PEZY-SC3: 高い電力効率を実現する MIMD メニーコアプロセッサ

初田 直也[1,a)]　角田 俊太郎[1]　内田 広平[1]　石谷 太一[1]　塩谷 亮太[2]　石井 敬[1]

概要：PEZY-SC3 は我々が開発した高い電力効率と面積効率を持つスーパーコンピュータ向けプロセッサであり，TSMC 7nm プロセス技術を用いて製造されている．PEZY-SC3 は高いスレッドレベル並列性を含むアプリケーションを対象としており，それらにおいて高い効率を実現するために MIMD メニーコアアーキテクチャ，細粒度マルチスレッディング，ノンコヒーレントキャッシュなどの要素を採用している．PEZY-SC3 は MIMD メニーコアアーキテクチャの採用により各コアが独立して動作するため，機能が限定された特殊なテンソルユニットや Wide-SIMD を採用した既存のプロセッサと比較して，高いプログラマビリティを持ちながら高電力効率を実現している．また，PEZY-SC3 の各コアはアウトオブオーダ実行や投機実行のような高コストな技術を一切導入せず，シンプルなパイプラインにより高電力効率と高スループットを両立している．さらに，独自のノンコヒーレントで階層的なキャッシュシステムにより，プログラマビリティを損なうことなくメニーコアにおける高いスケーラビリティを実現している．PEZY-SC3 を搭載したシステムの電力効率は 21.892 GFlops/W であり，スーパーコンピュータの電力効率を測定する Green500（2023 年 6 月）において 39 位となった．本論文ではこの PEZY-SC3 のアーキテクチャの概要と設計について説明する．

1. はじめに

PEZY-SC3 は TSMC 7nm プロセスで製造されたスーパーコンピュータ向けプロセッサであり，高い電力効率と面積効率を持つ．PEZY-SC3 は，PEZY Computing により開発された PEZY-SCx シリーズの 3 世代目のプロセッサである．PEZY-SCx シリーズはこれまでにさまざまな研究機関に展開され，大規模な科学技術計算に利用されてきた [1], [2], [3], [4], [5].

PEZY-SCx シリーズは高いスレッドレベル並列性を持つアプリケーションをターゲットとしたプロセッサである．PEZY-SC3 を含む PEZY-SCx シリーズの各世代に共通した特徴として，MIMD メニーコアアーキテクチャ，細粒度マルチスレッディング，ノンコヒーレントキャッシュなどの要素の採用がある．PEZY-SCx シリーズでは MIMD メニーコアアーキテクチャの採用により，特定計算向けに特化したテンソルユニットや Wide-SIMD [6] などを採用したプロセッサと比較して，より柔軟なプログラミングモデルを提供している．また，細粒度マルチスレッディングによる単純なパイプラインにより，アウトオブオーダー実行や分岐予測といった複雑で高コストな機構を用いずに高いスループットを達成している．さらに，ノンコヒーレントな階層キャッシュシステムにより，プログラマビリティを損なうことなくコア数をスケールさせている．

PEZY-SC3 を搭載したスーパーコンピュータシステムにおいて LINPACK ベンチマーク [7] を実行した際の電力効率は 21.892 GFlops/W であり，スーパーコンピュータの電力効率を測定している Green500 [8]（2023 年 6 月）において 39 位にランクされている．この Green500 の上位にランクされている各システムでは一般にアクセラレータ部が電力効率において支配的である．そのようなアクセラレータとしては，PEZY-SC3 より上位にランクされているシステムは全て NVIDIA H100，NVIDIA A100，AMD MI250X，PFN MN-Core のいずれかを採用しており，PEZY-SC3 はそれらに次いで 5 位の電力効率を持つと言える．また，上位の 4 アーキテクチャがいずれもテンソルユニットなど行列計算に特化した演算ユニットによって電力効率を向上させているのに対し，PEZY-SC3 はそれらの演算ユニットを採用しておらず，行列計算に限らず幅広いアプリケーションにおいて高い電力効率を発揮することができる．

本論文ではまず PEZY-SCx シリーズに共通するアーキテクチャについて説明し，PEZY-SC3 までの各世代の構成について述べる．次に設計・製造した PEZY-SC3 の実装

