SS研　HPCフォーラム　富士通報告

FUJITSU
shaping tomorrow with you

FXシリーズの今後の取り組みについて

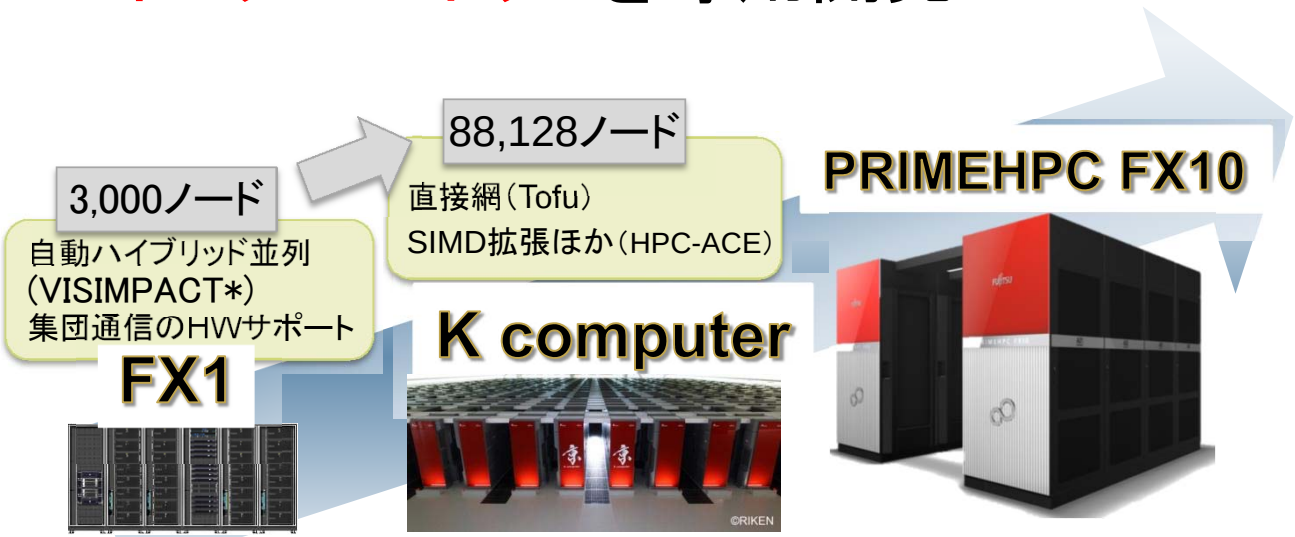
2013年8月27日

追永　勇次
富士通株式会社
次世代テクニカルコンピューティング開発本部

FXシリーズ・「京*1」の超並列への取り組み

FUJITSU

- ■マルチコアを想定したシングルCPU/ノード構成
- ■超並列を見据えたアーキを段階的に開発・導入
  - ■1千万コアを使いこなす（VISIMPACT*2、集団通信HW）
  - ■CPUとインターコネクトを専用開発



CY2008～	CY2010～	CY2012～
40GF, 4コアCPU	128GF, 8コアCPU	236.5GF, 16コアCPU

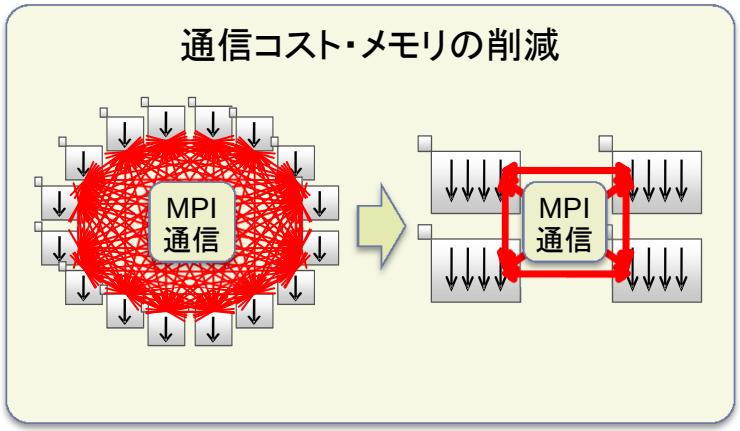
*1「京」は独立行政法人理化学研究所の登録商標です。
*1 スーパーコンピュータ「京」は、独立行政法人理化学研究所と富士通の共同開発です。
*2 VISIMPACT : Virtual Single Processor by Integrated Multi-core Parallel Architecture

