「ZettaScaler-1.5によるHPCシステム構築と、ZettaScaler-2.0構想」

2015年 12月18日

齊　藤　　元　章

（株式会社PEZY Computing/株式会社ExaScaler/UltraMemory株式会社）

11月発表のGreen500最新結果の顛末

- 本来は、Green500で1-4位独占を実現する目論見であった

- 7月のISCで、計測ルールがv2.0になることが予告された
（現行のv1.2ルールでの計測値改善には注力せず、v2.0対応作業のみ進めていた）

- 最後までv1.2維持の発表も、v2.0への移行も発表されず
（Shoubu, Suiren Blue, Suiren共に前回の値を更新せずに、最新結果を待つことに）

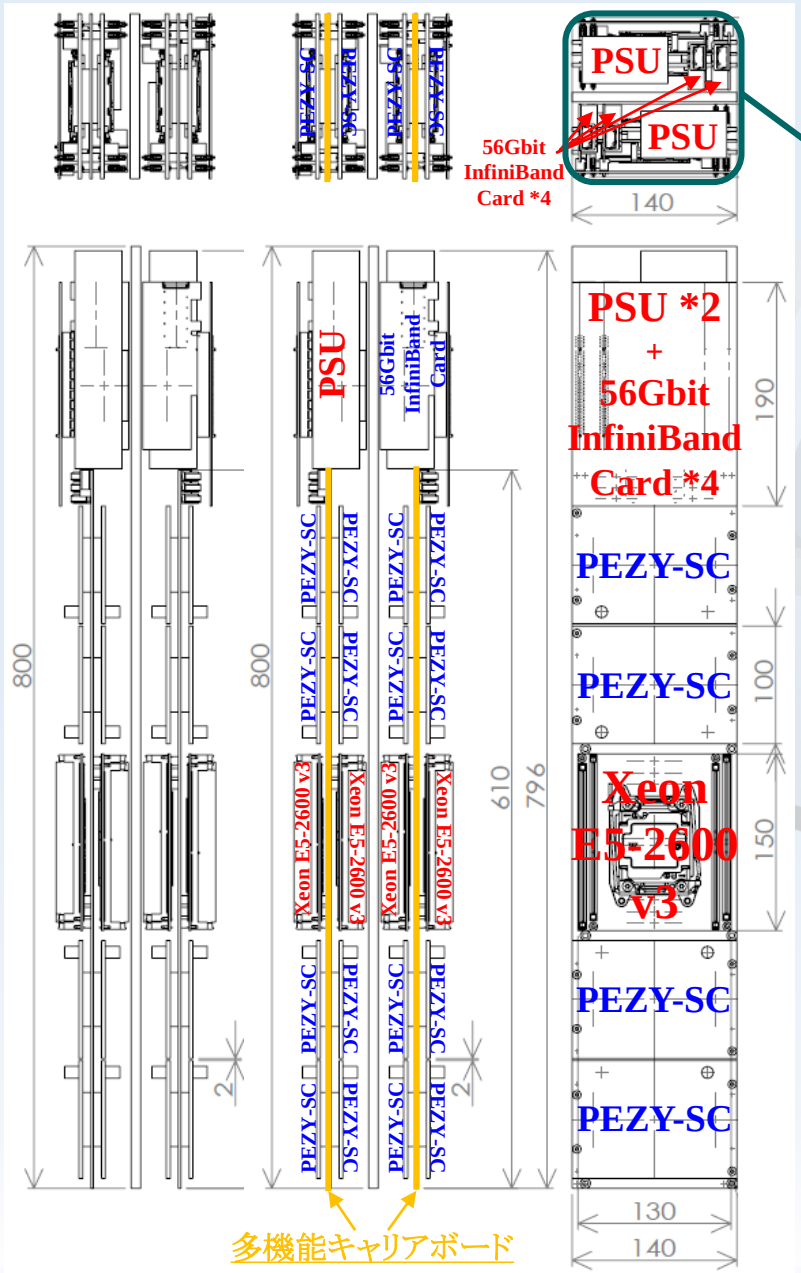
- 最新TOP500にSuiren BlueとSuiren（Green500計測用）が入れず
（4か月間で中国勢が38台から109台に大躍進し、36台の日本を大きく上回ったため）