[1] PEZY Computing
[2] 東京大学大学院 情報理工学系研究科
[a)] hatta@pezy.co.jp

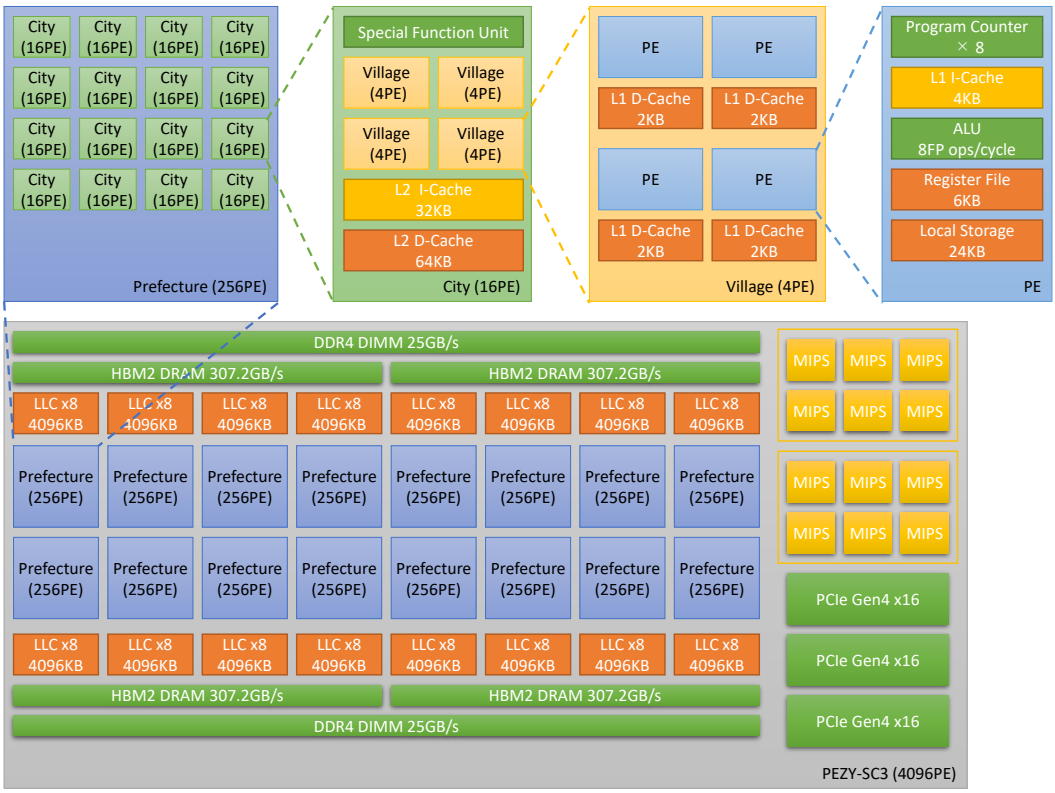

図 1 PEZY-SC3 のブロックダイアグラム

詳細とそれを搭載したスーパーコンピュータシステムの構成について説明し，その後に電力効率の測定結果について述べる．

2. PEZY-SCx のアーキテクチャ

PEZY-SCx シリーズは世代ごとにコア数や動作周波数などを向上させているが，基本的なアーキテクチャは共通している．ここでは図 1 に示す PEZY-SC3 のブロックダイアグラムを用いて PEZY-SCx のアーキテクチャについて述べる．PEZY-SCx は以下のユニットから構成される：

- Processor Element (PE)：PE は我々が独自に開発した RISC 命令セットを持つプロセッサであり，PEZY-SCx の主要な計算リソースである．32bit および 64bit の整数演算と，半精度・単精度・倍精度の浮動小数点数演算をサポートする．
- 管理プロセッサ：管理プロセッサは PE や PCIe インターフェースを制御するために使用される．PEZY-SC3 には MIPS64 命令セットを持つ管理プロセッサが 6 コア× 2 クラスター搭載されている．
- 外部メモリ：PE および管理プロセッサからアクセス可能な主記憶として外部メモリが接続される．PEZY-SC3 は外部メモリのインターフェースとして DDR4 と HBM2 の 2 つをサポートしている．
- 外部インターフェース：外部インターフェースはホストプロセッサが PEZY-SCx を制御するため，あるいは PEZY-SCx が外部の PCIe デバイスを制御するために使用される．PEZY-SC3 には外部インターフェースとして PCIe Gen4 が 48 レーン搭載されている．

