

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 Reedbushに おけるストレージシステム

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

発表の概要

- 背景
- Reedbushシステムの紹介
 - ストレージシステム
- ベンチマークによる評価
 - STREAM, HPL, HPCG
 - Intel MPI Benchmark
 - IOR
- まとめ

東大情報基盤センターのスパコン

2基の大型システム, 6年サイクル

FY

08

09

10

11

12

13

14

15

16

17

18

22

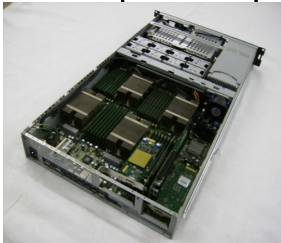
Hitachi SR11K/J2
IBM Power-5+
18.8TFLOPS, 16.4TB

Yayoi: Hitachi SR16000/M1
IBM Power-7
54.9 TFLOPS, 11.2 TB

Hitachi HA8000 (T2K)
AMD Opteron
140TFLOPS, 31.3TB

Oakleaf-FX: Fujitsu PRIMEHPC
FX10, SPARC64 IXfx
1.13 PFLOPS, 150 TB

Oakbridge-FX
136.2 TFLOPS, 18.4 TB



Peta

K

K computer

Post-K

2016/10/18

日本 Lustre ユーザ 会 2016

3

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システム：導入の背景

- ポストT2K

- 筑波大と共同で単一システムを導入
最先端共同HPC基盤施設 (Joint Center for Advanced HPC: JCAHPC)
- 2016年12月運用開始予定

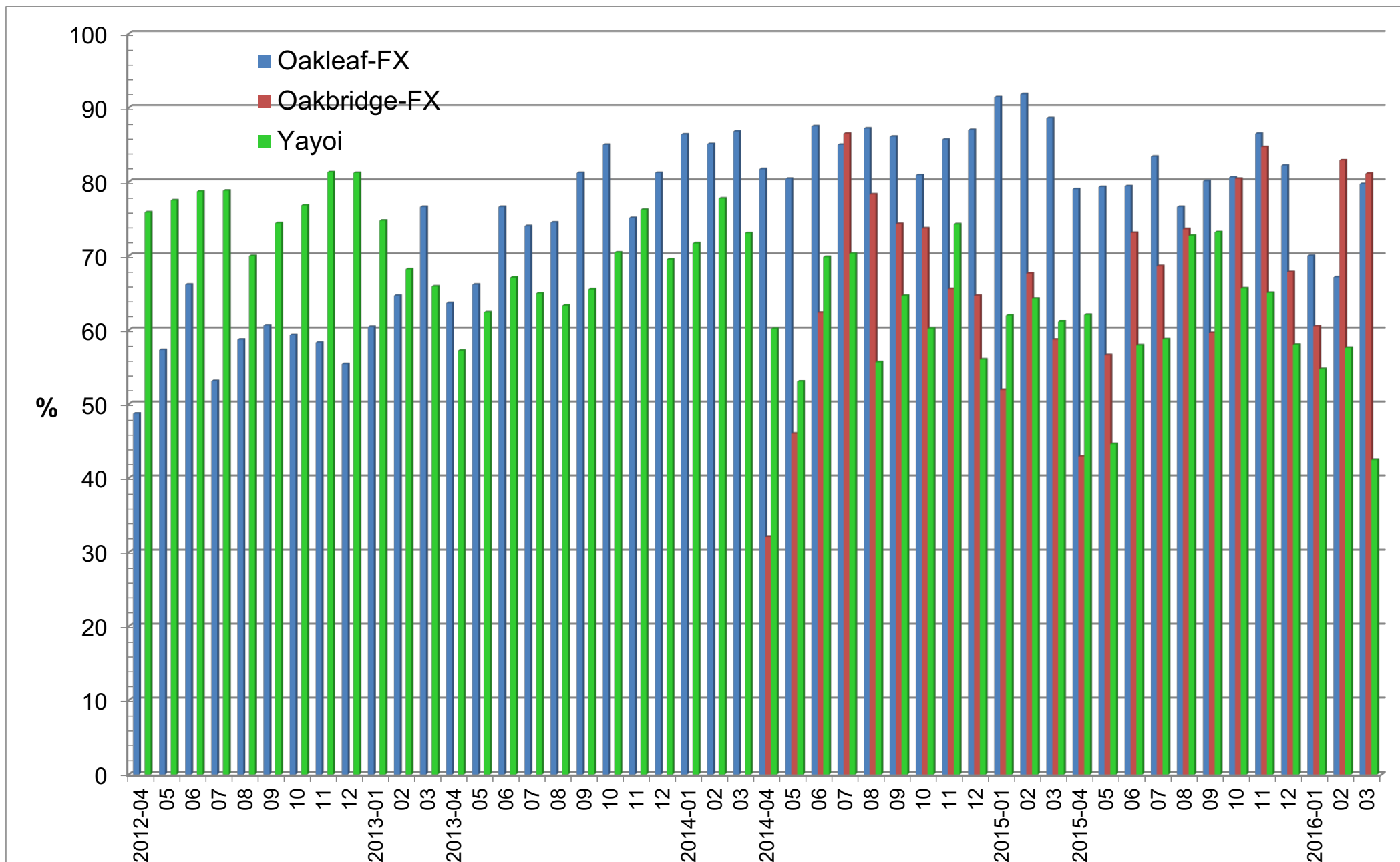


- 東大のみで導入

- Oakleaf/Oakbridge-FX (FX10) 混雑緩和
 - 2015年度平均が80%超
- ポストT2K移行のための中間的なシステム
- Post FX10 (2018年秋稼働開始予定) のパイロットシステム
 - ✓ データ解析, Deep Learningなど、新規利用分野開拓

