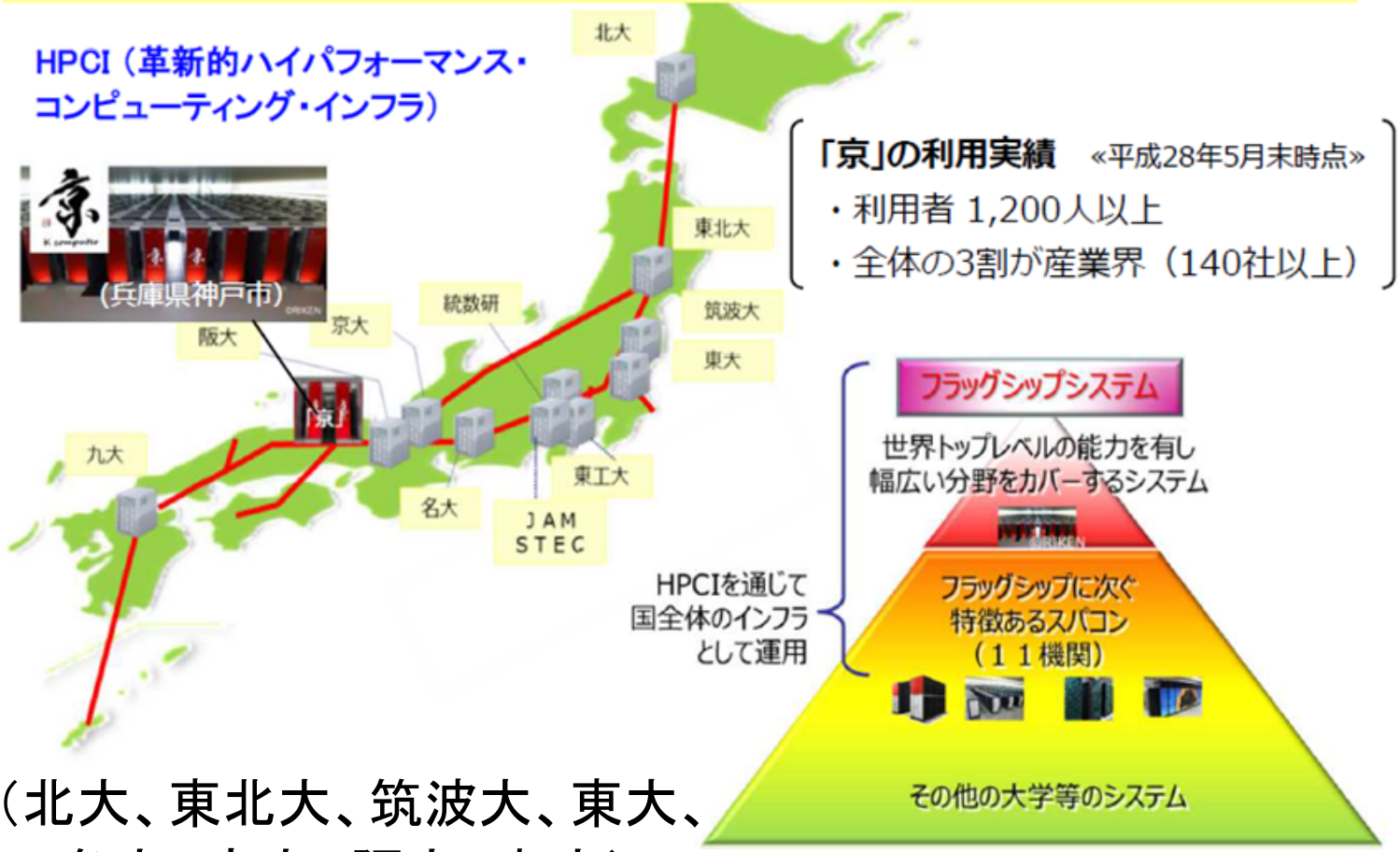
東京大学 情報基盤センター
最先端共同HPC基盤施設
(JCAHPC)