2.1 階層構成

PEZY-SCx は階層化された構造を持っており，上位階層から順に state，prefecture，city，village と呼ぶ．以下では例として PEZY-SC3 における構成を述べる．まず最上位の state はチップ全体を示し，各 state は 16 個の prefecture から構成される．同様に各 prefecture は 16 個の city から，各 city は 4 個の village から，各 village は 4 個の PE から構成される．チップ内の各 PE は独立してプログラムを実行するが，専用の同期命令により指定された範囲内で待ち合わせを行うことができる．たとえば，city を指定した同期命令を発行すると同一 city 内の 16PE で待ち合わせを行い，state を指定した同期命令ではチップ全体で同期を取ることができる．

2.2 PE のアーキテクチャ

図 2 に PE のブロックダイアグラムを示す．PE は細粒度マルチスレッディングをサポートしたプロセッサであり，プログラムカウンタとレジスタファイルを 8 スレッド分搭載している．L1 命令キャッシュは 4-KB の 8-way セットアソシアティブキャッシュであり，8B/cycle の命令を供給することができる．PE の命令長は 32bit であるため，1 サ

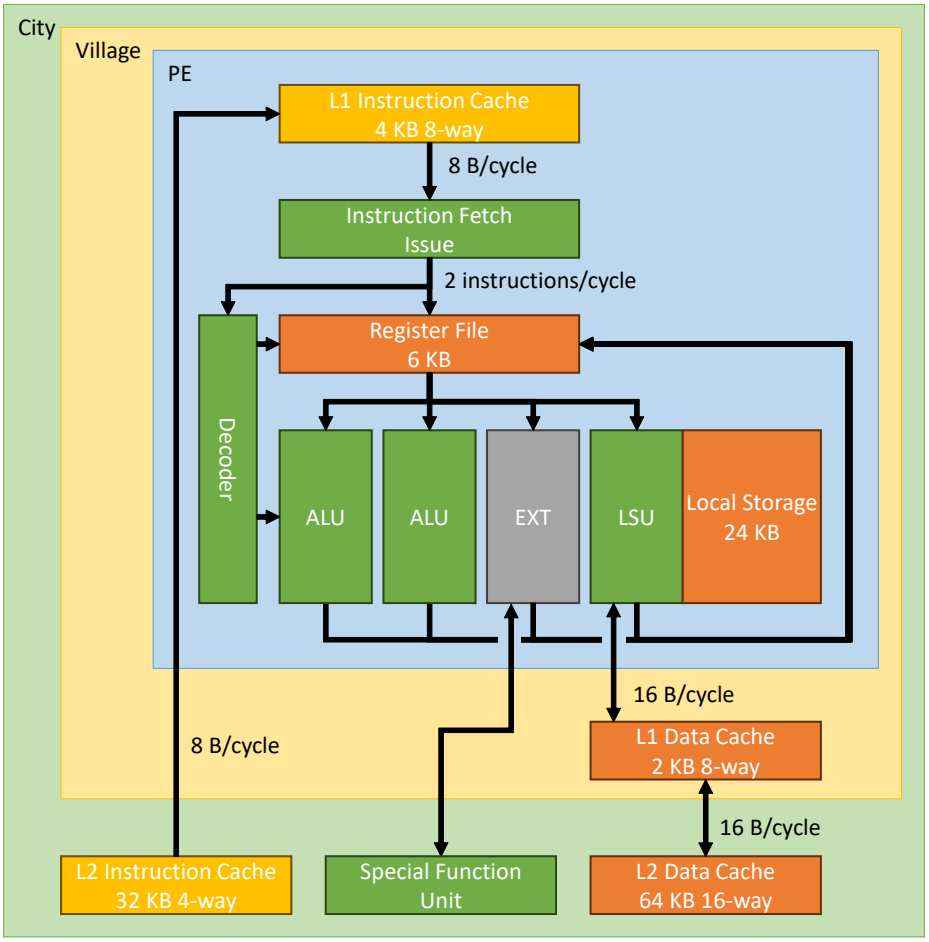

図 2 PE の構造

イクルに最大 2 命令まで発行可能となっている．

レジスタファイルはスレッドあたり 64bit の整数レジスタ 32 本，128bit の SIMD 浮動小数点数レジスタ 32 本を搭載し，8 スレッドの合計は 6-KB となる．実行ユニットは整数・浮動小数点数演算を行う ALU(Arithmetic Logic Unit) が 2 ユニット，ロードストアを行う LSU(Load Store Unit) が 1 ユニット搭載されている．除算や平方根など一部の特殊演算は city に 1 つずつ搭載された SFU(Special Function Unit) で実行される．LSU には 24-KB のローカルストレージが搭載されており，パイプラインストールを発生させることなくロードストアを行うことができる．