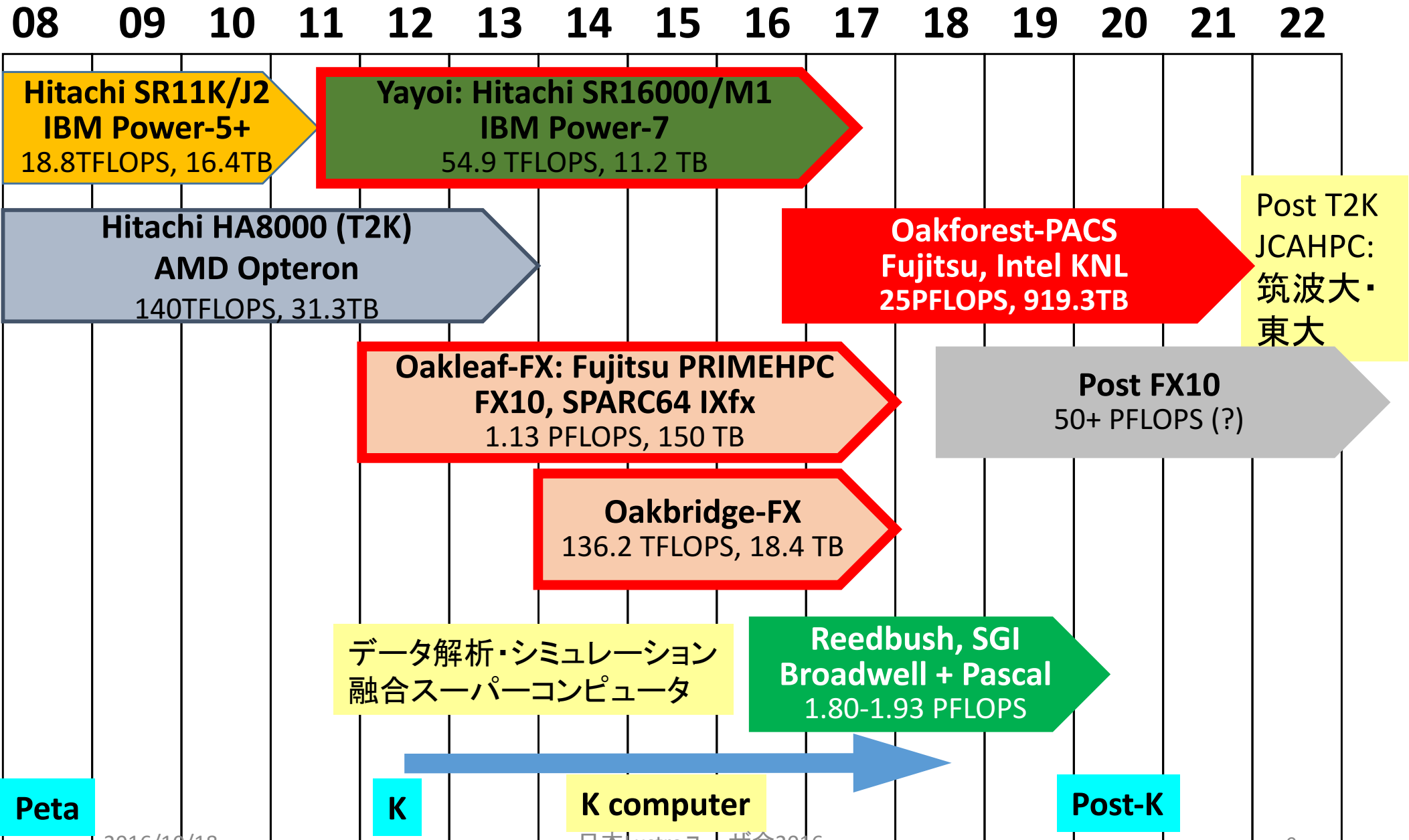
利用率

80+% Average
Oakleaf-FX + Oakbridge-FX



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
FY



ポストT2K: Oakforest-PACS

<http://www.cc.u-tokyo.ac.jp/system/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
- **最先端共同HPC 基盤施設(JCAHPC: Joint Center for Advanced High Performance Computing)**
 - 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>

GPUの導入

これまで:

- CUDAによるプログラミング: 複雑
- 2,000人を超えるユーザー
 - CUDAを勉強してもらうのは大変



- OpenACC

- OpenMPと類似したインタフェース: 使いやすいが性能悪かった
- 昨今の性能向上, CUDAとそれほど大きな差がなくなった

- データ科学, 深層学習 (Deep Learning)

- 従来の計算科学, 計算工学分野とは異なった分野の新規ユーザー開拓が急務
 - 東京大学ゲノム医科学研究機構
 - 東京大学病院
 - 医療画像処理への深層学習適用

Reedbushシステム (1/2)

- システム構成・運用：日本SGI => HPE?
- Reedbush-Uサブシステム：CPUのみ
 - Intel Xeon (Broadwell-EP) x 2ソケット
 - 256GB DDR4-2400 メモリ
 - 420 ノード, 508.0 TF
 - 2016年7月稼働開始
- Reedbush-Hサブシステム：GPU搭載
 - CPU・メモリ：Reedbush-U と同様
 - NVIDIA Tesla P100 (Pascal世代 GPU)
 - 120 ノード, 145.2 TF(CPU)+ 1.15~1.27 PF(GPU)= 1.30~1.42 PF
 - 2017年3月稼働開始

Reedbushシステム (2/2)

- ストレージ/ファイルシステム
 - 並列ファイルシステム (Lustre)
 - 5.04 PB, 145.2 GB/sec
 - 高速ファイルキャッシュシステム: Burst Buffer (DDN IME (Infinite Memory Engine))
 - SSD: 209.5 TB, 385.2 GB/sec
- インターコネクト: 100G bps / ノード
 - Reedbush-U、ストレージ: InfiniBand EDR (100G)
 - Reedbush-H: InfiniBand FDR (56G) x2
- 電力, 冷却, 設置面積
 - 空冷, 368.4 kW (冷却除く)
 - < 90 m²

Reedbushシステム外観



“Reedbush”とは



Blaise Pascal
(1623-1662)

- L'homme est un roseau pensant.
- Man is a thinking reed.
- 人間は考える葦である

Pensées (Blaise Pascal)



計算ノード: **1.795-1.926 PFlops**

Reedbush-U (CPU only) 508.03 TFlops

CPU: Intel Xeon E5-2695 v4 x 2 socket
(Broadwell-EP 2.1 GHz 18 core,
45 MB L3-cache)
Mem: 256GB (DDR4-2400, 153.6 GB/sec)

× 420

SGI Rackable
C2112-4GP3

InfiniBand EDR 4x
100 Gbps /node

Reedbush-H (w/Accelerators)

1287.4-1418.2 TFlops

CPU: Intel Xeon E5-2695 v4 x 2 socket
Mem: 256 GB (DDR4-2400, 153.6 GB/sec)
GPU: NVIDIA Tesla P100 x 2
(Pascal, SXM2, 4.8-5.3 TF,
Mem: 16 GB, 720 GB/sec, PCIe Gen3 x16,
NVLink (for GPU) 20 GB/sec x 2 brick)

× 120

SGI Rackable C1102-PL1

Dual-port InfiniBand FDR 4x
56 Gbps x2 /node

InfiniBand EDR 4x, Full-bisection Fat-tree

145.2 GB/s

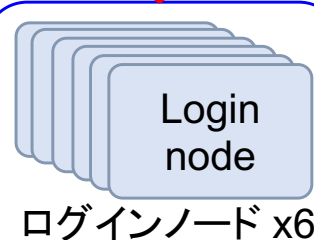
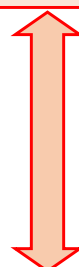
並列ファイル
システム
5.04 PB

