組込みマルチコアハードウェアの最新動向
Latest Trends in Embedded Multi-core Hardware

名古屋大学大学院情報学研究科
東京大学大学院工学系研究科 システムデザイン研究センター
荒川　文男

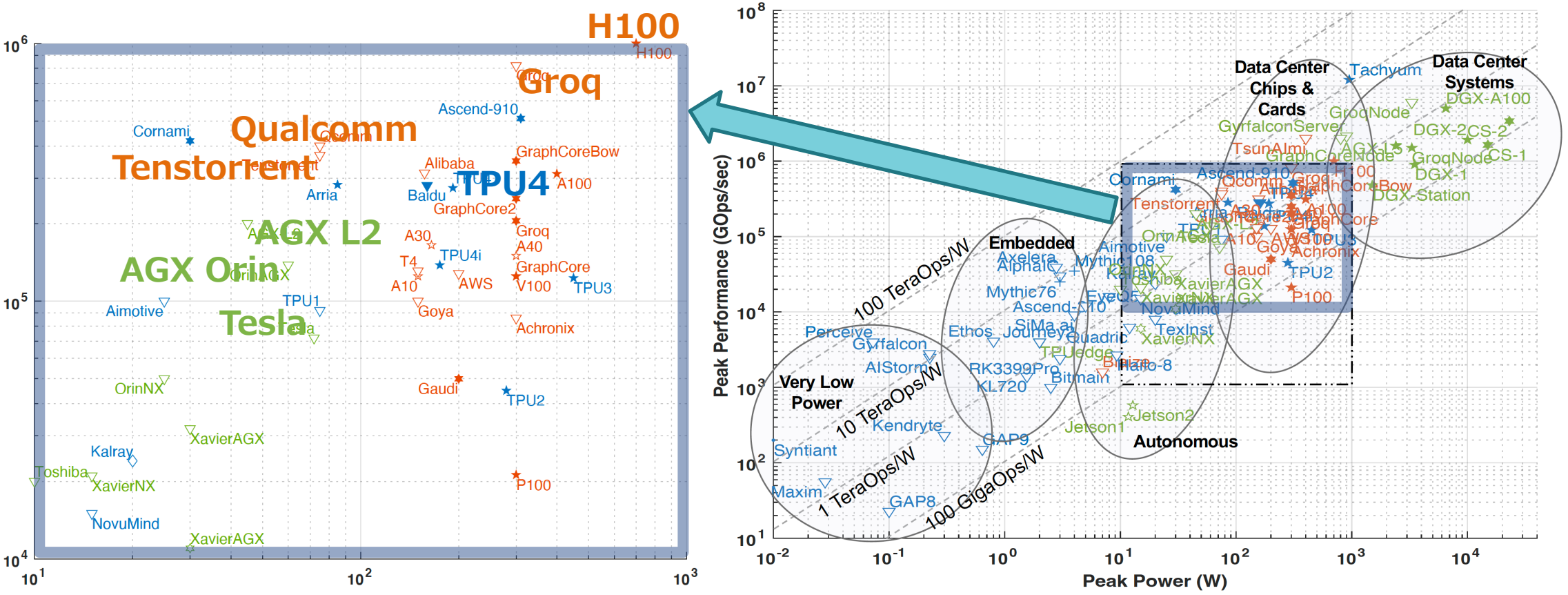
目次

プロセッサ研究開発の潮流

- 従来
  - 新アーキテクチャのメリット << 互換性維持のメリット
  - メジャーサプライヤーの漸進的改良で十分
- AI応用の急速な発展
  - AIに最適な仕様が従来と大幅に異なり、上記関係が逆転
  - 新アーキテクチャのアクセラレータが多数の機関から発表・製品化（次ページで紹介）
- コンパイラ
  - Multi-Level Intermediate Representation（**MLIR**）の標準化
  - 新アーキテクチャや複数アーキテクチャのサポートが容易に
- アーキテクチャ
  - RISC-Vの普及により**オープン化**が進展
  - ハイエンドでも電力、コスト制約が性能を決定
    - x86より高効率化し易いARM、RISC-Vがサーバー系にも進出
- **Chiplet**
  - サーバー系を中心にChiplet化が進展
  - Universal Chiplet Interconnect Express（**UCIe**）が標準化
- インタフェースとしてCompute Express Link（**CXL**）が台頭

