第93回　お試しアカウント付き
並列プログラミング講習会

Reedbush スパコンを用いたGPU ディープラーニング入門

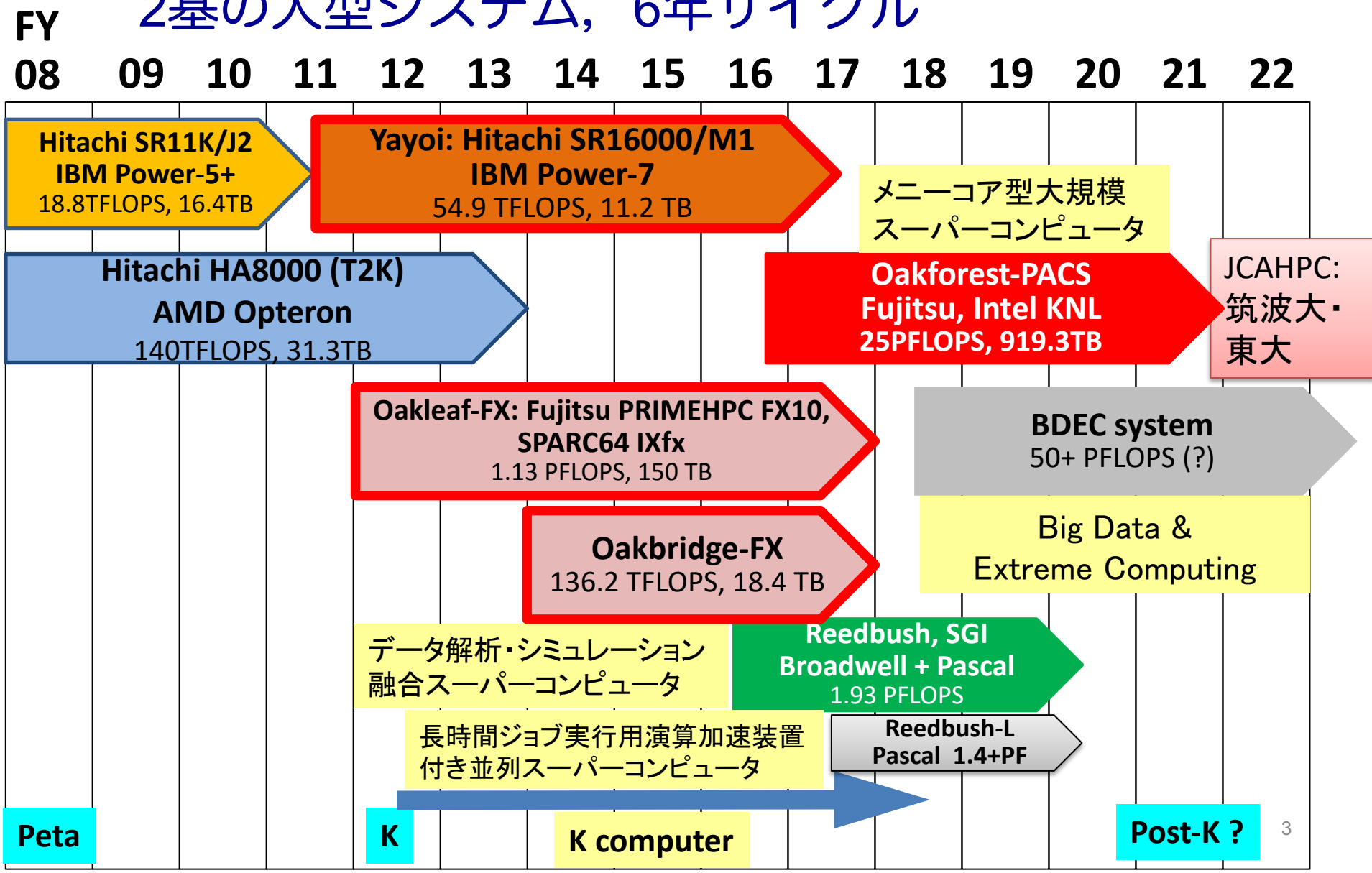
東京大学 情報基盤センター

担当：下川辺 隆史
shimokawabe @ cc.u-tokyo.ac.jp

東大情報基盤センター
スーパーコンピュータの概略

東大センターのスパコン

2基の大型システム，6年サイクル



4システム運用中

- Oakleaf-FX (富士通 PRIMEHPC FX10)
  - ✓ 1.135 PF, 京コンピュータ商用版, 2012年4月 ～ 2018年3月

- Oakbridge-FX (富士通 PRIMEHPC FX10)
  - ✓ 136.2 TF, 長時間実行用（168時間）, 2014年4月 ～ 2018年3月
- Reedbush (SGI, Intel BDW + NVIDIA P100 (Pascal))
  - ✓ 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
  - ✓ 3.361 PF, 2016年7月～ 2020年6月
  - ✓ 東大ITC初のGPUシステム (2017年3月より), DDN IME (Burst Buffer)

- Oakforest-PACS (OFP) (富士通、Intel Xeon Phi (KNL))
  - ✓ JCAHPC (筑波大CCS＆東大ITC)
  - ✓ 25 PF, TOP 500で6位 (2016年11月) (日本で1位)
  - ✓ Omni-Path アーキテクチャ, DDN IME (Burst Buffer)

