FUJITSU

shaping tomorrow with you

富士通のHPCに向けた取り組み

2015年8月28日

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
新庄　直樹

アウトライン

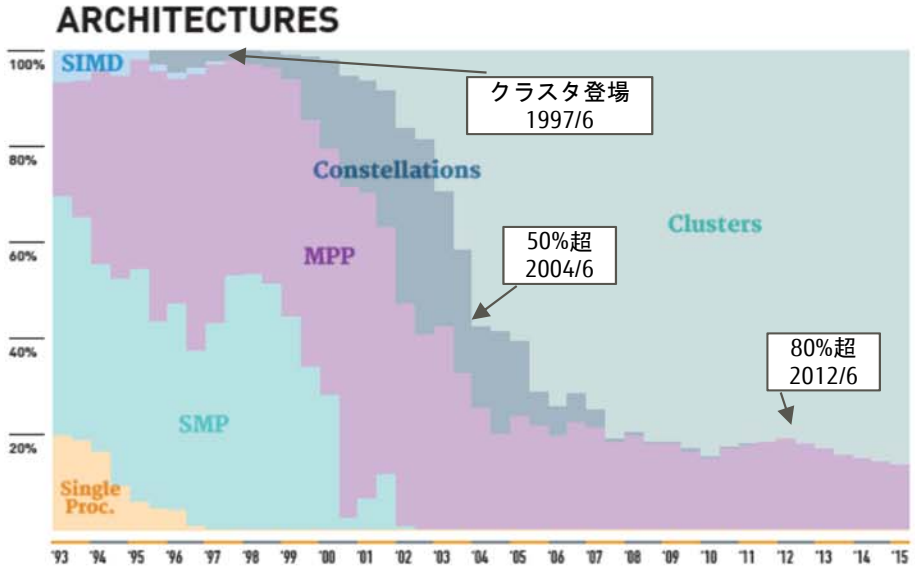
- HPCシステムの動向
- 富士通の取り組み
  - ハイエンドシステムPRIMEHPC FX100とポスト京への取り組み
  - エクサスケール時代を見据えてハード/ソフトからアプローチ
- PRIMEHPC FX100での評価とまとめ

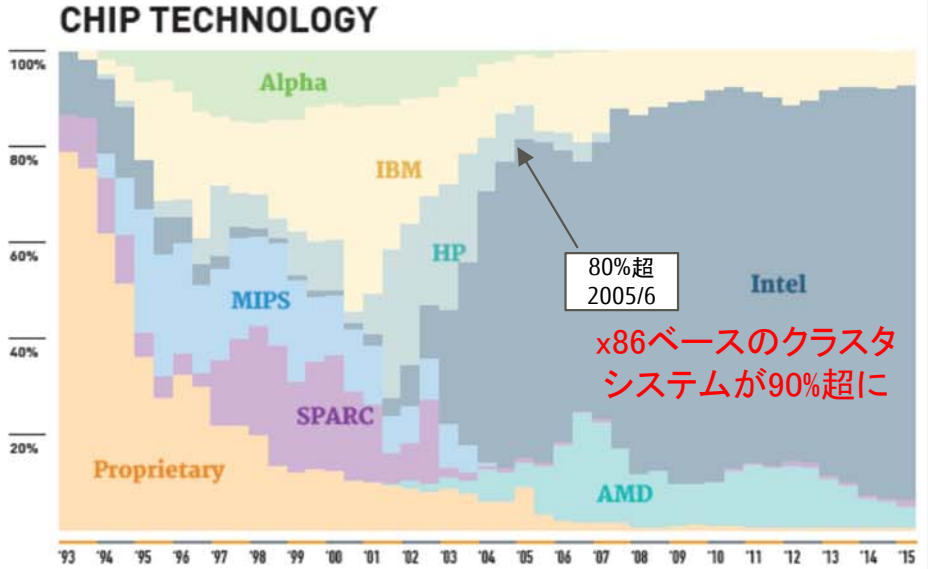
HPCシステムの動向(1/2)

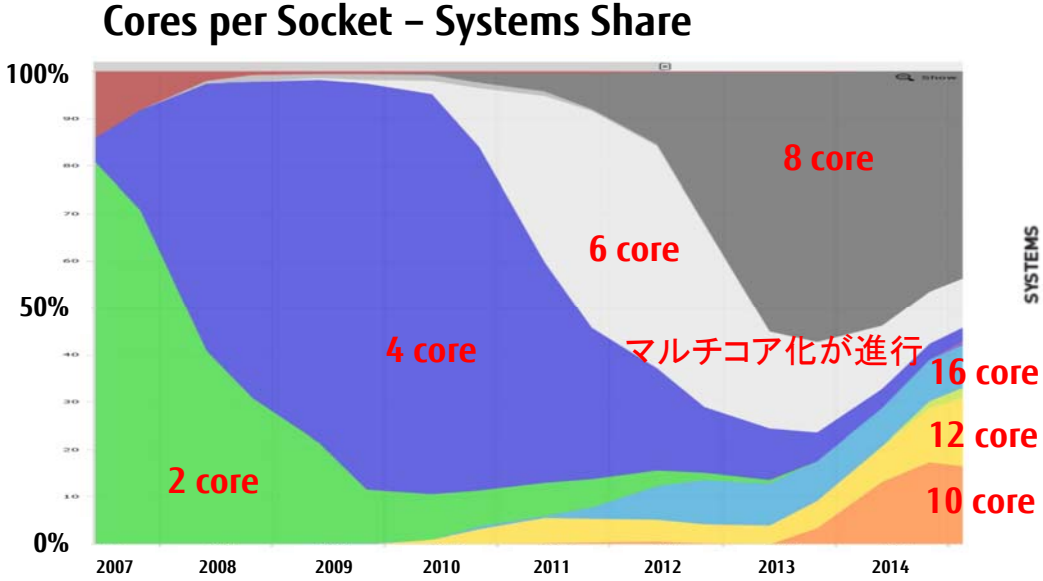
FUJITSU

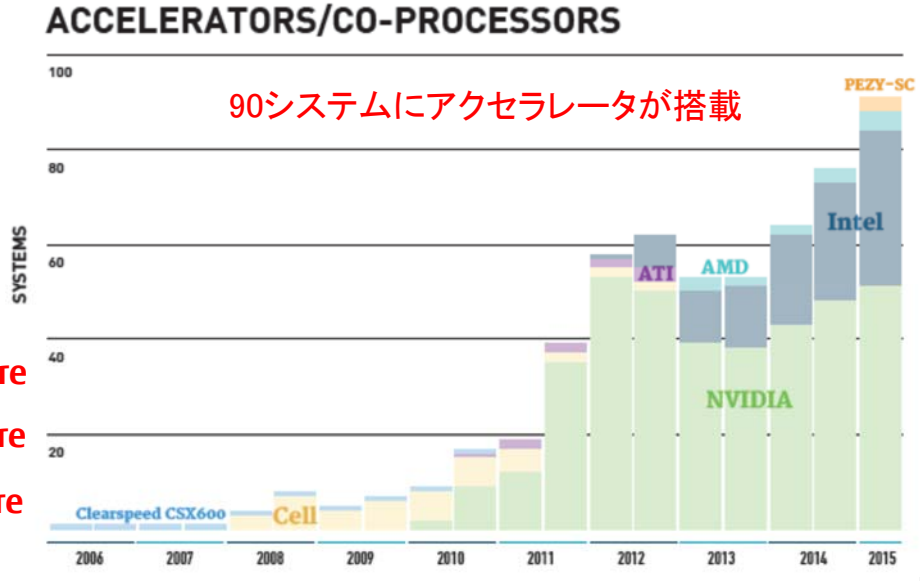
Top500に見るシステムの傾向-全体

- http://www.top500.org/lists/2015/06/download/TOP500_201506_Poster.pdf
- http://www.top500.org/statistics/overtime/









HPCシステムの動向(2/2)

FUJITSU

- Top500に見るシステムの傾向-トップ10
  - http://www.top500.org/lists/2015/06/download/TOP500_201506.xls
  - Rmax(HPL性能)シェアがトップ10で30%を占める
  - 4システム(40%)がアクセラレータを搭載(Top500全体では18%)
  - 9システムが専用インターコネクトを使用
- ハイエンドについては、特別な取り組みが必要

トップ10システムの概要 (2015年6月)