多数のAIアクセラレータ

◆ AI応用の急速な発展により、新アーキテクチャのアクセラレータが多数の機関から発表・製品化



出典：Albert Reuther, et al., "AI and ML Accelerator Survey and Trends," 2022 IEEE High Performance Extreme Computing (HPEC) Conference (Oct 2022), Fig.2 & Fig.3

組込みマルチコア

- 組込みプロセッサ
  - 電力、コストなどの厳しい制約から、組込み系はローエンド中心
  - ADAS、自動運転、ロボットなどの進化により、組込み系でも**AI応用**を中心に**高性能化**要求
- AI応用
  - 学習も推論もサーバー処理が中心でAI向けプロセッサはサーバー系から製品化
    - 組込み系はサーバー系からの**派生品**
    - NVIDIAはA100の派生品としてOrin AGXを製品化
    - サーバー系の技術が組込み系に波及
  - 組込み系は**推論中心**
  - サーバー系は多数の要求をこなすので、スループット重視
  - 組込み系は**リアルタイム処理**が多いので、**レスポンス重視**
  - 要求仕様が異なるので派生品より専用設計が望ましい

- **AI推論に適したレスポンス重視の高効率な組込みマルチコア**を研究開発すべき
- これに向けてサーバー系適用例も含めた**AI向けプロセッサ技術**を紹介

AI向けプロセッサ技術

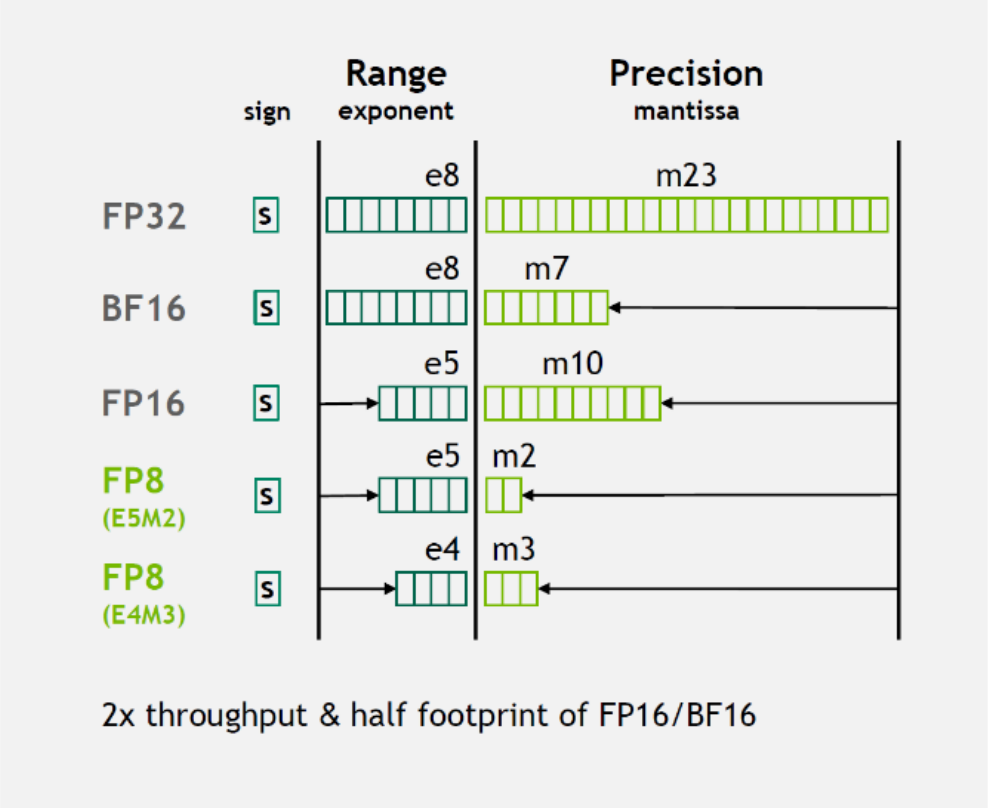
- AI応用からの技術的要求と対応
  - 演算の高性能化、演算量・データ量削減（CNN、RNN）
    - 高並列**低ビット**演算（INT4, INT8, FP8, BF16, LOG8）⇦ INT1, INT2は一般にAccuracy低下
    - **混合精度**演算（乗算ビット数を抑えて累算のみ高精度化）
    - **スパース性**活用
  - データ転送量削減
    - Processing Near-/In-Memory（PNM/PIM、転送時間>>演算時間の場合に有効）
    - メモリ階層の最適化、内蔵メモリの大容量化
  - **大規模データ供給**を含む総合的高性能化（Transformer、BERT、GPT、LLM）
    - 汎用プロセッサとAIアクセラレータの密結合
      - 汎用ハイエンドプロセッサは、大規模データ供給は得意なものの、低効率が問題
  - 今後も新アルゴリズムが普及する可能性大
    - 特定のアルゴリズムへの特化はリスク大 ⇨ **総合的高性能化**