塙　敏博

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

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

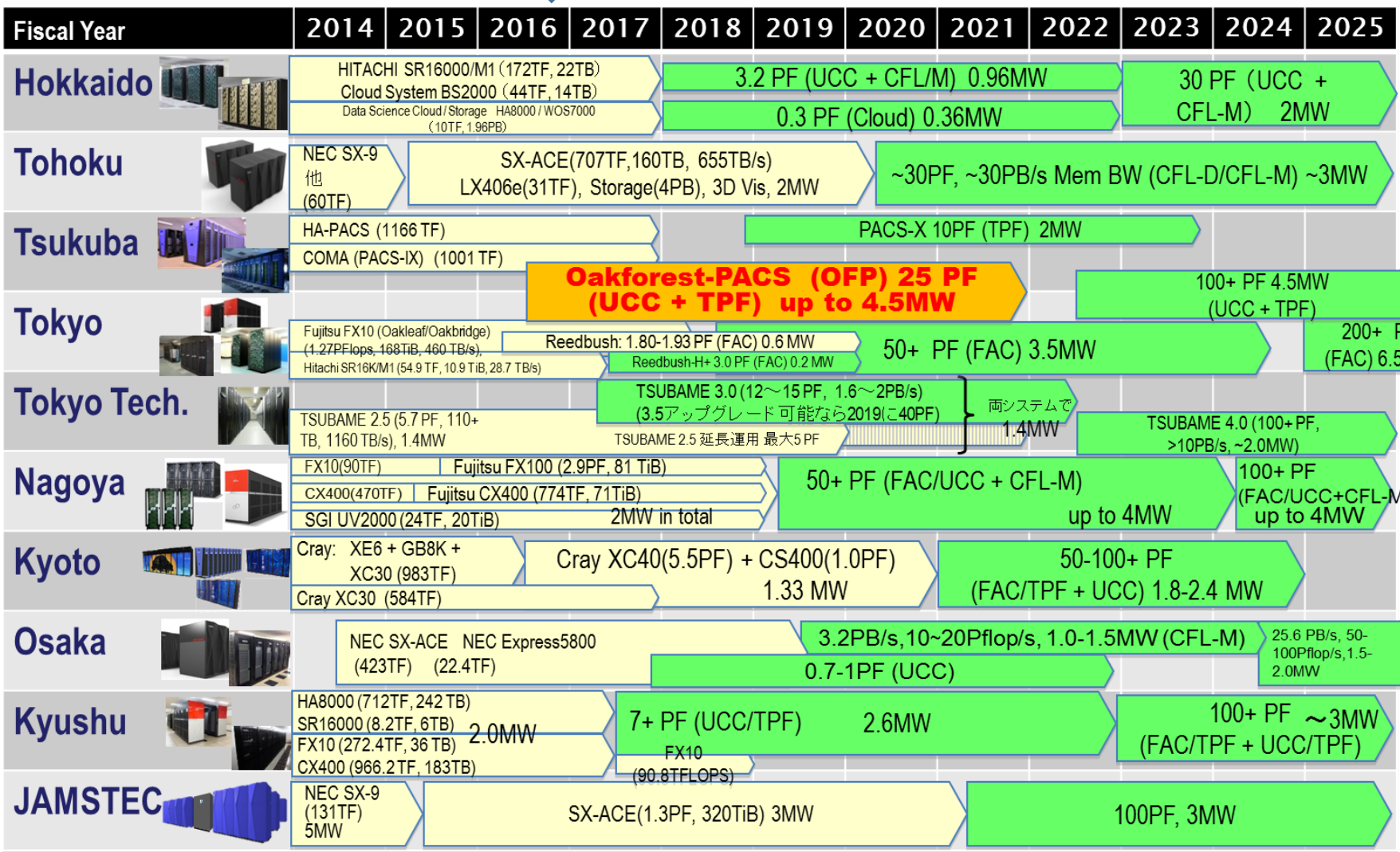


- 9大学（北大、東北大、筑波大、東大、東工大、名大、京大、阪大、九大）の情報基盤センター
- 海洋開発研究機構、統数研＋SINET

HPCI第2階層システムの開発・整備・運用計画

(2016年9月時点)

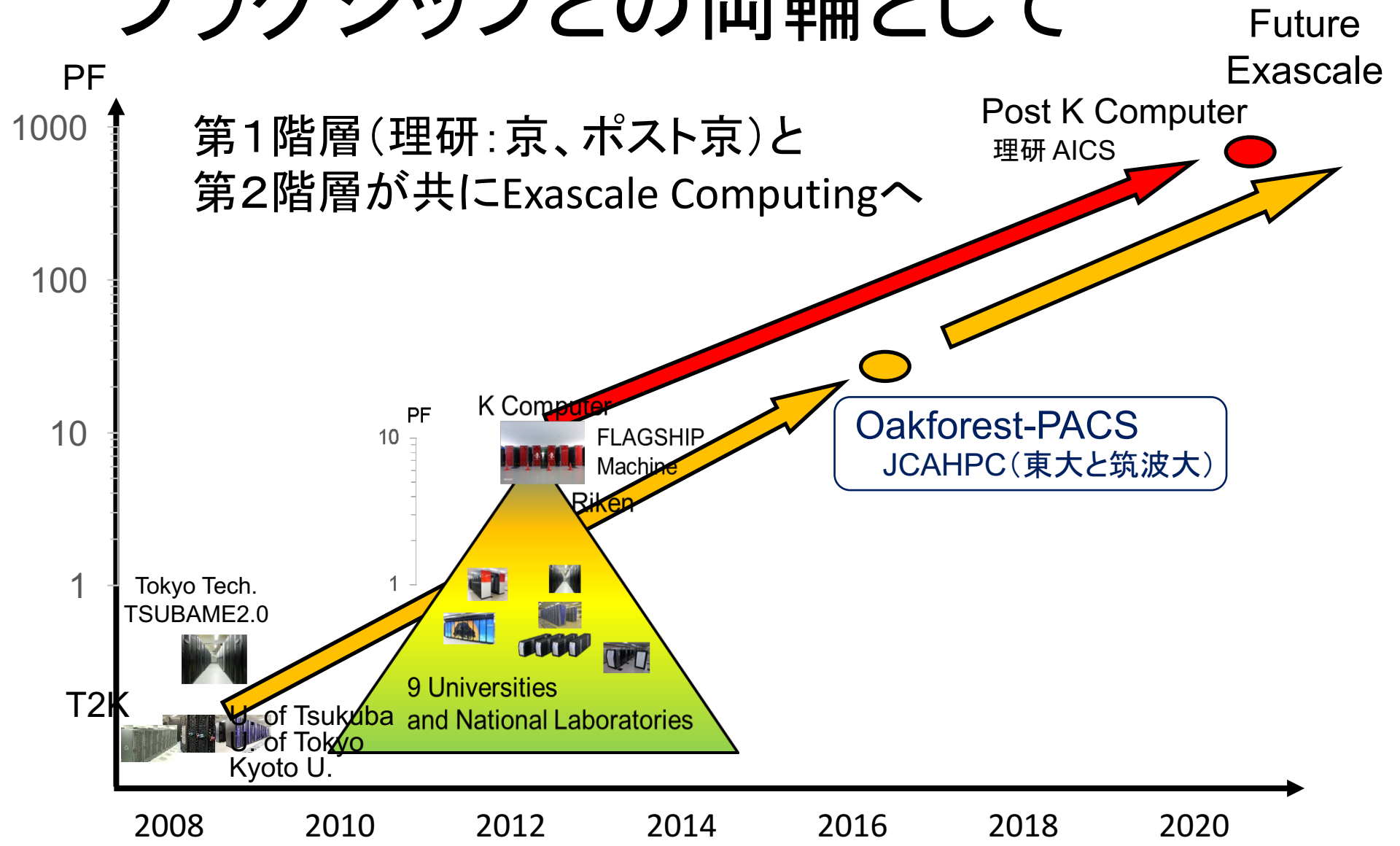
HPCIコンソーシアムのホームページ掲載 http://www.hpci-c.jp/