Lustre Filesystem
DDN SFA14KE x3

385.2 GB/s

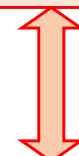
高速ファイル
キャッシュシステム
209 TB

DDN IME14K x6



Login
node

ログインノード x6



管理サーバ
群

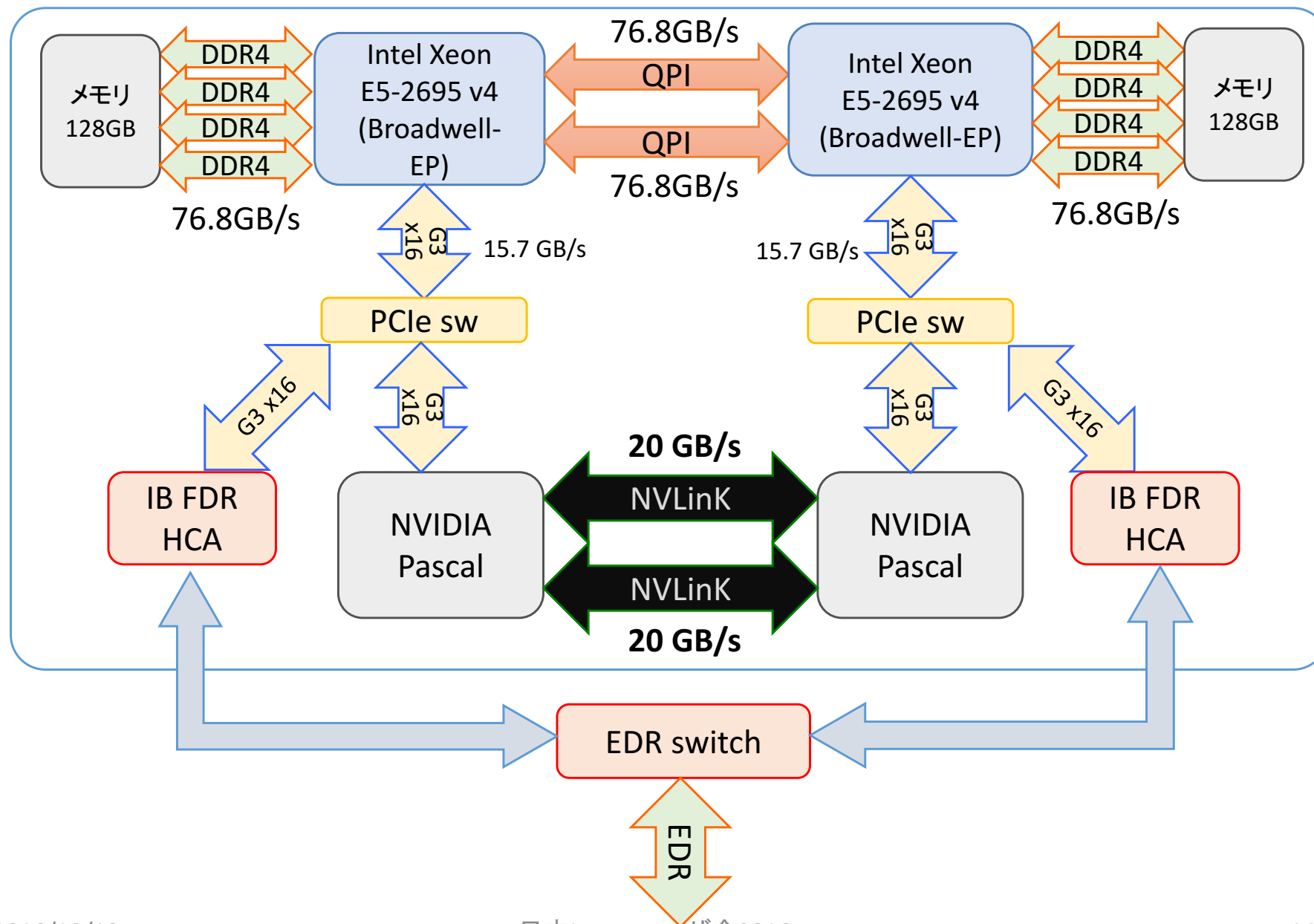
Mellanox CS7500
634 port +
SB7800/7890 36
port x 14