Copyright 2013 FUJITSU LIMITED

■自動並列化で容易にハイブリッド並列化を実現

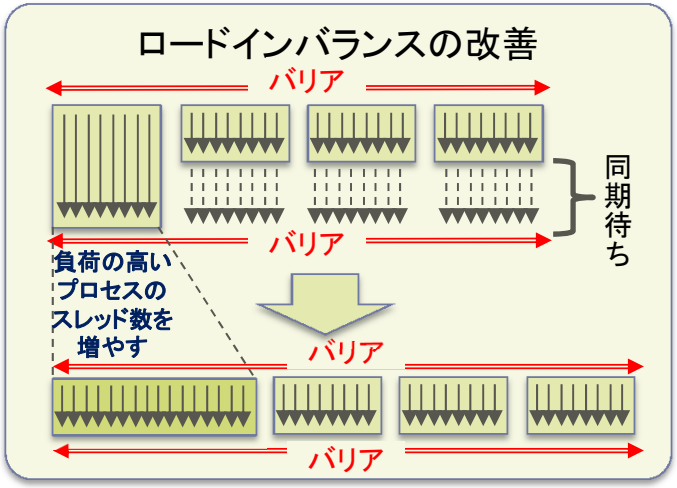
■当初の狙い

- ■高並列時の通信コスト、通信バッファの削減
- ■ユーザメモリの容量増によりアプリの移植性を向上



■次期FX

- ■ヘテロなハイブリッド並列のサポート
  - ロードインバランス対策
  - マルチフィジクス向け実行環境



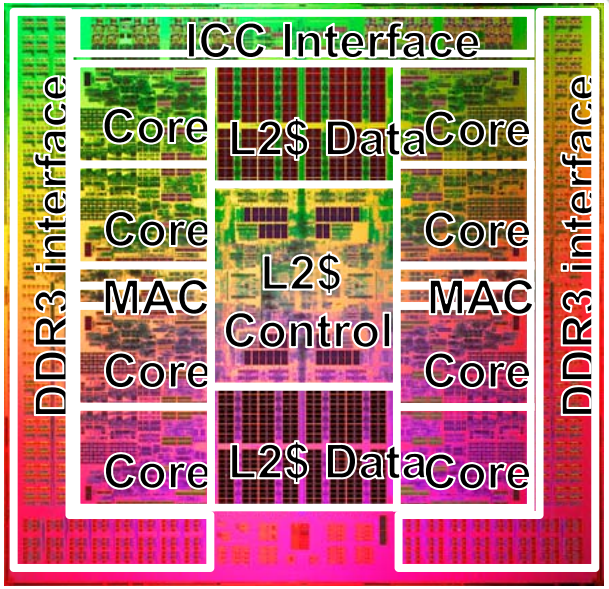
Copyright 2013 FUJITSU LIMITED

「京」のCPU SPARC64 VIIIfx

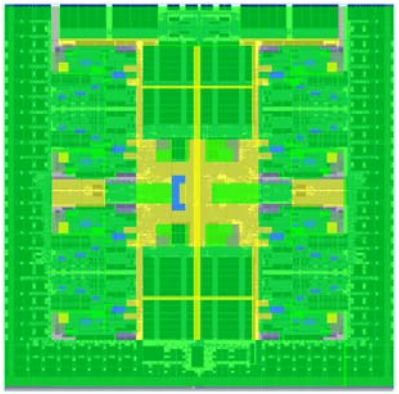
FUJITSU

■概要

テクノロジ	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



- ■Out-of-Orderスーパスカラ CPU
- ■HPC-ACE命令セット拡張
- ■低消費電力化の追求
- ■エンタープライズサーバに匹敵する高信頼性
  - ■メインフレームで培ったRAS機能の継承