Oakforest-PACS(OFP) JCAHPC

- 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
    - 両センターが共同で、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
    - 東京大学柏キャンパスの東京大学情報基盤センター内
- 2016年12月1日稼働開始
- 8,208 Intel Xeon/Phi (KNL)
- ピーク性能25PFLOPS
- **TOP 500　6位（国内1位），HPCG 3位（国内2位），Green 500 6位（国内2位）（2016年11月）**

フラグシップとの両輪として



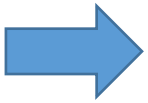
最先端共同HPC基盤施設 JCAHPC

- 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
  - 両センターが共同で、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
  - 東京大学柏キャンパスの東京大学情報基盤センター内

JCAHPC共同調達のポリシー ～2センターで共有したこと～

JCAHPC

- 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



Oakforest-PACS

- 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

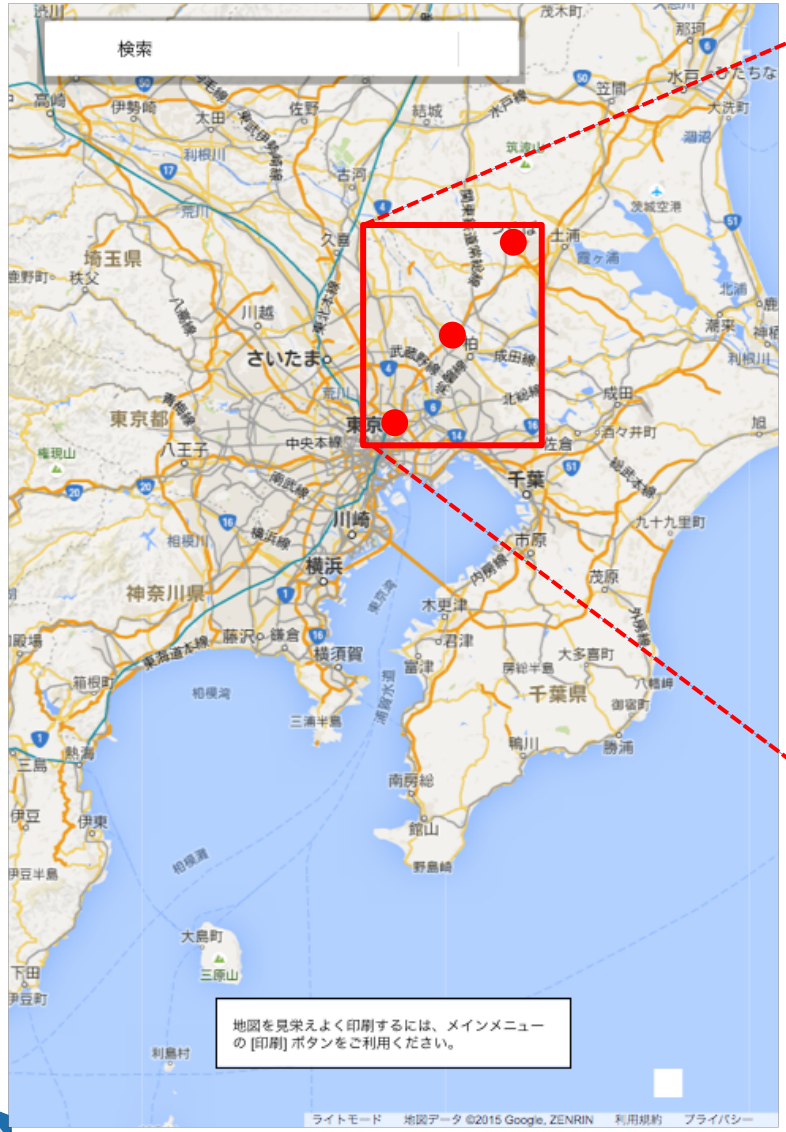
Oakforest-PACS 全景

時事通信
www.jiji.com

国内最高性能の新スパコン「オークフォレスト・パックス」の前で握手する
東大の中村宏情報基盤センター長（左）と筑波大の梅村雅之計算科学研究センター長
＝1日午後、千葉県柏市の東大柏キャンパス