UTnet

ユーザ

Reedbush-Hノードのブロック図



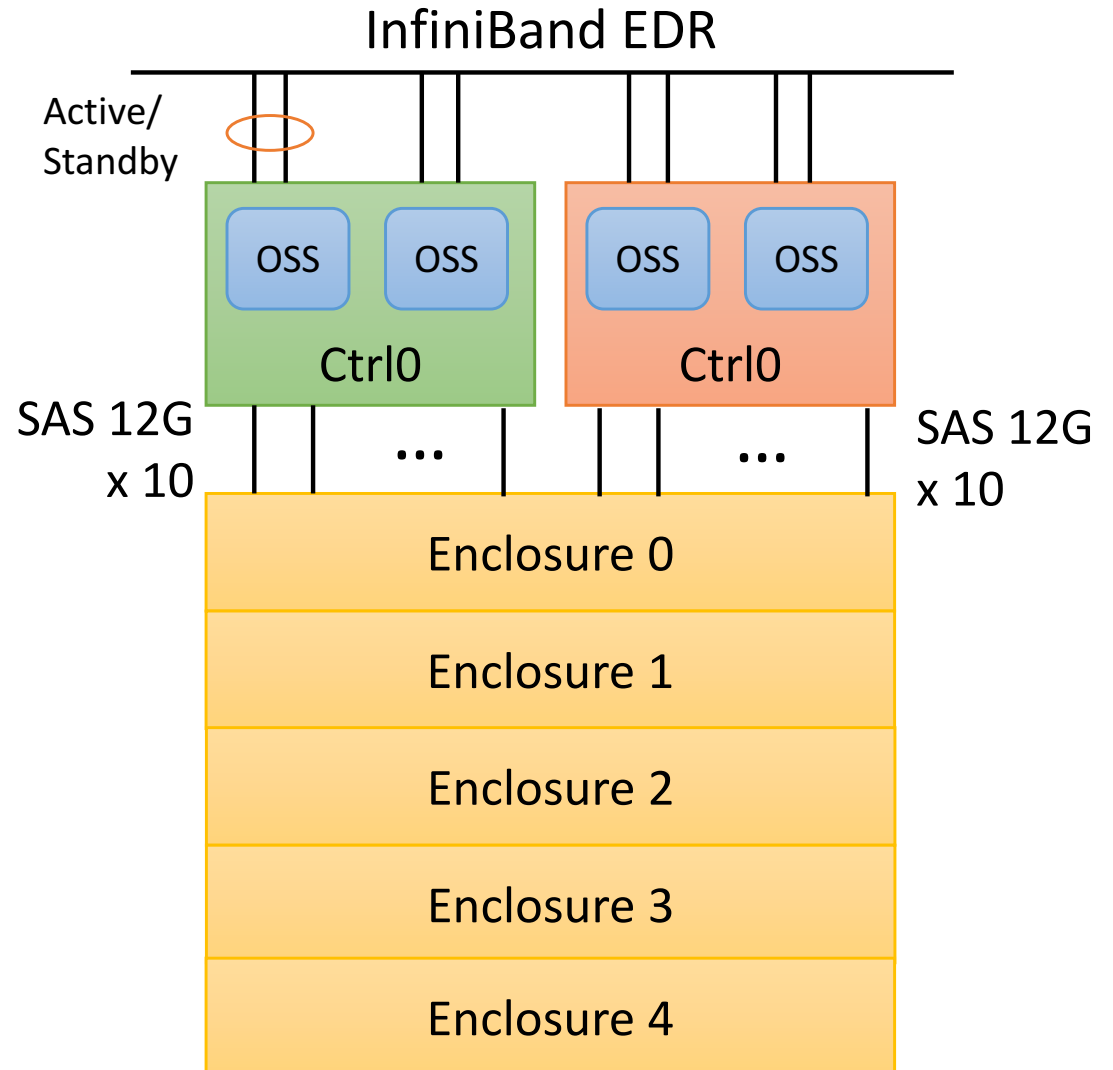
全体構成

項目			Reedbush-U	Reedbush-H
システム全体 (計算ノード)	総理論演算性能		508.03 TFlops	1297.15～1417.15 TFlops
	総ノード数		420	120
	総主記憶容量		105 TByte	30 TByte
	ネットワークトポロジー		Full-bisection BW Fat-Tree	
	並列ファイルシステム	システム名	Lustreファイルシステム	
		サーバ(OSS)	DDN SFA14KE	
		サーバ(OSS)数	3	
		ストレージ容量	5.04 PB	
		帯域幅	145.2 GB/sec	
	高速ファイルキャッシュシステム	サーバ	DDN IME14K	
		サーバ数	6	
		容量	209 TByte	
		帯域幅	385.2 GB/sec	

項目		Reedbush-U	Reedbush-H
マシン名		SGI Rackable C2112-4GP3	SGI Rackable C1100シリーズ (開発中)
CPU	プロセッサ名	Intel Xeon E5-2695v4 (Broadwell-EP)	
	プロセッサ数(コア数)	2 (36)	
	周波数	2.1 GHz (Turbo boost 時最大 3.3 GHz)	
	理論演算性能	1209.6 GFlops	
Memory	容量	256 GB	
	メモリ帯域幅	153.6 GB/sec	
GPU	プロセッサ名	None	NVIDIA Tesla P100 (Pascal)
	コア数(単体)		56 (SM)
	メモリ容量(単体)		16 GB
	メモリ帯域幅(単体)		720 GB/sec
	理論演算性能(単体)		4.8～5.3 TFlops
	搭載数		2
	CPU-GPU間接続		PCI Express Gen3 x16レーン (16 GB/sec)
	GPU間接続		NVLink 2 brick (20 GB/sec x2)
インターコネクト		InfiniBand EDR 4x (100 Gbps)	InfiniBand FDR 4x 2リンク (56 Gbps x2)

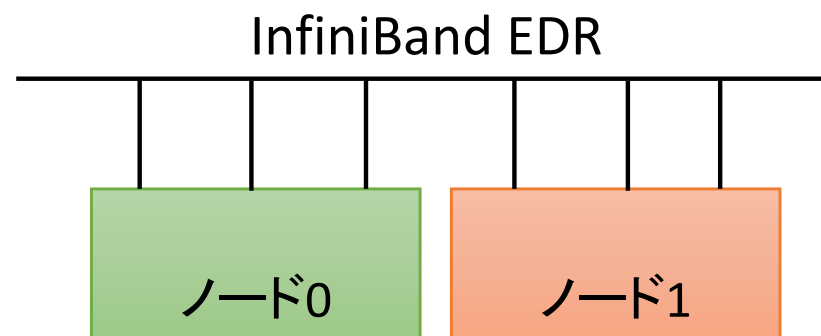
並列ファイルシステム

- DDN SFA14KE 3セット
 - OSSはコントローラ中の**仮想マシン**として動作
- 1セット 2コントローラ
 - HDD: 360 x 6TB NLSAS (35 RAID6 (8D2P) +10 スペア)
- コントローラ当たり
IB-EDR x 4ポート
(Active-Standby x 2 組)
- 容量: **5.04** PB
 - 6TB * 8本 * 35RAIDグループ * 3セット
- バンド幅: **145.2** GB/sec
 - (EDR:100G *2ポート)
* (64b66bエンコーディング)
* 2 コントローラ
* 3セット



ファイルキャッシュシステム

- DDN IME14K 6セット (12ノード)
- 1セット 2ノード
 - SSD: 48 x 800GB NVMe
- ノード当たり
IB-EDR x 3ポート
(うち1ポートはGen3 x8ポートの
ため 64Gbpsで律速)
- 容量: 約**209** TB
 - 1セット当たり 38.4 TB, erasure
coding 10D1Pで換算
- バンド幅: **385.2** GB/sec
 - ((EDR:100G * 2ポート)
* (64b66bエンコーディング)
+ (PCIe Gen3 x8: 64G * 1ポート)
* (128b130bエンコーディング))
* 2ノード
* 6セット



ストレージラックの構成(3ラック)



- 2 x IME, コントローラ (SFA14KE)+ 5 x エンクロージャ

SFA14KE背面



IME14K 背面



IME
1台分
(2ノード)

評価に用いたソフトウェア環境

- コンパイラ
 - Intel C/Fortran Compiler 16.0.3
- 数値ライブラリ
 - Intel Math Kernel Library 11.3.3
- MPI
 - Intel MPI 5.1.3
 - SGI MPT 2.14
 - Mellanox HPC-X 1.6.392
 - Open MPI 1.10.2
 - MVAPICH2 2.2rc1