- 更に、理研神戸AICS様に設置したAjisai（紫陽花）が選外に
（TOP500に2週間、Green500締め切りには3日間に合わず。**本来はGreen500で2位**）

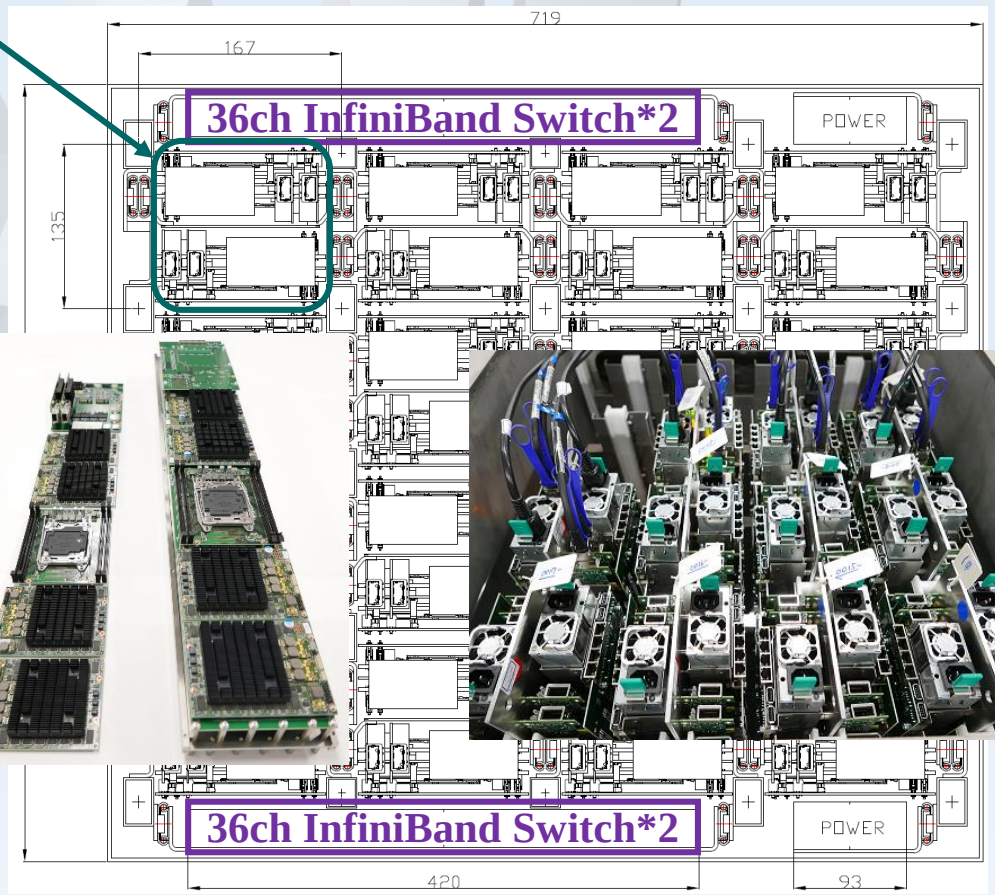
- 来年1月からv2.0対応が必要となるが、此方の準備は完了
（20%区間ルールの廃止、15ノード以上、IB Switchを含めるなど）