2.3 スレッド実行モデル

PE におけるスレッドの実行モデルを図 3 に示す．8 つのスレッドは 2 スレッドずつ 4 つのスレッドグループに分かれており，実行されるスレッドグループはクロックサイクル毎に切り替わる．各スレッドグループのうち 1 スレッドが実行状態，もう 1 スレッドが待機状態であり，専用命令により実行状態にあるスレッドを切り替えることができる．このスレッド切り替え命令により，長いメモリレイテンシを効率良く隠蔽することができる．また，各スレッドは 4 サイクルに 1 回しか実行されないため，次命令のフェッチまでに分岐先を確定させることができる．そのため分岐予測器を搭載することなく高い IPC を維持することが可能である．

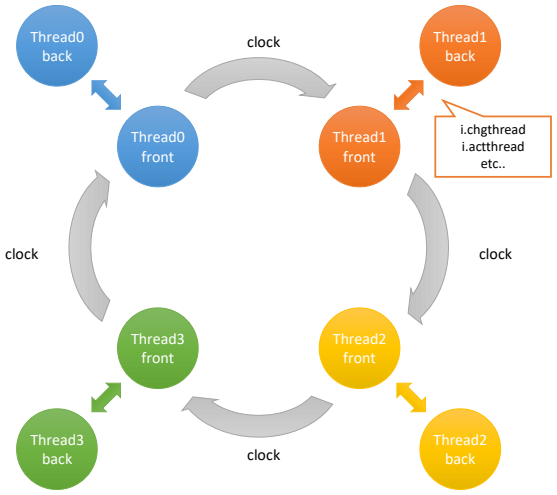

図 3 スレッド実行モデル

2.4 キャッシュ階層

PEZY-SCx が搭載している L1/L2/LLC の各キャッシュはキャッシュコヒーレンシのための機構を持っておらず，

表 1　PEZY-SCx シリーズの構成

		PEZY-SC3s	PEZY-SC3	PEZY-SC2	PEZY-SC
Processor Element	命令セット コア数 動作周波数	Custom ISA 512 1.0 GHz	Custom ISA 4096 1.2 GHz	Custom ISA 2048 1.0 GHz	Custom ISA 1024 733 MHz
管理プロセッサ	命令セット コア数 動作周波数	RISC-V 1 1.0 GHz	MIPS64 6 × 2cluster 1.5 GHz	MIPS64 6 × 1cluster 1.0 GHz	ARM v5TEJ 2 733 MHz
ピーク演算性能	倍精度 単精度 半精度	2.0 TFlops 4.0 TFlops 8.0 TFlops	19.7 TFlops 39.3 TFlops 78.6 TFlops	4.1 TFlops 8.2 TFlops 16.4 TFlops	0.75 TFlops 1.5 TFlops 3.0 TFlops
外部メモリ (DDR)	構成 帯域		DDR4-3200 2ch 51.2 GB/s	DDR4-3200 4ch 102.4 GB/s	DDR4-2400 8ch 153.6 GB/s
外部メモリ (HBM)	構成 帯域	HBM2 2.4 Gbps 2ch 614.4 GB/s	HBM2 2.4 Gbps 4ch 1.2 TB/s		
外部インターフェイス	構成 帯域	PCIe Gen4 4lane 8 GB/s	PCIe Gen4 48lane 96 GB/s	PCIe Gen4 32lane 64 GB/s	PCIe Gen3 32lane 32 GB/s
製造プロセス		TSMC 7nm	TSMC 7nm	TSMC 16nm	TSMC 28nm
リリース		2022/04	2021/04	2017/04	2014/08