東京大学情報基盤センター スパコン（1/3）

Fujitsu PRIMEHPC FX10 （FX10スーパーコンピュータシステム）

Total Peak performance	: 1.13 PFLOPS
Total number of nodes	: 4,800
Total memory	: 150TB
Peak performance per node	: 236.5 GFLOPS
Main memory per node	: 32 GB
Disk capacity	: 2.1 PB
SPARC64 IXfx 1.848GHz	

2012年7月~2018年3月（予定）

Oakbridge-FX

：長時間ジョブ用のFX10
ノード数：24～576
制限時間：最大168時間
（1週間）

東京大学情報基盤センター スパコン（2/3）

Reedbush（SGI Rackable クラスタシステム）

Reedbush-U (2016/7/1 ~)
- 理論性能：508TFlops
- ノード数：420
- ノード構成：Intel Xeon Broadwell x2

Reedbush-H (2017/3/1 ~)
- 理論性能：1418TFlops
- ノード数：120
- ノード構成：Intel Xeon Broadwell x2 + **NVIDIA P100 GPU x2**

Reedbush-L (**2017/10/1 ~**)
- 理論性能：1435TFlops
- ノード数：64
- ノード構成：Intel Xeon Broadwell x2 + **NVIDIA P100 GPU x4**

東京大学情報基盤センター スパコン（3/3）

筑波大学計算科学研究センターと共同運用

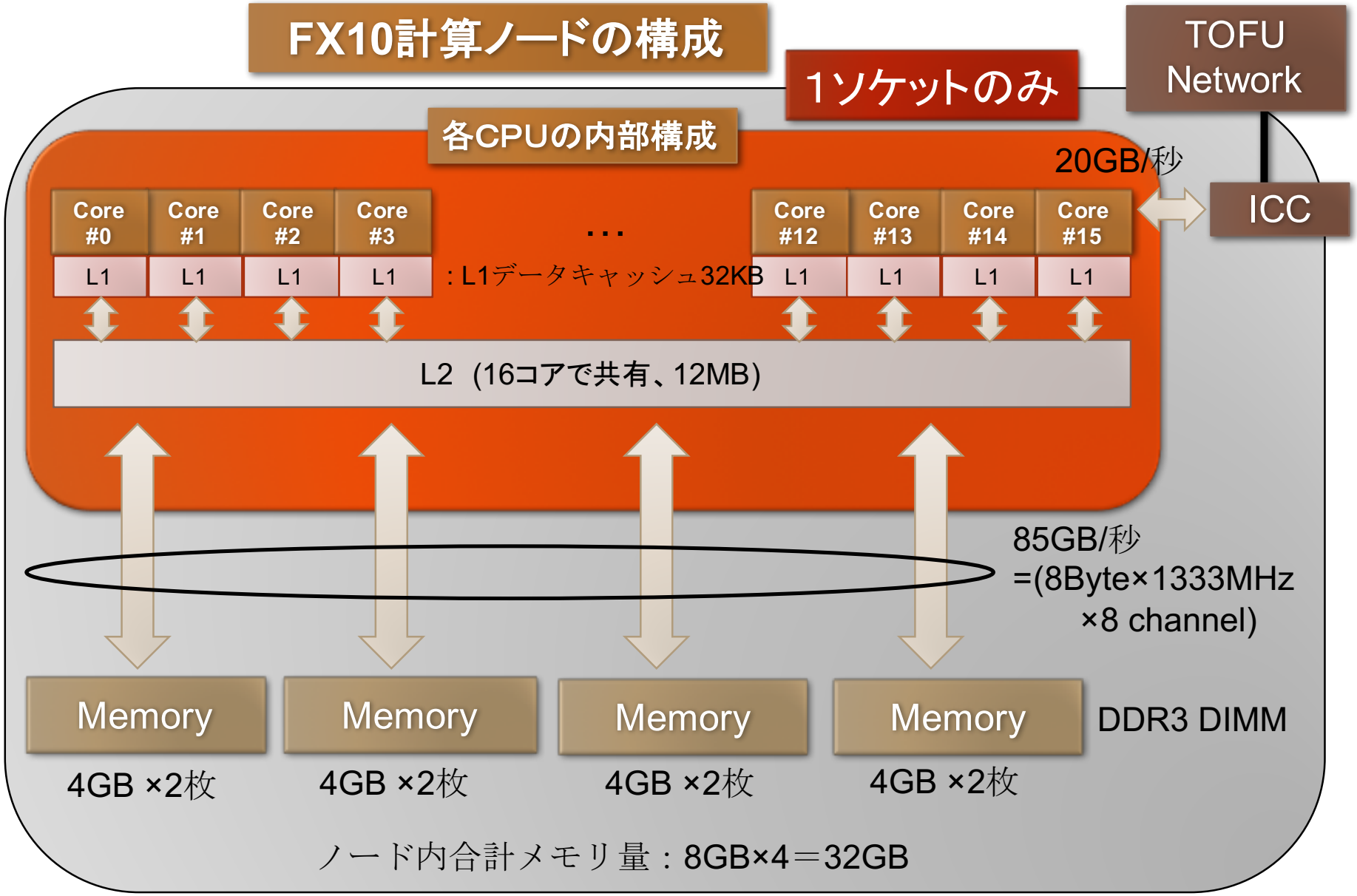
Oakforest-PACS （Fujitsu PRIMERGY CX600）

Total Peak performance	: 25 PFLOPS
Total number of nodes	: 8,208
Total memory	: 897.7 TB
Peak performance per node	: 3.046 TFLOPS
Main memory per node	: 96 GB (DDR4)
	+ 16 GB(MCDRAM)
Disk capacity	: 26.2 PB
File Cache system (SSD)	: 960 TB