- 来年6月は、5台目を含めてGreen500で1-5位独占としたい

第2世代HPCシステム「ZettaScaler-1.4」

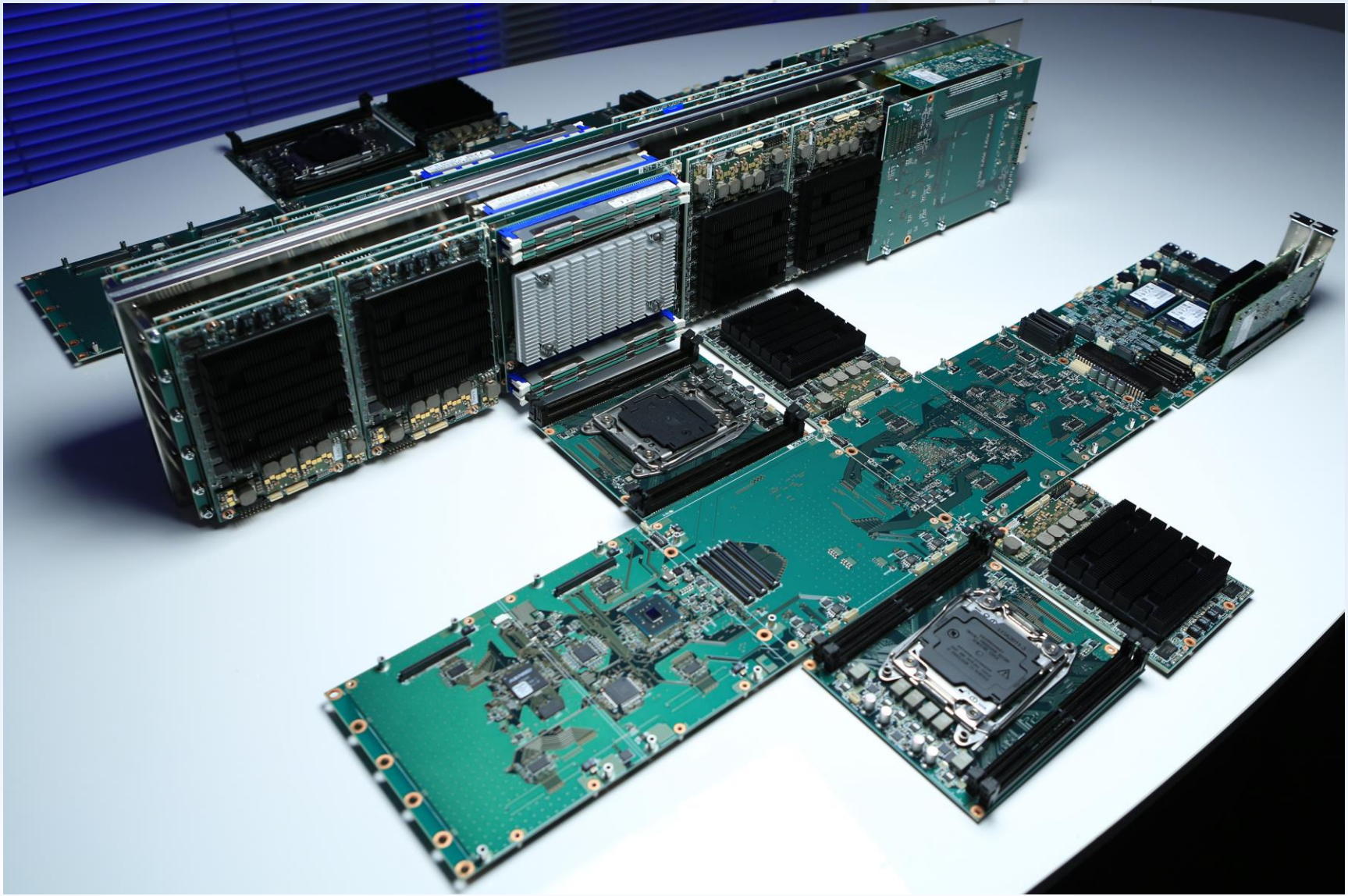


新型液浸槽「ESLiC-32」上面図
(Xeon*64, PEZY-SC*256, Switch*4)



←新型液浸槽専用に新たに開発する、
モジュール＋キャリアボード基板構成

第2世代、液浸冷却専用システムを開発
（液浸冷却専用の基板複合体“Brick”）



理化学研究所の「Shoubu（菖蒲）」

最新のZettaScaler-1.4による5台構成の2 PetaFLOPS級スパコンを、「Shoubu（菖蒲）」として、理化学研究所和光の情報基盤センターに設置して頂く（全系の安定動作を得るのに、予想を超える問題点が生じており対応中）

