

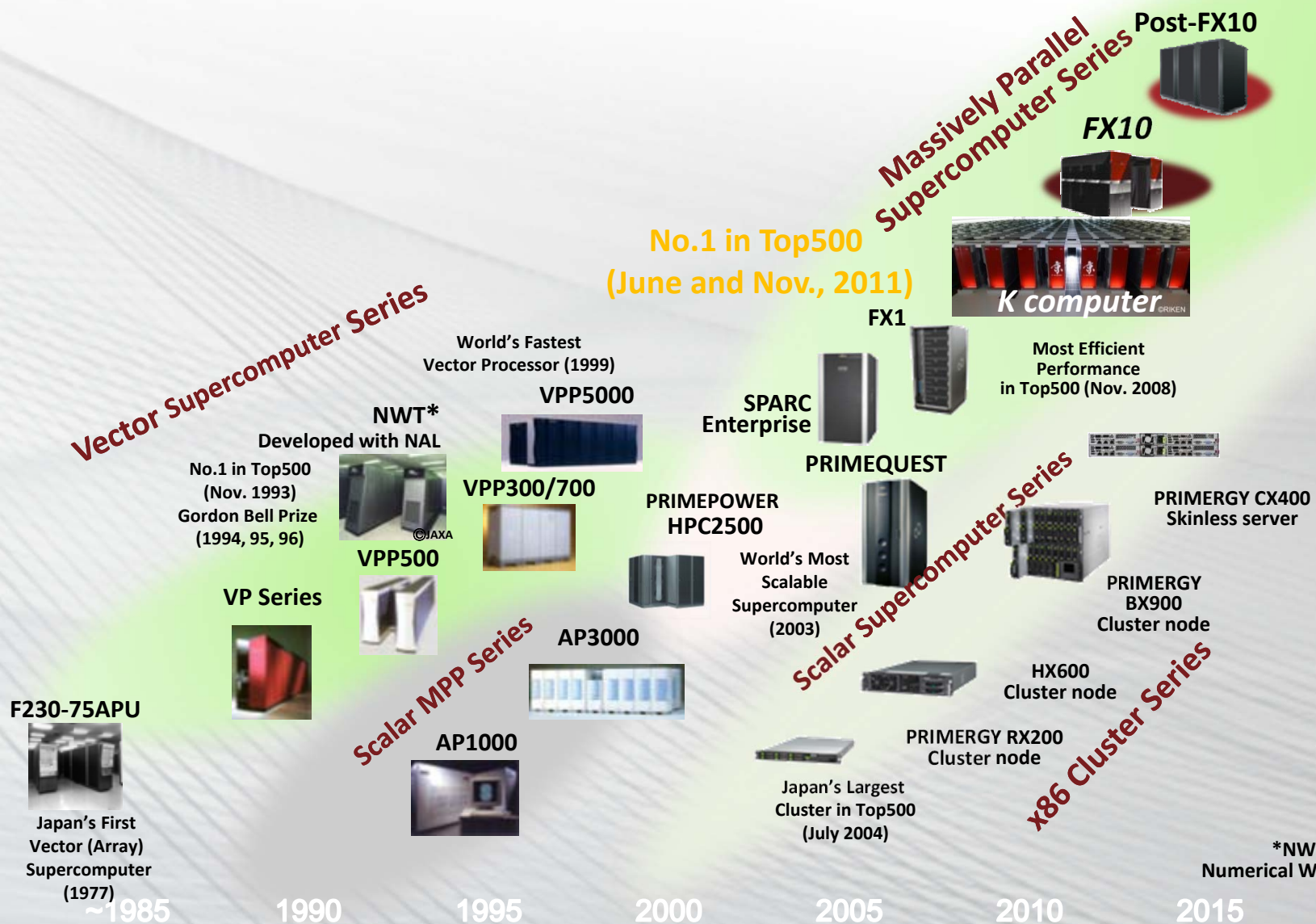
PRIMEHPC FX10後継機 開発の取り組み

2014.8.26.

新庄 直樹
富士通株式会社

- 富士通のスパコン開発の歴史とロードマップ
 - 「京」とPRIMEHPC FX10
- Post-FX10: PRIMEHPC FX10後継機
 - デザインコンセプトと取組み
 - システムの特長
 - システム構成
 - 性能評価
- まとめ

富士通のスパコンの歴史



スパコン開発とExascaleへの取組



■ 「京」、PRIMEHPC FX10

- ・多くのアプリケーションが創出
- ・産業分野含む多くの成果

■ Post-FX10

- ・CPU、インターコネクトなど「京」、FX10のアーキを継承

■ Exascaleへ向けた研究開発

- ・高性能、省電力テクノロジーの開発
- ・国家プロジェクトへの参画

「京」、PRIMEHPC FX10

「京」の概要

■ システム概要

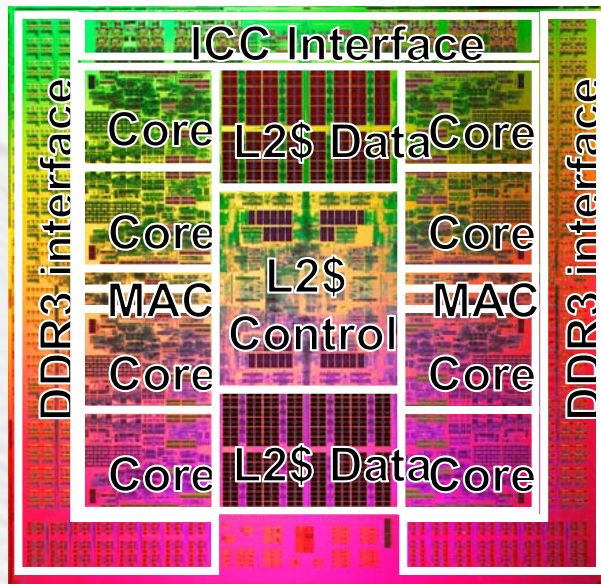
- 88,128 ノード
- 理研との共同開発
- Top500 #1(2011年6月、11月)
- ACM Gordon Bell賞受賞(2011、2012)



「京」

システム	理論ピーク性能： 11.28 petaflops LINPACK性能： 10.51petaflops CPU数： 88,128 総メモリ容量： 1.26 petabytes
CPU	SPARC64™ VIIIfx (8 core, 128 gigaflops)
インターコネクト	6次元メッシュ/トーラストポロジ (Tofu)