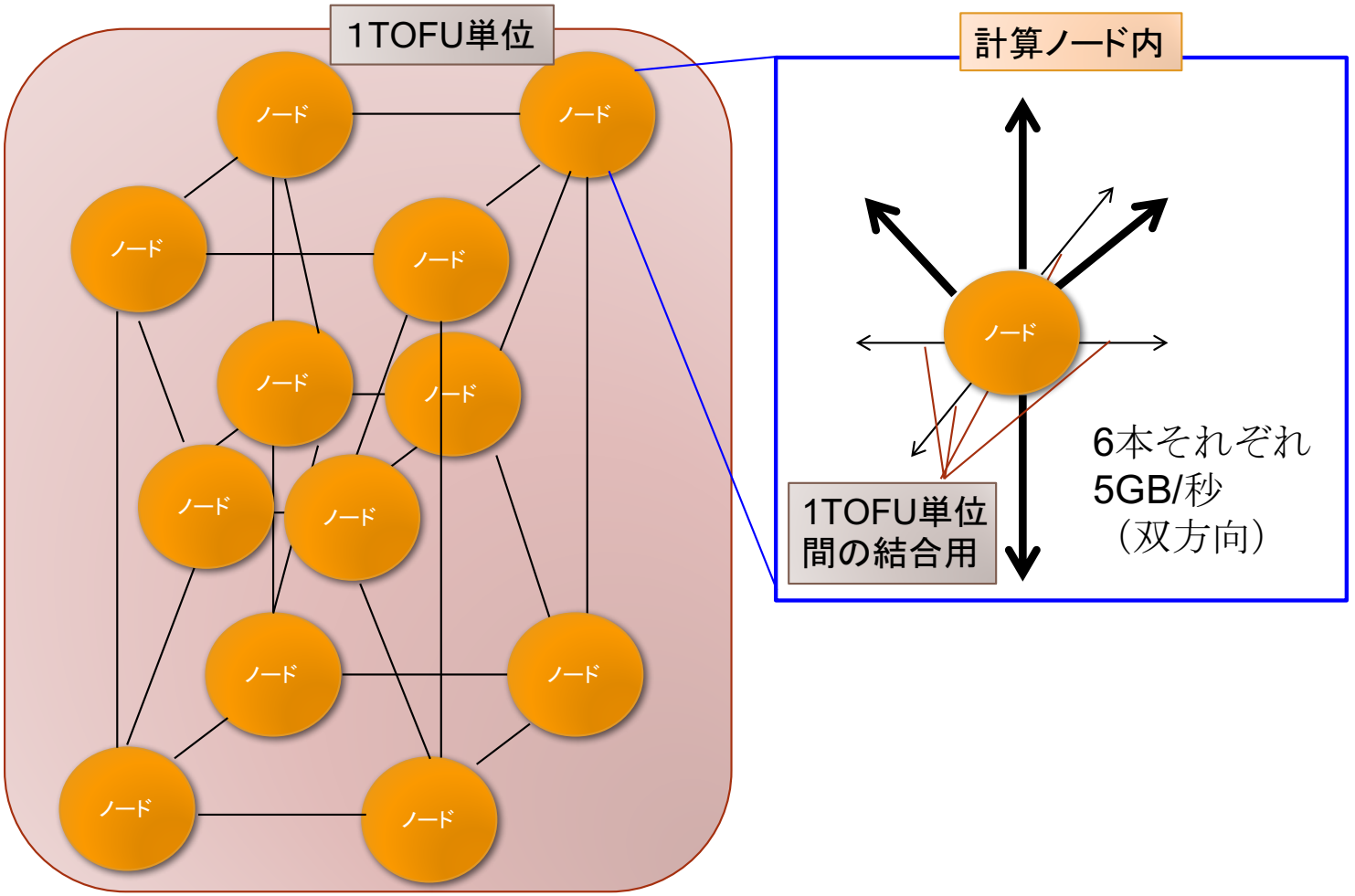
Intel Xeon Phi 7250 1.4 GHz 68 core x1 socket