ハードによるエラー検出
＋自己修復可能領域

ハードによるエラー検出領域

エラーが実運用に影響しない領域

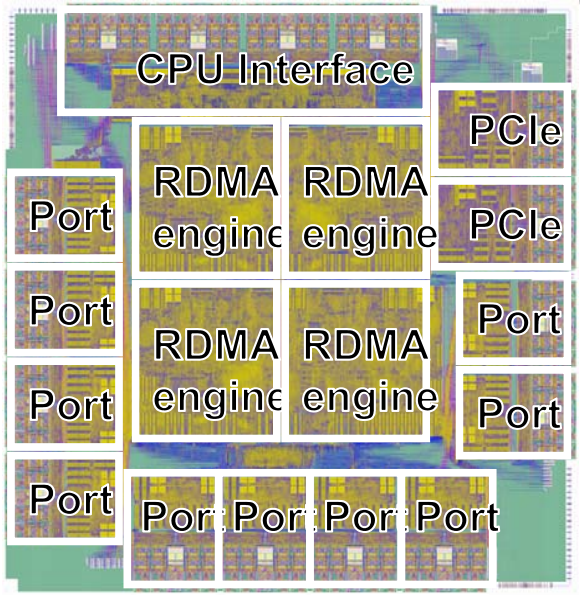
Copyright 2013 FUJITSU LIMITED

「京」のインターコネクト Tofu

FUJITSU

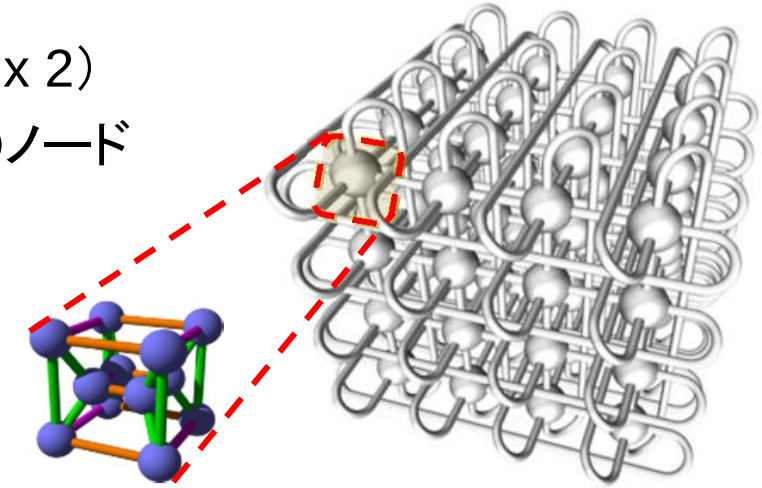
■LSI（ICC）概要

テクノロジ	65nm
同時転送数	送信 x 4 + 受信 x 4
リンクバンド幅	5+5GB/s X 10ポート
PCIe	16 lane Gen2
トランジスタ数	200M



■直接網6D mesh/torusトポロジー

- （X, Y, Z, A, B, C）=（32 x 32 x 32 x 2 x 3 x 2）
  - 京：（24 x 18 x 17 x 2 x 3 x 2）、Z=0がIOノード
  - X軸、Z軸、B軸がトーラス
- 平均ホップ数、バイセクション帯域を改善



■集団通信HWサポート

- データバスは共用

Copyright 2013 FUJITSU LIMITED

「京」・FX10のOS・システムソフト

FUJITSU

- Linux（Redhat）をベースにHPC機能（ラージページ、ジッター対策）を追加
- OS・システムソフト・言語全体（14Mstep）を自前でサポート
  - システム運用管理 / コンパイラは自社開発
  - ファイルシステム / MPIは、オープンソースを移植・カスタマイズ（成果をコミュニティへもフィードバック）
- PCクラスタとのハイブリッドシステムにも対応

システム管理ポータル / HPCポータル

Technical Computing Suite

システム運用管理
- システム管理ソフト
  - 1システムイメージ
  - 1タッチIPL
  - フェールセーフ化
- ジョブ管理ソフト
  - 高効率スケジューラ

ファイルシステム
- FEFS
  - Lustreベース
  - 高スケーラビリティ（数千IOサーバに対応）
  - 高速IO性能（1.4TB/s）

プログラミング環境
- コンパイラ
  - Fortran, XPF, C, C++
  - 自動並列化/SIMD化
- MPI
  - OpenMPIベース
- ツール、数学ライブラリ

OS（Linux + HPC機能）

Copyright 2013 FUJITSU LIMITED

FXシリーズ開発で得られた成果（1）

FUJITSU

■「京」の成果

- ■超大規模システム（8万ノード超）をセンターマシンとして安定運用
- ■超高並列（8万並列超）の並列アプリで高い実行性能を達成
- ■ベクトル機でカバーできなかったライフサイエンス等、新分野でも高い実行効率

名称	アプリ概要	並列数	効率
LINPACK	連立一次方程式求解	88,128	93%
NICAM	全球雲解像大気大循環モデル	81,920	8%
Seism3D	地震波伝播・強震動シミュレーション	82,944	18%
PHASE	平面波展開第一原理電子状態解析	82,944	20%
RSDFT	実空間第一原理電子状態解析	82,944	52%
FrontFlow/Blue	Large Eddy Simulation (LES)に基づく非定常流体解析	82,944	3%
Lattice QCD	格子QCDシミュレーションによる素粒子・原子核研究	82,944	15%
ZZ-EFSI	Euler型流体構造連成による血栓シミュレーション	82,944	43%

理化学研究所 南様、次世代生命体統合シミュレーションソフトウェアの研究開発プロジェクト様 ご提供

Copyright 2013 FUJITSU LIMITED

FXシリーズ開発で得られた成果（2）

FUJITSU

- プロセッサとインタコネクト双方の技術を取得、SoCにより高密度化、低コスト化が可能 ⇒ エクサに向けて必須な技術
- プロセッサとインタコネクト双方の技術を持つのは富士通とIBMのみ

ベンダー		富士通	IBM	CRAY	Intel	NVIDIA
HPCシステム		「京」・FX10	BG/Q	XC30		
プロセッサ	ベース	SPARC64	PowerPC A2	Intel	Pentium	GK104
	HPC拡張	HPC-ACE	QPX	Intel / NVIDIA	Xeon Phi	GK110
インタコネクト	名称	Tofu inter-connect	BG/Q network	Aries	Crayの技術を買収	
	トポロジー	6D mesh/torus	5D torus	Dragonfly		