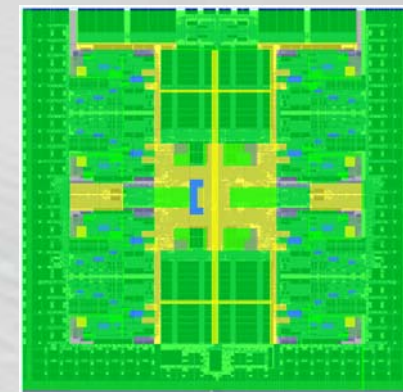
SPARC64 VIIIfx



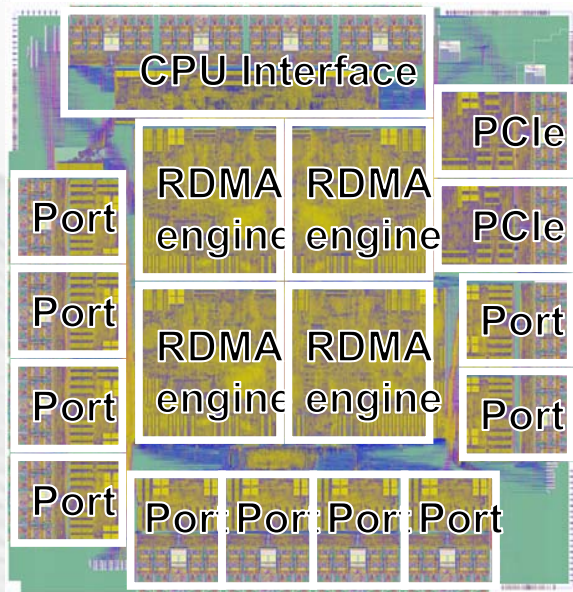
テクノロジー	45nm
浮動小数点性能	128GFLOPS
メモリバンド幅	64GB/s
消費電力	58W
トランジスタ数	760M

- 8 core Out-of-orderスーパースcalarCPU
- HPC-ACE*命令サポート
- VISIMPACTハイブリッド実行モデルサポート
- 低消費電力
- メインフレームの高信頼設計

*HPC- Arithmetic Computational Extensions



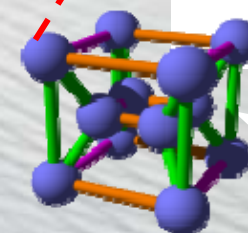
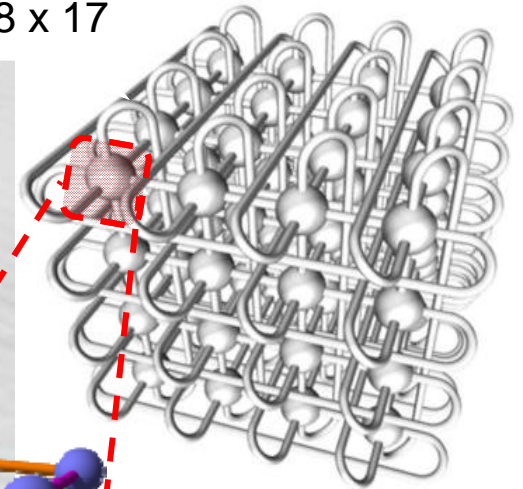
- Error detect by hardware and recover automatically
- Error detect by hardware
- No affect to system operation



テクノロジー	65nm
DMA Engine	Send x 4 + recv. x 4
リンクバンド幅	5+5GB/s X 10 ports
PCIe	16 lane Gen2
トランジスタ数	200M

- 6次元メッシュ/トーラス網
K: $(24 \times 18 \times 17) \times (2 \times 3 \times 2)$
低ホップ数、高バイセクションバンド幅
- 仮想3次元トーラス
- ハードウェア集団通信サポート
- GAP挿入による輻輳制御

3D torus
24 x 18 x 17



3D mesh/torus
 $2 \times 3 \times 2$

「京」からPRIMEHPC FX10へ



■「京」のアーキテクチャを継承し、性能向上

- 40nm 16コアCPU
- 性能、メモリバンド幅を向上
- 「京」アプリとのバイナリコンパチビリティ

	「京」	PRIMEHPC FX10	Note
CPU	SPARC64 VIIIfx	SPARC64 IXfx	SPARC V9 + HPC-ACE
ピーク性能	128 GFLOPS	236.5 GFLOPS	
コア数	8	16	
メモリ容量	16GB	32GB/64GB	2GB/core~
バンド幅	64GB/s	85GB/s	
インターコネクト	6D mesh/torus	←	Tofu
システムサイズ	X x Y x 17	X x Y x 9	Z=0 is I/O node
リンクバンド幅	5GB/s x 双方向	←	

Post-FX10

■ デザインコンセプト

- 「京」やPRIMEHPC FX10とのアプリケーション互換
- 高いノード性能、スケーラビリティと可用性
- 高い電力性能効率と信頼性

■ 取組み

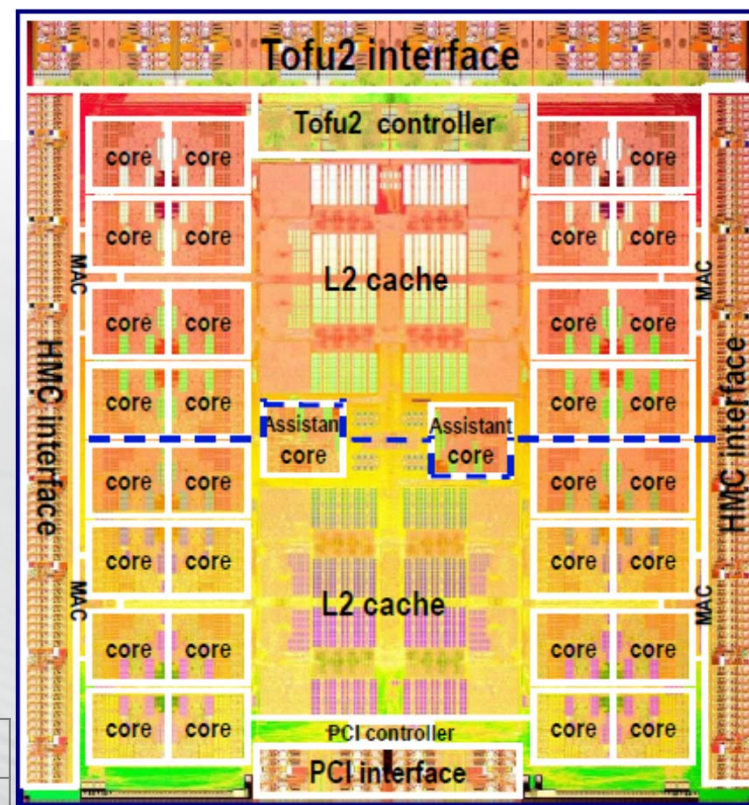
- 1 CPU/ノード
- HPC向け命令、uArchの強化
- Tofu2内蔵
- 高速積層メモリ、光サポート

Post-FX10の特長(1/2)

■ SPARC64 XIfx

- L1キャッシュ、Wayを2倍に
- アウトオブオーダー資源、分岐予測などを強化
- 256 bit幅SIMD
 - 2倍単精度モード、8バイト整数命令
- アシスタントコアx2
 - OSノイズ低減、演算と通信のオーバーラップ
- HMCサポート
- Tofuインタコネクト2内蔵

アーキテクチャ	SPARC V9 + HPC-ACE2
コア数	32 コア + 2 アシスタントコア
演算器	FMA x 2 (256 bit wide SIMD)
キャッシュ	L1 inst. cache: 64 KB / core L1 data cache: 64 KB / core L2 cache: 24 MB / node
メモリ容量	32 GB / ノード
メモリバンド幅	240 GB/s (Read) + 240 GB/s (Write)



Post-FX10の特長(2/2)