2016年12月1日試験運転開始

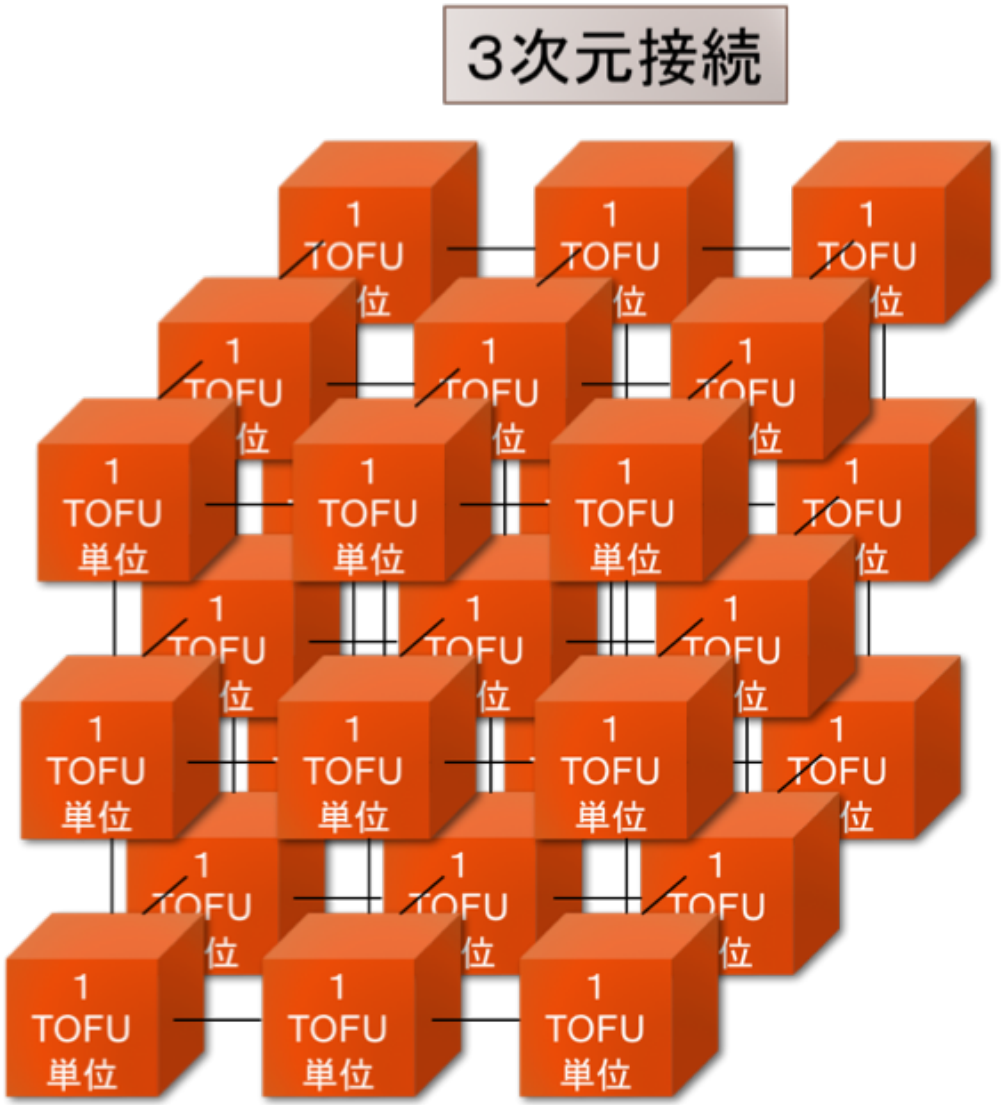
2017年4月3日正式運用開始


FX10計算ノードの構成
1ソケットのみ
TOFU Network
各CPUの内部構成
20GB/秒
ICC
Core #0
Core #1
Core #2
Core #3
…
Core #12
Core #13
Core #14
Core #15
L1
: L1データキャッシュ32KB
L2（16コアで共有、12MB)
85GB/秒
=(8Byte×1333MHz
×8 channel)
Memory
DDR3 DIMM
4GB ×2枚
ノード内合計メモリ量：8GB×4＝32GB

FX10の通信網


1TOFU単位
ノード
ノード
ノード
ノード
ノード
ノード
ノード
ノード
ノード
ノード
ノード
ノード
計算ノード内
ノード
1TOFU単位
間の結合用
6本それぞれ
5GB/秒
（双方向）

