「Post-Exa時代のHPCシステムアーキテクチャに関する妄想」

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）

2025年に向けて普通にスケールすると、、

	ExaScaler-1.0	ExaScaler-1.6	ExaScaler-2.0	ExaScaler-3.0	ExaScaler-4.0	ExaScaler-5.0	ExaScaler-6.0
	2014年10月	2016年5月	2017年	2019年	2021年	2023年	2025年
システム消費電力性能	5 GFLOPS/W	10 GFLOPS/W	20-25 GFLOPS/W	40-50 GFLOPS/W	60-75 GFLOPS/W	80-100 GFLOPS/W	100-125 GFLOPS/W
主演算プロセッサ	PEZY-SC (ES)	PEZY-SC (パッケージ改版)	PEZY-SC2	PEZY-SC3	PEZY-SC4	PEZY-SC5	PEZY-SC6
製造プロセス	28nm Planar	←	14-16nm FinFET	10 or 7nm FinFET	7 or 5nm TunnelFET	5nm (CNT?)	TBD
MIMDコア数	1,024	←	4,096	8,192	16k	32k	64k
駆動周波数	660MHz	833MHz	1.0GHz	1.25GHz	1.5GHz	2.0GHz	2.5GHz
倍精度演算性能	1.35TFLOPS	1.70TFLOPS	8.19TFLOPS	20TFLOPS	48TFLOPS	128TFLOPS	320TFLOPS
搭載メモリ	DDR3@1,333MHz	DDR4@2,133MHz	TCI-3DS-DRAM Gen1	TCI-3DS-DRAM Gen2	TCI-3DS-DRAM Gen3	TBD	TBD
メモリ容量	32GB	32GB	32-64GB	128-256GB	256-512GB	512GB-1TB	1-2TB
メモリ帯域	85.3GB/s	136.5GB/s	4.1TB/s	10.2TB/s	25TB/s	64TB/s	160TB/s
Byte/FLOP	0.063	0.080	0.5	0.5	0.5	0.5	0.5
単体消費電力効率	25GFLOPS/W	←	40-50GFLOPS/W	80-100GFLOPS/W	120-150GFLOPS/W	160-200GFLOPS/W	200-250GFLOPS/W
プロセッサ単体消費電力	50W	75W	160W	200W	320W	640W	1,200W
汎用CPU							
CPU種別	Xeon E5-2600 v2	Xeon E5-2600 Lv3	64bit CPU	←	TBD	TBD	TBD
実装形態	外付け別システム	←	同一Die上に内蔵	←	TBD	TBD	TBD
接続方法	PCIe Gen2*16	PCIe Gen3*8	内部ローカルバス	←	TBD	TBD	TBD
搭載メモリ / 容量	DDR3 / 128GB	DDR4 / 128GB	主演算プロセッサと共有	←	TBD	TBD	TBD
Network Switch							
Inteconnect種別	InfiniBand FDR	←	InfiniBand EDR (TBD)	独自TCI-3DS-Switch	←	TBD	TBD
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	TBD	TBD	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	第2世代Brick	第3世代Brick	TBD	TBD	TBD
冷却システム							
冷却方法	単純液浸冷却	2重合液浸冷却	3重合液浸冷却	←	TBD	TBD	TBD
体積当たり性能							
サーバーラック体積性能	250TeraFLOPS	1PetaFLOPS	8PetaFLOPS	20PetaFLOPS	50PetaFLOPS	TBD	TBD
ExaFLOPSシステム構成							
サーバラック筐体数	4,000台相当	1,000台相当	125台相当	50台相当	20台相当	TBD	TBD
消費電力	200MW	100MW	50MW	25MW	15MW	TBD	TBD