: 自社技術で対応

Copyright 2013 FUJITSU LIMITED

FXシリーズ開発で得られた成果（3）

FUJITSU

- システムソフト全体を開発できる技術力を維持
- システムソフトを全て自社技術で対応できるのは、富士通、IBM、NECのみ

ベンダー	富士通	IBM	日立	CRAY	SGI	NEC
HPCシステム	「京」・FX10	BG/Q	SR16000	XC30	UV	SX-9
OS	○	○	IBM	○	RedHat, SuSE	○
コンパイラ	○	○	○	Intel	Intel	○
MPI	○	○	IBM	○	○	○
システム管理	○	○	IBM	○	○	○
ジョブ管理	○	○	IBM	PBS Pro etc.	PBS Pro etc.	○
ファイルシステム	○ (FEFS)	○ (GPFS)	○ (HSFS)	Lustre	○ (CXFS)	○ (GFS)

○ : 自社技術で対応

Copyright 2013 FUJITSU LIMITED

しかしながら・・・

Copyright 2013 FUJITSU LIMITED

「京」・FX10へのご指摘

FUJITSU

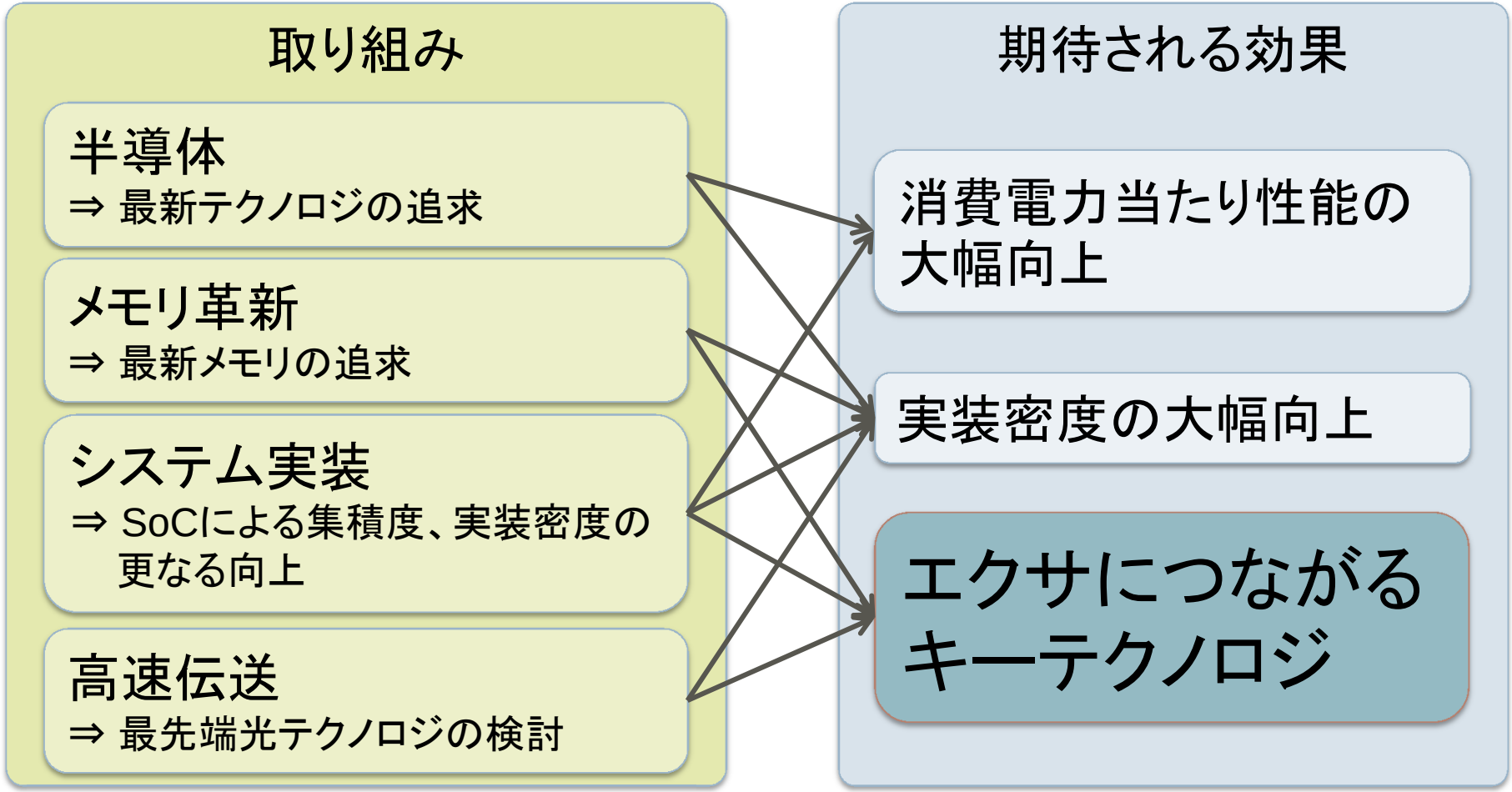
- 消費電力当たりの性能をもっと上げて欲しい
- 実装密度をもっと上げて欲しい
- 1コア性能を引き出すのが難しい（手間がかかる）
- C/C++プログラムの性能をもっと上げて欲しい