設置場所：東京大学柏キャンパス

Google マップ　https://www.google.com/maps/@?dg=dbrw&newdg=1



筑波大学

東京大学
柏キャンパス

東京大学本郷キャンパス

設置場所：東京大学柏キャンパス

第2総合研究棟　2階

写真はマイナビニュースより
http://news.mynavi.jp/news/2016/12/02/035/

JCAHPC

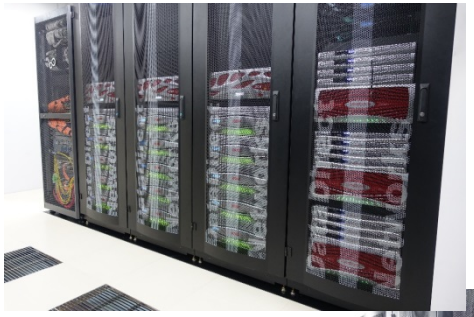
Oakforest-PACSの特徴 (1/2)

- 計算ノード
  - Intel Xeon Phi (Knights Landing)
  - 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
  - Intel Omni-Path Architecture
  - 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

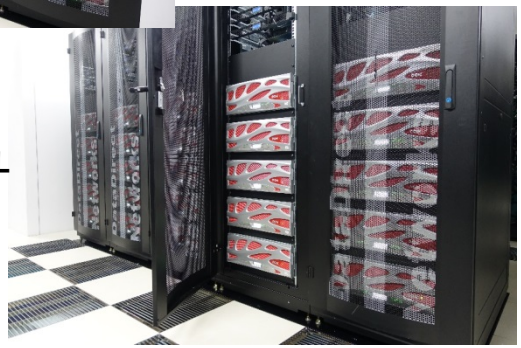
JCAHPC

Oakforest-PACS の特徴(2 / 2)

- ファイルI/O
  - 並列ファイルシステム:
    Lustre 26PB
  - ファイルキャッシュシステム
    (DDN IME):
    1TB/secを超える実効性能,
    約1PB
    - 計算科学・ビッグデータ解析・機械学習にも貢献
- 消費電力
  - Green 500でも世界6位
  - Linpack: 2.72 MW
    - 4,986 MFLOPS/W(OFP)
    - 830 MFLOPS/W(京)



並列ファイル
システム



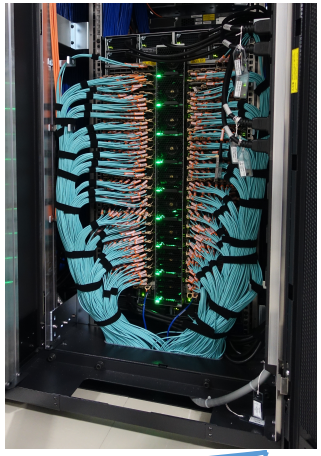
ファイルキャッシュ
システム

ラック当たり120ノードの
高密度実装

リアドア冷却

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

JCAHPC



768 port Director Switch
12台
(Source by Intel)

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

JCAHPC

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.

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

48th TOP500 List (November, 2016)

	Site	Computer/Year Vendor	Cores	R_{max} (TFLOPS)	R_{peak} (TFLOPS)	Power (kW)
1	National Supercomputing Center in Wuxi, China	**Sunway TaihuLight** , Sunway MPP, Sunway SW26010 260C 1.45GHz, 2016 NRCPC	10,649,600	93,015 (= 93.0 PF)	125,436	15,371
2	National Supercomputing Center in Tianjin, China	**Tianhe-2**, Intel Xeon E5-2692, TH Express-2, Xeon Phi, 2013 NUDT	3,120,000	33,863 (= 33.9 PF)	54,902	17,808
3	Oak Ridge National Laboratory, USA	**Titan** Cray XK7/NVIDIA K20x, 2012 Cray	560,640	17,590	27,113	8,209
4	Lawrence Livermore National Laboratory, USA	**Sequoia** BlueGene/Q, 2011 IBM	1,572,864	17,173	20,133	7,890
5	DOE/SC/LBNL/NERSC USA	**Cori**, Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Cray Aries, 2016 Cray	632,400	14,015	27,881	3,939
6	Joint Center for Advanced High Performance Computing, Japan	**Oakforest-PACS**, PRIMERGY CX600 M1, Intel Xeon Phi Processor 7250 68C 1.4GHz, Intel Omni-Path, 2016 Fujitsu	557,056	13,555	24,914	2,719
7	RIKEN AICS, Japan	**K computer**, SPARC64 VIIIfx , 2011 Fujitsu	705,024	10,510	11,280	12,660
8	Swiss Natl. Supercomputer Center, Switzerland	**Piz Daint** Cray XC30/NVIDIA P100, 2013 Cray	206,720	9,779	15,988	1,312
9	Argonne National Laboratory, USA	**Mira** BlueGene/Q, 2012 IBM	786,432	8,587	10,066	3,945
10	DOE/NNSA/LANL/SNL, USA	**Trinity**, Cray XC40, Xeon E5-2698v3 16C 2.3GHz, 2016 Cray	301,056	8,101	11,079	4,233

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

http://www.top500.org/

HPCG Ranking (SC16, November, 2016)