■ Tofu2

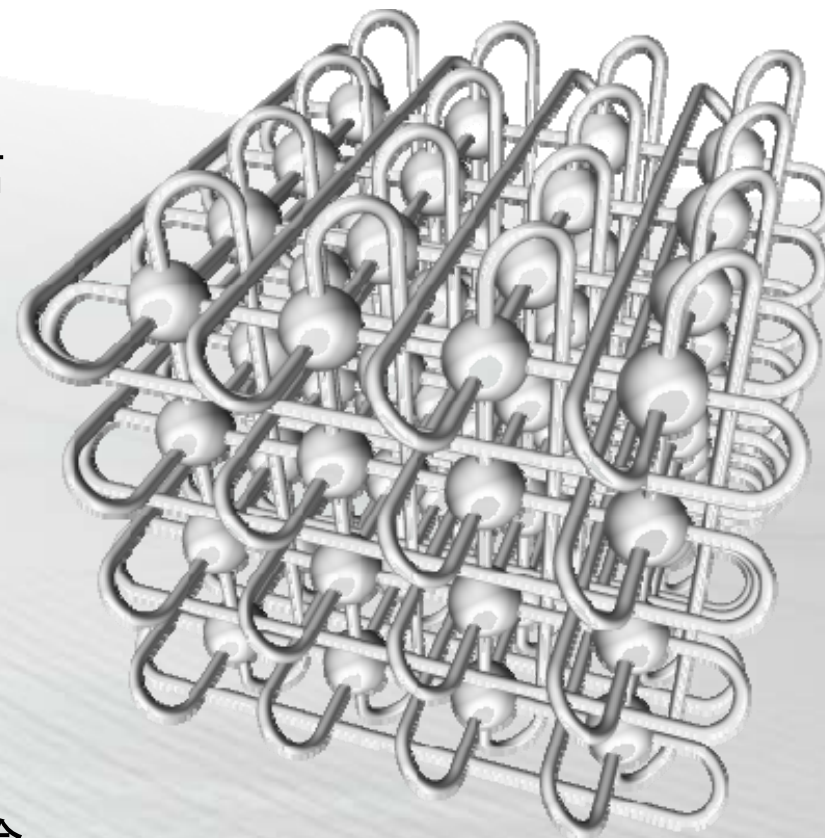
- 「京」互換のトポロジ、通信方式
- 複数RDMAエンジンによる高速集団通信
- バリアハードウェアサポート

■ 19インチラックマウント型シャーシ

- 12ノード / 2U
- シャーシ間光接続

■ 水冷

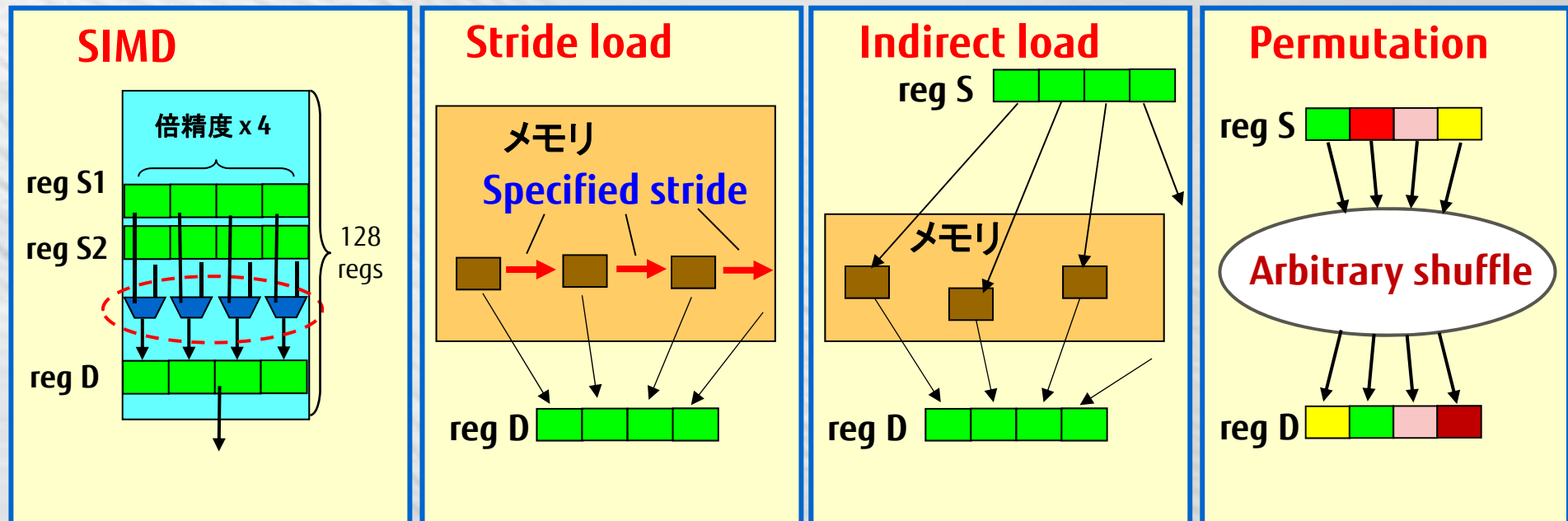
- メモリ、光モジュールに対しても直接水冷
 - ・ 水冷率90%



SIMD拡張 (HPC-ACE2)

■ 256bit wide SIMD

- 倍精度浮動小数点演算 x 4 または 単精度 x 8 または 整数 x 4
- スライドLoad/Store,
- インダイレクト(list) Load/Store
- Permutation, Concatenate