- 仮に5nm利用可能でも、プロセッサ単体電力が1,000W越え

- 64kコアも2次元チップ内に集積した場合には、チップ内のバス構造、キャッシュ構成が破綻することが目に見えている

2025年に向けて普通にスケールすると、、

	ExaScaler-3.0	ExaScaler-4.0	ExaScaler-5.0	ExaScaler-6.0
	2019年	2021年	2023年	2025年
システム消費電力性能	**40-50 GFLOPS/W**	**60-75 GFLOPS/W**	**80-100 GFLOPS/W**	**100-125 GFLOPS/W**
主演算プロセッサ	PEZY-SC3	PEZY-SC4	PEZY-SC5	PEZY-SC6
製造プロセス	10 or 7nm FinFET	7 or 5nm TunnelFET	5nm (CNT?)	TBD
MIMDコア数	**8,192**	**16k**	**32k**	**64k**
駆動周波数	1.25GHz	1.5GHz	2.0GHz	2.5GHz
倍精度演算性能	**20TFLOPS**	**48TFLOPS**	**128TFLOPS**	**320TFLOPS**
搭載メモリ	**TCI-3DS-DRAM Gen2**	**TCI-3DS-DRAM Gen3**	**TBD**	**TBD**
メモリ容量	128-256GB	256-512GB	512GB-1TB	1-2TB
メモリ帯域	**10.2TB/s**	**25TB/s**	**64TB/s**	**160TB/s**
Byte/FLOP	**0.5**	**0.5**	**0.5**	**0.5**
単体消費電力効率	**80-100GFLOPS/W**	**120-150GFLOPS/W**	**160-200GFLOPS/W**	**200-250GFLOPS/W**
プロセッサ単体消費電力	**200W**	**320W**	**640W**	**1,200W**
汎用CPU				
CPU種別	**64bit CPU**	**TBD**	**TBD**	**TBD**
実装形態	**同一Die上に内蔵**	**TBD**	**TBD**	**TBD**
接続方法	**内部ローカルバス**	**TBD**	**TBD**	**TBD**
搭載メモリ / 容量	**主演算プロセッサと共有**	**TBD**	**TBD**	**TBD**
Network Switch				
Inteconnect種別	**独自TCI-3DS-Switch**	**←**	**TBD**	**TBD**
Inteconnect速度	**TBD**	**TBD**	**TBD**	**TBD**
システムボード				
ボード種別	第3世代Brick	**TBD**	**TBD**	**TBD**
冷却システム				
冷却方法	**3重合液浸冷却**	**TBD**	**TBD**	**TBD**
体積当たり性能				
サーバーラック体積性能	**20PetaFLOPS**	**50PetaFLOPS**	**TBD**	**TBD**
ExaFLOPSシステム構成				
サーバラック筐体数	**50台相当**	**20台相当**	**TBD**	**TBD**
消費電力	**25MW**	**15MW**	**TBD**	**TBD**

現状での認識（MIMD型メニーコアの将来）

・高集積度には駆動電圧の継続的低電圧化は不可欠（○）
（0.4V駆動レベルまでは、期待出来そう）

・1,000W超のチップ単体冷却手法が不可欠（○）
（ハイブリッド液浸冷却で1,000W程度までは行けそう）

・プロセス毎のIP対応（期間、開発工数、費用）が困難に（○）
（GP-TCI活用で、ペリフェラルIPは最先端プロセスの必要はなくなる）

・16kレベル以降は、ロジックも3次元積層の検討が必要（△）
（積層自体は16-32層程度は恐らく可能だが、積層間冷却が問題に）

・チップ内インターコネクト、チップ間インターコネクト双方での革新（トポロジ、配線、実装、消費電力）が必要（△）
（磁界結合による無線インターコネクトがどこまで活用出来るかが鍵）

・5nm以降のプロセス・テクノロジーが不明（×）
（CNTが利用可能に？光トランジスタに期待？）

そもそもプロセッサの作り方を変える？

・並行して進める、規模を追求する汎用人工知能開発では、無線接続による大規模“3D Connection Array”を追求

・汎用人工知能では、最終的にコアやメモリを持たない方針

・HPCプロセッサ用に100万コアを想定した時、コア間配線やキャッシュ階層（Coherency無しでも）の設計は想像不可能

・であれば、大規模（100兆規模）“3D Connection Array”の中にHPCコアをどれだけ埋め込めるかを考えた方が早い？
（1億コア位までなら入りそう。メモリ容量、キャッシュ構造などは未検討）

・しかし、1GHzでは消費電力と発熱（冷却）問題は非現実的

・CNT以降のプロセス技術と省電力（低電圧）化については、2019年頃までには利用可能となっているはずのExaFLOPSスパコンを使用して問題解決が図られることを期待