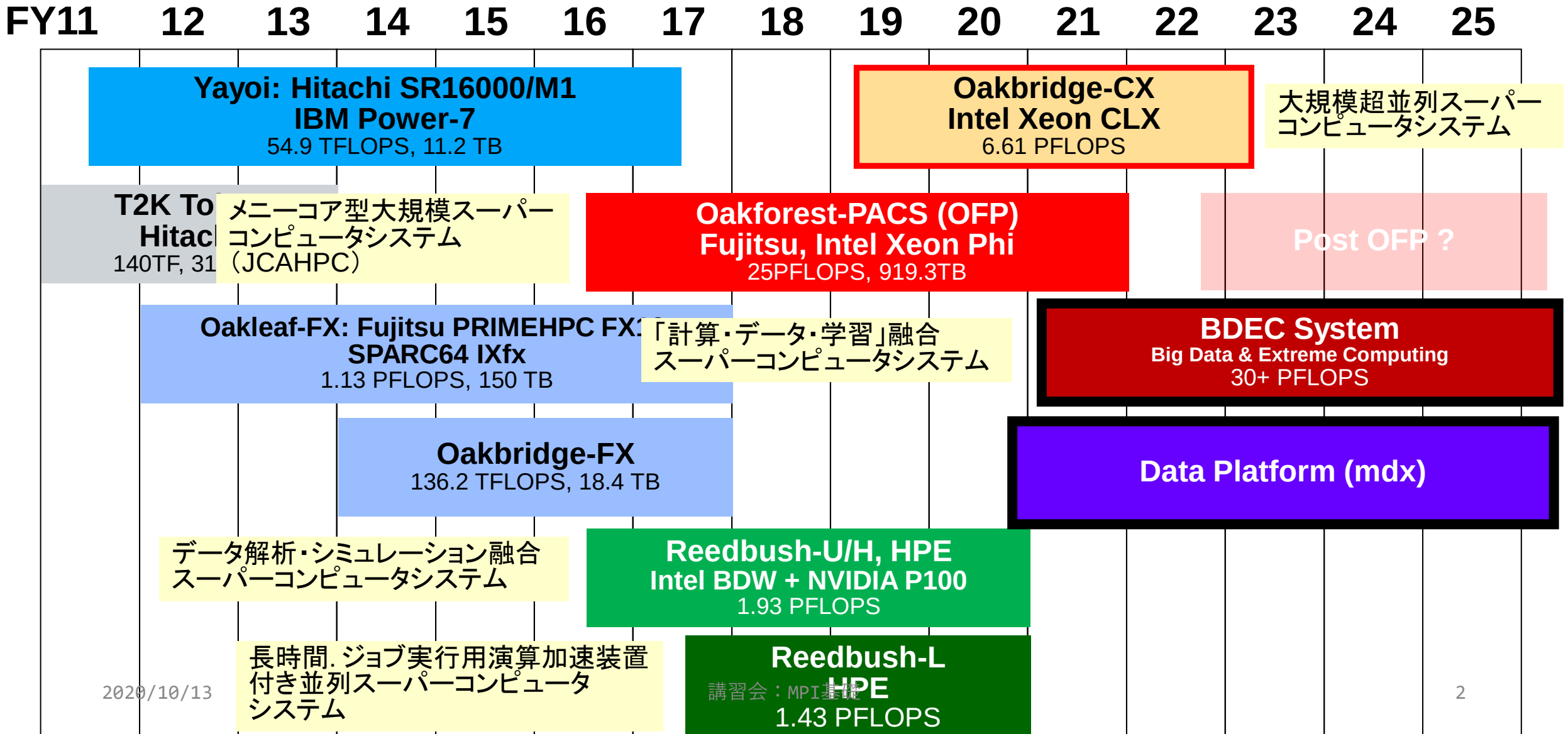
第141回 お試しアカウント付き 並列プログラミング講習会 「MPI基礎：並列プログラミング入門」

東京大学 情報基盤センター
三木 洋平

東大情報基盤センターのスパコン



3システム：利用者2600+，学外55+%

- Reedbush (HPE, Intel BDW + NVIDIA P100 (Pascal))（本郷）
  - データ解析・シミュレーション融合スーパーコンピュータ
  - 3.36 PF, 2016年7月～ 2021年3月末（予定）
    - Reedbush-U（CPU only，2020年6月30日で退役）
    - Reedbush-H（2GPU's/n），Reedbush-L（4GPU's）
  - 東大ITC初GPUクラスタ (2017年3月より), DDN IME (Burst Buffer)
- Oakforest-PACS (OFP) (富士通, Intel Xeon Phi (KNL))（柏）
  - JCAHPC (筑波大CCS＆東大ITC)
  - 25 PF, TOP500で18位（日本3位）（2020年6月）
  - Omni-Path アーキテクチャ, DDN IME (Burst Buffer)
- Oakbridge-CX (富士通, Intel Xeon Platinum 8280)（柏）
  - 大規模超並列スーパーコンピュータシステム
  - 6.61 PF, 2019年7月 ～ 2023年6月，TOP500で60位（2020年6月）
  - 全1,368ノードの内128ノードにSSDを搭載

Reedbush

- GPUクラスタ

NVIDIA Tesla P100 SXM2

SGI C1102-GP8

外部接続ルータ 1Gigabit/10Gigabit Ethernet Network

x6

x9

ログインノード群
SGI Rackable C1110-GP2
6nodes
E5-2680v4 2.4GHz
14core,256GiB Mem

管理サーバ群
SGI Rackable C1110-GP2
9nodes
E5-2680v4 2.4GHz
14core,128GiB Mem

Reedbush-U
SGI Rackable C2112-4GP3
420 nodes, 508.03TFLOPS
・CPU : E5-2695v4 2.1GHz 18core