Rank	Name	Rmax	Rmax Share	Processor	Accelerator	Interconnect
1	Tianhe-2	33,862,700	9.3%	Intel Xeon E5-2692v2 12C 2.2GHz	Intel Xeon Phi 31S1P	TH Express-2
2	Titan	17,590,000	4.9%	Opteron 6274 16C 2.2GHz	NVIDIA K20x	Cray Gemini interconnect
3	Sequoia	17,173,224	4.7%	Power BQC 16C 1.6GHz	None	Custom Interconnect
4	K computer	10,510,000	2.9%	SPARC64 VIIIfx 8C 2GHz	None	Custom Interconnect
5	Mira	8,586,612	2.4%	Power BQC 16C 1.6GHz	None	Custom Interconnect
6	Piz Daint	6,271,000	1.7%	Xeon E5-2670 8C 2.6GHz	NVIDIA K20x	Aries interconnect
7	Shaheen II	5,536,990	1.5%	Xeon E5-2698v3 16C 2.3GHz	None	Aries interconnect
8	Stampede	5,168,110	1.4%	Xeon E5-2680 8C 2.7GHz	Intel Xeon Phi SE10P	InfiniBand FDR
9	JUQUEEN	5,008,857	1.4%	Power BQC 16C 1.6GHz	None	Custom Interconnect
10	Vulcan	4,293,306	1.2%	Power BQC 16C 1.6GHz	None	Custom Interconnect
Top10 total		114,000,799	31.4%	50%	40%	90%

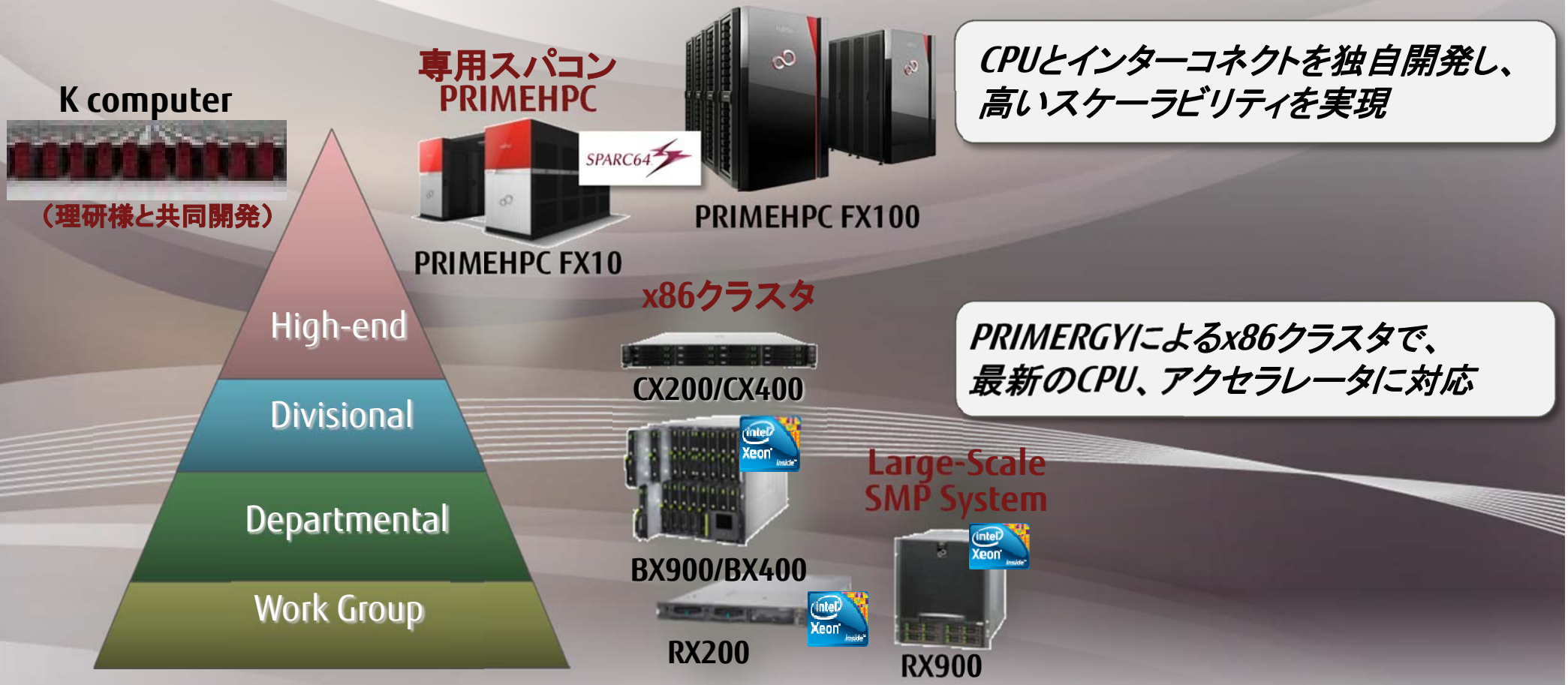
FUJITSU

ハイエンドシステムPRIMEHPC FX100とポスト京への取り組み

富士通の取り組み

FUJITSU

- お客様のニーズに合わせたHPCソリューションを提供
  - 独自CPU搭載の専用スパコンとx86クラスタシステムの両面サポート
  - シングルシステムイメージ運用を実現するシステムソフトの開発・提供
  - 高性能、高可用性、高信頼性の実現



システムソフト(TCS: Technical Computing Suite)

FUJITSU

HPCアプリケーションに最適化したLinux OS

・ラージページサポート、OSジッタ最適化

自社開発ソフトウェアとOpen Source Software

・自社開発：システム管理ソフトとコンパイラ
・OSSベース：ファイルシステム(FEFS)、MPI（コミュニティにフィードバック）

PRIMEHPCとx86クラスタとのシングルシステムイメージ運用

システムマネージメントポータルとHPCポータル

Technical Computing Suite(TCS)

Management
- System management
  - Single system image
  - Single action IPL
  - Fail safe capability
- Job management
  - Highly efficient scheduler

File system (FEFS)
- Lustre based
- Higher scalability (thousands of IO servers)
- Higher IO performance (1.4 TB/s)

Programing environment
- Compiler
  - Fortran, XPF, C, C++
  - Automatic parallelization
  - SIMD support
- MPI: Open MPI based
- Tools and math libraries

ロードマップ – エクサスケールへ

FUJITSU

「京」　：稼働中
PRIMEHPC FX10 : 稼働中

科学/技術分野の多数のアプリが開発され稼動している

PRIMEHPC FX100 :出荷中

「京」のアーキテクチャコンセプトを引き継ぎ、CPUとインターコネクトの性能を向上

エクサスケールに向けて

理研プロジェクトにて、ポスト「京」コンピュータの基本設計に参画

PRIMEHPC FX100の特長

FUJITSU

広範な実アプリで高性能を実現する独自開発CPU

高いスケーラビリティを持つインターコネクト

	FX100	FX10	K computer
Double Flops / CPU	Over 1 TF	235 GF	128 GF
Single Flops / CPU	Over 2 TF	235 GF	128 GF
Max. # of cores	32	16	8
Memory / CPU	32 GB	32 GB/64 GB	16 GB
SIMD width	256 bit	128 bit	128 bit
Byte per flop	0.4 ~ 0.5		
Interconnect	Tofu 6D mesh/torus		
Interconnect BW	12.5 GB/s	5 GB/s	5 GB/s