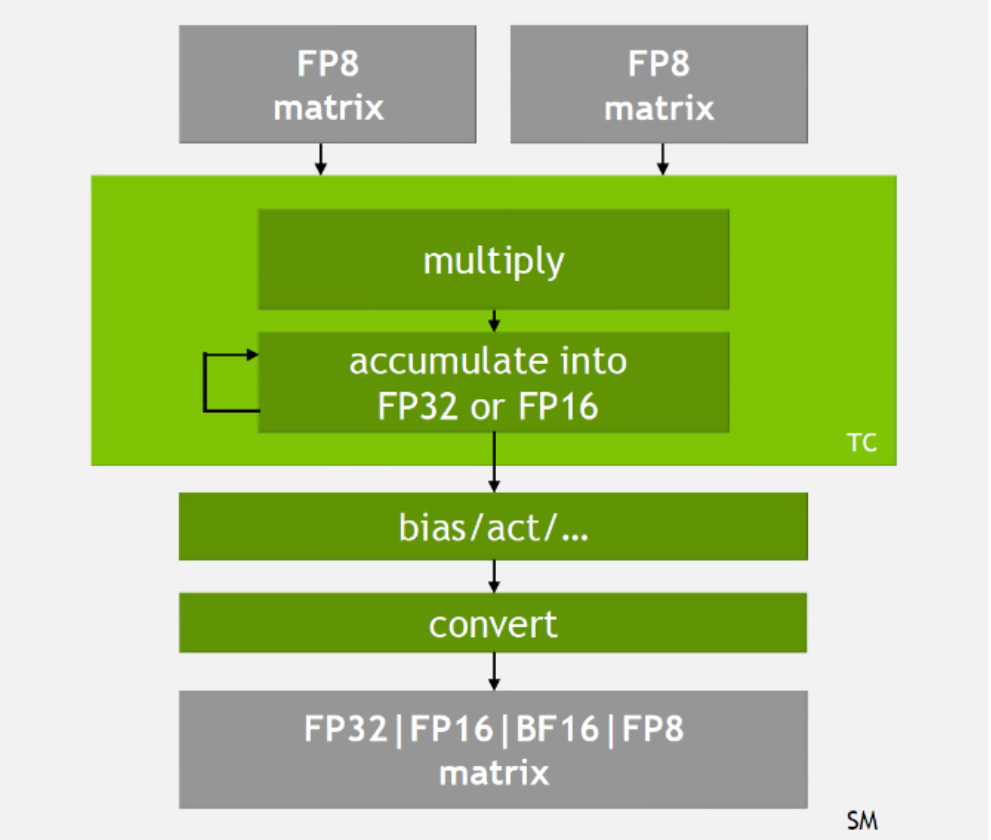
BERT: Bidirectional Encoder Representations from Transformers
GPT: Generative Pretrained Transformer
LLM: Large Language Models