Reedbush-H
SGI Rackable C1102-GP8
120 nodes, 240GPUs, 1.418PFLOPS
・CPU : E5-2695v4 2.1GHz 18core
・GPU : NVIDIA Tesla P100 SXM2 x2/node

Reedbush-L
SGI Rackable C1102-GP8
64 nodes, 256GPUs, 1.434PFLOPS
・CPU : E5-2695v4 2.1GHz 18core
・GPU : NVIDIA Tesla P100 SXM2 x4/node

x420

x120

x64

ライフ／管理ネットワーク 1Gigabit/10Gigabit Ethernet Network

GbE SW

NFS Filesystem
16TB

NAS Storage
24TB

x4

x6

x9

x2

x2(for PBS)

Reebush-U
x420

Reedbush-H
x240 (FDRx2/node)

Reedbush-L
x128(EDR x2/node)

InterConnect(4x EDR InfiniBand)

InterConnect(4x EDR InfiniBand)

x12

x4
(MDS:x 2)

x24
(OSS(VM):x 12 x2)

x36(IME:6x6)

x16(IME:8x2)

Management port

Lustre Filesystem
DDN SFA14KE x3set
5.04PB

高速キャッシュ
DDN IME14K x6set
209TB

管理コンソール
Mac Pro

電力管理サーバ

電力計器

高速キャッシュ
DDN IME240 x8set
153.6 TB

管理用補助
サーバ
SGI Rackable
C1110-GP2 x2

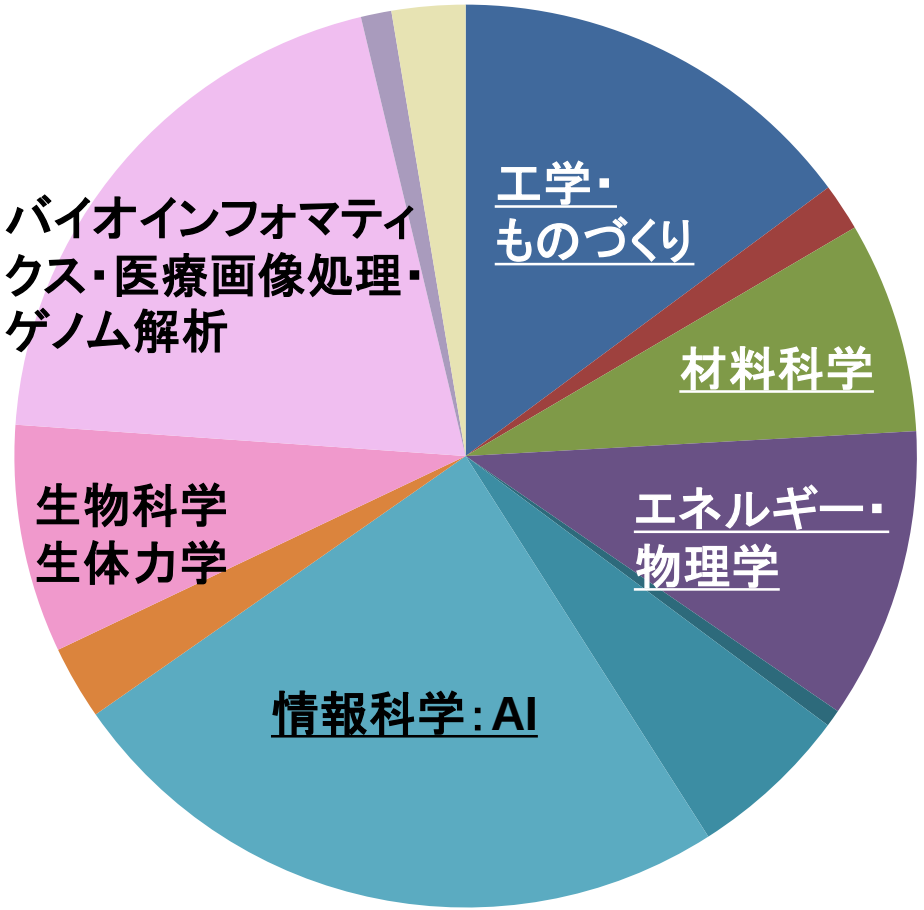
x10(Ctrl:8,MDS:2)

x12

x8

x2

研究分野別利用CPU時間割合（2019年度）



- 工学・ものつくり
- 地球科学・宇宙科学
- 材料科学
- エネルギー・物理学
- 情報科学：システム
- 情報科学：アルゴリズム
- 情報科学：AI
- 教育
- 産業利用
- 生物科学・生体力学
- バイオインフォマティクス
- 社会科学・経済学
- データ科学・データ同化

従来からの工学・材料・物理学の研究に加えて、機械学習・AI，医用画像解析・ゲノム解析などの研究にも多く利用

Reedbush-H
Intel BDW+ NVIDIA P100 x2 / node

最先端共同HPC基盤施設 JCAHPC
Joint Center for Advanced High Performance Computing

- 2013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
  - 筑波大学計算科学研究センター
  - 東京大学情報基盤センター