HMC(Hybrid Memory Cube)の採用

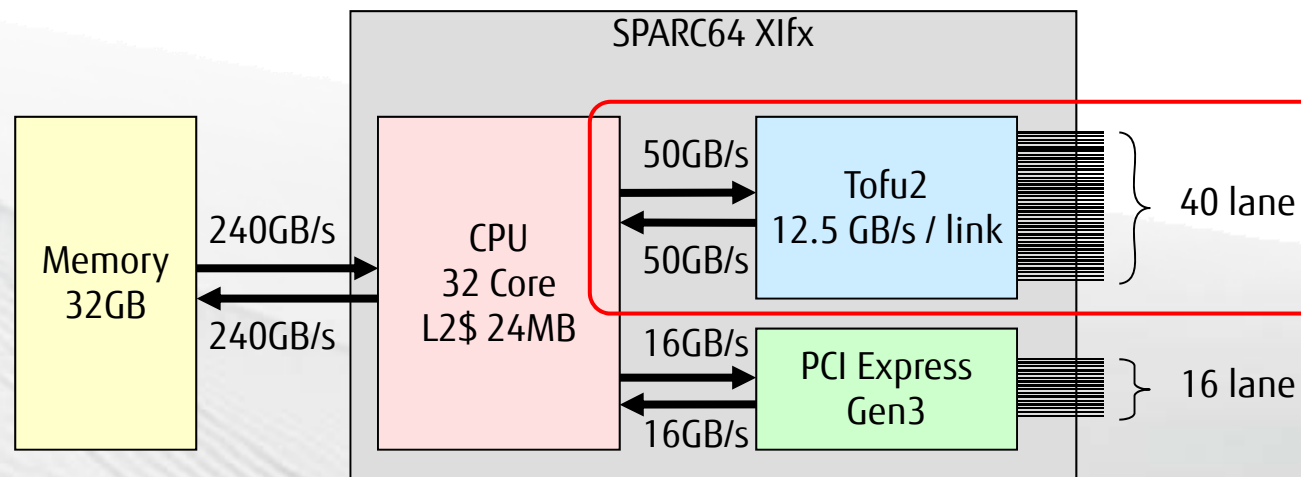
■ HMCの特長

- バンド幅あたり高実装密度
- パッケージあたり大容量、高バンド幅
- バンド幅あたり低消費電力

■ HMCの採用により、「京」、FX10と同等の容量、バンド幅比を実現

1 CPUあたり	容量	バンド幅
HMC x8	32GB	480GB/s
DDR4-DIMM x8	32~128GB	154GB/s
GDDR5 x16	8GB	320GB/s

Tofu2



	Tofu	Tofu2
システム	「京」および FX10	Post-FX10
対応CPU	SPARC64 VIIIfx/Ix fx	SPARC64 Xlfx
CPUへの内蔵	(別LSI(ICC)で実現)	内蔵
トポロジ	6次元メッシュ/トーラス	←
リンクバンド幅	5 GB/s (6.25 Gbps x 8 lanes x 10 dirs)	12.5 GB/s (25 Gbps x 4 lanes x 10 dirs)
ノードバンド幅	20 GB/s x in/out	50 GB/s x in/out
その他	-	キャッシュインジェクション、アトミック 光接続: シャーシ間接続(全体の2/3)を光化

Post-FX10の構成

Fujitsu designed SPARC64™ XIfx

- ◆ 32 + 2 core CPU
- ◆ HPC-ACE2 support
- ◆ Tofu2 integrated

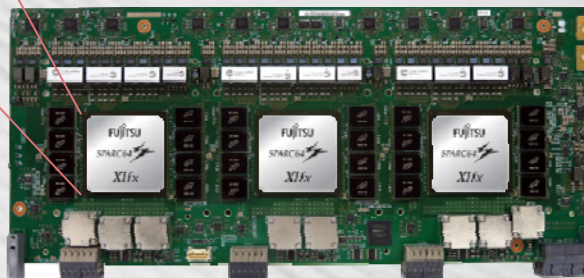


Tofu Interconnect 2

- ◆ 12.5 GB/s×2(in/out)/link
- ◆ 10 links/node
- ◆ Optical technology

Chassis (12 CPUs)

- ◆ 1 CPU/1 node
- ◆ 12 nodes/2U Chassis
- ◆ Water cooled



CPU Memory Board

- ◆ CPU x 3
- ◆ 3 x 8 Micron's HMCs
- ◆ 8 Finisar's opt modules, BOA, for inter-chassis connections



Cabinet

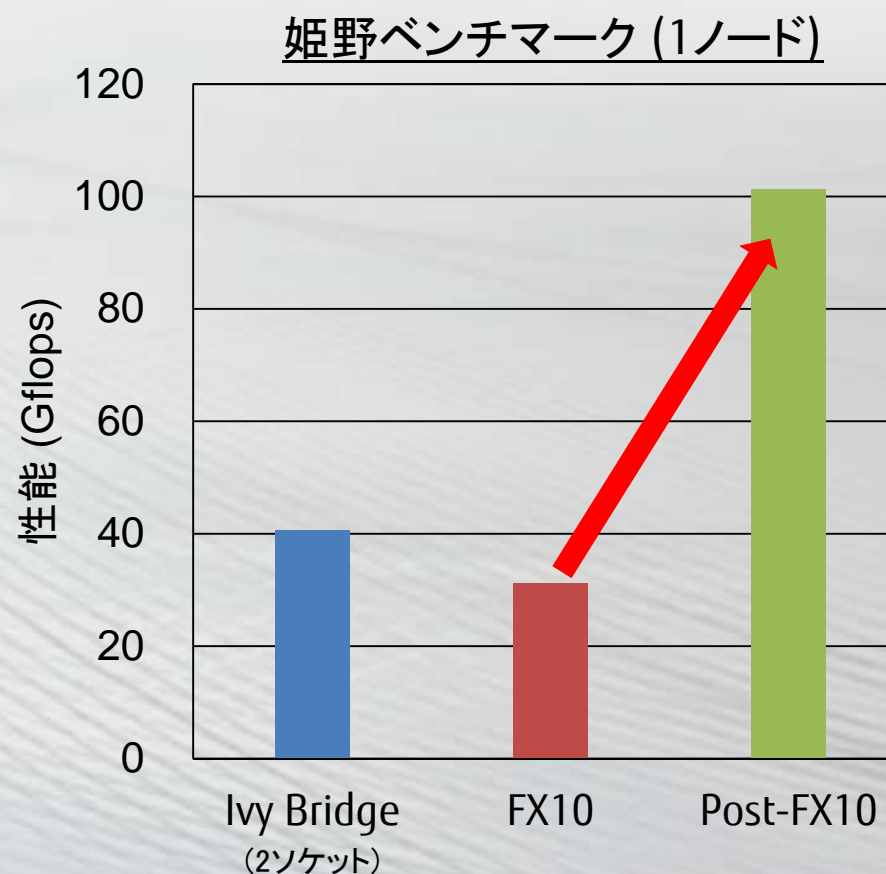
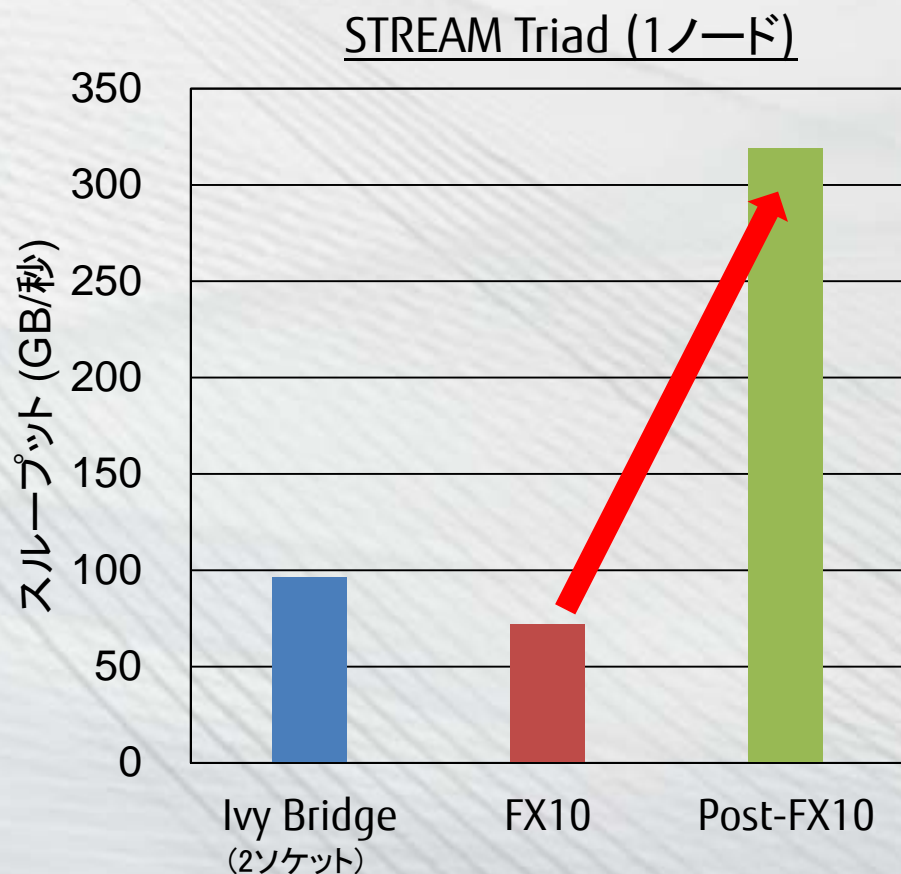
- ◆ 216 nodes/cabinet
- ◆ High-density
- ◆ 100% water cooled with EXCU (option)