CNN: Convolutional Neural Network
RNN: Recurrent Neural Network

高並列低ビット演算・混合精度演算

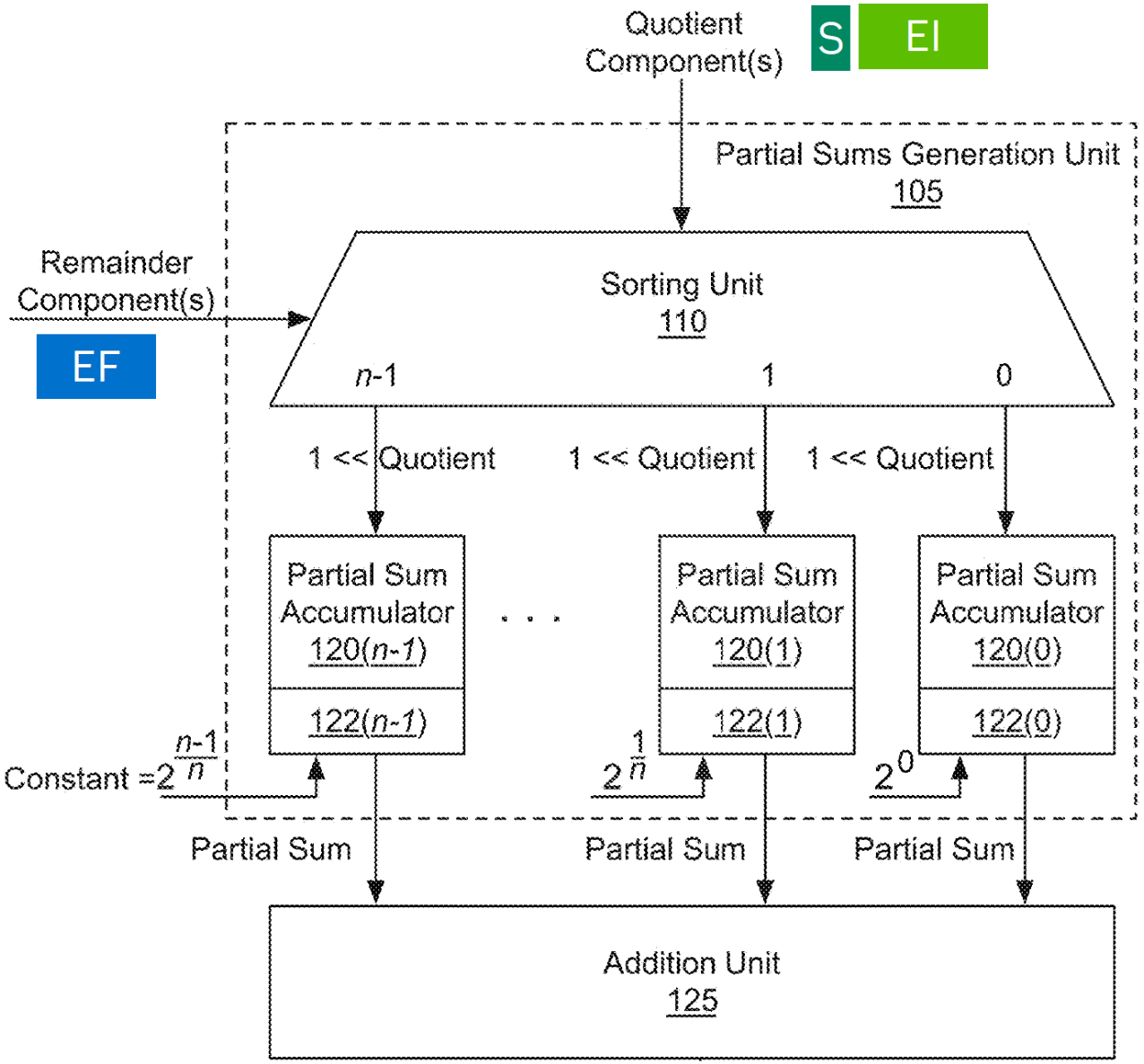
- NVIDIA H100の例
  - FP8 (E5M2, E4M3) を定義実装 ⇐ E: 指数部、M: 仮数部
  - FP8で乗算し、FP32/16で累算して、効率化（スループット2倍、フットプリント1/2）
  - FP8変換エンジンが、入力データを数値範囲分析に基づいてスケーリングと型変換処理





出典：JACK CHOQUETTE, "NVIDIA HOPPER GPU: SCALING PERFORMANCE," HOT Chips2022 (Aug 2022)