「京」及びPRIMEHPC FX10との互換性

バイナリコンパチビリティによりアプリ資産の容易な移行を実現
リコンパイルとライブラリにより性能改善・新機能が享受可能

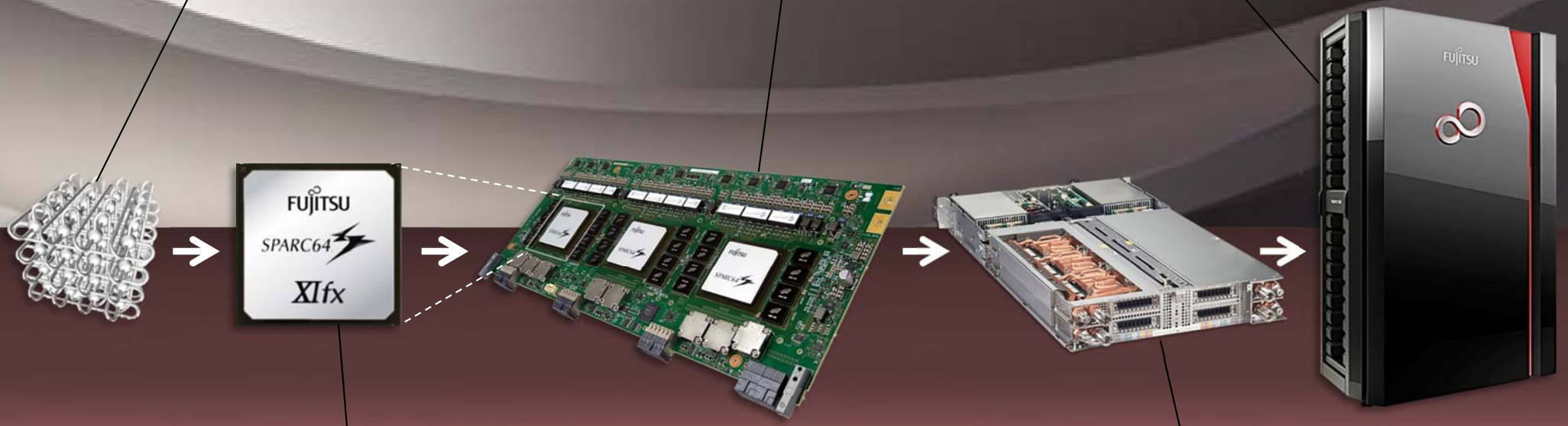
PRIMEHPC FX100の構成と機能

FUJITSU

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

CPU Memory Board
- Three CPUs
- 3 x 8 Micron's HMCs
- 8 opt modules, for inter-chassis connections

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



Fujitsu designed SPARC64 XIfx
- 1TF~(DP)/2TF~(SP)
- 32 + 2 core CPU
- HPC-ACE2 support
- Tofu2 integrated

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

FUJITSU

エクサスケール時代を見据えて ハードウェアからアプローチ

スパコン用CPUのトレンドと富士通の方針

FUJITSU

■ポストムーア時代に向かうトレンド

- ポスト京の時代以降、半導体プロセスの微細化は限界に近づく
- その後の性能向上は3次元スタックに向かう（あるいは新デバイス？）
- いずれにせよトランジスタは今後も増加、メニーコア化のトレンドは継続
- スパコン用メニーコアCPU開発で想定されるアプローチは2通り
  - ① 一定の性能を有する、ある程度の大きさのコアを並べる
  - ② 徹底的に軽量化した小さなコアを大量に並べる

■ 富士通の取り組み

- 広範なアプリケーションが動作するプラットフォームとして社会に貢献するため、既存システムに対して継続性があり、汎用性の高い①を選択
  ②のアプローチでは、汎用性に限界があり、十分に使命を果たせない
- その上で、以下の開発を目標とする
  - ✓ テクノロジトレンドに合致する適切な面積、性能、電力を備えたコア
  - ✓ コア数に応じてスケーラブルな性能を持つメニーコアチップ

ポスト京のCPUで目指したいこと

FUJITSU

■ ポストムーア時代まで通用する、*スケーラブルメニーコア技術*の確立

- スケーラブルメニーコア技術を支える3つの柱:
  - 計算コア
  - アシスタントコア
  - コアメモリグループ(CMG)

	単体コア性能	電力性能	汎用性	高いスケーラビリティの実現性
Xeon	◎	×	◎	×(メニーコア化に限界)
GPGPU	△	○	△	×(プログラミングモデルが未成熟)
ポスト京CPU	○	△※1	◎	◎(京のスケーラビリティ※2を継承 ＋スケーラブルメニーコア技術)

- 単体コア性能: アプリケーションを高性能で実行できること
- 電力性能: 性能当たりの電力がreasonableであること
- 汎用性: 多様なアプリケーションに対応できること
- スケーラビリティ: メニーコア化に伴い、性能がスケーラブルに向上すること

※1: 電力制御などの技術開発により、他の長所を損なわない工夫で実効電力の低減を図る
※2: VISIMPACT、Tofuバリアなどの独自技術

スケーラブルメニーコア技術を支える三本の柱

FUJITSU

■計算コア

- 多様なアプリケーションを高性能に実行
  - ✓ OOO機能を備えた汎用スーパースカラ計算エンジン
- メニーコア化が容易な、reasonableな面積で実現
  - ✓ 面積当たり性能でXeonを凌駕

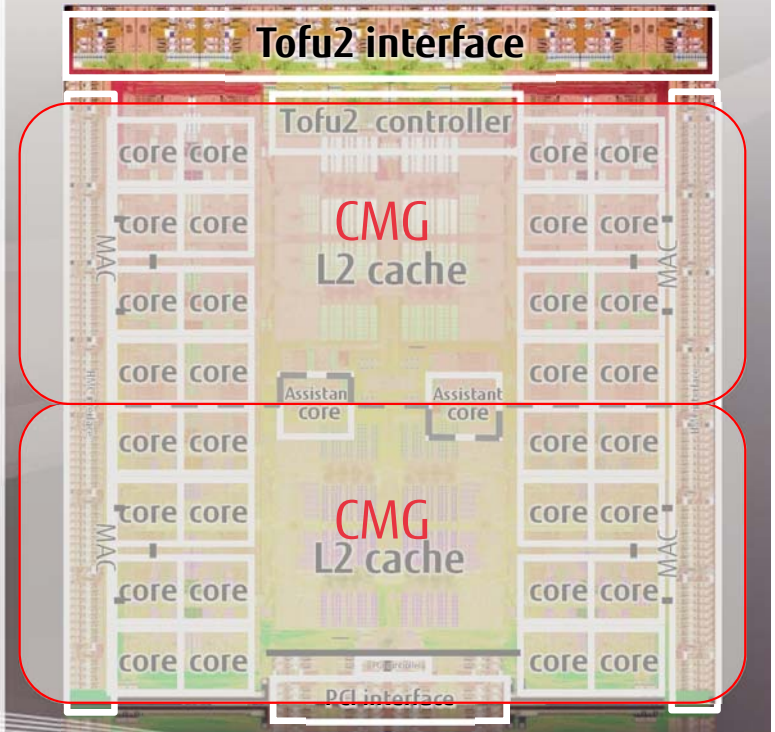
■アシスタントコア

- 計算コアを、OS等のアプリ実行以外のオーバヘッドから解放

■コアメモリグループ (CMG)

- L2キャッシュを共有するコアグループ
- CMG毎に直結されたメモリを高BW、低レイテンシでアクセス
- CMG間はディレクトリによるコヒーレント管理
  - ✓ CMG増に伴うハードウェアオーバヘッドを抑え、スケーラビリティを確保

FX100 CPUから導入済



FUJITSU

エクサスケール時代を見据えて システムソフトからのアプローチ

エクサスケール向けて6つの取り組み

FUJITSU

- 性能
  - ハイエンドシステムにふさわしい単体性能とシステム性能の達成に取り組む
  - 様々な観点での性能向上、処理時間短縮にこだわる
- ★省リソース　　★追加分
  - アシスタントコアでのIO処理(FX100)：IOノード削減
  - 省電力、省メモリ、省時間、省スペース
- エンドユーザの使い勝手・継続性
  - オープンソース・市販ツール対応の拡大
  - 既存環境・ユーザ資産継承
- ★柔軟性
  - 計算科学ユーザに加え、計算機科学・データ処理ユーザへの対応
- 信頼性
  - 安定動作、即時故障検出、短時間復旧によるサービス停止時間の最小化
- 保守性
  - 運用中のシステムアップデート・調査資料取得実現による保守時間の最小化