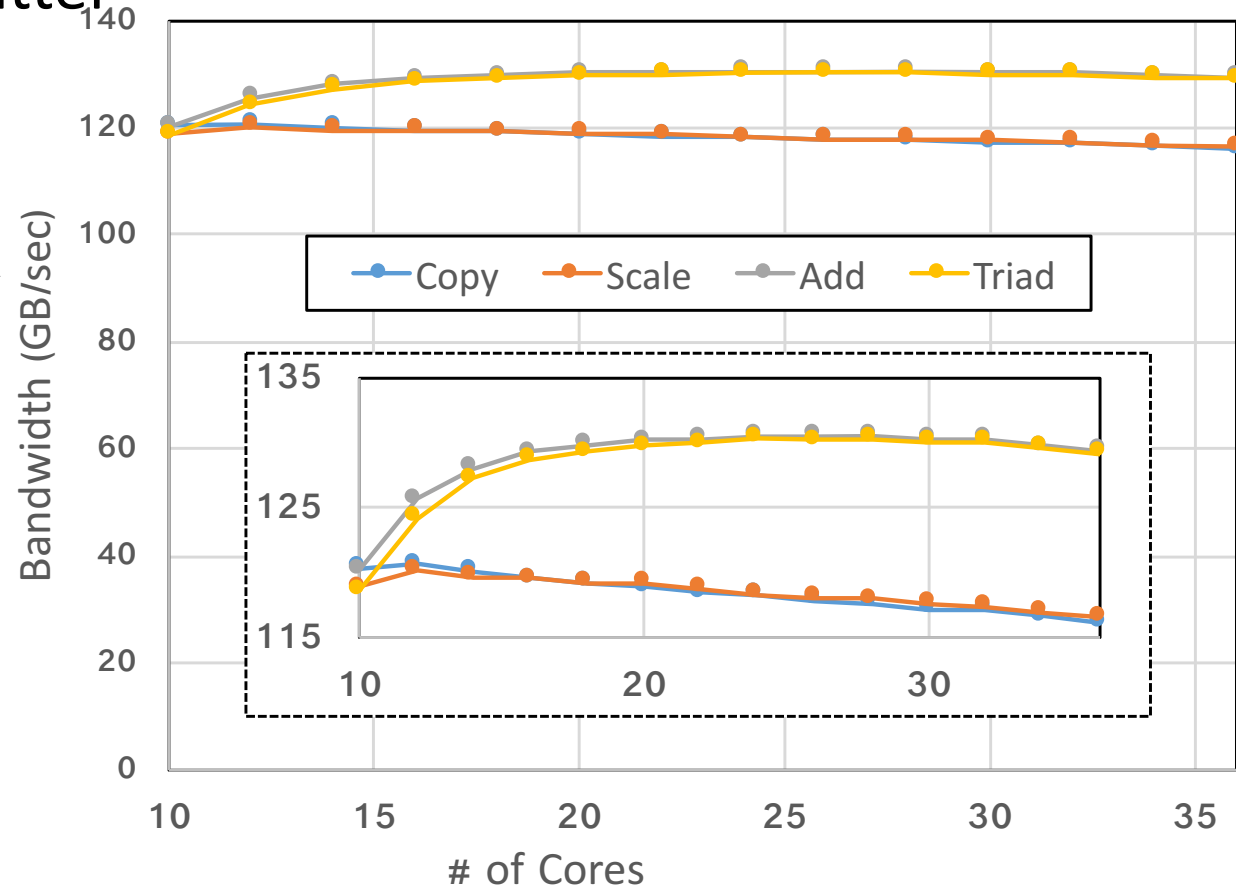
ベンチマーク評価

今回は Reedbush-U のみ

- STREAM
- HPL : ノード単体、全系
- HPCG: ノード単体、全系
- Intel MPI Benchmark
- (mdtest)
- IOR

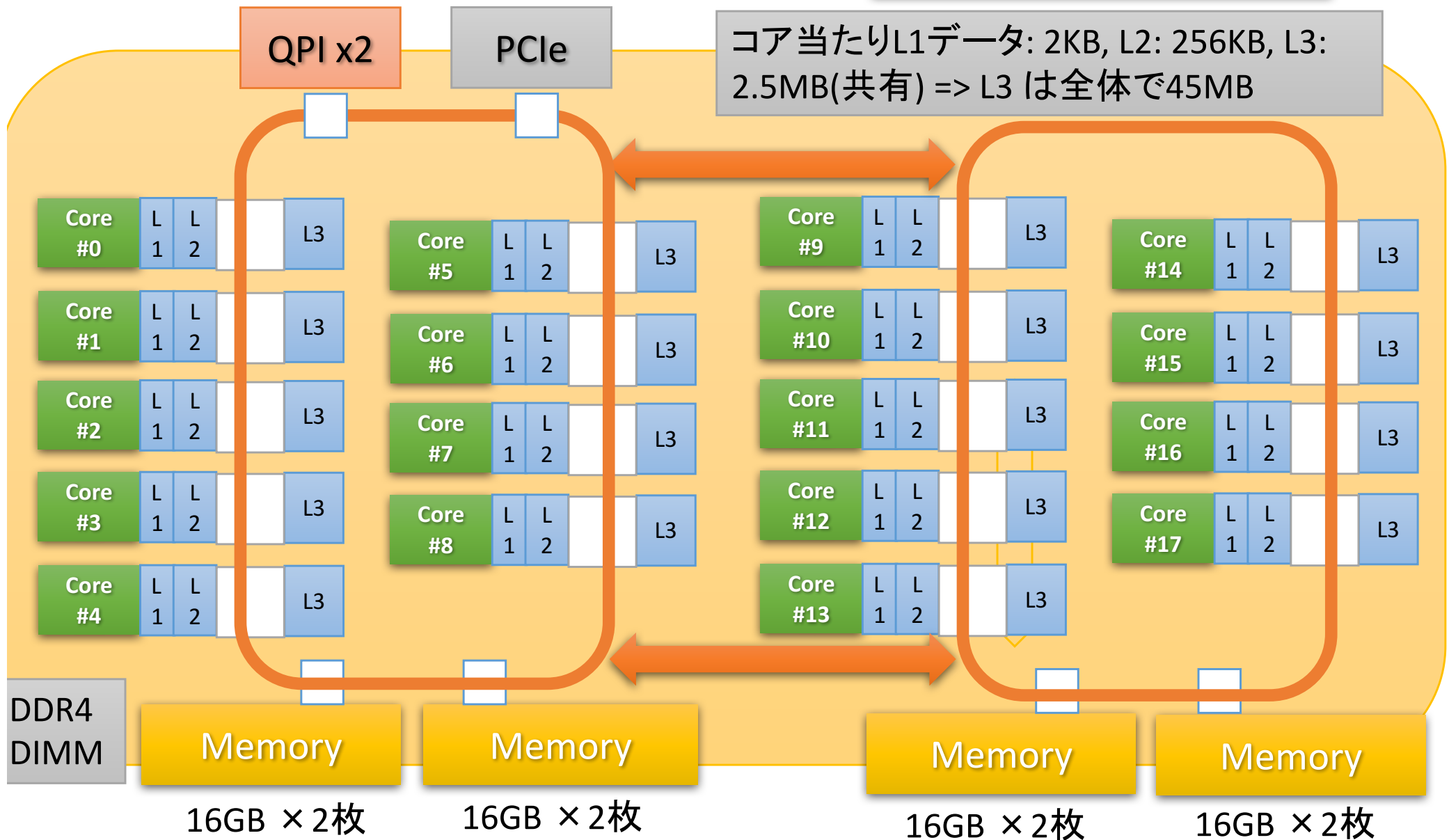
STREAM

- コンパイルオプション: -xCORE-AVX2 -O3
-qopt-streaming-stores always
-DSTREAM_ARRAY_SIZE=400000000
- KMP_AFFINITY=scatter
- FX10 1ノード
(65GB/s)
= BDW 1ソケット
- 130GB/s: 理論
ピーク 85%