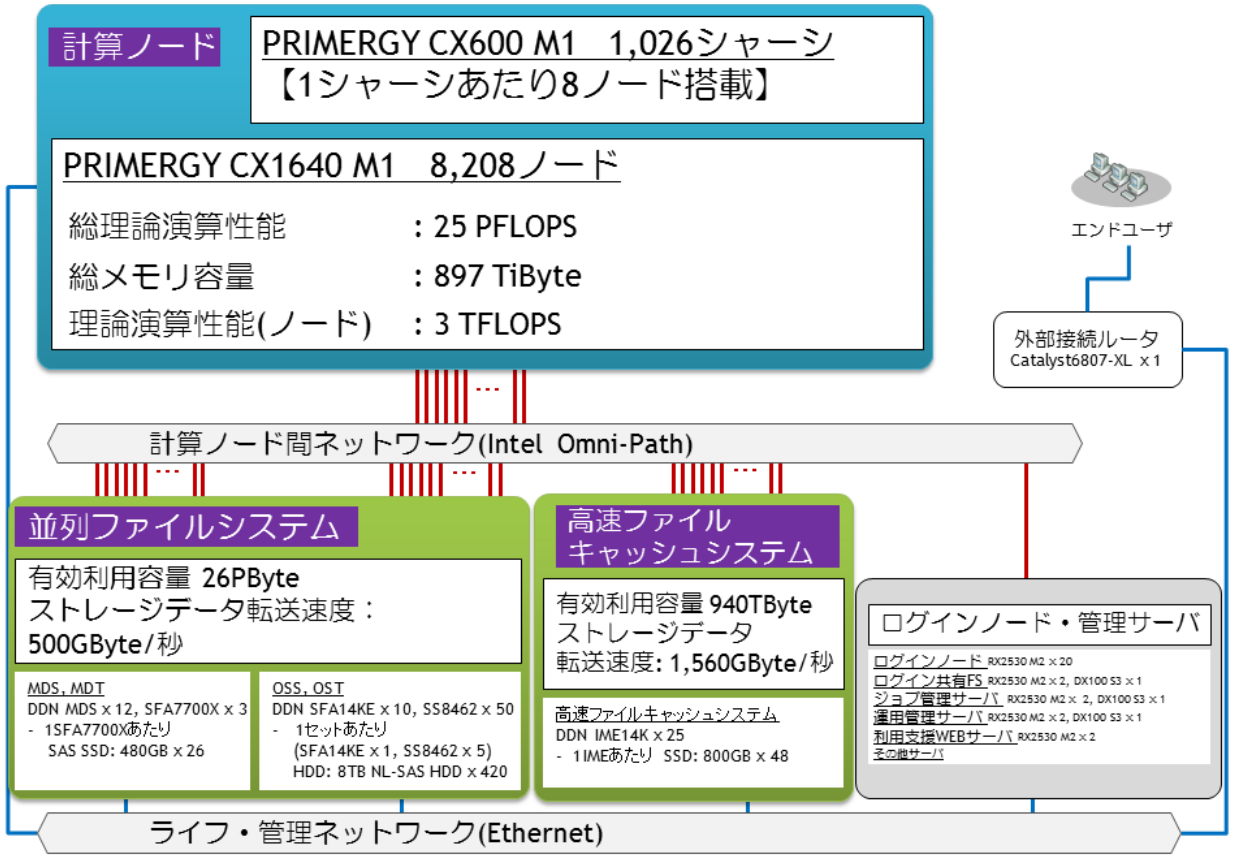
- 東京大学柏Iキャンパスの東京大学情報基盤センター内に，両機関の教職員が中心となって設計するスーパーコンピュータシステムを設置し，最先端の大規模高性能計算基盤を構築・運営するための組織
  - http://jcahpc.jp

Oakforest-PACS (OFP)

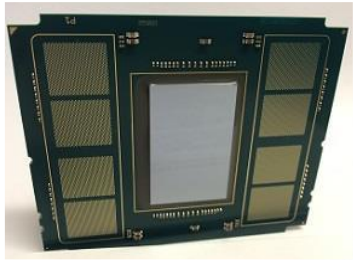
- Top500: 世界第18位
- HPCG: 世界第11位
  （2020年6月現在）

Fujitsu
PRIMERGY CX1640 M1

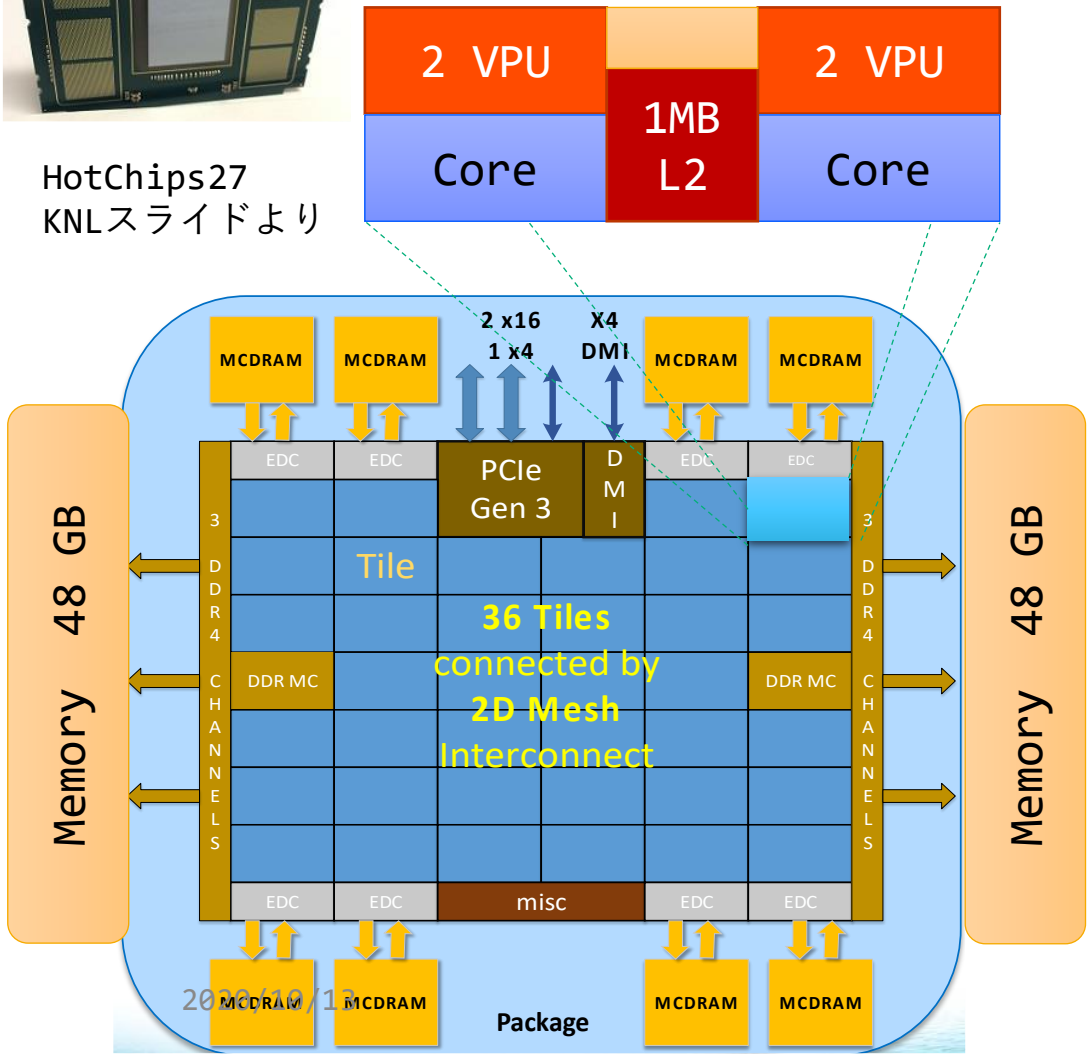
Fujitsu PRIMERGY CX600 M1
シャーシ当たりCX1640 M1 ×8搭載