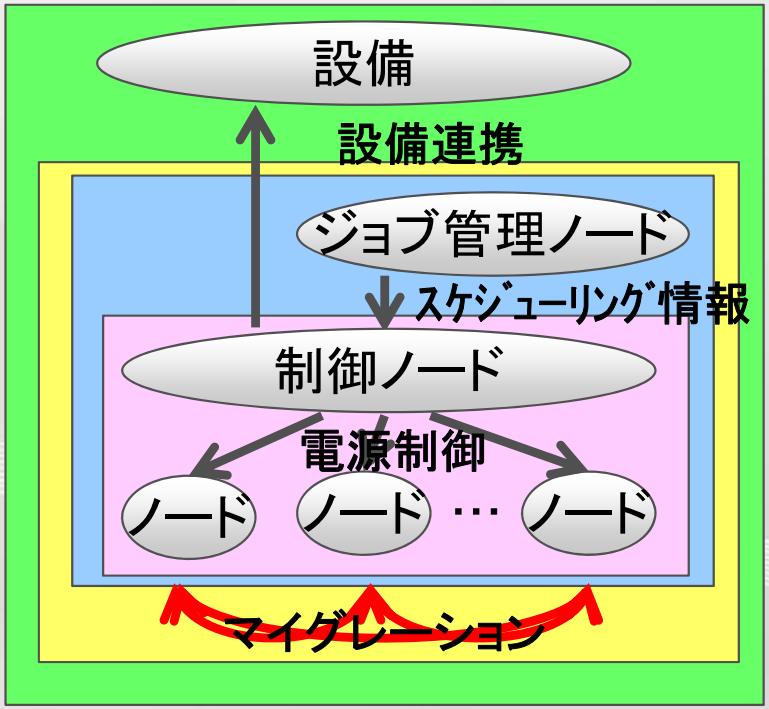
省電力課題への取り組み

FUJITSU

要素技術の段階的提供により進化を図る

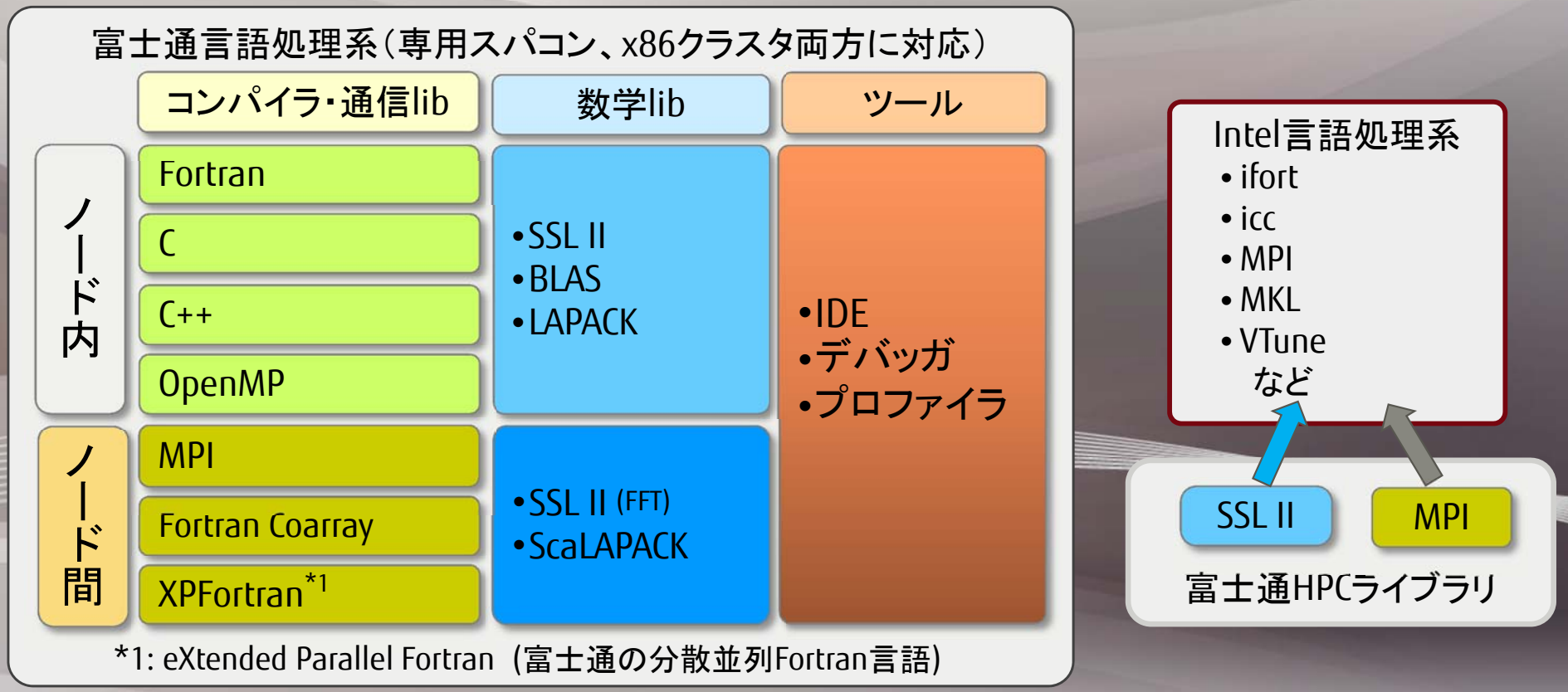
STEP1	STEP2	STEP3	STEP4
計算ノード電力制御	スケジューラ連携	きめ細かな節電	センタ全体の電力管理
（電力制御APIの提供）	（スケジューラ連携API）	（ジョブの移動による電力効率改善）	（設備連携し電力を制御）

- **計算ノード電力制御**
  - ー電源制御/パワーキャップインターフェイス
- **スケジューラ連携**
  - ースケジューラ連携のためのインターフェイス
  - ー使用電力/ノード稼動状況の可視化
  - ー未使用ノードの省電力化、パワーキャップ制御
- **きめ細かな節電**
  - ーマイグレーションと組み合わせたジョブの局所化
  - ージョブの片寄せによる空調、未使用インターコネクトの停止
- **センタ全体の電力管理**
  - ー設備連携のためのインターフェイス
  - ーシステム稼動状況と連動した空調設備などの制御



言語処理系での取り組み

- FX100・ポスト京向けに強化した規格・機能をx86向けにもタイムリーに移植
  - 新規格サポート、最新x86クラスタへも対応
- 京・ポスト京との親和性と既存のソース資産の移行性を確保
- 富士通の強みを活かすべく、数学libとMPIはIntel言語処理系にも対応