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)

HPL

- Intel MKL 11.3.3 に含まれているバイナリ
xhpl_intel64_dynamic を使用
- ノード単体
 - 2MPI x 18スレッド
 - N=98112, NB=192
 - P=1,Q=2
- 全系 420ノード
 - (1MPI x 36スレッド) x 420
 - N=1538688,NB=192
 - P=20,Q=21
- 結果: 1149.6 GF
(ピーク比 95.0%)
- 結果: 446.3 TF
(ピーク比 87.8%)

HPCG 3.0

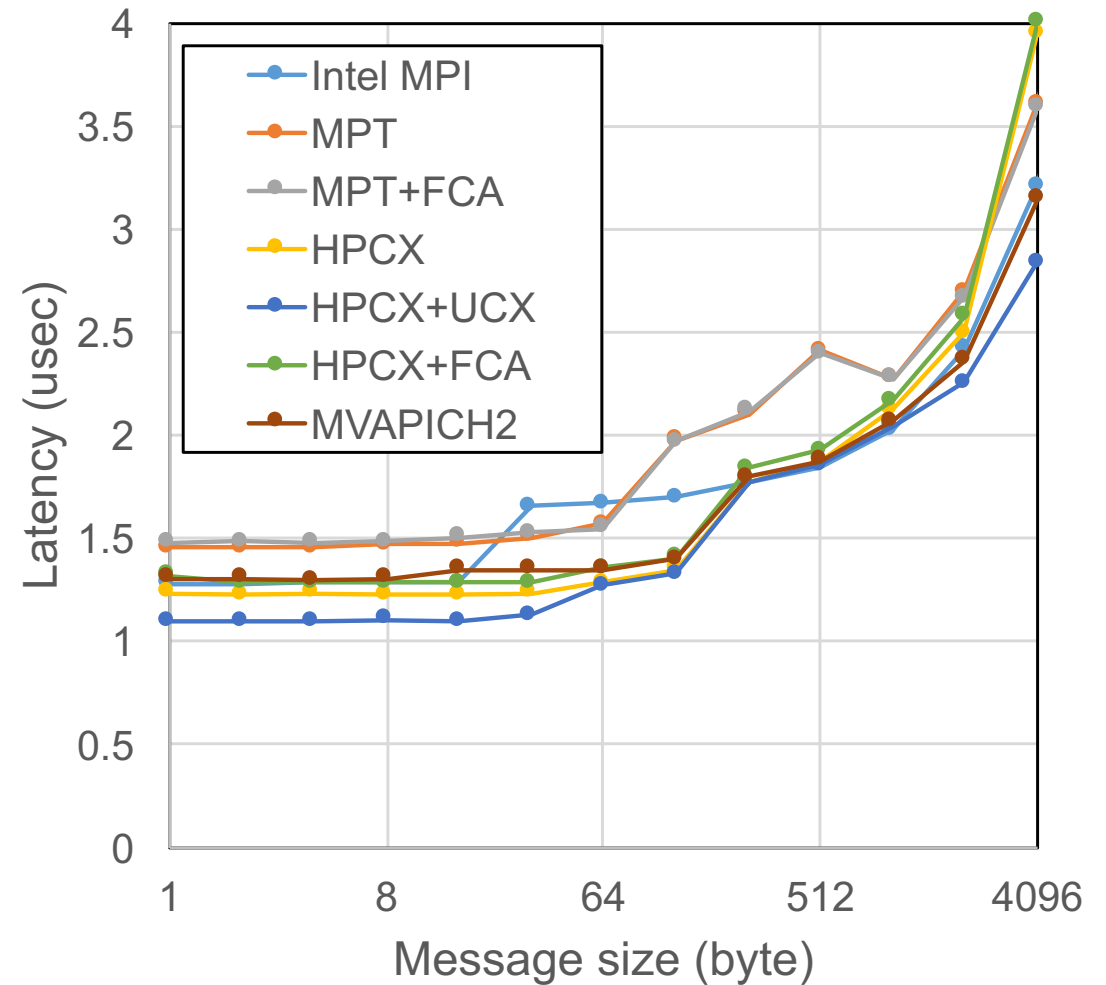
- Intel MKL 11.3.3 に含まれているバイナリ xhpcg_avx2 を使用
- ノード単体
 - 2MPI x 18スレッド
 - nx=ny=nz=144
- 結果: 21.9 GF
(ピーク比 3.6 %)
- 全系 420ノード
 - (2MPI x 18スレッド) x 420
 - nx=ny=nz=144
- Intel MPIでコンパイル
=> SGI MPTを使うためのラッパを利用
- 結果: 8457.78 TF
(ピーク比 1.66%)

Intel MPI Benchmark

- MPI実装を比較
 - Intel MPI 5.1.3
 - SGI MPT 2.14
 - Mellanox HPC-X 1.6.392 ... Open MPI 1.10.3rc4
 - (Open MPI 1.10.2)
 - MVAPICH2 2.2rc1
- Mellanox CS7500 (Switch-IB2搭載)の持つCollective Offloading機能
 - FCA (Fabric Collective Accelerator) 3.5 ... SGI MPT, HPC-X
 - SHArP (Scalable Hierarchical Aggregation Protocol) ... HPC-X
- 軽量プロトコル OpenUCX (United Communication X) ... HPCX