Oakforest-PACS計算ノード



HotChips27
KNLスライドより

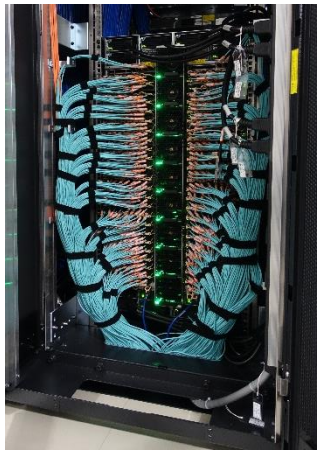


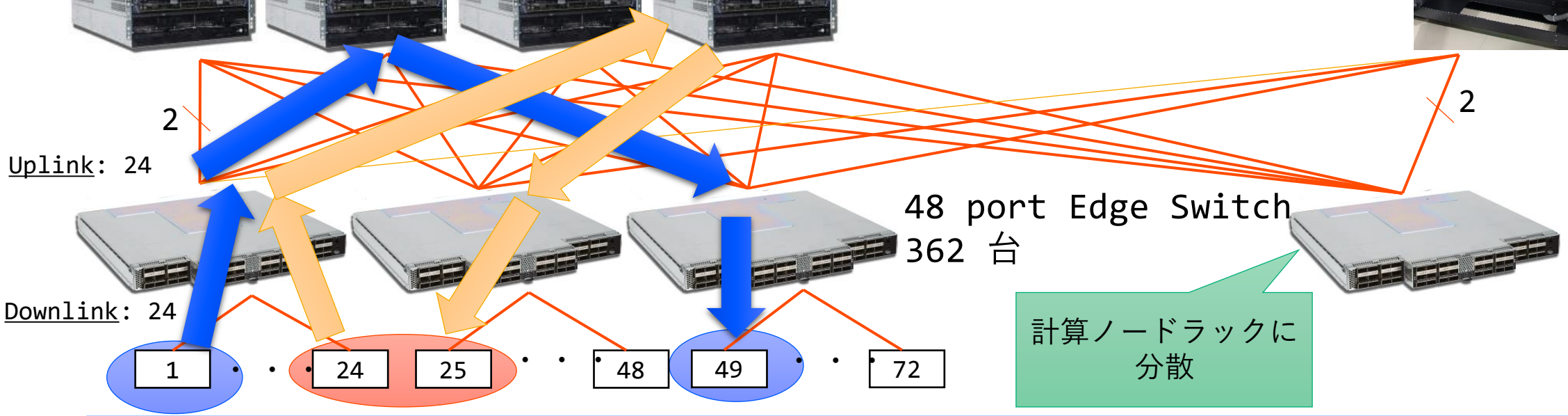
Intel Xeon Phi 7250 (Knights Landing)

- 1CPU（1ノード）当たり68コアのメニーコアプロセッサ
- 1.4 GHz，クロック当たり32回の倍精度実数演算（Double Precision, DP）
  - コア当たり最大性能⇒1.4×32= 44.8 GFLOPS（1秒間に448億回の倍精度実数演算）
  - 1CPU 56コア，3,046.4 GFLOPS= 3.046 TFLOPS（1秒間に3兆464億回演算）
- MCDRAM: オンパッケージ高バンド幅メモリ搭載 16 GB, 490 GB/秒以上(実測)
  - DDR4 96 GB, 85 GB/秒（実測）
- 全系 8,208ノード： 25.004 PFLOPS (1秒間に2京5,004兆回演算)

Intel Omni-Path Architectureによるフルバイセクションバンド幅Fat-tree網

768 port Director Switch
12台
(Source by Intel)





フルバイセクションバンド幅を維持

- フルバイセクションバンド幅：全ての計算ノードから同時に異なる計算ノードに通信しても性能が低下しない

→ 計算ノードの物理配置を気にせずに必要な数の計算ノードを確保できる

ファイルシステム（ストレージ）

並列ファイルシステム：Lustreファイルシステム

- 容量：26.2 PB
- バンド幅性能：500 GB/秒
  - サーバ当たり 50 GB/秒 x 10台
- 全10セット＋メタデータ（管理）サーバで構成
  - 1セット当たり、サーバ兼コントローラ+5エンクロージャ
- 合計 4,200個の 8TB ハードディスク，20%冗長化 (8D2P)