Copyright 2013 FUJITSU LIMITED

次期FXでのチャレンジ

FUJITSU

- 消費電力当たり性能の大幅向上
- 実装密度の大幅向上



Copyright 2013 FUJITSU LIMITED

次期FXでの逐次性能の向上

FUJITSU

■逐次性能はアプリ性能の基本、次期FXで取り組む

逐次性能の課題

チューニングしないと性能がでない
- 少回転ループ
- メモリアクセスが多いコード
- 間接参照が多いコード
- IF文の多いコード

ピーク性能に比べて性能が低い
- FMA化されないコード
- 単精度のコード
- 整数系のコード

C/C++プログラム性能

L1キャッシュ容量不足

マイクロアーキ
- Out-of-order資源不足
- 最新分岐予測技術の未対応

命令セット
- FMA命令採用
- 単精度倍幅SIMDの未サポート
- 整数SIMDの未サポート

コンパイラの機能不足
- プリフェッチ、マスク機能の標準オプション未サポート
- 最新最適化技術の未対応

Copyright 2013 FUJITSU LIMITED

プロセッサの比較

FUJITSU

- SPARC64 IXfx
  - ピーク性能比通りの性能が出ない
    - FMAをサポート
    - 単精度倍幅SIMD、整数SIMDを未サポート
- 今後のHPC向けプロセッサは長SIMDがカギ

		SPARC64 IXfx	**Sandybridge**	Xeon Phi (KNC)	GPGPU (Kepler)
周波数（GHz）		1.849	2.9	1.1	0.732
ピーク性能（GFlops）		236.7	185.6	1073.6	1311.7
構成	演算器構成	**FMA×2**	**（FM or FA）×2**	FMA×1	FMA×2
	SIMD幅	2	4	8	32
	コア数	16	8	61	14
SIMD機能	単精度倍幅	**×**	**○**	○	○
	整数SIMD	**×**	**○**	○	○

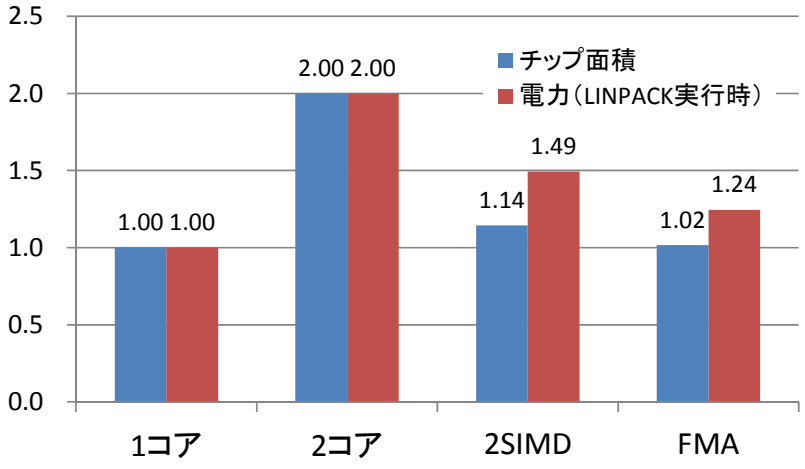
Copyright 2013 FUJITSU LIMITED

FMA演算（性能を2倍にする方式の比較）

FUJITSU

- 絶対性能の向上は、コア2倍、2SIMDに比べて低い
- チップ面積当たりの性能向上は最良
  - SPEC fpでもチップ面積当たりの性能向上が良好
  - LINPACKでは2倍の性能向上を実現
    - 「京」のLINPACKで93%の効率を達成

実行命令の比率（SPEC CFP2006）				チップ面積増
浮動小数点命令	47.59%	(演算命令27.62%)		ー
SIMD命令	18.77%	(演算命令12.07%)		14%
FMA命令	**11.87%**	非SIMD	6.77%	**2%**
		SIMD	5.10%	