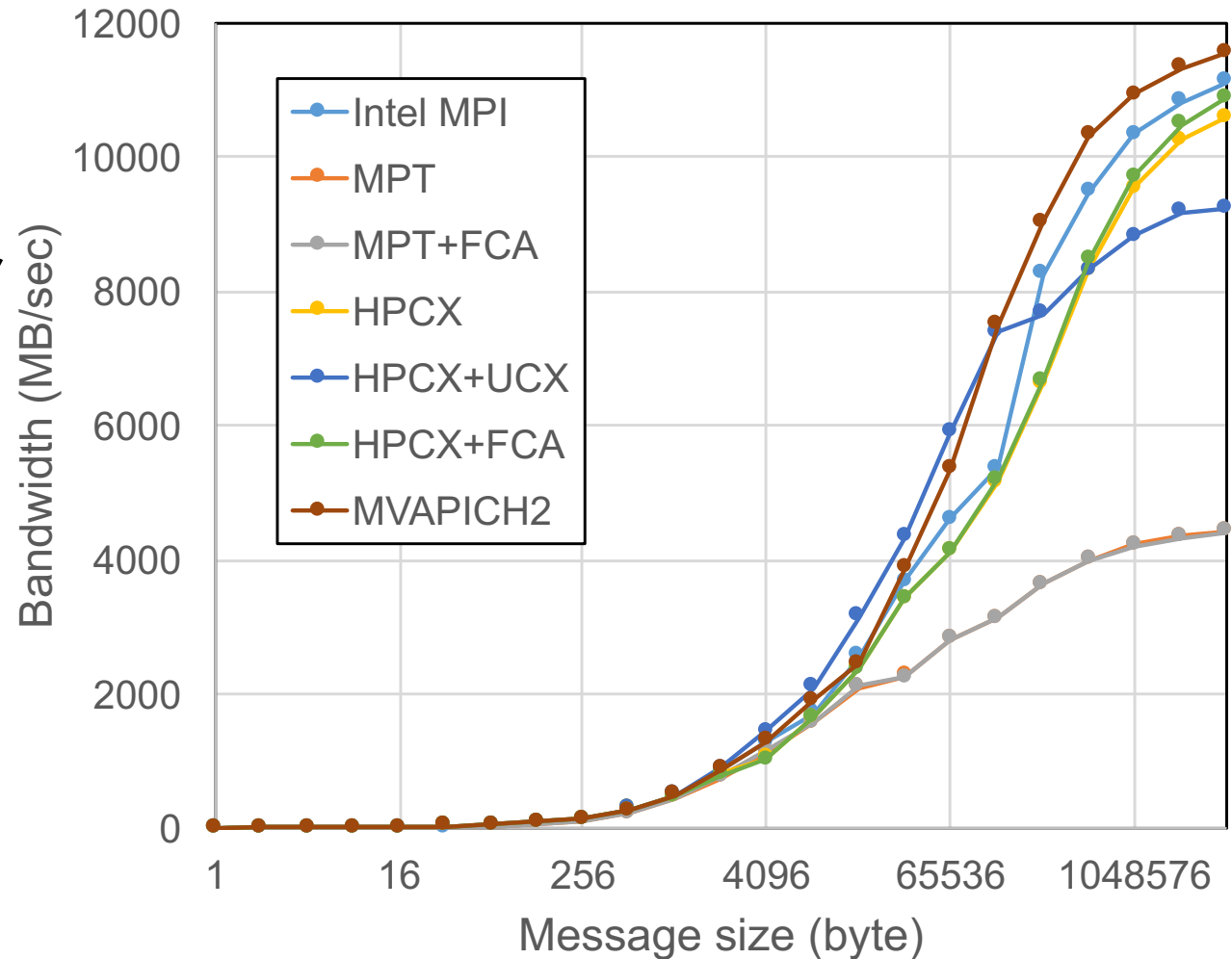
Pingpong (レイテンシ)

- 最小: HPCXにおいてUCXを使った場合: 1.1 us



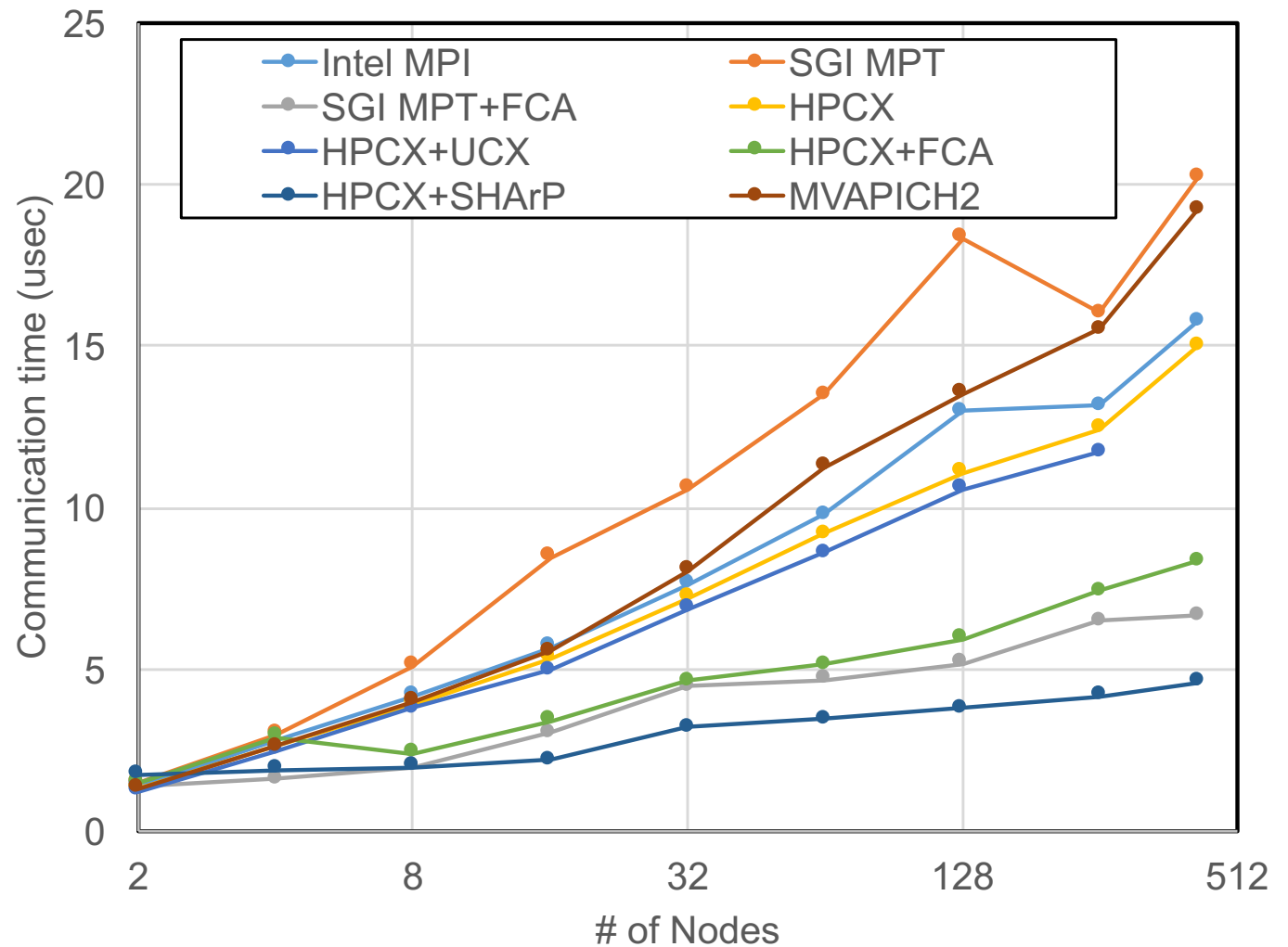
Pingpong (バンド幅)