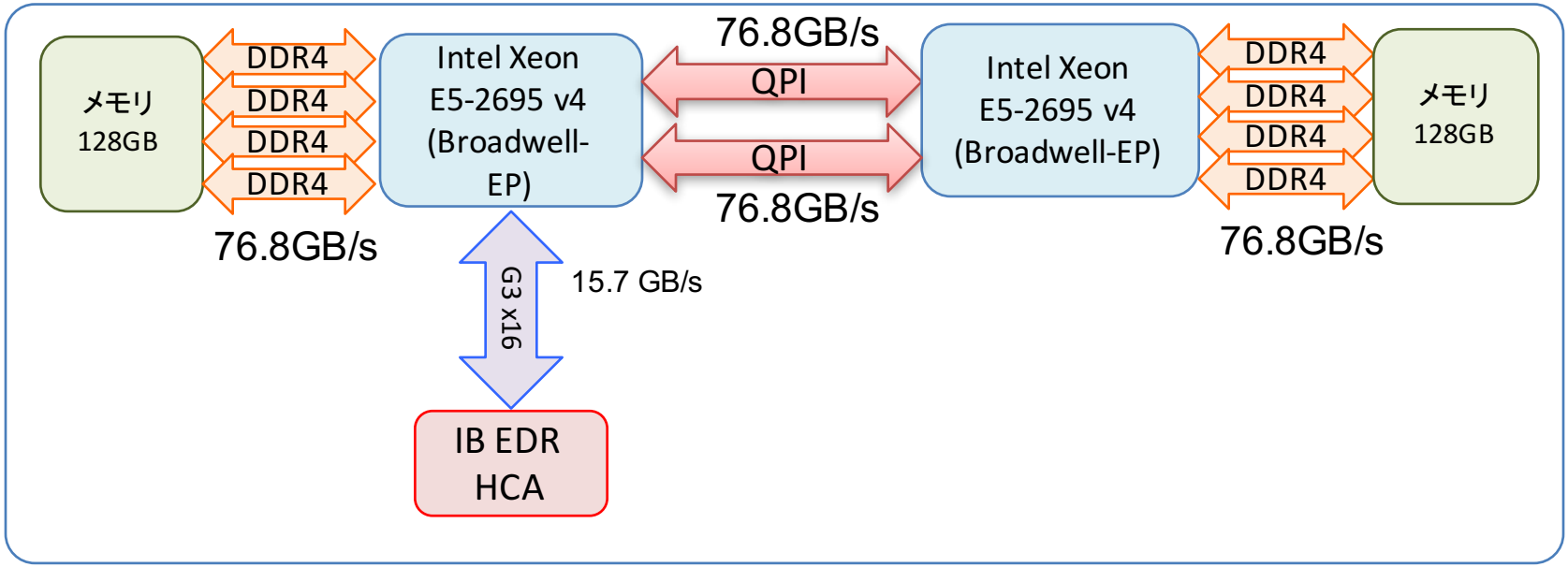
FX10の通信網（１TOFU単位間の結合）



- ユーザから見ると、X軸、Y軸、Z軸について、奥の1TOFUと、手前の１TOFUは、繋がってみえます（３次元トーラス接続）
- ただし物理結線では
  - X軸はトーラス
  - Y軸はメッシュ
  - Z軸はメッシュまたは、トーラス
    になっています