キャッシュの一貫性制御は各 PE がフラッシュ命令を発行することで行う．フラッシュ命令は引数に同期範囲を取る命令であり，指定された範囲の全スレッドがその命令で待ち合わせた後，範囲内のキャッシュを下位階層に書き戻すことで一貫性を保つ．これによりチップ全面に張り巡らされたキャッシュ間ネットワークの構成が簡素化される．

各階層キャッシュはクロスバスイッチを介して接続されており，全 PE がメモリアクセスを行った場合でも十分な帯域を確保している．

3. PEZY-SCx の構成

表 1 に PEZY-SCx シリーズの各世代の構成を示す．PEZY-SC では倍精度浮動小数点数演算の発行性能がサイクル当たり 1 演算であったが，PEZY-SC2 で積和演算命令 (MAD) の追加により 2 演算となり，PEZY-SC3 では SIMD MAD 命令により 4 演算となった．これに加えて PE のコア数と動作周波数を向上させることにより各世代で 4 倍以上の性能向上を達成している．また，演算性能の向上にあわせてローカルストレージの容量も 3 世代で 16-KB, 20-KB，24-KB と増加させている．

PEZY-SC および PEZY-SC2 は液浸冷却技術 [9] を用いた大規模なスーパーコンピュータシステムをターゲットとしてきたが，PEZY-SC3 ではそれに加えて比較的小規模な空冷システム [10] においても利用可能となっている．さらに PEZY-SC3 を 1/8 に縮小した PEZY-SC3s もリリースし，より幅広い領域での利用を目指している [11].

4. PEZY-SC3 の実装と評価

表 2 に実装した PEZY-SC3 のスペックをまとめる．PEZY-SC3 は TSMC 7 nm プロセスで製造され，そのダイサイズは 25.7 mm × 30.6 mm である．図 4 に製造された PEZY-SC3 の外観を，図 5 に LSI 製造用のマスクデータを画像化したものを示す．チップ中央部に PE が，左右辺に HBM メモリコントローラと LLC が，上辺に DDR メモリコントローラが，下辺に PCIe コントローラがそれぞれ配置されている．

表 2　PEZY-SC3 の実装

製造プロセス	TSMC 7 nm FinFET
ダイサイズ	25.7 mm × 30.6 mm
ゲート数	3300M ゲート
メモリビット数	2300M ビット
消費電力	470 W (Max)

図 4　PEZY-SC3 のパッケージ外観

PEZY-SC3 では設計時にシミュレーションベースの電力効率の推定を行っている．設計したゲートネットリストに対し，倍精度の行列積演算を行うテストパターンを用い，RTL シミュレーションにより各ゲートの動作波形を取得した．次に設計データより配線とゲートの容量を抽出し，取得した動作波形から消費電力を推定した．使用したツールは以下の通りである．

- RTL シミュレータ: Synopsys VCS

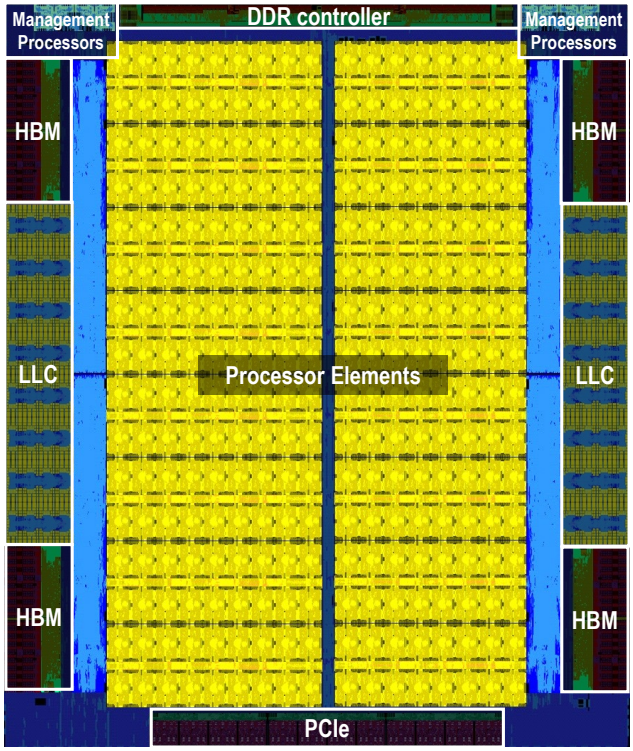

図 5 PEZY-SC3 のチップレイアウト

- 容量抽出: Synopsys StarRC
- 消費電力推定: Synopsys PrimeTime PX

使用したテストパターンは 800 MHz 動作時のものであり，このときの演算性能は 8546.38 GFlops，消費電力は 300.4 W となった．従って電力効率は 28.45 GFlops/W である．