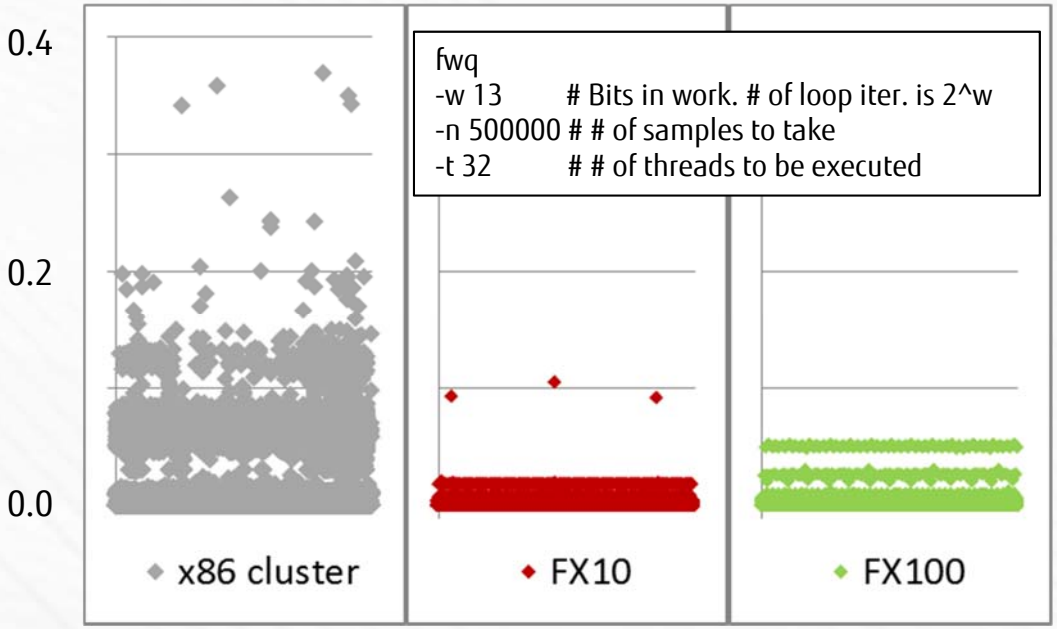
アシスタントコアによるOSジッタの低減

FUJITSU

デーモン、IO処理等をアシスタントコアで実行することでOSジッタを大幅に低減

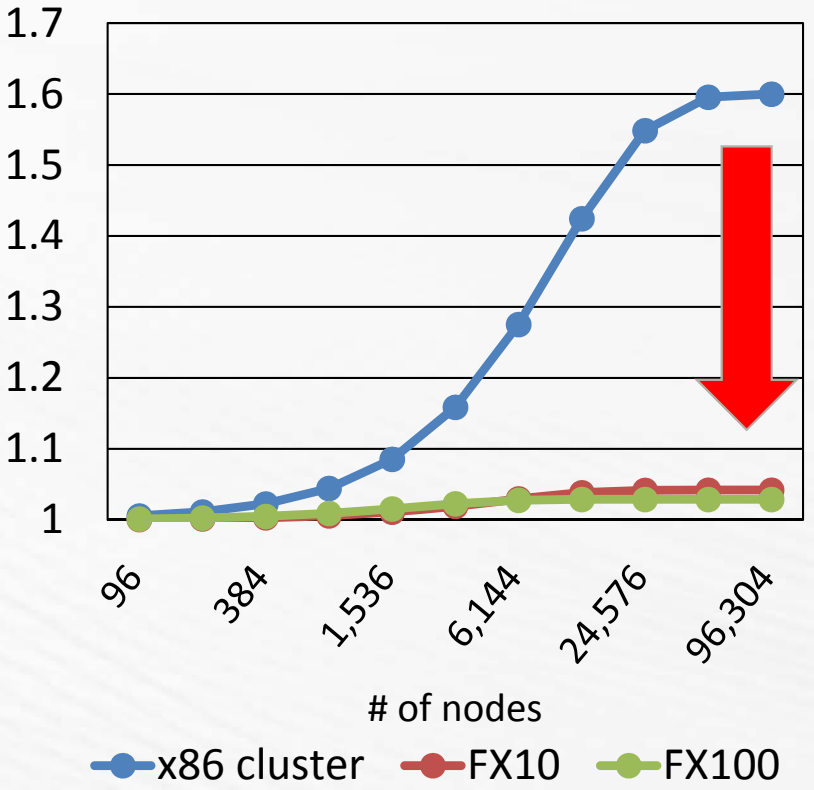
OSジッタによる計算時間のばらつき評価

x86はより大きいノイズあり



表示区間150秒

OSジッタ低減によるスケーラビリティ向上見積もり(comm. interval=1ms)

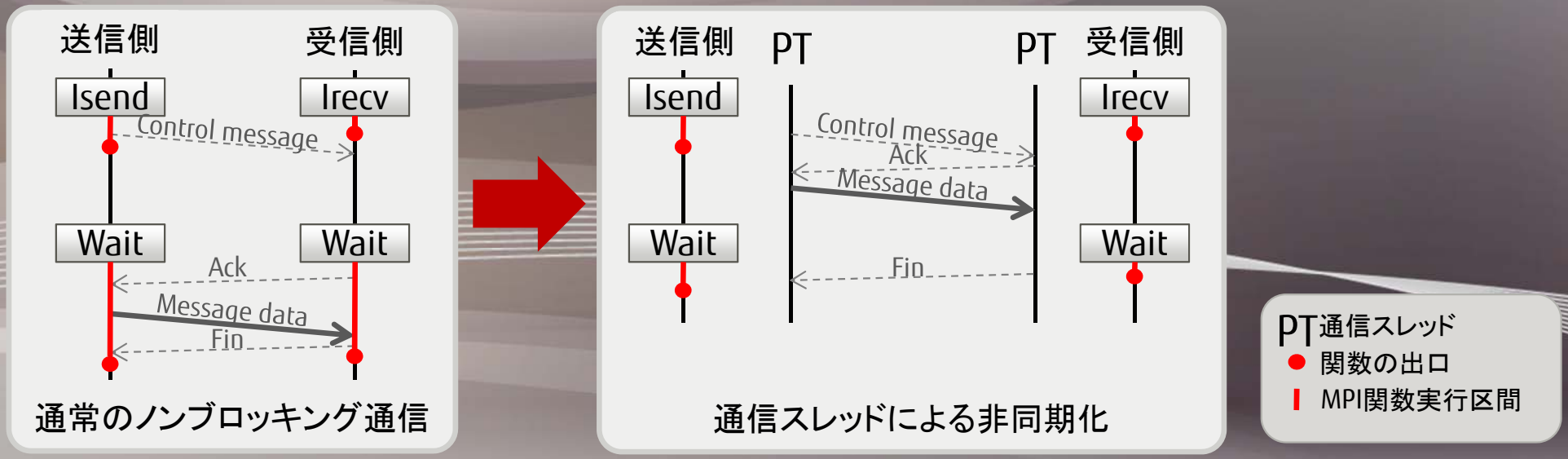


MPIライブラリの強化

FUJITSU

■ Open MPIをベースに強化・最適化を推進

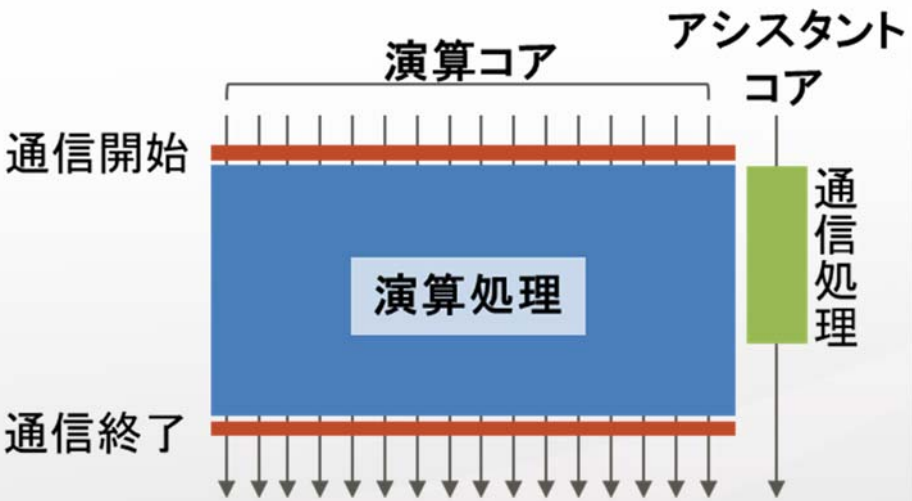
システム (対象インターコネクト)	PRIMEHPC向け (Tofu)	X86クラスタ (InfiniBand)
非同期通信の計算処理とのオーバラップ	アシスタントコアを活用した通信スレッド	通信スレッド (対応予定)
中・長メッセージ向け集団通信	複数DMAエンジン 多次元軸活用	フルRDMA化 (対応予定)
短メッセージ向け集団通信 (Barrier, Bcast, Reduce, Allreduce)	Tofuバリア活用	-



アシスタントコアによる通信と計算のオーバラップ

FUJITSU

- 動作イメージ
  アシスタントコアがノンブロッキング通信処理の大半を実行
- 利用方法
  - 実行時オプションでprogress threadモードを選択
  - さらに高速化を行うためには、オーバラップ通信の対象区間をユーザーが明に指定
- 特長
  - コード書き換えなしにオーバラップ通信が可能
  - 簡易な区間指定でスレッド排他制御オーバヘッドも抑制可能



対象区間を明示することでクリティカルセクションを限定

```
MPI_IRECV(...);
FJMPI_Progress_start();
calc(...);
FJMPI_Progress_stop();
MPI_WAITALL(...);
```