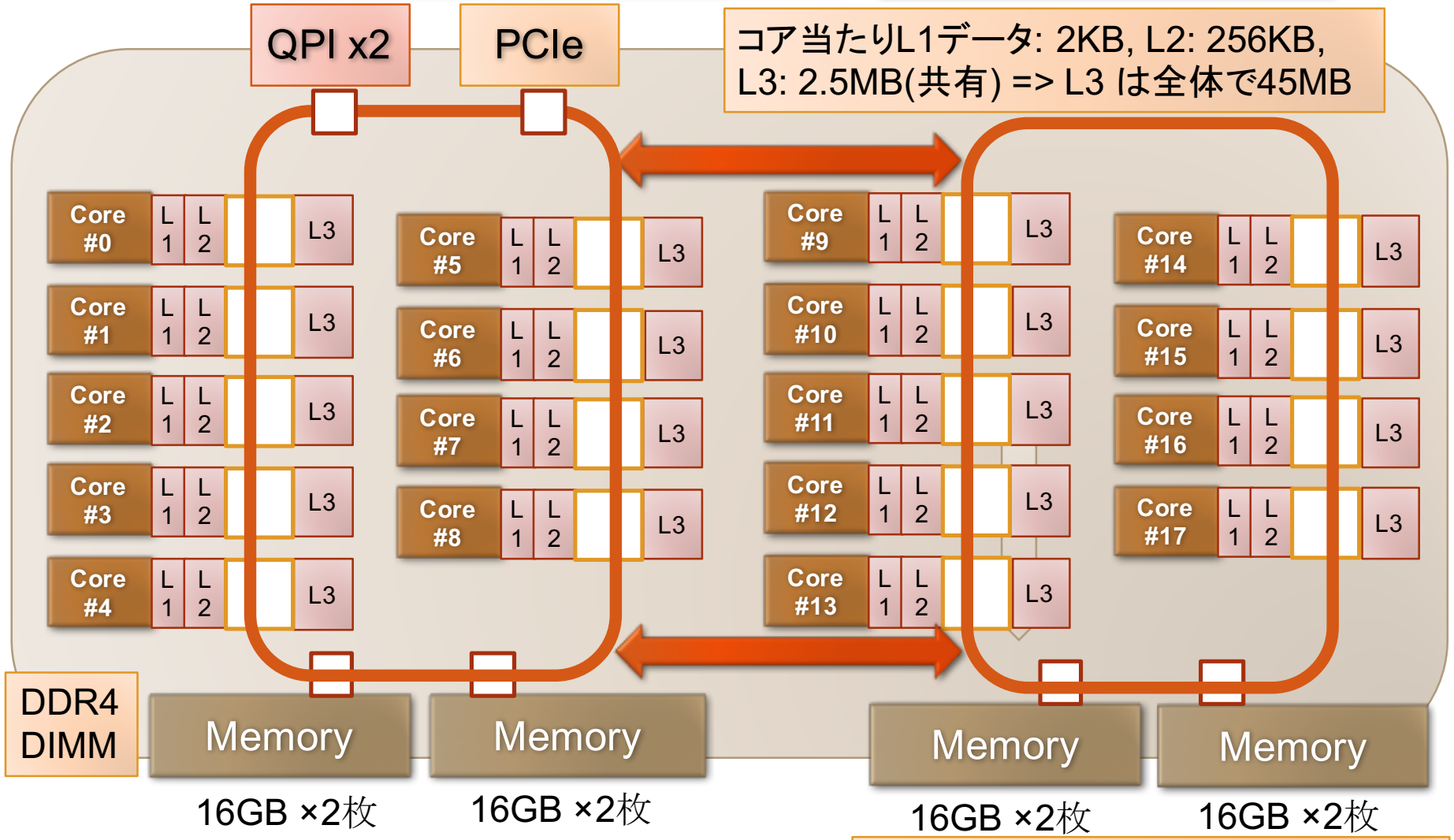
Reedbush-Uノードのブロック図

- メモリのうち、「近い」メモリと「遠い」メモリがある
  => NUMA (Non-Uniform Memory Access)
  (FX10はフラット)