ZettaScaler-1.4から1.5への変更点

・ZettaScaler-1.5からは、液浸冷却の新プラットフォーム全体のシステム構成に

・Brickサイズを、2cm拡張して14cm×16cmの断面形状に

・Multi-Xeon Brick, Storage Node Brickを追加し混成可能に（Multi-GPGPU Brickも準備中）

・PEZY-SC BrickはXeon用DIMMスロット数を倍増し、128GBまで搭載可能に（32GB VLP DIMMでは256GBまで）

・1,600W電源をBrick当たり4台、2 Node当たり2台の冗長構成としつつ、AC/DC変換効率が最も高い負荷区間で使用可

・Node内の4個のPEZY-SC間でPCIe Gen3 x8で1対3の双方向プロセッサ間通信が可能に

液浸冷却スパコンの消費電力性能

PEZY Computing/ExaScaler社開発の4台のスパコン

システム名	設置サイト	システム構成	Rmax（TFLOPS） Green500申請用	消費電力性能（GFLOPS/W）	
				v1.2ルール	v2.0ルール
Shoubu（菖蒲）	理化学研究所和光ACCC様	ZettaScaler-1.4	353.8	7.03	n/a
Suiren Blue（青睡蓮）	高エネルギー加速器研究機構様	ZettaScaler-1.4	193.3	6.84	n/a
Ajisai（紫陽花）	理化学研究所神戸AICS様	ZettaScaler-1.5	214.9	6.52	5.92
Suiren（睡蓮）	高エネルギー加速器研究機構様	ZettaScaler-1.0	202.6	6.22	n/a
Suiren Blue（青睡蓮）	高エネルギー加速器研究機構様	ZettaScaler-1.5	194.7	n/a	5.47

Green500足切り数値: 206.4

参考（Green500: 2-4位）

TSUBAME-KFC/DL	東京工業大学	Xeon/Tesla K80	272.6	5.33	n/a
The L-CSC cluster	GSI Helmholtz Center	Xeon/FirePro S8150	301.3	6.01	5.27
Sugon Cluster W780I	IMP, Chinese Academy of Science	Xeon/Tesla K80	310.6	n/a	4.78

ZettaScaler-1.5の現状

・ボードデバッグが間に合わずに、InfiniBandカードはPCIe Gen2での接続に留まる

・プロセッサ間通信は未使用

・電圧設定が不十分で、消費電力効率が最適化段階にない

・PEZY-SCパッケージの電圧降下問題が未解消

・PEZY-SCパッケージのDDR4クロック問題が未解消

・PEZY-SCモジュールのDDR4 DRAMを倍容量化したものの、パラメータ調整が不十分で低速での通信しか行えていない

・PEZY-SCパッケージとPEZY-SCモジュールの改版作業中

・上記の問題を全て解決できれば、v2.0で10 GFLOPS/Wも

次世代HPCシステム開発へ

・今後のHPCシステム向け独自要素技術開発項目

1）MIMD型プロセッサの圧倒的な超々メニーコア化

2）低消費電力・大容量の積層DRAMを独自開発

3）プロセッサ・DRAM間の無線による超広帯域接続

4）Switch Chipを独自開発し、ファットツリーを1チップ化

5）Brick内Interconnectを、無線化、無ケーブル化

6）3重合液浸冷却で超高集積化・小型化・低消費電力化

7）上記全てを2年で開発し、2020年までに2世代分進化

・ZettaScaler-2.0では、このうちの1）、2）、3）、6）を実現予定

ZettaScaler-2.0の開発構想

・第3世代MIMDプロセッサ「PEZY-SC2」
（4,096コア, 8TFLOPS, 16nm FinFET, 4TB/sメモリ帯域, **64bit CPU**内蔵）