FMA、2SIMD、2コアの比較

Copyright 2013 FUJITSU LIMITED

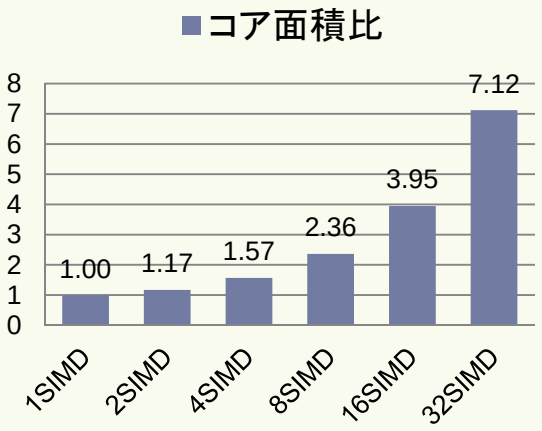
SIMDの特徴

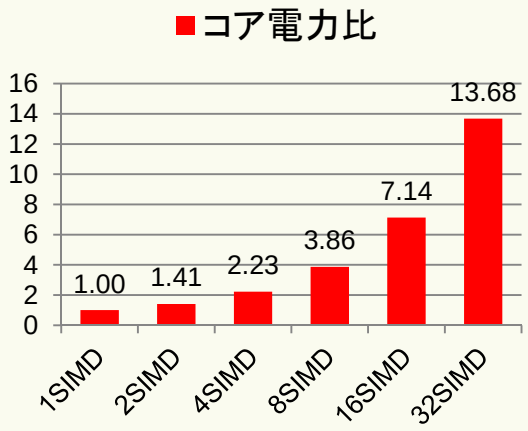
FUJITSU

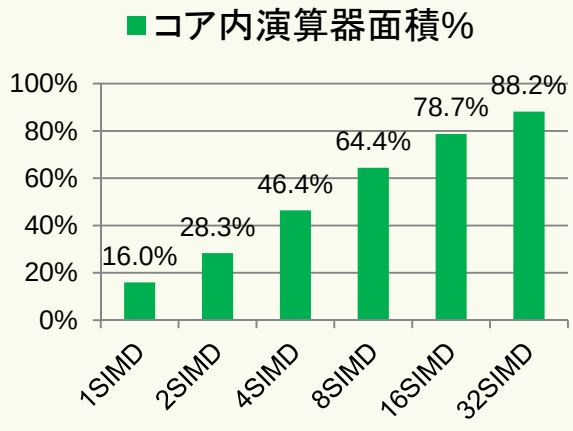
- コア数増よりSIMD幅増の方が、面積および電力効率が良好
- 8SIMDを超えると演算器の面積が半分以上を占める
- 汎用プロセサもSIMD化
  - SSE(128bit)からAVX、AVX2(256bit)に至るSIMD幅拡大
  - “IntelはAVX-512をKnights Landingで採用し、後続のXeonにも展開”

http://software.intel.com/en-us/blogs/2013/07/10/avx-512-instructions

SPARC64 IXfxコアをSIMD拡張したときの特性





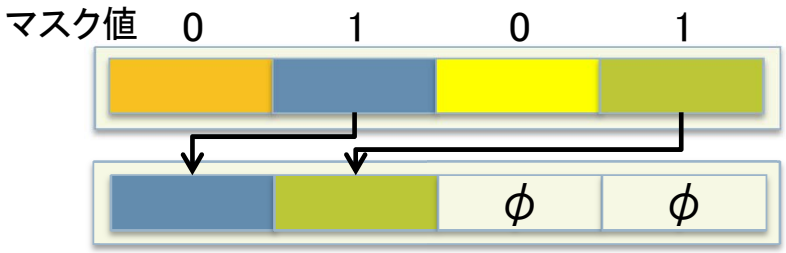


Copyright 2013 FUJITSU LIMITED

次期FX ― プロセッサのチャレンジ

FUJITSU

- L1キャッシュ強化
- マイクロアーキ改善
  - Out-of-order資源強化
  - 分岐予測改善
    - ローカル履歴ベース予測 ⇒ グローバル履歴を併用したハイブリッド予測
- SIMD強化
- その他の強化
  - IF文対策
    - マスクに加えて、低真率向けに
      リストアクセス・圧縮命令サポート



圧縮命令の動作

Copyright 2013 FUJITSU LIMITED

次期FX ー コンパイラの強化(1)

FUJITSU

- ソフトウェアパイプライニング(SWPL)の更なる徹底
  - 128個の浮動小数点レジスタを有効活用
  - アンローリングとの連携強化
  - SWPLが効かない少回転ループの最適化強化

- 並列化技術の強化
  - ループ並列化とSIMD化に共通する技術
  - FX10で自動ベクトル化と同等、次期FXで自動ベクトル化を超える

ANL BMT (135ループ)	VPP	FX10	次期FX
並列化ループ数	85	85	102

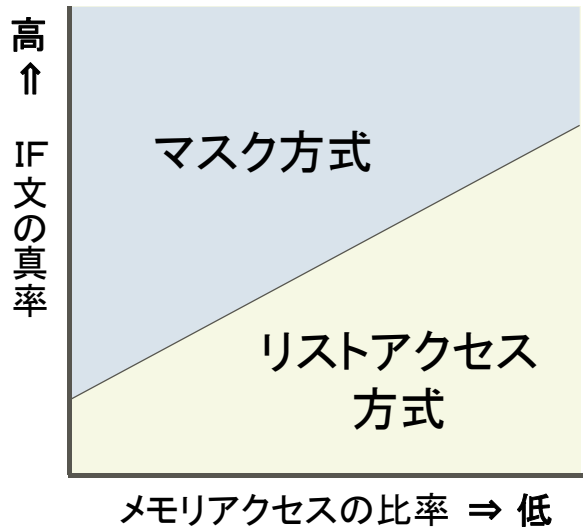
- 標準オプション指定時の最適化機能の拡大
  - 連続アクセス以外のプリフェッチ生成、IFのマスク化
  - 多彩なプリフェッチ生成アルゴリズムはオプションで提供