Post-FX10の性能

メモリスループットの向上

■ 高速メモリ(HMC)の採用により、 ノードあたりメモリスループットがFX10の3~4倍に向上

■ 姫野ベンチマーク: 流体解析コードの性能評価。メモリスループットの影響大

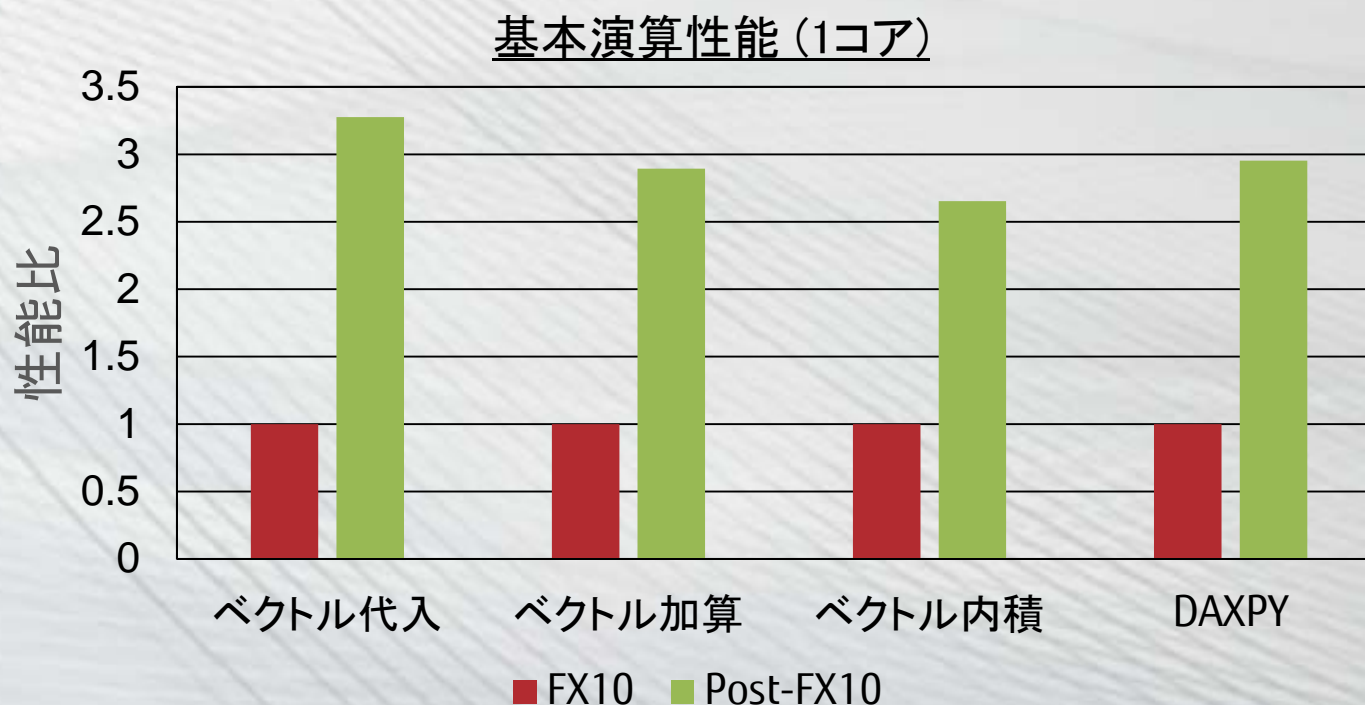


演算性能の向上 (1/3)

■ SIMD幅の倍増(256ビット)により、
コアあたり演算性能がFX10の2.7~3.3倍に向上

【他の強化ポイント】

- L1キャッシュ制御を改良
- CPU周波数を向上



ベクトル代入:

$$y(i) = x(i)$$

ベクトル加算:

$$y(i) = x1(i) + x2(i)$$

ベクトル内積:

$$s = s + x1(i) * x2(i)$$

DAXPY:

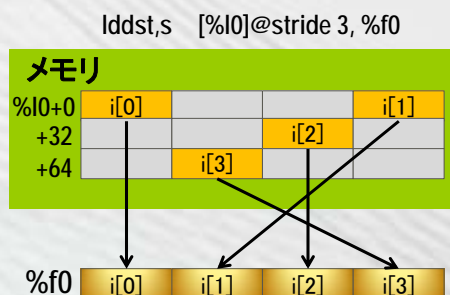
$$y(i) = y(i) + s * x(i)$$

演算性能の向上 (2/3)

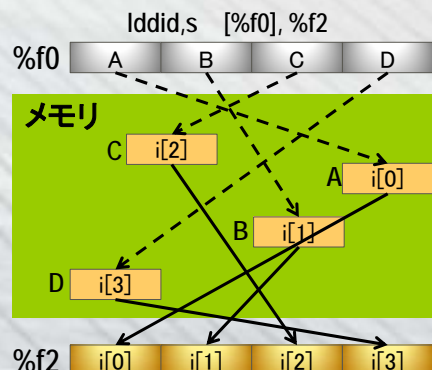
■ スライドやインダイレクトのロード/ストア命令を追加し、SIMD化対象の処理を拡大

- 対応するスライド幅: 2~7
- インダイレクトアクセスのアドレス計算もSIMD化対象

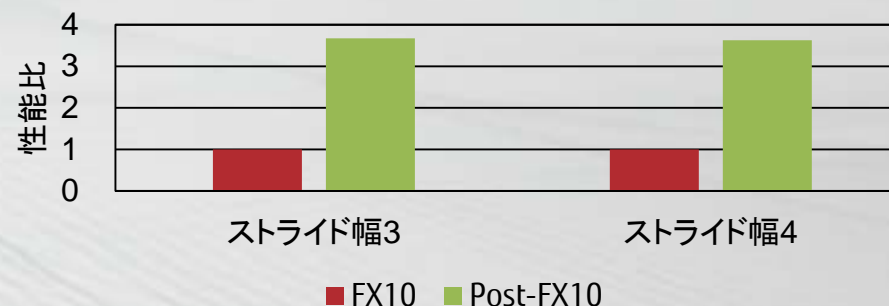
スライド幅3のロード命令



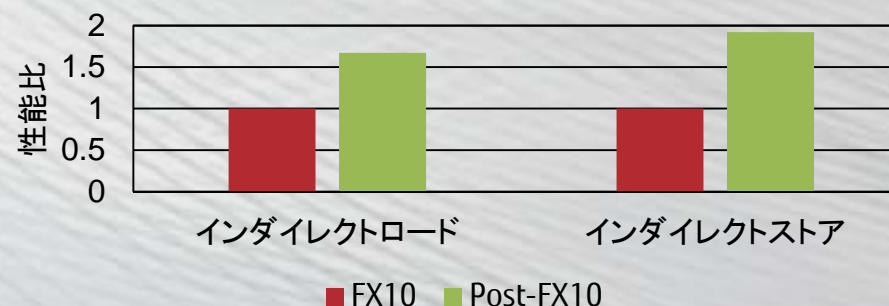
インダイレクトロード命令



スライドロード性能 (1コア)