オーバラップ通信の対象区間

FUJITSU

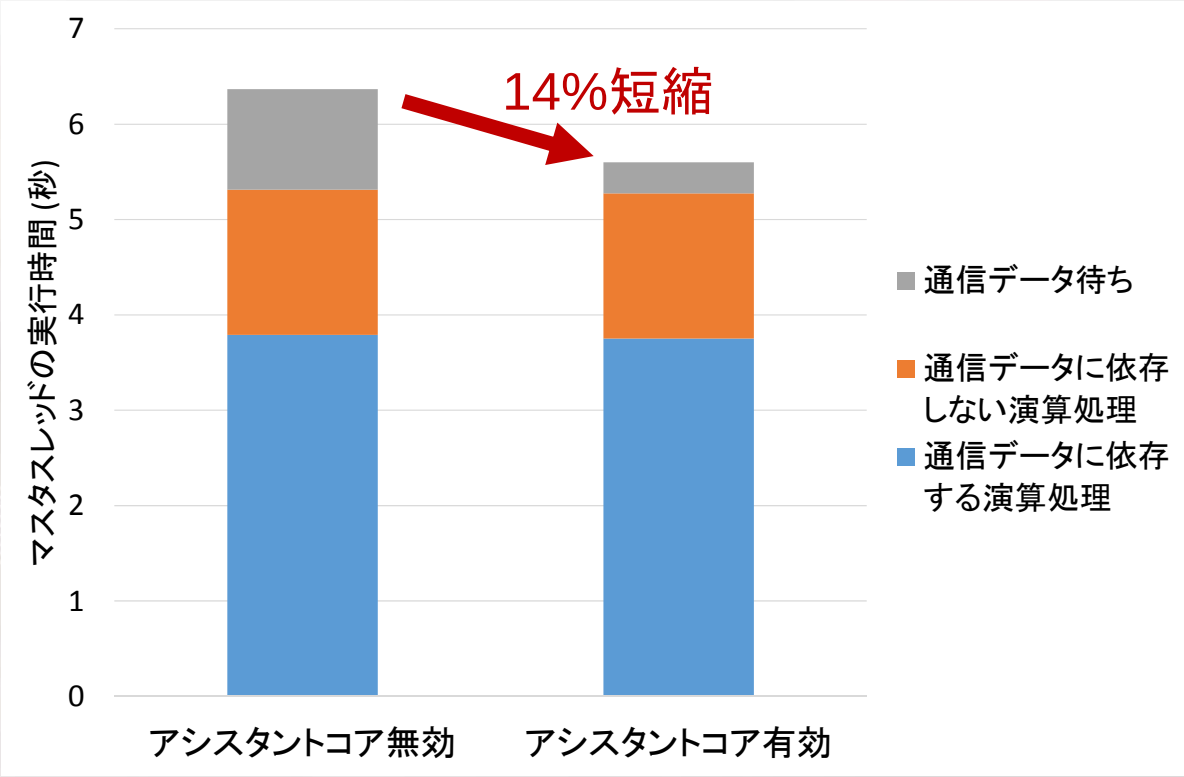
FX100での評価とまとめ

アシスタントコアによる通信と演算のオーバラップ (GT5D)

- 通信データに依存しない演算処理を、MPI_Waitallの前で実行
  ⇒演算中にアシスタントコアが通信を制御
- 性能向上と可搬性/保守性を両立
  - 従来はOpenMPで通信と演算のオーバラップを実装して強スケーリングを達成していたが、アシスタントコアの利用によって特殊な実装なしでオーバラップが可能

通信/演算オーバラップのコード例

```
   MPI_Isend
   MPI_Irecv
!$OMP PARALLEL DO
   do i=…
      通信データに依存しない演算処理
   enddo
!$OMP END PARALLEL DO
   MPI_Waitall
   do i=…
      通信データに依存する演算処理
   enddo
```

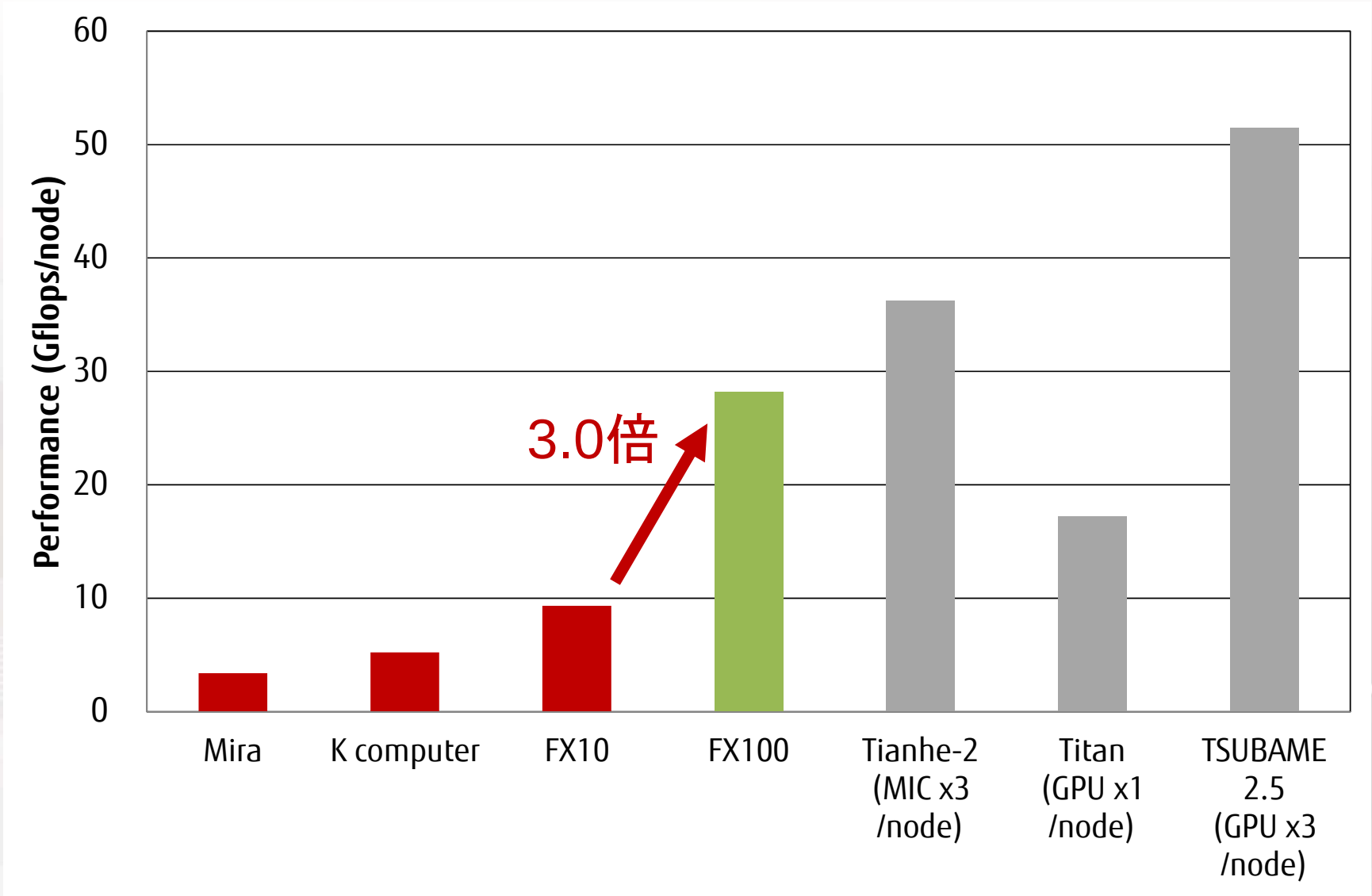


問題サイズ:256×256×64×128、並列数:16スレッド×64プロセス、評価対象区間:l4dx、通信促進:自動区間モード

HPCGベンチマークによる評価

FUJITSU

- メモリスループットの向上により、FX10の3倍のノードあたり性能
  - 汎用プロセッサの使いやすさを維持しつつ、性能を向上



HPCGベンチマークへの取り組み

FUJITSU

HPCGの主要演算

- 疎行列 A とベクトル v の積

$$\begin{pmatrix} 1.0 & 2.0 & & \\ & 3.0 & 4.0 & \\ 5.0 & 6.0 & 7.0 & \\ & 8.0 & 9.0 & 10.0 \end{pmatrix} \begin{pmatrix} v_0 \\ v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 1.0v_0 + 2.0v_2 \\ \cdots \\ \cdots \\ \cdots \end{pmatrix}$$

(A, v)