高並列低ビット演算・混合精度演算



NVIDIAのLOG8実装例

- $v = -1^{s}2^{ei.ef}$

S EI EF

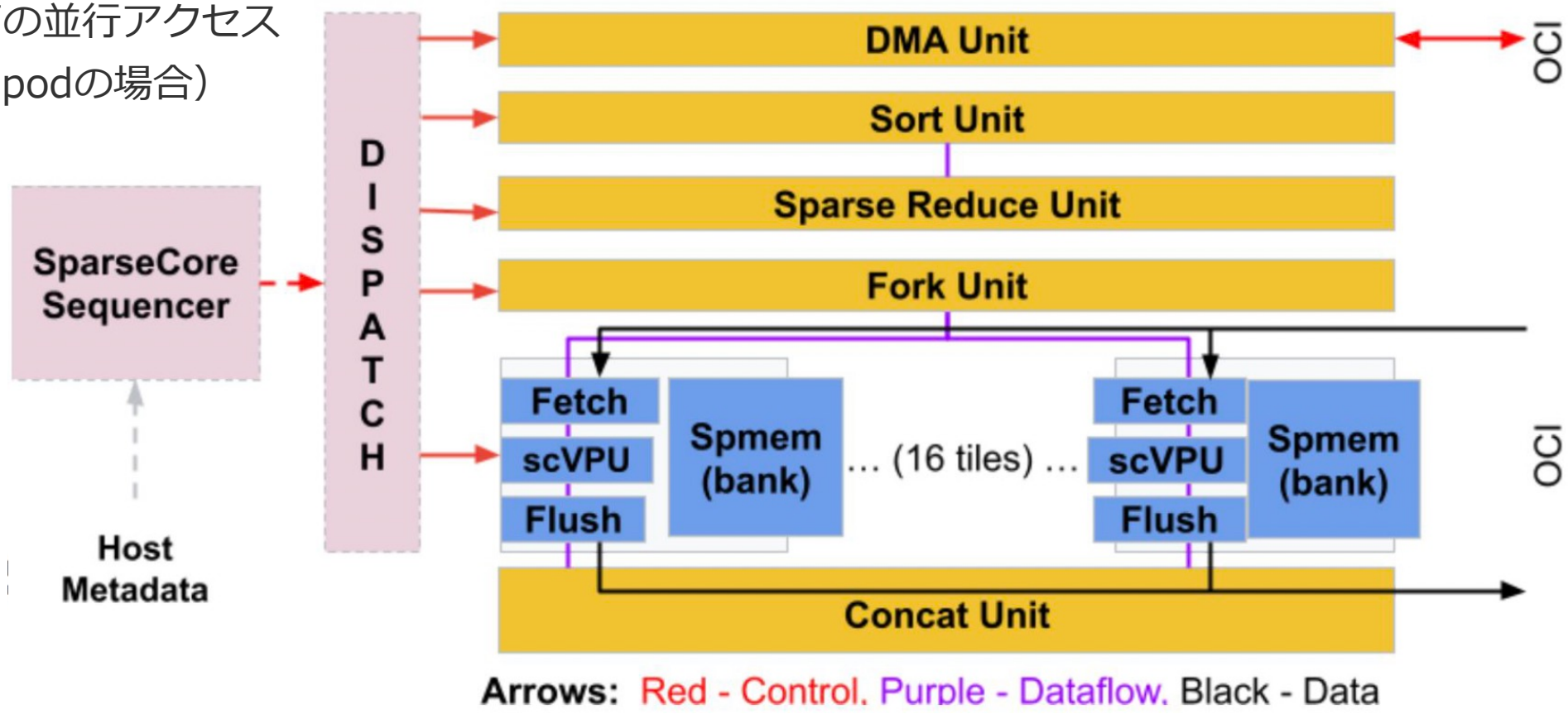
符号 1b, 整数部 4b, 小数部 3b

- 乗算はLOGの加算に
- 加算のための変換・逆変換回数を削減
- 小数部3bの2^3=8通りをそれぞれ加算してからテーブル参照
- 整数部はシフト、同じ値の加算は個数の乗算

出典：Bill Dally, "Hardware for Deep Learning," HOT Chips 2023 (Aug 2023)