Broadwell-EPの構成

1ソケットのみを図示

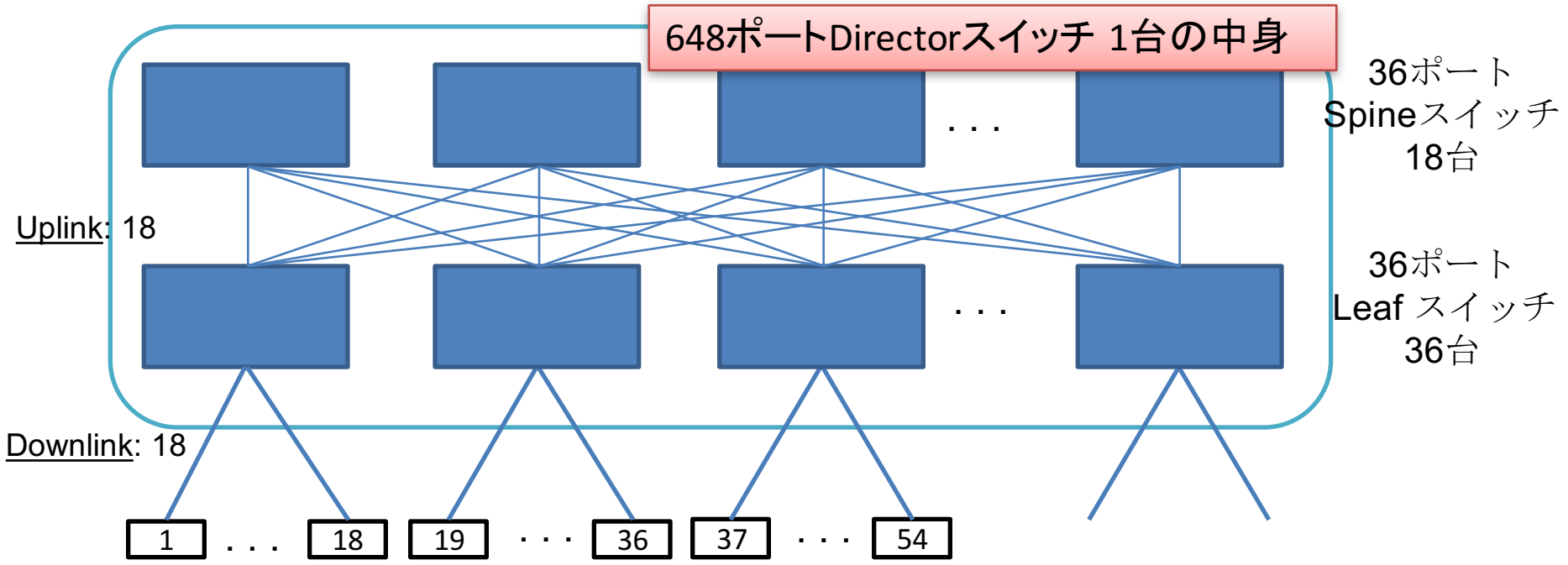


ソケット当たりメモリ量：16GB × 8 = 128GB

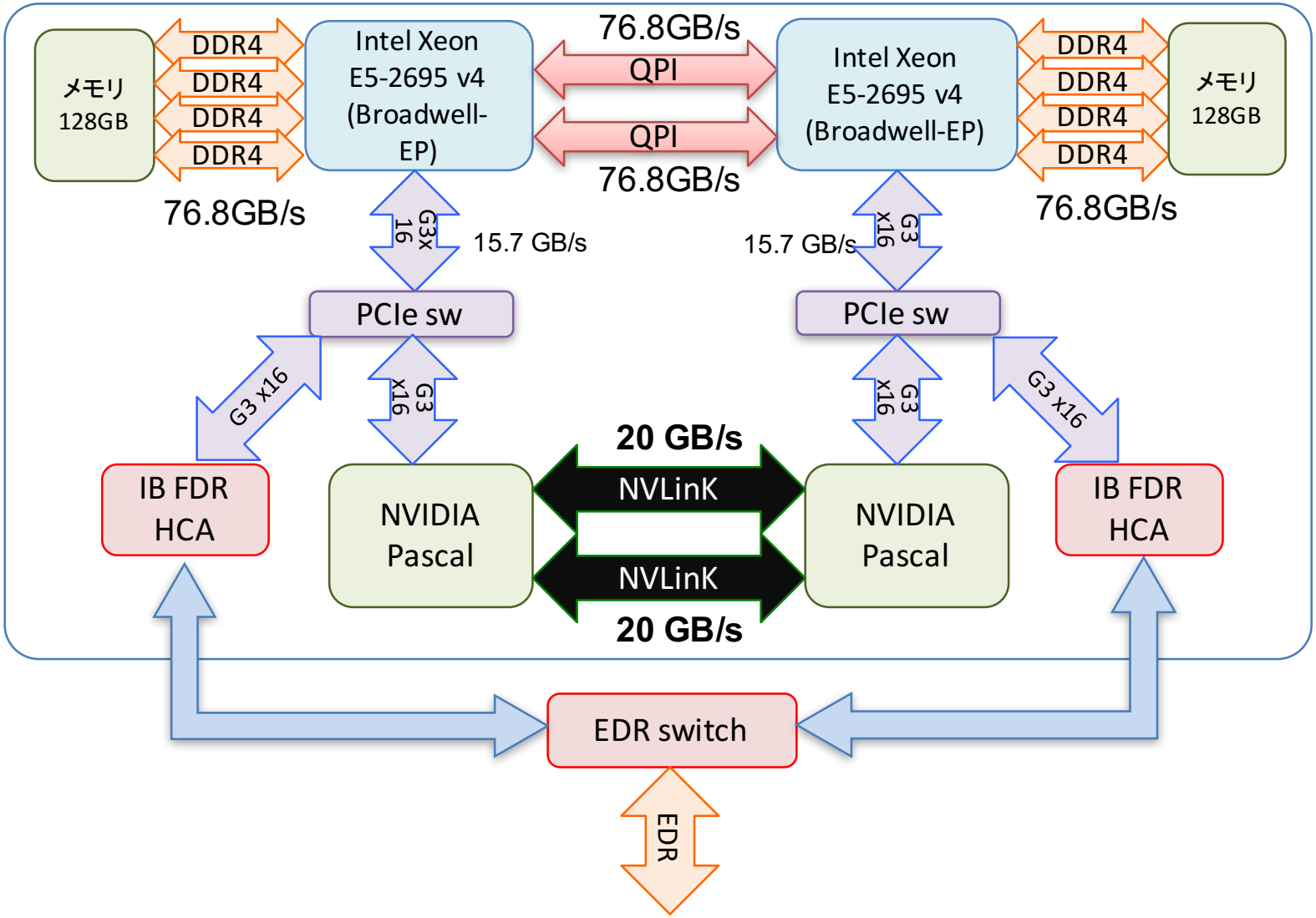
76.8 GB/秒
=(8Byte × 2400MHz × 4 channel)

Reedbush-Uの通信網

- フルバイセクションバンド幅を持つFat Tree網
  - ✓ どのように計算ノードを選んでも互いに無衝突で通信が可能
- Mellanox InfiniBand EDR 4x CS7500: 648ポート
  - ✓ 内部は36ポートスイッチ (SB7800)を (36+18)台組み合わせたものと等価