インダイレクトアクセス性能 (1コア)

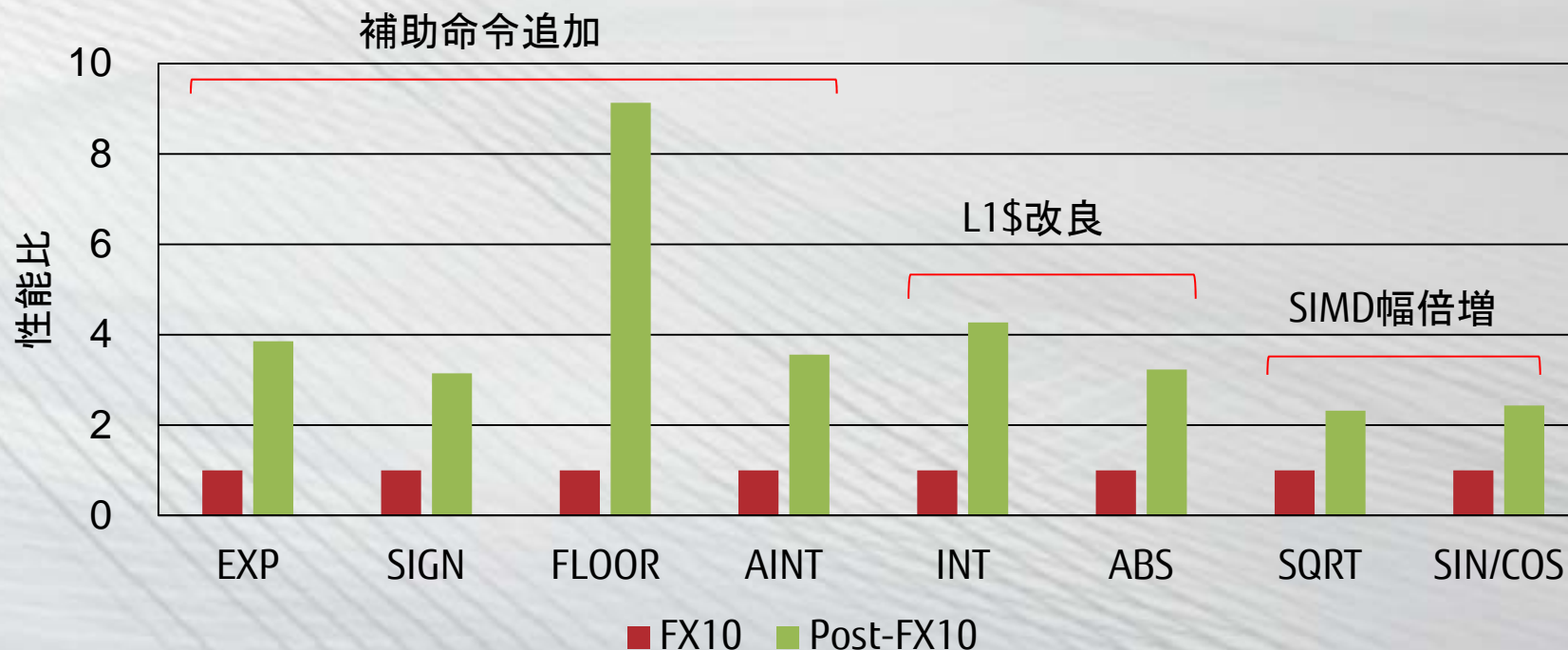


演算性能の向上 (3/3)

■ 演算補助命令の追加やSIMD幅倍増により、組み込み関数の処理性能も向上

- EXP/SIGN/FLOOR/AINT: 補助命令を新たに追加
- INT/ABS: L1キャッシュ制御を改良
- SQRT/SIN/COS: SIMD幅倍増により性能向上

組み込み関数の処理性能 (1コア)