A の非零要素に対応した v の要素との積和演算

A の圧縮格納形式: 長SIMDアーキ向きの**Sliced-ELL**を採用

格納形式の比較

CRS

1.0	2.0	
3.0	4.0	
5.0	6.0	7.0
8.0	9.0	10.0

vs

ELL

1.0	2.0	**0.0**
3.0	4.0	**0.0**
5.0	6.0	7.0
8.0	9.0	10.0

vs

Sliced-ELL

1.0	3.0	5.0	8.0
2.0	4.0	6.0	9.0
0.0	**0.0**	7.0	10.0

連続アクセス方向

行単位で圧縮
（例）HPCG
リファレンスコード

非零要素数／行を揃える
最内ループ長が定数となり、
ループ最適化が容易となる
（例）K（理研）

ELLをn行単位でスライス
＆転置格納（この例は4行）
効率的なSIMD命令生成が容易

今後の課題

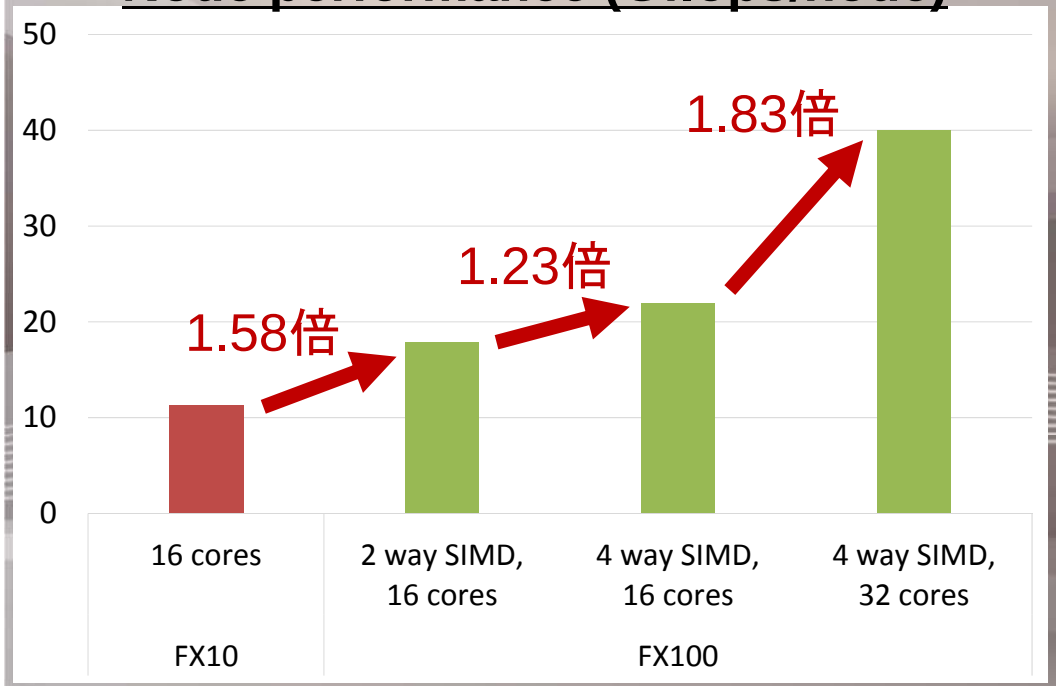
- 演算性能の向上 ⇒ 前処理と疎行列ベクトル積のループ融合
- 通信コスト削減 ⇒ MPI_Allreduceと計算とのオーバラップ

4way SIMD、32演算コアの効果 (NPB-FT)

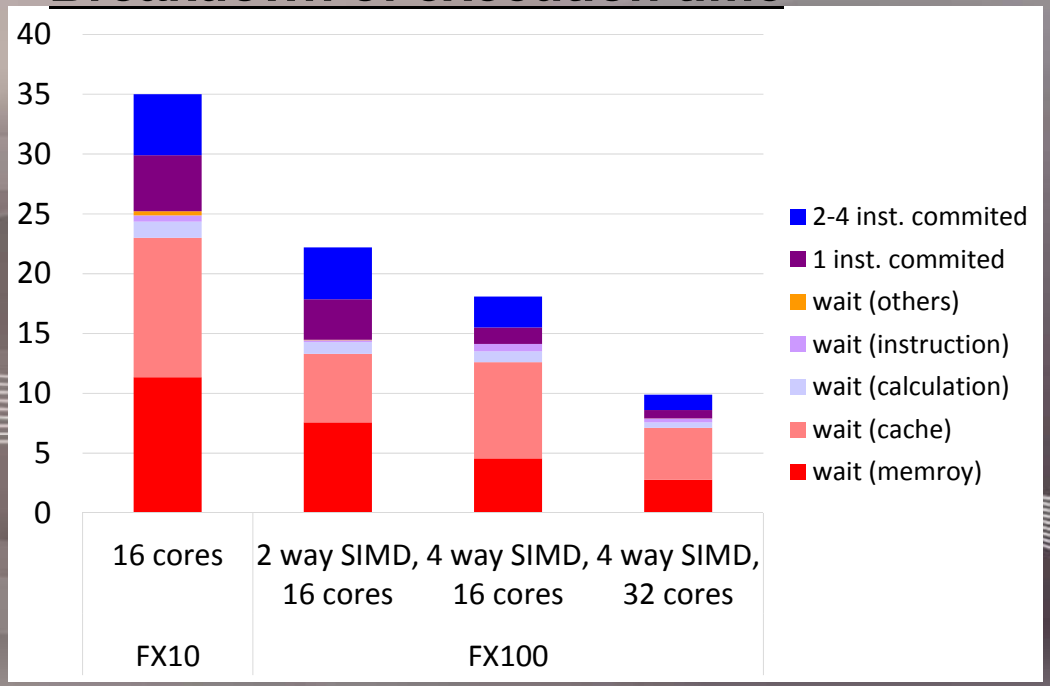
FUJITSU

- FFT演算カーネルによる評価
  - FX10用2way SIMDバイナリを実行すると、1.58倍高速化
    - メモリ&キャッシュスループット向上による効果
  - 4way SIMDバイナリに再翻訳すると、1.23倍高速化
    - 実行命令数43%削減による効果
  - 2倍の演算コアを使うと、1.83倍高速化
    - 並列化効率91%の良好なスケーラビリティ

Node performance (Gflops/node)



Breakdown of execution time



使用コード: NAS Parallel Benchmarks Ver. 3.3.1 OpenMP版 クラスC

セクタキャッシュ機能の効果 (CCS QCD)

FUJITSU

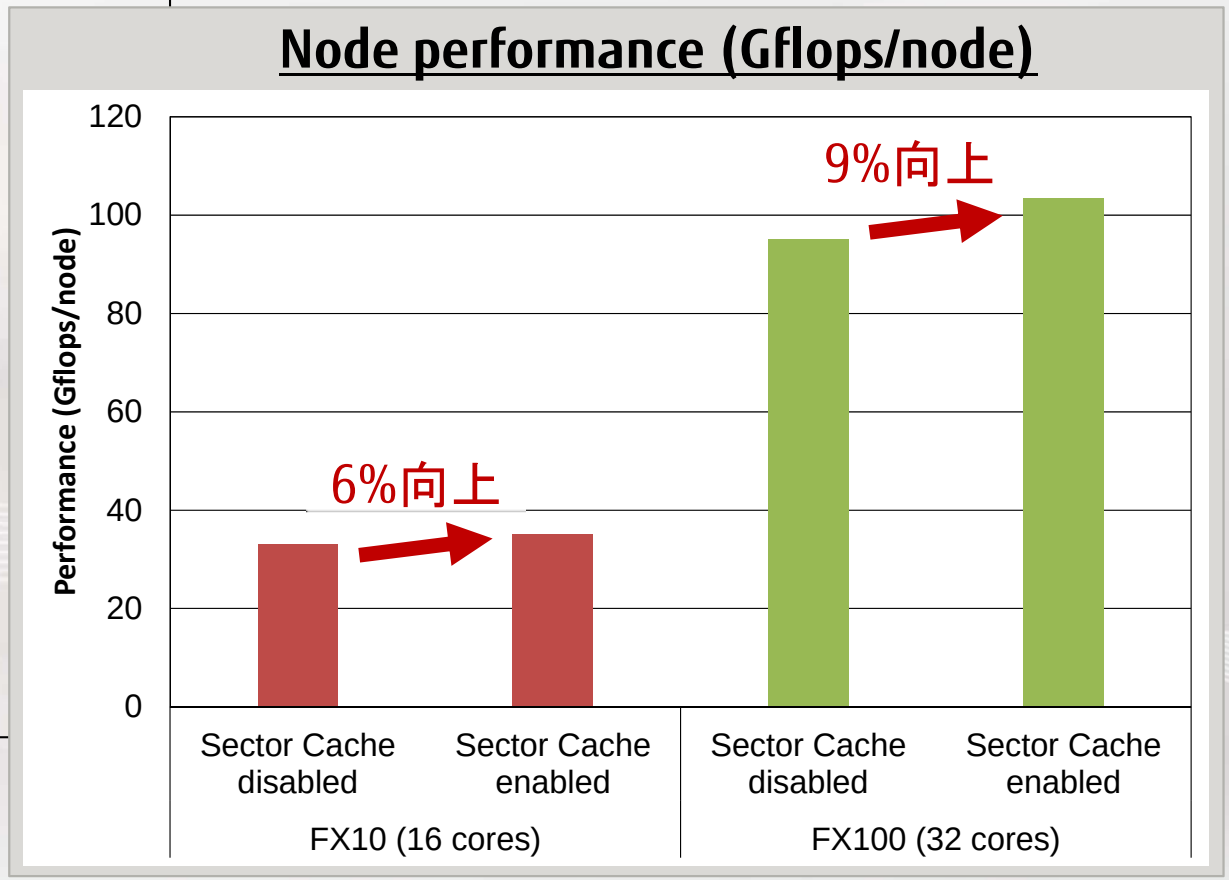
- 高いメモリスループットを、無駄なくさらに活用
  - セクタキャッシュを用いて再利用するデータをL2$に維持