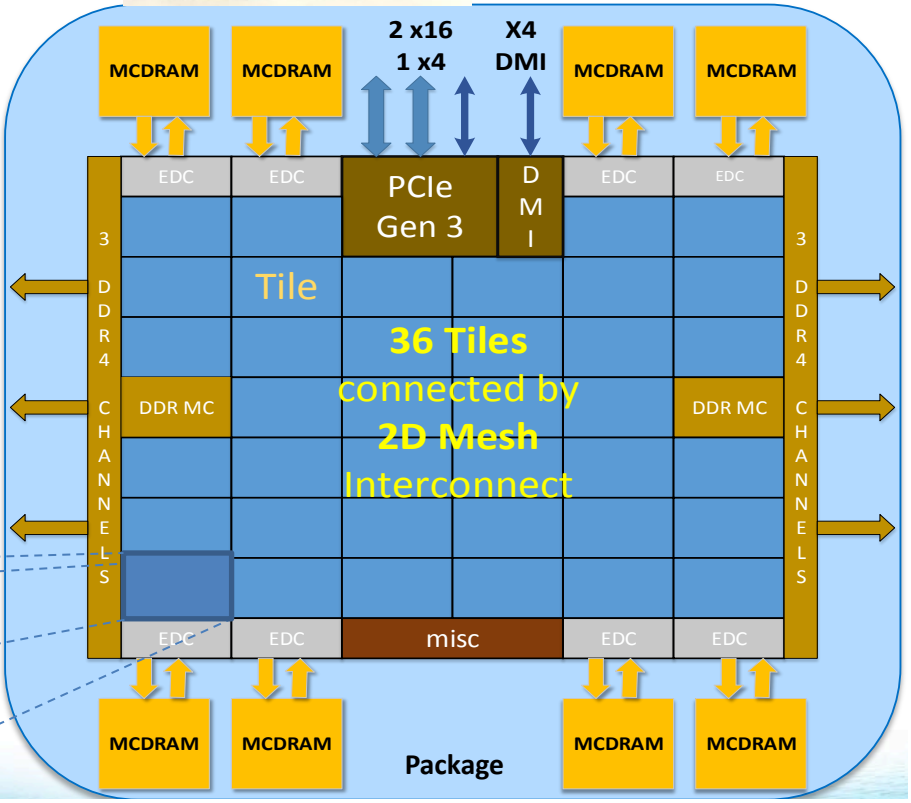
Reedbush-Hノードのブロック図



Oakforest-PACS 計算ノード

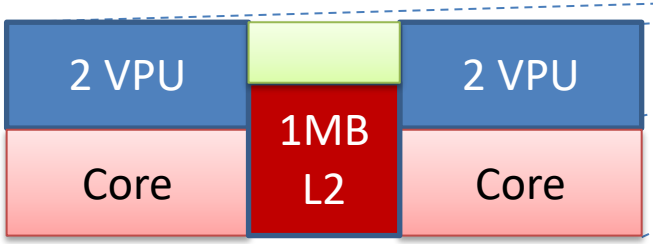
- Intel Xeon Phi (Knights Landin
  - 1ノード1ソケット
- MCDRAM: オンパッケー
  の高バンド幅メモリ16G
  + DDR4メモリ

HotChips27
KNLスライドより



ソケット当たりメモリ量：16GB×6＝96GB

MCDRAM: 490GB/秒以上（実測）
DDR4: 115.2 GB/秒
=(8Byte×2400MHz×6 channel)



Oakforest-PACS: 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)

2

2

Uplink: 24

48 port Edge Switch
362 台

Downlink: 24

1 . . . 24 25 . . . 48 49 . . . 72

計算ノード
ラックに分散

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

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（2017年4月1日）

- パーソナルコース（年間）
  - コース１：　100,000円　：　8ノード(基準)、最大16ノードまで
  - コース２：　200,000円　：　16ノード(基準)、最大64ノードまで
- グループコース
  - 400,000円 (企業 480,000円)　：1口 8ノード（基準）、最大128ノードまで
- 以上は、「トークン制」で運営
  - 申し込みノード数×360日×24時間の「トークン」が与えられる
  - 基準ノードまでは、トークン消費係数が1.0
  - 基準ノードを超えると、超えた分は、消費係数が2.0になる
  - 大学等のユーザはFX10、Reedbushとの相互トークン移行も可能

東大情報基盤センターReedbushスーパーコンピュータシステムの料金表（2017年4月1日）

- パーソナルコース（年間）
  - 150,000円　：　RB-U: 4ノード（基準）、最大16ノードまで
    RB-H: 1ノード（基準）、最大2ノードまで
- グループコース
  - 300,000円：1口 4ノード（基準）、最大128ノードまで、
    RB-H：1ノード（基準）、最大32ノードまで（トークン係数はUの2.5倍）
  - RB-Uのみ 企業 360,000円　：1口 4ノード（基準）、最大128ノードまで
  - RB-Hのみ 企業 216,000円　：1口 1ノード（基準）、最大32ノードまで
- 以上は、「トークン制」で運営
  - 申し込みノード数×360日×24時間の「トークン」が与えられる
  - 基準ノードまでは、トークン消費係数が1.0
  - 基準ノードを超えると、超えた分は、消費係数が2.0になる
  - 大学等のユーザはFX10, Oakforest-PACSとの相互トークン移行も可能
  - ノード固定もあり

東大情報基盤センターFX10スーパーコンピュータシステムの料金表（2017年4月1日）

- パーソナルコース（年間）
  - コース１：　90,000円　：　12ノード(基準)、最大24ノードまで
  - コース２：　180,000円　：　24ノード(基準)、最大96ノードまで
- グループコース
  - 360,000円 (企業 432,000円)　：1口、12ノード、最大1440ノードまで
- 以上は、「トークン制」で運営
  - 申し込みノード数×360日×24時間の「トークン」が与えられる
  - 基準ノードまでは、トークン消費係数が1.0
  - 基準ノードを超えると、超えた分は、消費係数が2.0になる
  - 大学等のユーザはReedbush, Oakforest-PACSとの相互トークン移行も可能

JPY (=Watt)/GFLOPS Rate

Smaller is better (efficient)

System	JPY/GFLOPS
Oakleaf/Oakbridge-FX (Fujitsu) (Fujitsu PRIMEHPC FX10)	125
Reedbush-U (SGI) (Intel BDW)	62.0
Reedbush-H (SGI) (Intel BDW+NVIDIA P100)	17.1
Oakforest-PACS (Fujitsu) (Intel Xeon Phi/Knights Landing)	16.5

トライアルユース制度について

- 安価に当センターのOakleaf/Oakbridge-FX, Reedbush-U/H, Oakforest-PACS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）（FX10: 最大24ノード、最大96ノード,
      RB-U: 最大16ノード, RB-H: 最大2ノード、OFP: 最大16ノード, 最大64ノード）
      講習会いずれかの受講が必須、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（FX10:最大1,440ノード、RB-U: 最大128ノード、RB-H: 最大32ノード、OFP: 最大2048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

スーパーコンピュータシステムの詳細

- 以下のページをご参照ください
  - 利用申請方法
  - 運営体系
  - 料金体系
  - 利用の手引

  などがご覧になれます。

  http://www.cc.u-tokyo.ac.jp/system/ofp/
  http://www.cc.u-tokyo.ac.jp/system/reedbush/
  http://www.cc.u-tokyo.ac.jp/system/fx10/