Copyright 2013 FUJITSU LIMITED

コンパイラの強化(2): リストアクセス方式

FUJITSU

- IFのストレート化でマスクに加えてリストアクセスをサポート
- 真率が低いIFもストレート化
  - ベクトル化技術を元に、真率、アクセス頻度、IF配下命令数等から適切な方式を選択
- 気象コード(IFSのsrtm_reftra)
  - 放射過程の計算(雲の有無によるIF)
  - リストアクセス方式で1.7倍に高速化




```
do i=1,N
  if (A(i)>0) then
    B(i)=C(i)+...
  endif
enddo
```

リストアクセス化

```
do i=1,N
  if (A(i)>0) then
    L(j)=I
    j=j+1
  endif
enddo
```

前処理: マスクを生成し、圧縮命令で真値のリストを作成

```
do i=1,j
  B(L(i))=C(L(i))+...
enddo
```

本処理: 真値のリストでリストアクセス

Copyright 2013 FUJITSU LIMITED

コンパイラの強化(3): C/C++コンパイラの最適化

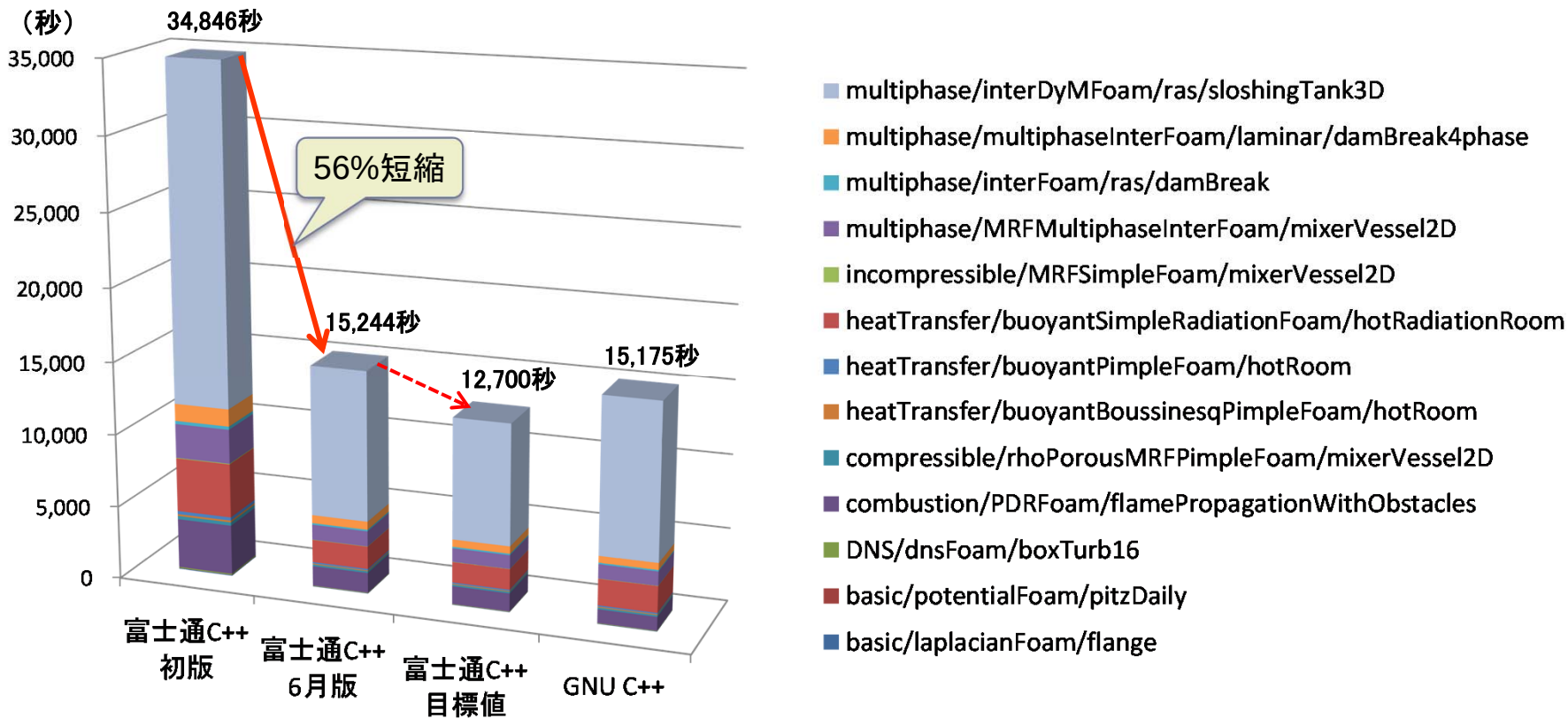
FUJITSU

- 次期FXに向け強化を実施中
  - C++向け最適化を全面的に見直し、リニューアル
  - 目標:GNU C++の2割アップ
- GNU C++同等性能を達成、6月末に「京」、FX10に提供