IO500 9位 (2018年11月), 15位 (2019年6月)

高速ファイルキャッシュ（バーストバッファ）：Infinite Memory Engine (IME)

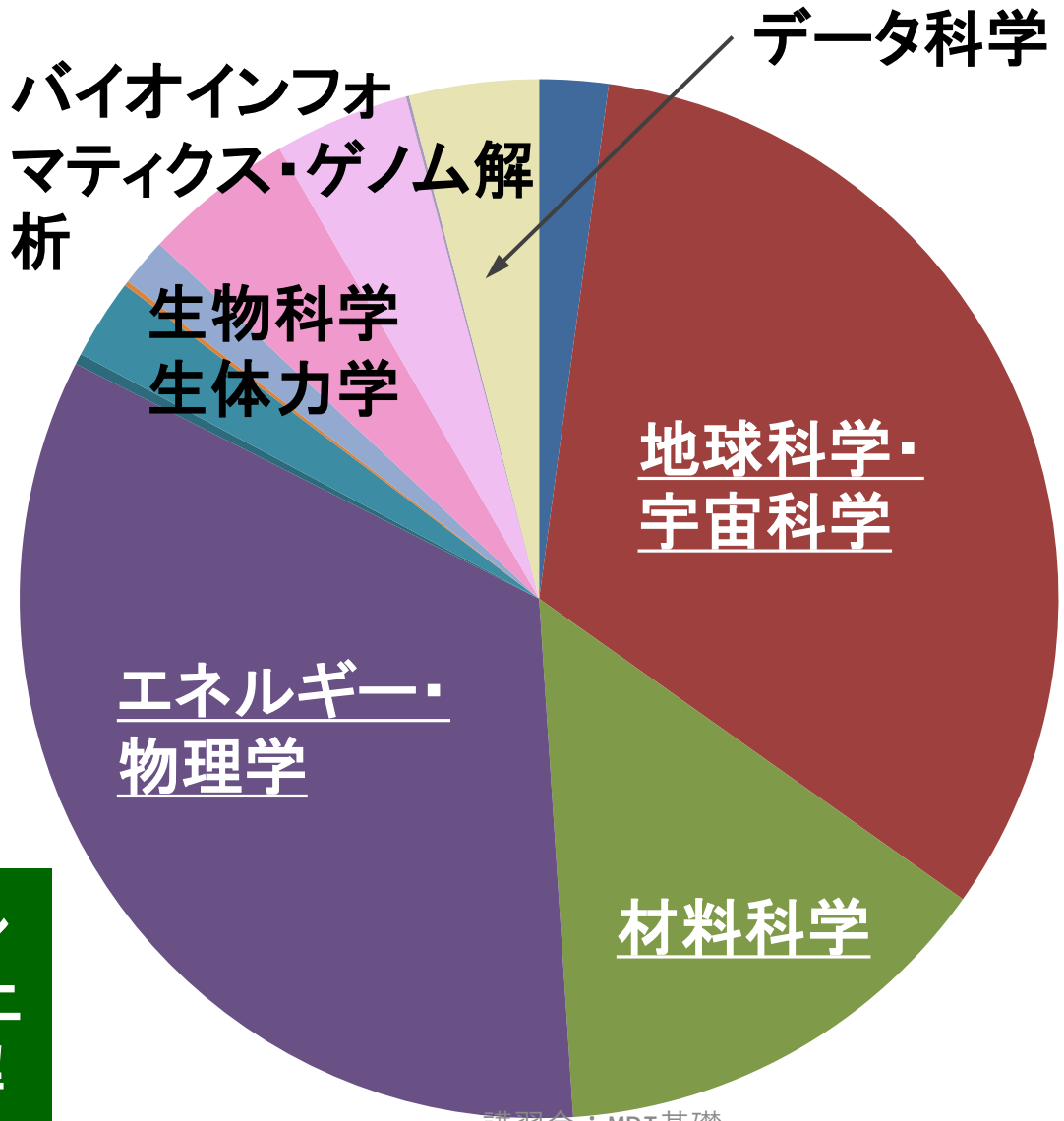
- 容量： 940 TB
- バンド幅性能: 1.56 TB/秒 = 1,560 GB/秒
- NVMe接続SSDを使用
  - 合計1,200枚の 800 GB SSD, 冗長化 (Erasure coding)
  - 単体 1.3 GB/s×48本×25台 = 1,560 GB/s

IO-500 1位（2017年11月、2018年6月）
4位 (2018年11月), 3位 (2019年6月)

Oakforest-PACS全景

- 全体： 102ラック
  （6列：計算x4列 + 他x2列）
  - 計算ノード： 69ラック
    - 2ラックで1単位
      （1箇所だけ3ラックで1単位）
  - インタコネクト(OmniPath) Director スイッチ： 12ラック
  - ファイルシステム： 16ラック
    - メタデータ（管理情報）：1ラック
    - 並列ファイルシステム：10ラック
    - ファイルキャッシュ： 5 ラック
  - その他： 5ラック
    - 管理ネットワーク機器： 2ラック
    - ログインノード+プリポスト+Webポータル：1ラック
    - 管理サーバ： 1ラック
    - 外部接続ルータ： 1ラック
- 消費電力：3.37 MW
  冷却含め　4.24 MW
  - 水冷 3.0 MW分程度,
    残り 0.4 MW分程度は空冷

研究分野別利用CPU時間割合（2019年度）



- 工学・ものつくり
- 地球科学・宇宙科学
- 材料科学
- エネルギー・物理学
- 情報科学：システム
- 情報科学：アルゴリズム
- 情報科学：AI
- 教育
- 産業利用
- 生物科学・生体力学
- バイオインフォマティクス
- 社会科学・経済学
- データ科学・データ同化