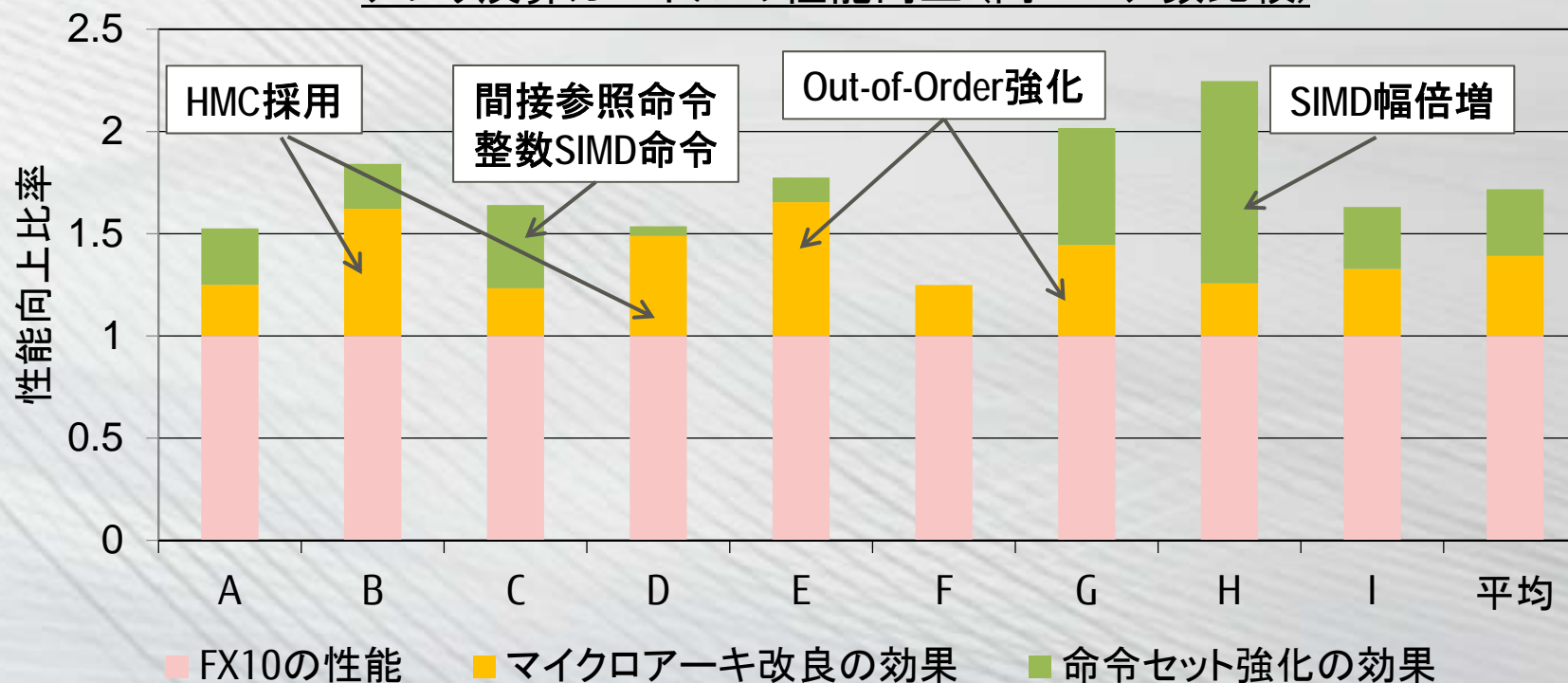


アプリ演算カーネルによる評価

■ CPUマイクロアーキ改良と命令セット強化により、コアあたり演算性能がFX10の1.7倍に向上

- 主に流体系コードによる評価
- マイクロアーキ改良で平均40%、命令セット強化で平均33%の向上

アプリ演算カーネルの性能向上（同一コア数比較）

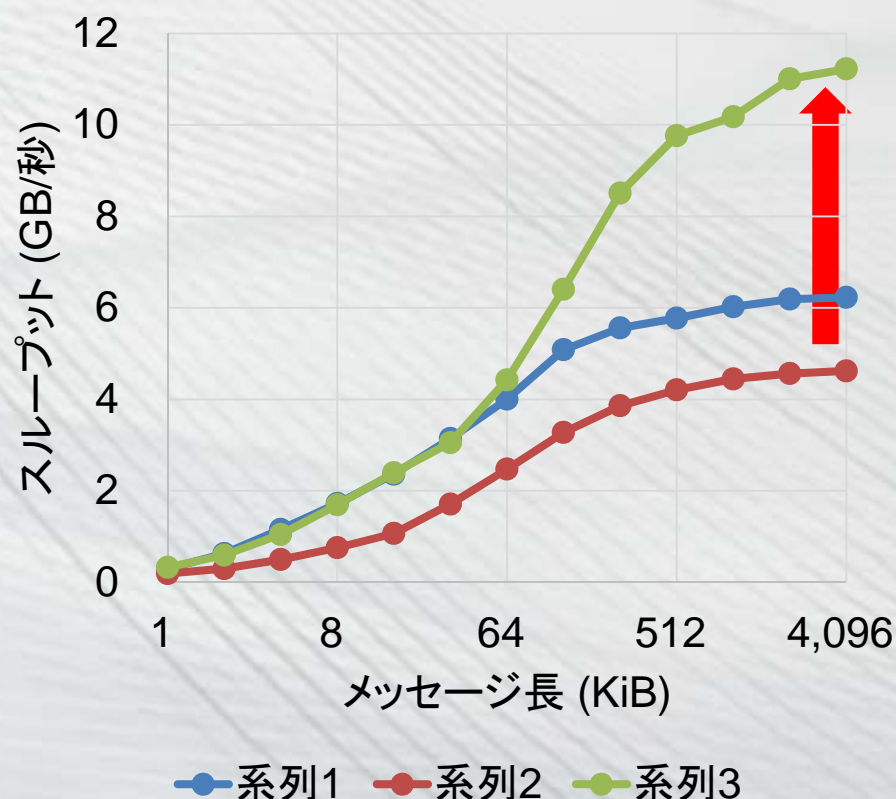


通信スループットの向上

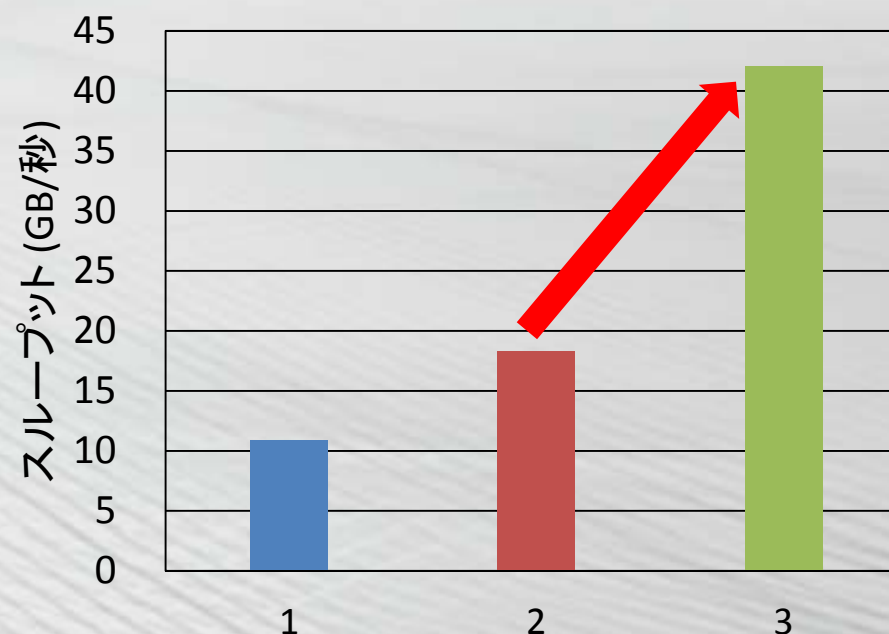
■ Tofuリンクの高速化(25Gbps)により、
リンクあたりスループットがFX10の2.4倍に向上

■ 特に多方向同時通信では、圧倒的なスループットを実現

IMB PingPongスループット (ノード間)