	Site	Computer	Cores	HPL Rmax (Pflop/s)	TOP500 Rank	HPCG (Pflop/s)	HPCG/ HPL (%)
1	RIKEN AICS, Japan	K computer	705,024	10.510	7	0.6027	5.73
2	NSCC / Guangzhou, China	Tianhe-2	3,120,000	33.863	2	0.5800	1.71
3	JCAHPC, Japan	Oakforest-PACS	557,056	13.555	6	0.3855	2.84
4	National Supercomputing Center in Wuxi, China	Sunway TaihuLight	10,649,600	93.015	1	0.3712	.399
5	DOE/SC/LBNL/NERSC USA	Cori	632,400	13.832	5	0.3554	2.57
6	DOE/NNSA/LLNL, USA	Sequoia	1,572,864	17.173	4	0.3304	1.92
7	DOE/SC/ Oak Ridge National Laboratory, USA	Titan	560,640	17.590	3	0.3223	1.83
8	DOE/NNSA/ LANL/SNL, USA	Trinity	301,056	8.101	10	0.1826	2.25
9	NASA / Mountain View, USA	Pleiades: SGI ICE X	243,008	5.952	13	0.1752	2.94
10	DOE/SC/ Argonne National Laboratory, USA	Mira: IBM BlueGene/Q,	786,432	8.587	9	0.1670	1.94

http://www.hpcg-benchmark.org/


HPCG
PRESENTED AT
SC16
NOVEMBER 15, 2016
NUMBER 3
3
SYSTEM
Oakforest-PACS
Joint Center for Advanced
High Performance Computing
JAPAN
ACHIEVED
0.385
Pflop/s
ICL
INNOVATIVE
Sandia
National
Laboratories
HPCG
3

Green 500 Ranking (SC16, November, 2016)

	Site	Computer	CPU	HPL Rmax (Pflop/s)	TOP500 Rank	Power (MW)	GFLOPS/W
1	NVIDIA Corporation	DGX SATURNV	NVIDIA DGX-1, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100	3.307	28	0.350	9.462
2	Swiss National Supercomputing Centre (CSCS)	Piz Daint	Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100	9.779	8	1.312	7.454
3	RIKEN ACCS	Shoubu	ZettaScaler-1.6 etc.	1.001	116	0.150	6.674
4	National SC Center in Wuxi	Sunway TaihuLight	Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway	93.01	1	15.37	6.051
5	SFB/TR55 at Fujitsu Tech. Solutions GmbH	QPACE3	PRIMERGY CX1640 M1, Intel Xeon Phi 7210 64C 1.3GHz, Intel Omni-Path	0.447	375	0.077	5.806
6	JCAHPC	Oakforest-PACS	PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path	1.355	6	2.719	4.986
7	DOE/SC/Argonne National Lab.	Theta	Cray XC40, Intel Xeon Phi 7230 64C 1.3GHz, Aries interconnect	5.096	18	1.087	4.688
8	Stanford Research Computing Center	XStream	Cray CS-Storm, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, Nvidia K80	0.781	162	0.190	4.112
9	ACCMS, Kyoto University	Camphor 2	Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect	3.057	33	0.748	4.087
10	Jefferson Natl. Accel. Facility	SciPhi XVI	KOI Cluster, Intel Xeon Phi 7230 64C 1.3GHz, Intel Omni-Path	0.426	397	0.111	3.837

http://www.top500.org/

運用

- 2017年度3月末までは無料（但し停止期間等あり）
- 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

来年度の利用募集中!!
- 筑波大：大規模一般利用（予定）
- 東大：パーソナル・グループコース、等

- 両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

HPCIへの資源提供

- 平成29年度課題募集におけるハードウェア資源一覧
  - http://www.hpci-office.jp/pages/h29_boshu_hpci_resource?parent_folder=23

筑波大学 計算科学研究センター	COMA(PACS-IX) ▶資源提供元	Xeon +Phi(KNC)	計算ノード：90ノード(約230TFLOPS) 1,800コア (＋メニーコアアーキテクチャ 10,980コア) 資源量：756,000ノード時間積 ストレージ：300TB
最先端共同HPC基盤施設 (JCAHPC)※	Oakforest-PACS ▶資源提供元	Xeon Phi(KNL)	計算ノード：1,600ノード(4,872 TFLOPS) 108,800コア 資源量：13,824,000ノード時間積 ストレージ：3,000TB
東京大学 情報基盤センター	スーパーコンピュータ FX10 ▶資源提供元	SPARC	計算ノード：1,500ノード (354.78TFLOPS) 24,000コア 資源量：12,960,000ノード時間積 ストレージ：500TB
	Reedbush-U ▶資源提供元	Xeon	計算ノード：60ノード(72.58TFLOPS) 2,160コア 資源量：518,400ノード時間積 ストレージ：60TB
	Reedbush-H ▶資源提供元	Xeon +Tesla	計算ノード：18ノード (193.11-212.73TFLOPS) コア数 648 + GPU 36枚 資源量：155,520ノード時間積 ストレージ：18TB