地球・宇宙科学とエネルギー・物理学で半分以上を占める一方、データ解析にも利用

Oakbridge-CX

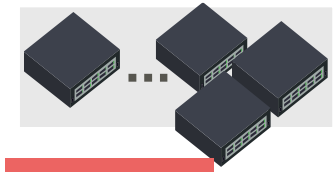
- 世界第60位
  （2020年6月現在）

Fujitsu PRIMERGY CX2550 M5

Fujitsu PRIMERGY CX400 M1
シャーシ当たりCX2550 M5×4搭載

計算ノード

Chassis: PRIMERGY CX400 M4 x342 <4node / Chassis>
Node: PRIMERGY CX2550 M5 x1,240, CX2560 M5 x128



x1,368 node

全体性能

理論演算性能: 6.61PF
主記憶容量: 256.5TiB
メモリバンド幅: 385.1TB/s
ラック数: 21ラック
SSD搭載: 128ノード

ノード単体

理論演算性能: 4.8384 TF
手記憶容量: 192GiB
メモリバンド幅: 281.6GB/s

計算ノード間ネットワーク (Omni-Path Architecture)
通信性能 100Gbps

ログインノード

FUJITSU Server
PRIMERGY CX2560 M5 x 10

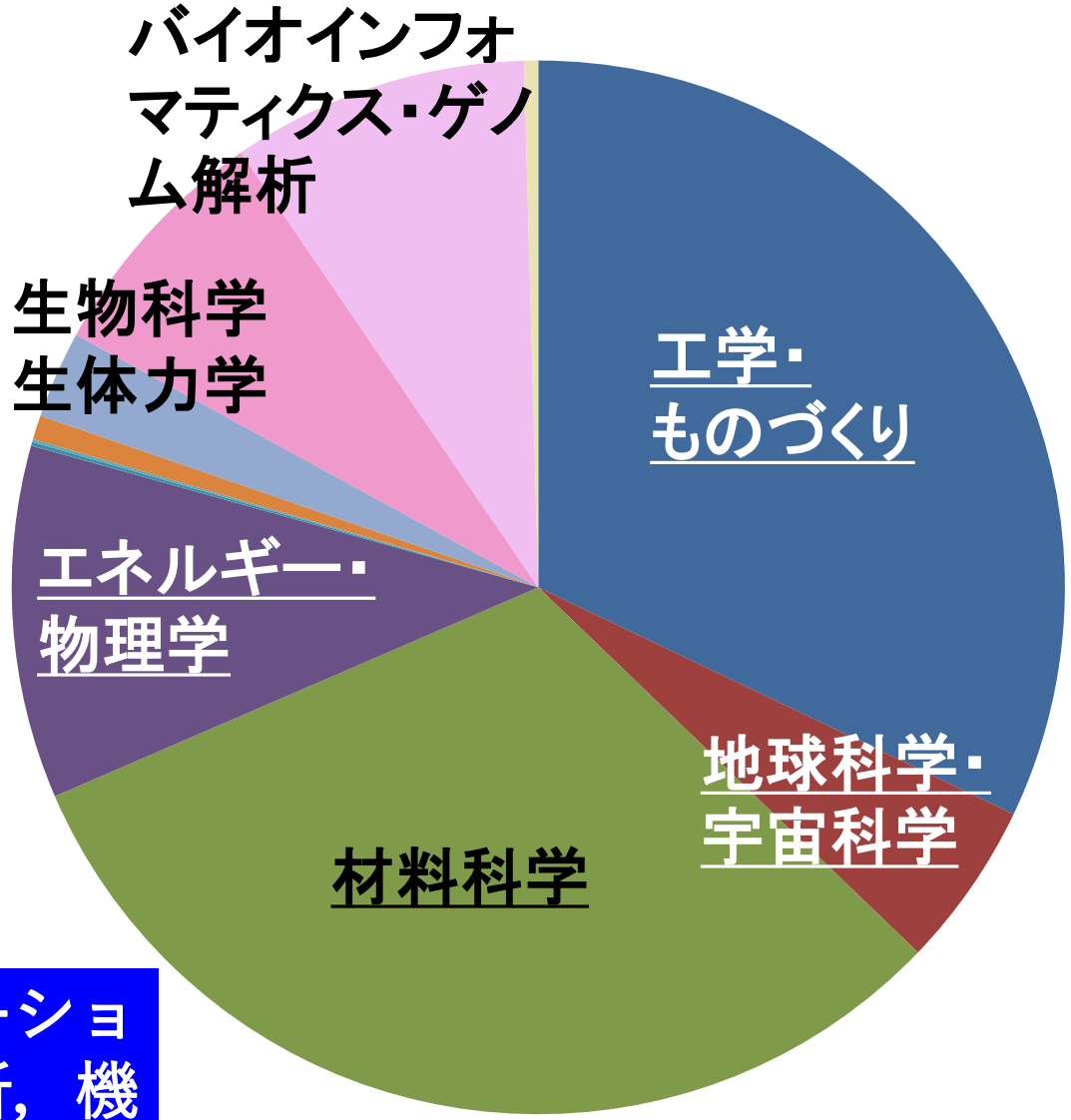
管理サーバ群

FUJITSU Server
PRIMERGY RX2530 M4 x 15
(ジョブ、運用、認証、Web、
セキュリティログ保存)

並列ファイルシステム

ストレージ装置: DDN ES18KE x2セット
ファイルシステム: DDN ExaScaler
(Lustreベースファイルシステム)