・**64bit CPUを内蔵**することでチップ外のホストCPUを不要として、絶対性能と消費電力性能の大幅な向上

・**超広帯域・大容量、3次元TCI（磁界結合）積層メモリ**
（**プロセッサパッケージに同梱する、超広帯域接続の独自積層メモリ**）

・沸騰冷却を組み合わせた、3重合液浸冷却システム
（冷却電力部を含めたシステム消費電力の低減と、冷却能力の強化）

開発を進める「PEZY-SC2」の仕様（更新版）

Processor		PEZY-SC	PEZY-SC2
製造プロセス		TSMC 28HPM（28nm）	TBD（14-16nm FinFET）
	ダイサイズ	412mm2	400-500mm2
コア性能	動作周波数	733MHz	1GHz
	キャッシュ	L1: 1MB, L2: 4MB, L3: 8MB	50MB in total (TBD)
周辺回路	動作周波数	66MHz	66MHz
IPs	内蔵CPU	ARM926 x 2 管理・デバッグ用	64bit CPU (MIPS) 汎用演算用
	PCIe	PCIe Gen3 x 8Lane 4Port (8GB/s x 4 = 32GB/s)	PCIe Gen3/4 x 8Lane 6Port (48-96GB/s)
	DRAM	DDR4 64bit 2,400MHz 8Port (19.2GB/s x 8 = 153.6GB/s)	Custom Stacked DRAM 8Port (500GB/s x 8 = 4.0TB/s)
コア（PE）数		1,024 PE	4,096 PE
演算性能		3.0T Flops (単精度浮動小数点) 1.5T Flops (倍精度浮動小数点)	16.4T Flops (単精度浮動小数点) 8.2T Flops (倍精度浮動小数点)
消費電力		60W (Leak: 10W, Dynamic: 50W)	100W (Leak: 10W, Dynamic: 90W)
パッケージ		47.5*47.5mm (2,112pin)	Multi-Die Package (TBD)

次世代、次々世代システムの開発構想

	ExaScaler-1.0	ExaScaler-1.4	ExaScaler-1.6	ExaScaler-2.0	ExaScaler-3.0
	2014年10月	2015年6月	2016年10月	2017年5月？	2019年5月？
システム消費電力性能	**5 GFLOPS/W**	7 GFLOPS/W	**10 GFLOPS/W**	**20 GFLOPS/W**	**40 GFLOPS/W**
主演算プロセッサ	PEZY-SC (ES)	PEZY-SC（プロセス修正）	PEZY-SC (パッケージ改版)	PEZY-SC2	PEZY-SC3
製造プロセス	28nm Planar	←	←	14-16nm FinFET	10nm FinFET
MIMDコア数	**1,024**	←	←	**4,096**	**8,192**
駆動周波数	660MHz	690MHz	833MHz	1.0GHz	1.25GHz
倍精度演算性能	**1.35TFLOPS**	1.41TFLOPS	**1.70TFLOPS**	**8.19TFLOPS**	**20.46TFLOPS**
搭載メモリ	DDR3@1,333MHz	DDR4@1,600MHz	DDR4@2,133MHz	**TCI-3DS-DRAM Gen1**	**TCI-3DS-DRAM Gen2**
メモリ容量	32GB	16GB	32GB	32-64GB	128-256GB
メモリ帯域	85.3GB/s	102.4GB/s	136.5GB/s	**4.1TB/s**	**10.2TB/s**
Byte/FLOP	**0.063**	0.073	**0.080**	**0.5**	**0.5**
単体消費電力効率	**25GFLOPS/W**	←	←	**40-50GFLOPS/W**	**80-100GFLOPS/W**
汎用CPU					
CPU種別	Xeon E5-2600 v2	Xeon E5-2600 Lv3	←	**64bit CPU (TBD)**	←
実装形態	外付け別システム	←	←	**同一Die上に内蔵**	←
接続方法	PCIe Gen2*16	PCIe Gen2*8	PCIe Gen3*8	**内部ローカルバス**	←
搭載メモリ / 容量	DDR3 / 128GB	DDR4 / 64GB	DDR4 / 128GB	**主演算プロセッサと共有**	←
Network Switch					
Inteconnect種別	InfiniBand FDR	←	←	InfiniBand EDR (TBD)	**独自TCI-3DS-Switch**
Inteconnect速度	7Gbit/主演算プロセッサ	14Gbit/主演算プロセッサ	←	25Gbit/主演算プロセッサ	**TBD**
システムボード					
ボード種別	空冷用汎用マザーボード	液浸冷却専用独自Brick	←（改版版）	第2世代Brick	第3世代Brick
冷却システム					
冷却方法	単純液浸冷却	←（4倍密）	2重合液浸冷却	**3重合液浸冷却**	←
体積当たり性能					
サーバーラック体積性能	**250TeraFLOPS**	800TeraFLOPS	**1PetaFLOPS**	**8PetaFLOPS**	**20PetaFLOPS**
ExaFLOPSシステム構成					
サーバラック筐体数	**4,000台相当**	1,250台相当	**1,000台相当**	**125台相当**	**50台相当**
消費電力	**200MW**	143MW	**100MW**	**50MW**	**25MW**

マルチダイで自在な組み合わせを可能とするプロセッサ構成手法