Oakforest-PACSのバッチジョブキュー

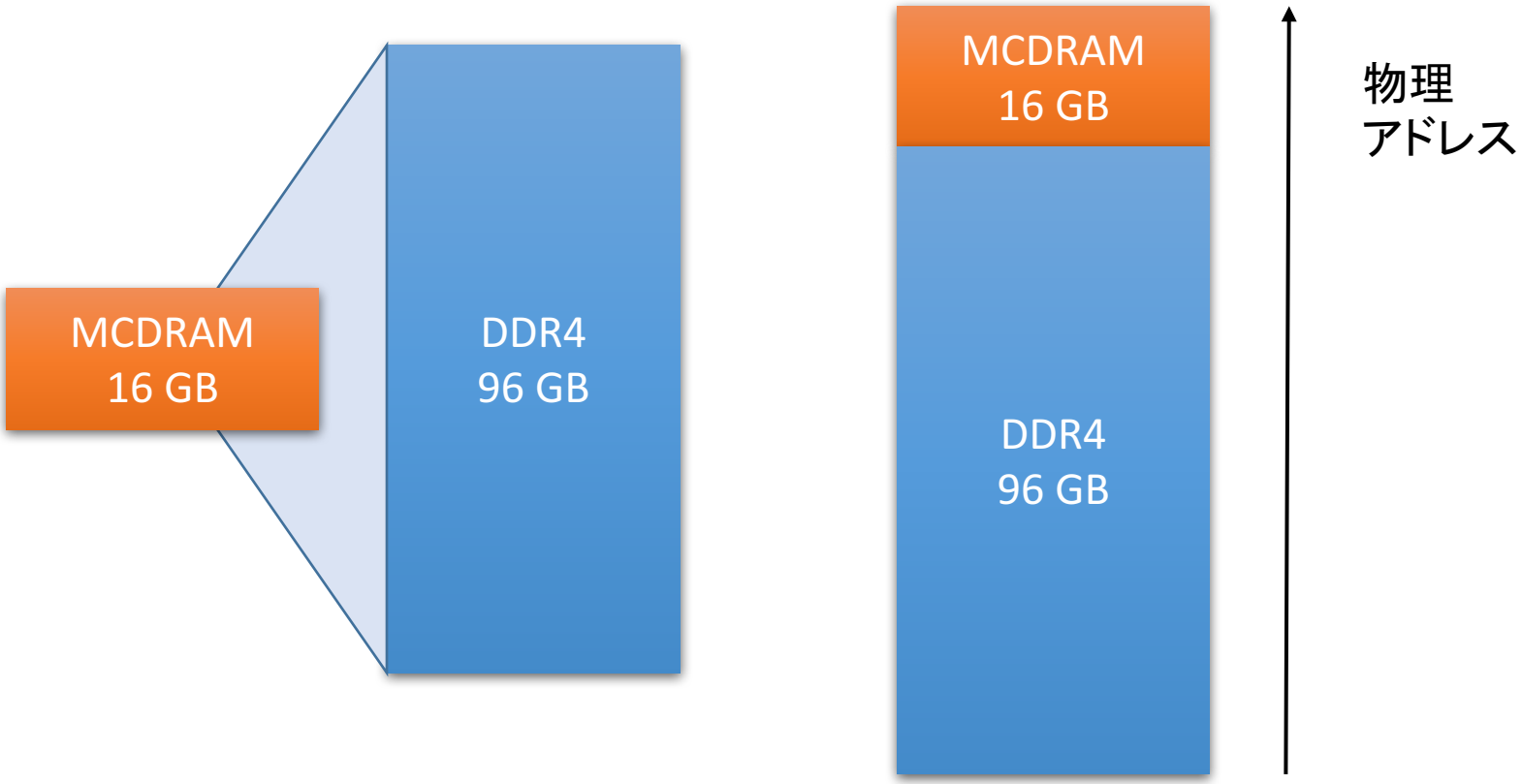
JCAHPC

- 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

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	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 medium large x-large	1-128 129-512 513-1024 1024-2048	48 h 48 h 48 h 24 h	82 (-cache)
				96 (-flat)
prepost	prepost	1	6 h	222 (Xeon)