C++ 技術	最適化機能	対応状況・予定		
		初版	6月版	1年後
クラス	コンストラクタ/デストラクタ分岐削除	×	△	○
例外処理	unwindライブラリ	○	○	○
自動ｲﾝﾗｲﾝ、ｲﾝﾗｲﾝ属性	ユーザ指示インライン展開	×	△	○
STL	SPARCアーキ対応	×	△	○
多重継承、テンプレート、オペレータ	インライン展開、再帰インライン展開、SSA最適化、ローカルスケジューリング	×	△	○
仮想関数、フレンド関数	特殊属性インライン展開	×	×	○
ポインタ、構造体	ポインタ解析、P.V.R	×	△	○

Copyright 2013 FUJITSU LIMITED

OpenFOAMの性能改善



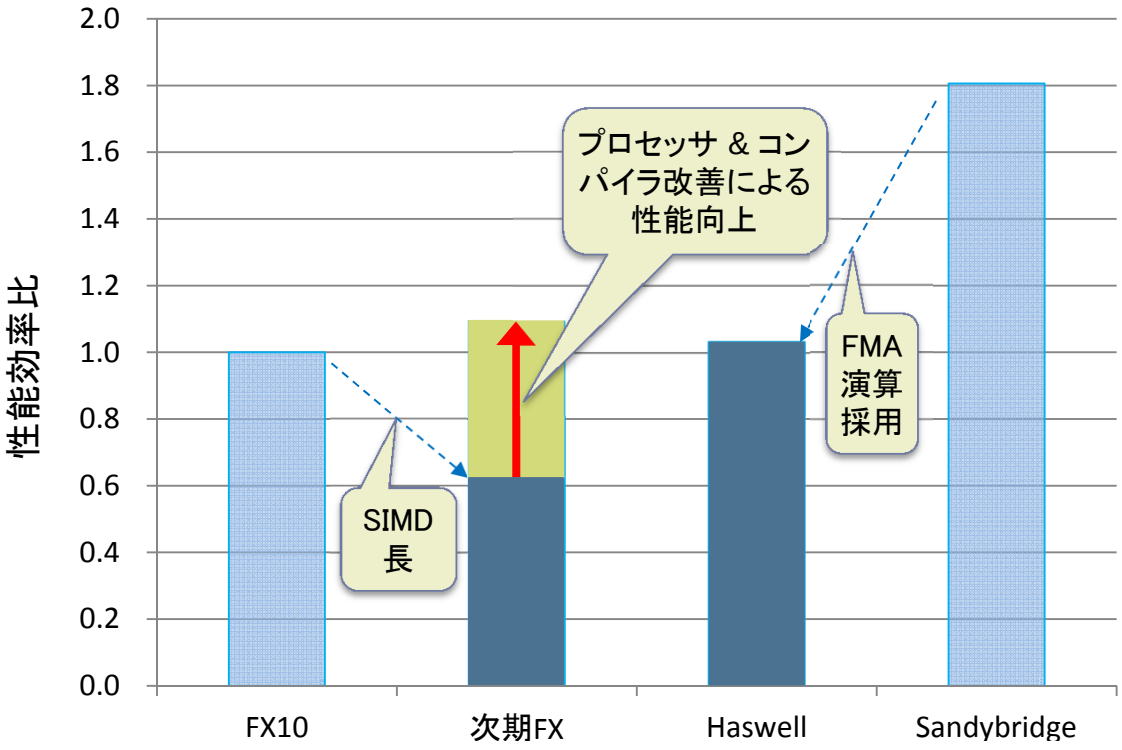
- OpenFOAM付属のテストプログラムから代表的な13本を測定

Copyright 2013 FUJITSU LIMITED

まとめ

FUJITSU

- 次期FXでは「京」・FX10の課題に真摯に対応する
- SPEC fpにおいて，最新のCPU中，ナンバーワンのコア性能効率を目指す



SPEC CFP2006性能効率比較

＊ HaswellはFMA ＆ 4SIMD
（Intel64 and IA-32 Architectures Optimization Reference Manual）
＊ Haswellの性能効率はPC版からの弊社推定

Copyright 2013 FUJITSU LIMITED

FUJITSU

shaping tomorrow with you