セクタキャッシュ利用のコード例

```
!OCL CACHE_SECTOR_SIZE(19,5)
!OCL CACHE_SUBSECTOR_ASSIGN(ue,uo,yde,fclinve)
!$OMP PARALLEL DO SCHEDULE(STATIC,1)
 do ix=1,NX
  do iy=1,NY
   do iz=1,NZ
    …
    gy11=yo(…,iy+1,…)+…
    …
    gy11=yo(…,iy-1,…)+…
    …
   enddo
  enddo
 enddo
!$OMP END PARALLEL DO
!OCL END_CACHE_SUBSECTOR
!OCL END_CACHE_SECTOR_SIZE
```

セクタ1にL2$ 2.5MBを確保

再利用しない配列をセクタ1に割り当て
（再利用するデータはセクタ0に入る）

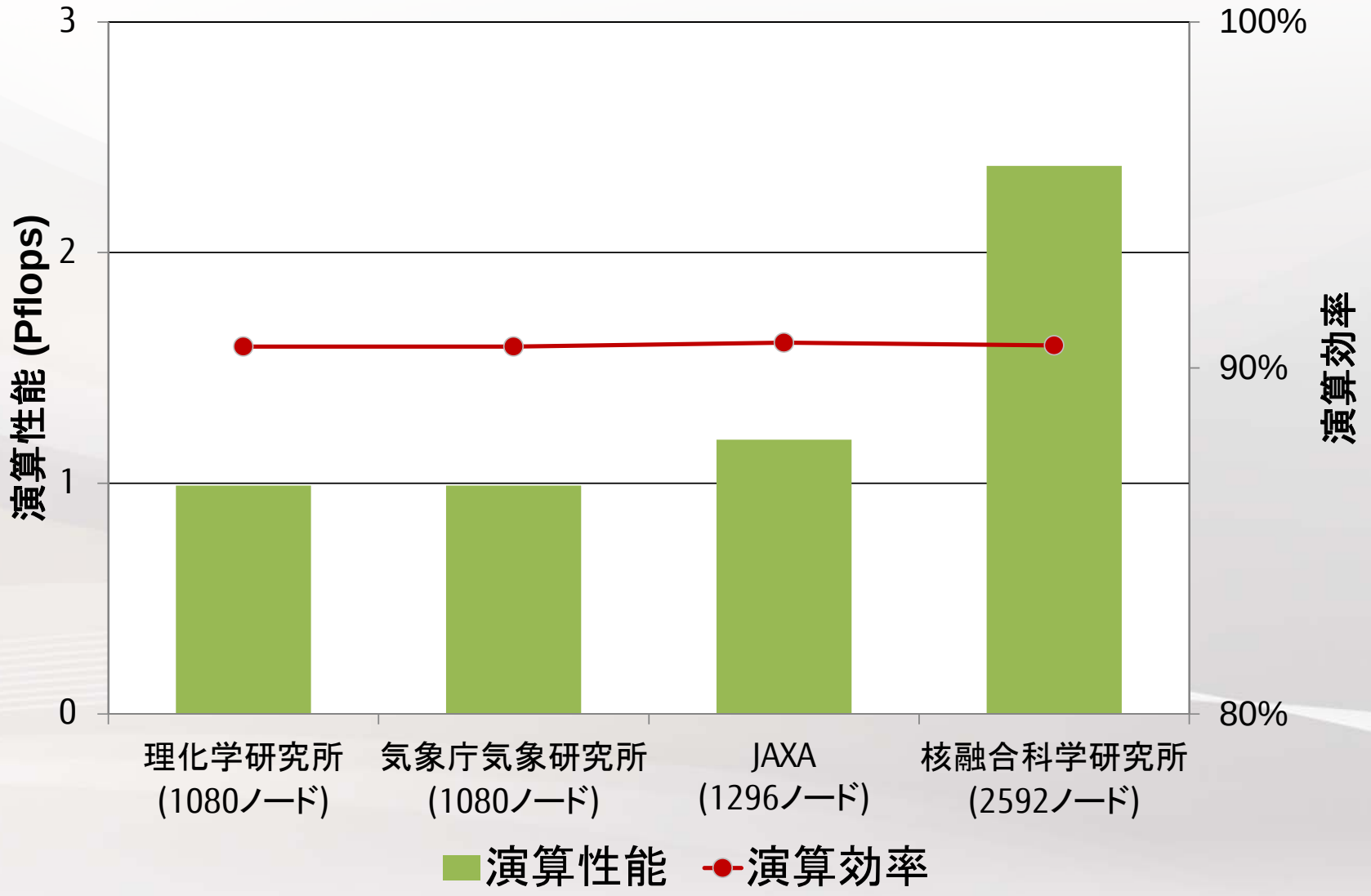


使用コード: CCS QCD Miniapp、問題サイズ: 32^4
https://github.com/fiber-miniapp/ccs-qcd

FX100導入サイトとLINPACK性能

FUJITSU

- 京、FX10に引き続き、90%超の演算効率を実現
  - ノードあたり約1Tflopsの演算性能



京、FX10、FX100、ポスト京、エクサスケール…

FUJITSU

エクサスケールを見据えてポスト京の基本設計中

- アプリケーションの高効率実行が鍵
- 「京」、FX10、FX100のアーキテクチャを継承、革新

ポスト京

PRIMEHPCシリーズ

© RIKEN

K computer
VISIMPACT
SIMD extension HPC-ACE
Direct network Tofu

CY2010~
128GF, 8-core/CPU

FX10
VISIMPACT
HPC-ACE
Direct network Tofu

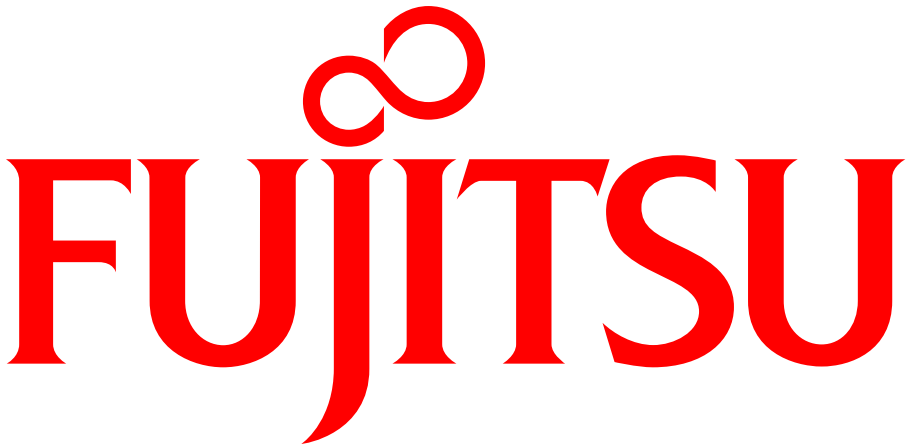
CY2012~
236.5GF, 16-core/CPU

FX100
VISIMPACT
HPC-ACE2
Tofu interconnect 2
HMC & Optical connections

CY2015~
1TF~, 32-core/CPU

FUJITSU

日本を代表するIT企業として、
国家プロジェクトに貢献するとともに、
お客様のニーズに応えるHPC環境を
提供していきます。



shaping tomorrow with you