メモリモード

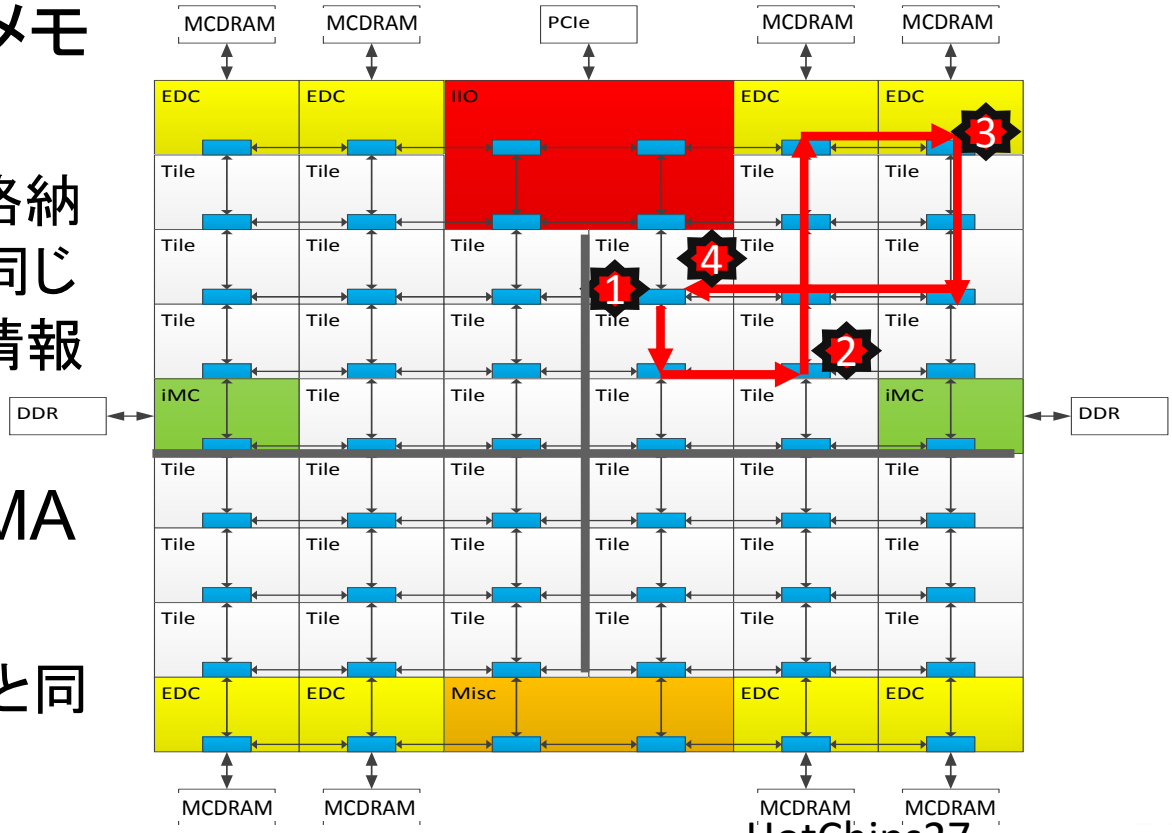
- Cacheモード
  - L3キャッシュとして動作
- Flatモード
  - 明示的に使い分け



クラスタリングモード

- Quadrant: ソフトウェアからはフラットなメモリ空間に見える
  - 内部では4分割、格納先のMCDRAMと同じ領域にキャッシュ情報（タグ）を配置
- SNC-4: 4つのNUMAドメインに見える
  - 4ソケットある場合と同じ

現在OFPではQuadrantのみ



HotChips27
KNLスライドより

利用実績

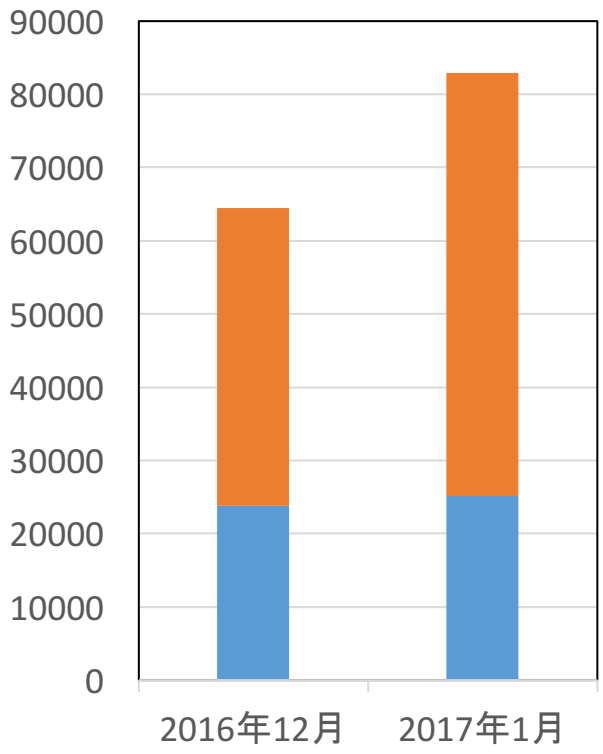
- 稼働率 = B / A
- 利用率 = C / B
  - A: 全計算ノード数 x 24時間 x 日数
  - B: 計算ノードが利用可能であった延べノード時間
  - C: ユーザジョブが動いていた延べノード時間

	稼働率	利用率
2016年12月	96.4 %	68.8%
2017年1月	87.1 %	81.7%

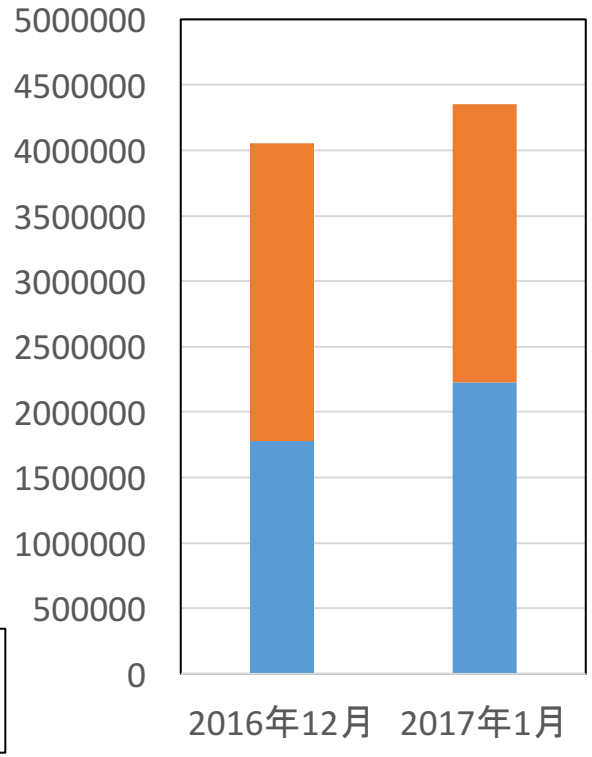
- 1月からの一般試験利用開始に伴うユーザ登録、ジョブキュー構成の変更等で、1月の稼働率はやや低い
- 1月の利用率は、システム導入直後としては非常に高い