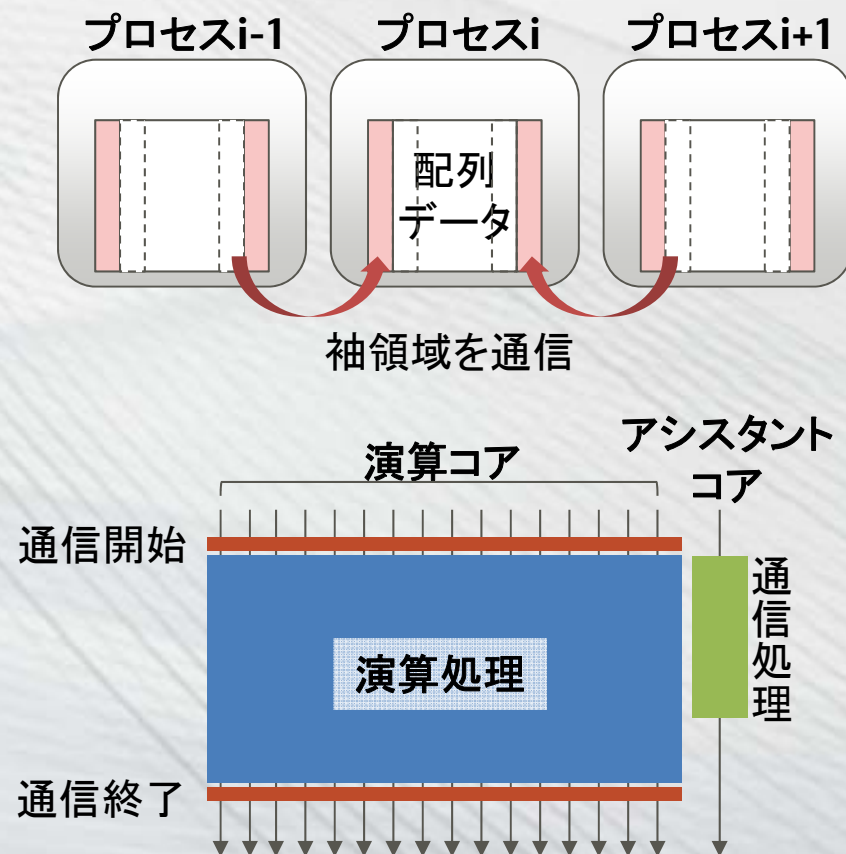
2方向同時通信時の送受信スループット



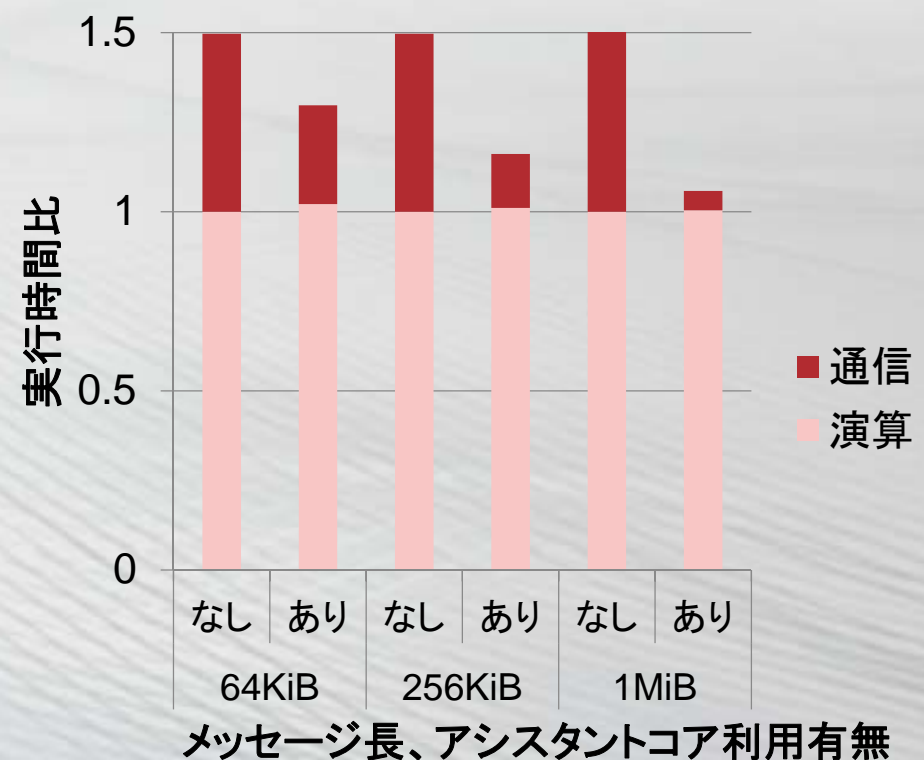
ノンブロッキング通信のオーバーラップ実行

■ アシスタントコア(2コア/ノード)により、 演算と通信のオーバーラップ実行が容易に

■ ステンシル計算の袖領域通信に有効



アシスタントコア利用による通信オーバーラップ

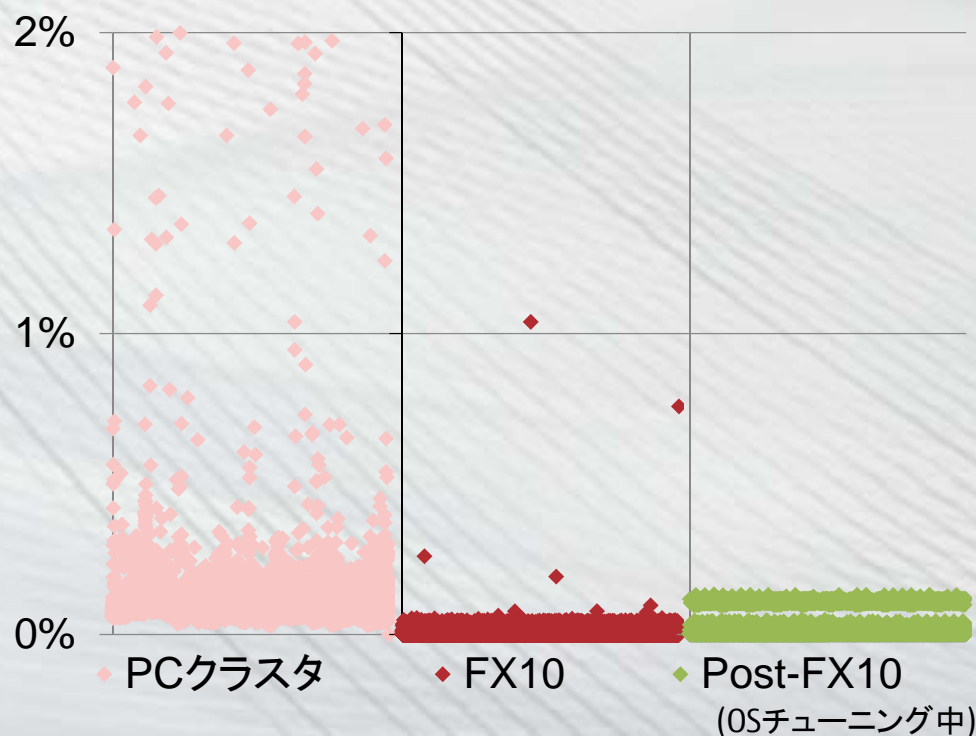


OSノイズのさらなる低減

■ OSカスタマイズとアシスタントコア活用により、 高並列時の性能ロスをFX10からさらに低減

- デーモンやIO割り込みの処理を演算コアから排除
- エクサ時代では通信間隔が短くなるため、いっそう重要な対策に

OSノイズによる演算時間のバラつき



OSノイズによる性能ロス(通信間隔1ms)



■ Post-FX10 (PRIMEHPC FX10後継機)

■ CPUを独自開発

- HPC向けに性能強化を実施
- インターコネクトを統合
- 「京」、PRIMEHPC FX10のアーキテクチャを継承

■ メモリにHMCを採用

- 高いメモリバンド幅を実現

■ システムソフトウェアも開発

- OS、コンパイラ、スケジューラ、ライブラリ

■ エクサスケールに向けた研究開発



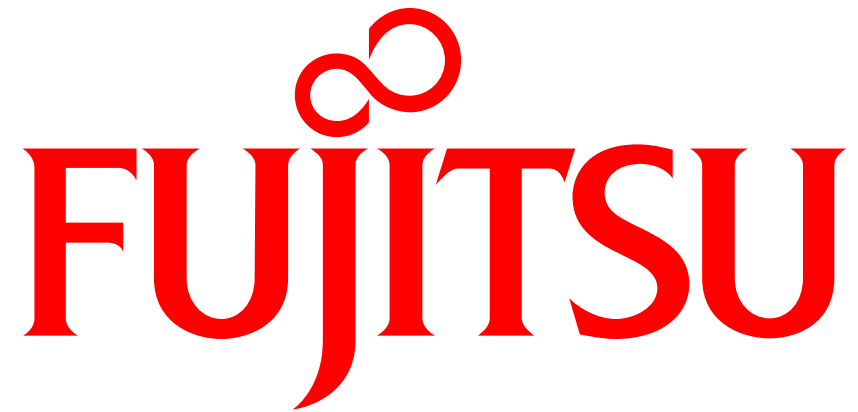
2010

2015

2020

Post-FX10

Exascale system



shaping tomorrow with you