研究分野別利用CPU時間割合（2019.10～2020.9）

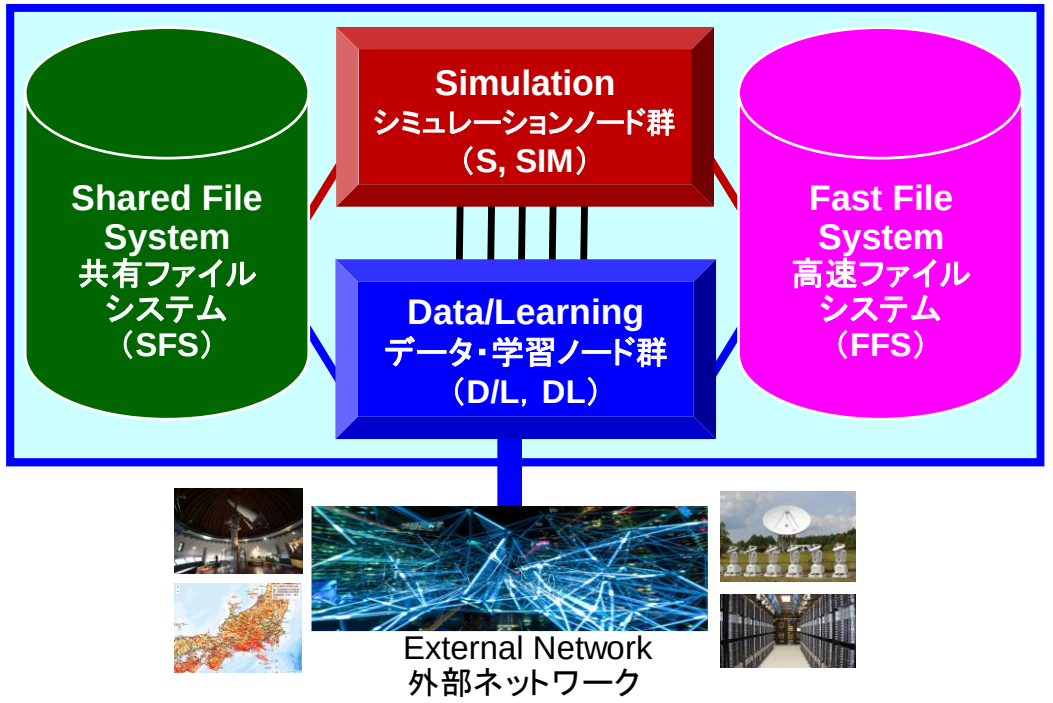


- 工学・ものつくり
- 地球科学・宇宙科学
- 材料科学
- エネルギー・物理学
- 情報科学：システム
- 情報科学：アルゴリズム
- 情報科学：AI
- 教育
- 産業利用
- 生物科学・生体力学
- バイオインフォマティクス
- 社会科学・経済学
- データ科学・データ同化

計算科学シミュレーションの他，データ解析，機械学習分野にも利用

BDEC: Big Data & Extreme Computing

- **BDEC (Big Data & Extreme Computing)**
  - 2021年5月（以降）運用開始予定
  - 30+ PF, ~4.5 MVA（空調込み）, ~360m^2
  - （D+L）による（S）の高度化　AI for HPC
- 「シミュレーション・データ・学習（S+D+L）」融合，2種類のノード群
  - シミュレーションノード群（S, SIM）
    - 従来のスパコン
    - CPU　with HBM, 25+PF, 2+PB/sec （最大9.0）
  - データ・学習ノード群（D/L, DL）
    - データ解析，機械学習
    - GPU Cluster, 5+PF, 575TB/sec
    - データ・学習ノード群の一部は外部リソース（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
  - Hierarchical, Hybrid, Heterogeneous (h3)
- ファイルシステム：共有（大容量）＋高速



55th TOP500 List (June 2020)

https://www.top500.org/lists/top500/

	Name	Computer Site	Cores	Rmax (Tflop/s)	Rpeak (Tflop/s)	Power (kW)
1	**Supercomputer Fugaku**	Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D RIKEN Center for Computational Science	7299072	**415530 (=415.5 PF)**	513854.67	**28334.5**
2	**Summit**	IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband DOE/SC/Oak Ridge National Laboratory	2414592	**148600 (=148.6 PF)**	200794.88	**10096**
3	**Sierra**	IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband DOE/NNSA/LLNL	1572480	**94640 (= 94.6 PF)**	125712	**7438.28**
4	**Sunway TaihuLight**	Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway National Supercomputing Center in Wuxi	10649600	**93014.59 (= 93.0 PF)**	125435.9	**15371**
5	**Tianhe-2A**	TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 National Super Computer Center in Guangzhou	4981760	**61444.5 (= 61.4 PF)**	100678.66	**18482**
6	**HPC5**	PowerEdge C4140, Xeon Gold 6252 24C 2.1GHz, NVIDIA Tesla V100, Mellanox HDR Infiniband Eni S.p.A.	669760	**35450 (= 35.5 PF)**	51720.76	**2252.17**
7	**Selene**	DGX A100 SuperPOD, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband NVIDIA Corporation	272800	**27580 (= 27.6 PF)**	34568.6	**1344.19**
8	**Frontera**	Dell C6420, Xeon Platinum 8280 28C 2.7GHz, Mellanox InfiniBand HDR Texas Advanced Computing Center/Univ. of Texas	448448	**23516.4 (= 23.5 PF)**	38745.91	
9	**Marconi-100**	IBM Power System AC922, IBM POWER9 16C 3GHz, Nvidia Volta V100, Dual-rail Mellanox EDR Infiniband CINECA	347776	**21640 (= 21.6 PF)**	29354	**1476**
10	**Piz Daint**	Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 Swiss National Supercomputing Centre (CSCS)	387872	**21230 (= 21.2 PF)**	27154.3	**2384.24**
18	**Oakforest-PACS**	PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path Joint Center for Advanced High Performance Computing	556104	**13554.6 (= 13.6 PF)**	24913.46	**2718.7**