メモリモードの使い分け

- ジョブ回数

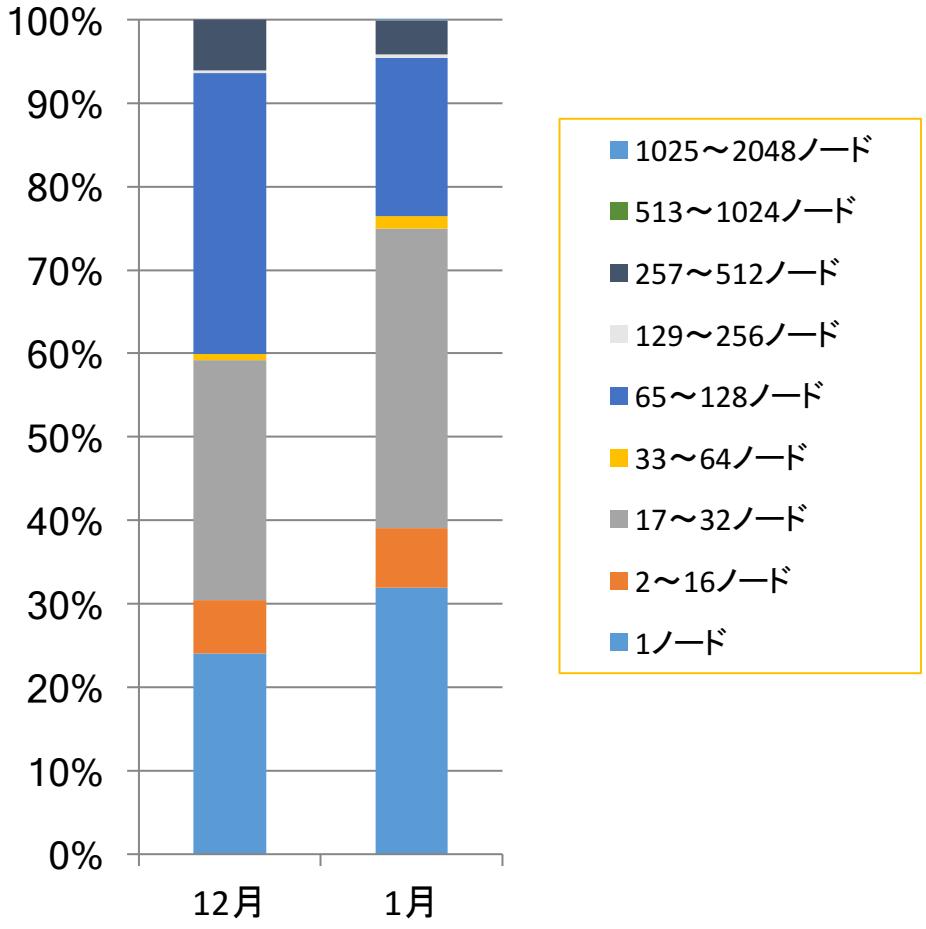


- 実行ノード時間

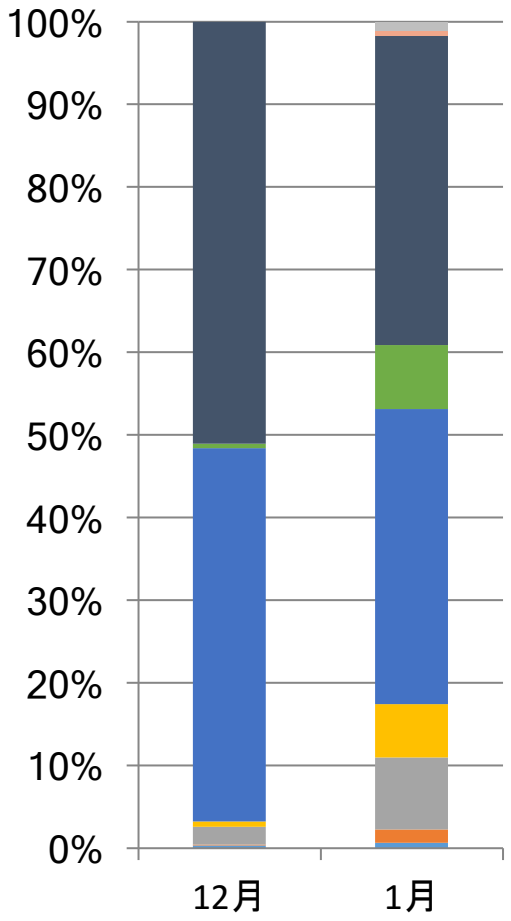


利用ノード数分布

・ジョブ回数



・実行ノード時間



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して設置
- 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
  - 2016/12から全系システム稼働、すでに高い利用率
  - HPCI資源を含め、通常の資源提供は2017/4から
- 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

JCAHPC

参考：Oakforest-PACSに関する情報

- 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=3954674183946240 1
  - 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