- 最大: MVAPICH2
11.56 GB/sec
- 128KB以下は
HPCX+UCXがベスト
- SGI MPTが悪い？



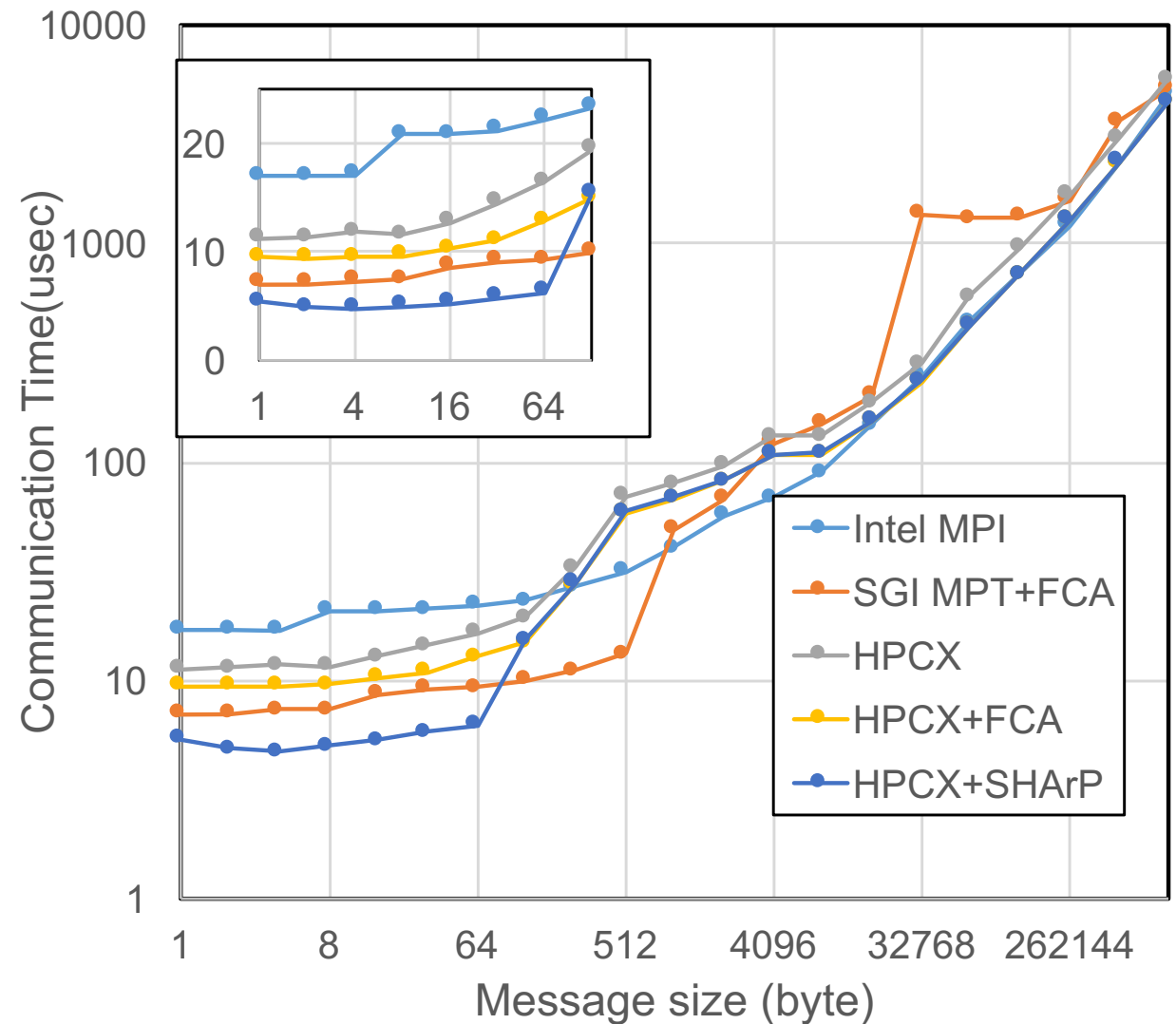
Barrier

- FCAにより3倍 (SGI MPT), 1.86倍(HPCX)高速
- SHArPにより3.2倍高速
- 420ノードでも4.6usしかかからない(!)



Allreduce

- 64byteまではSHArPでは6us程度



IOR (個別ファイル)

- プロセスごとに異なるファイルにアクセス
 - Intel MPI, -b 1m -t 1m -s 12000 (Lustre) -s 6000 (IME)

	I/O	向き	420ノード x1 プロセス	420ノード x2 プロセス	420ノード x4 プロセス
並列ファイルシステム (Lustre)	MPIIO	Write	76519	76194	77962
		Read	77484	78568	79715
	POSIX	Write	74308	77653	76774
		Read	79585	79944	80092
高速ファイルキャッシュ (IME)	MPIIO	Write	217422	220005	
		Read	196172	204259	
	POSIX	Write	218212	220852	
		Read	196849	208592	

Lustre: 理論ピーク性能 145.2 GB/sec の 50~55%

IME: 理論ピーク性能 385.2 GB/sec の 51~57%

IOR (単一ファイル)

- 単一ファイル中の別オフセットにアクセス
 - Intel MPI, -b 1m -t 1m -s 7000 (Lustre) -s 12000 (IME)

	I/O	向き	420ノード x1 プロセス	420ノード x2 プロセス	420ノード x4 プロセス
並列ファイルシステム (Lustre)	MPIIO	Write	20456	25549	
		Read	27225	35997	
	POSIX	Write	19775	20211	22419
		Read	28796	43939	33505
高速ファイルキャッシュ (IME)	MPIIO	Write	180496	173619	129778
		Read	247192	252049	245318
	POSIX	Write	149260	146783	133707
		Read	240619	243813	246661

Lustre: 理論ピーク性能 145.2 GB/sec の 13~30%, 個別ファイルに比べ性能低下
IME: 理論ピーク性能 385.2 GB/sec の 33~65%, 個別ファイルと同等かそれ以上

Burst Bufferに期待すること

- 大量の共有データの(比較的)高速な参照
 - ゲノム処理におけるデータベース
 - 大規模な密行列データ
 - ...
- 大規模アプリ連成のCoupler
 - 地震+建物
 - 大気+海洋
 - ...
- (Oakforest-PACSのテストベッド)

まとめ

- Reedbush
 - Reedbush-U: CPUのみ
2016年7月稼働開始、9月正式運用開始
 - Reedbush-H: GPU (NVIDIA Tesla P100)搭載
2017年3月稼働開始、4月正式運用開始
ストレージ等はU, Hで共有
- 最新のコモディティ技術を導入することで、高い性能を得るとともに短期間で安定稼働
 - Broadwell-EP, InfiniBand EDR, Burst Buffer (IME)
 - (NVIDIA Tesla P100)
 - 日本SGI, DDN Japanの皆様に感謝 (_o_)
- Burst Bufferについてはまだ性能評価途上