続いて実際のスーパーコンピュータシステムにおける消費電力測定について述べる．表 3 に示す構成のシステムを用い，Top500 [12] のレギュレーションに従い LINPACK ベンチマークを実行して測定を行った．システム全体の総演算性能は 2.93 PFlops/s であり，LINPACK 実行時の実際の演算性能は 1.95 PFlops/s であった．この時システム全体の消費電力は 88.94 kW であり，電力効率は 21.892 GFlops/W となった．

実測した電力効率は設計時に推定したものより 23%ほど低下しているが，これはメモリデバイス・ホストプロセッサ・インターコネクト・冷却システム・電源系といったチップ以外の電力消費によるものと考えられる．

表 3 システム構成

ノード数	70 ノード
ホストプロセッサ	AMD EPYC 7702P × 1 (ノード当たり)
プロセッサ	PEZY-SC3 × 4 (ノード当たり)
コア数	1,151,360 コア
インターコネクト	EDR Infiniband
Rmax (PFlops/s)	1.95
Rpeak (PFlops/s)	2.93

5. まとめ

PEZY-SC3 は TSMC 7nm プロセスで製造され高い電力効率をもつスーパーコンピュータ向けのプロセッサである．高いスレッドレベル並列性を持つプログラムに特化することでプログラマビリティと高い電力効率を両立させている．実際のスーパーコンピュータシステムにおける測定では 21.892 GFlops/W の電力効率を達成した．

参考文献

[1] Hosono, N. et al.: Implementation of SPH and DEM for a PEZY-SC Heterogeneous Many-Core System, *Proceedings of the International Conference on Computational & Experimental Engineering and Sciences*, pp. 709–715 (2020).
[2] Hishinuma, T. et al.: pzqd: PEZY-SC2 Acceleration of Double-Double Precision Arithmetic Library for High-Precision BLAS, *Proceedings of the International Conference on Computational & Experimental Engineering and Sciences (ICCES)*, pp. 717–736 (2020).
[3] Matsumoto, K. et al.: Effectiveness of Performance Tuning Techniques for General Matrix Multiplication on the PEZY-SC2, *Proceedings of the International Symposium on Highly-Efficient Accelerators and Reconfigurable Technologies (HEART)*, pp. 1–6 (2019).
[4] Tanaka, H. et al.: Automatic Generation of High-Order Finite-Difference Code with Temporal Blocking for Extreme-Scale Many-Core Systems, *IEEE/ACM International Workshop on Extreme Scale Programming Models and Middleware (ESPM2)*, pp. 29–36 (2018).
[5] Iwasawa, M. et al.: Implementation and Performance of Barnes-Hut N-body algorithm on Extreme-scale Heterogeneous Many-core Architectures, *The International Journal of High Performance Computing Applications*, Vol. 34, No. 6, pp. 615–628 (2020).
[6] Sato, M. et al.: Co-Design for A64FX Manycore Processor and “Fugaku”, *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC20)*, pp. 1–15 (2020).
[7] Dongarra, J. J., Luszczek, P. and Petitet, A.: The LINPACK Benchmark: past, present and future., *Concurr. Comput. Pract. Exp.*, Vol. 15, No. 9, pp. 803–820 (2003).
[8] TOP500 project: Green500, `https://www.top500.org/lists/green500/`.
[9] PEZY Computing, K.K.: ZettaScaler 2.0, `https://www.pezy.co.jp/products/zettascaler-2-0/`.
[10] PEZY Computing, K.K.: ZettaScaler 3.0, `https://www.pezy.co.jp/products/zettascaler-3-0/`.
[11] 大友広幸，坂本　亮：PEZY-SC3s プロセッサを用いた Full-state 量子回路シミュレーション，技術報告 14，東京工業大学, PEZY Computing (2022).
[12] TOP500 project: TOP500, `https://www.top500.org`.