Green500 List (June 2020)

https://www.top500.org/lists/green500/

	Name	Accelerator/Co-Processor	GFlops/Watts	Power (kW)	Rmax (TFlop/s)	TOP500 Rank
1	MN-3	MN-Core	**21.10807292**	76.8	1621.1	393
2	Selene	NVIDIA A100	**20.51793273**	1344.19	27580	7
3	NA-1	PEZY-SC2 700Mhz	**18.433**	80.17	1303.22	468
4	A64FX prototype		**16.87626604**	118.48	1999.5	204
5	AiMOS	NVIDIA Volta GV100	**16.28456491**	512.08	8339	26
6	HPC5	NVIDIA Tesla V100	**15.74037484**	2252.17	35450	6
7	Satori	NVIDIA Tesla V100 SXM2	**15.57446809**	94	1464	421
8	Summit	NVIDIA Volta GV100	**14.71870048**	10096	148600	2
9	Supercomputer Fugaku		**14.66516085**	28334.5	415530	1
10	Marconi-100	Nvidia Volta V100	**14.66124661**	1476	21640	9
20	Reedbush-L	NVIDIA Tesla P100	**10.167**	79	805.6	June'18
28	Reedbush-H	NVIDIA Tesla P100	**8.576**	94	802.4	June'18

（理論）ピーク性能

- OFPに搭載されているXeon Phi 7250のピーク性能：
  - コア数： 68コア
  - コア当たりのAVX-512 ユニット： 2
  - AVX-512ユニット当たりの同時演算数（倍精度）： 8 (= 512 / 64)
  - 積和演算 (Fused Multiply Add: FMA) : 2 に換算
  - クロック周波数: 1.40 GHz
- ノード当たりピーク性能： 68 * 2 * 8 * 2 * 1.40 = 3046.4 GFlop/s
- しかし、AVX-512ユニットは実は1.40 GHzでは動作しない（より低い周波数）
  - ピークに近い性能が得られるはずのもの（OFPでの実測値）
    - DGEMM (倍精度の行列積)： 2200 GFlop/s (ピーク比：72%)
    - HPL: 2000 GFlop/s (ピーク比： 66%)
  - Top500におけるOFPの登録値は理論ピークの54.4%
- 「ピーク性能」の定義を正しく把握しておくことが重要！！
  - CPUメーカーによって定義の仕方も異なる

利用制度の紹介

- 一般利用
  - 大学・公共機関に在籍の方（大学院生は代表者としては申し込めません）
  - 電気代相当料金の利用負担金支払いが必要
- 企業利用
  - 企業に在籍の方
  - 利用負担金は一般利用の約1.2倍
  - 書面・ヒアリング審査あり，成果報告（公開）義務あり
- 若手・女性利用
  - 大学・公共機関に在籍の方
  - 4月1日現在40歳以下の若手，または女性，または学生
  - 利用負担金なし
  - 書類審査あり，成果報告義務あり
- 学際大規模情報基盤共同利用・共同研究拠点（JHPCN）への課題申請
- HPCI課題への課題申請

東大情報基盤センターOakforest-PACSスーパーコンピュータシステムの料金表（2020年4月1日）

- https://www.cc.u-tokyo.ac.jp/guide/application/
  - 「利用負担金」を参照
- パーソナルコース
  - 基本セット 50,000円： 6セットまで、最大2048ノードまで
    トークン：1ノード x 24時間 x 360日分 = 8640ノード時間（1セット）
- グループコース
  - 基本セット 50,000円 (企業 60,000円)： 最大2048ノードまで
    トークン：1ノード x 24時間 x 360日分 = 8640ノード時間 （1セット）
- 以上は、「トークン制」で運営
  - 大学等のユーザはOakbridge-CX, Reedbushとの相互トークン移行も可能

GFLOPS（ピーク性能換算）あたり負担金（~W）

System	JPY/GFLOPS
Reedbush-U (HPE) (Intel BDW)	61.9
Reedbush-H (HPE) (Intel BDW+NVIDIA P100x2/node)	15.9
Reedbush-L (HPE) (Intel BDW+NVIDIA P100x4/node)	13.4
Oakforest-PACS (Fujitsu) (Intel Xeon Phi/Knights Landing)	16.5
Oakbridge-CX (Fujitsu) (Intel Cascade Lake (CLX))	20.7

トライアルユース制度について

- 安価に当センターのシステムが使える「無償トライアルユース」および「有償トライアルユース」制度があります
  - アカデミック利用
    - パーソナルコース、グループコースの双方（1ヶ月～3ヶ月）
  - 企業利用
    - パーソナルコース（１ヶ月～３ヶ月）（RB-H，L：最大4ノード、OFP：最大16ノード、OBCX：最大8ノード）
      本講習会の受講が必須、審査無
    - グループコース
      - 無償トライアルユース：（１ヶ月～３ヶ月）：無料（ RB-H：最大32ノード、RB-L：最大16ノード、OFP：最大2048ノード、OBCX:最大256ノード）
      - 有償トライアルユース：（１ヶ月～最大通算９ヶ月）、有償（計算資源は無償と同等）
      - スーパーコンピュータ利用資格者審査委員会の審査が必要（年2回実施）
    - 双方のコースともに、簡易な利用報告書の提出が必要
- トライアルユース(無料体験):利用申請するシステムを初めて使う場合
  - 1ヶ月間、講習会アカウントと同条件
  - 一般利用、有償トライアルユースへの移行も可能

スーパーコンピュータシステムの詳細

- https://www.cc.u-tokyo.ac.jp/guide をご覧ください
  - 利用申請方法
  - 運営体系
  - 料金体系
  - 利用の手引き