スパース性活用

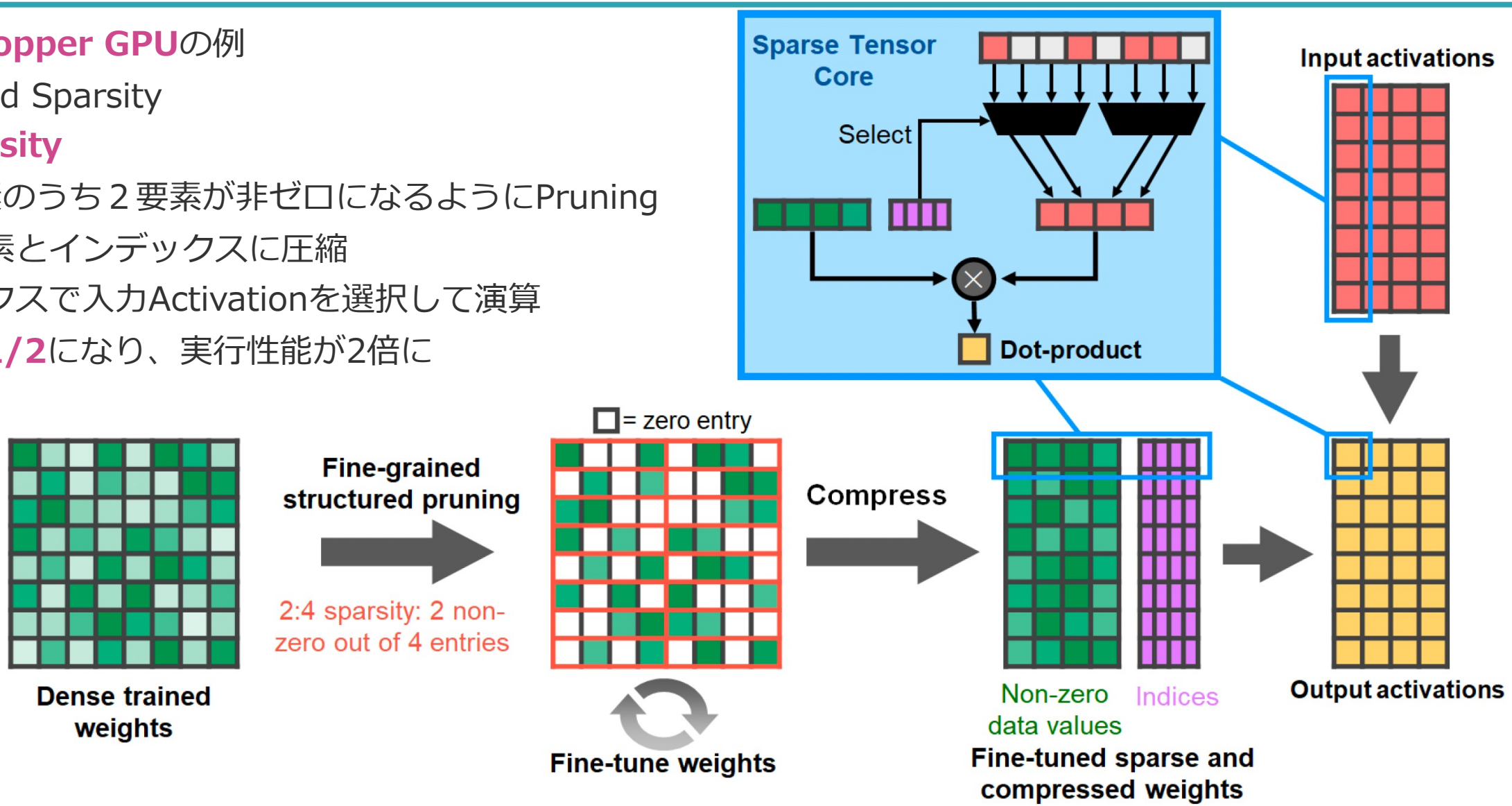
- Google TPUv4の例
  - HBM2メモリコントローラと同数の４つの第３世代SparseCore（インテリジェントなDMAコントローラ）
  - SparseCoreがpod全体の共有メモリを活用
  - 全ノード対象の数百万の並行アクセス
    （4096チップのTPUv4 podの場合）



出典：Norman Jouppi & Andy Swing, "A Machine Learning Supercomputer With An Optically Reconfigurable Interconnect and Embeddings Support," HOT Chips 2023

スパース性活用

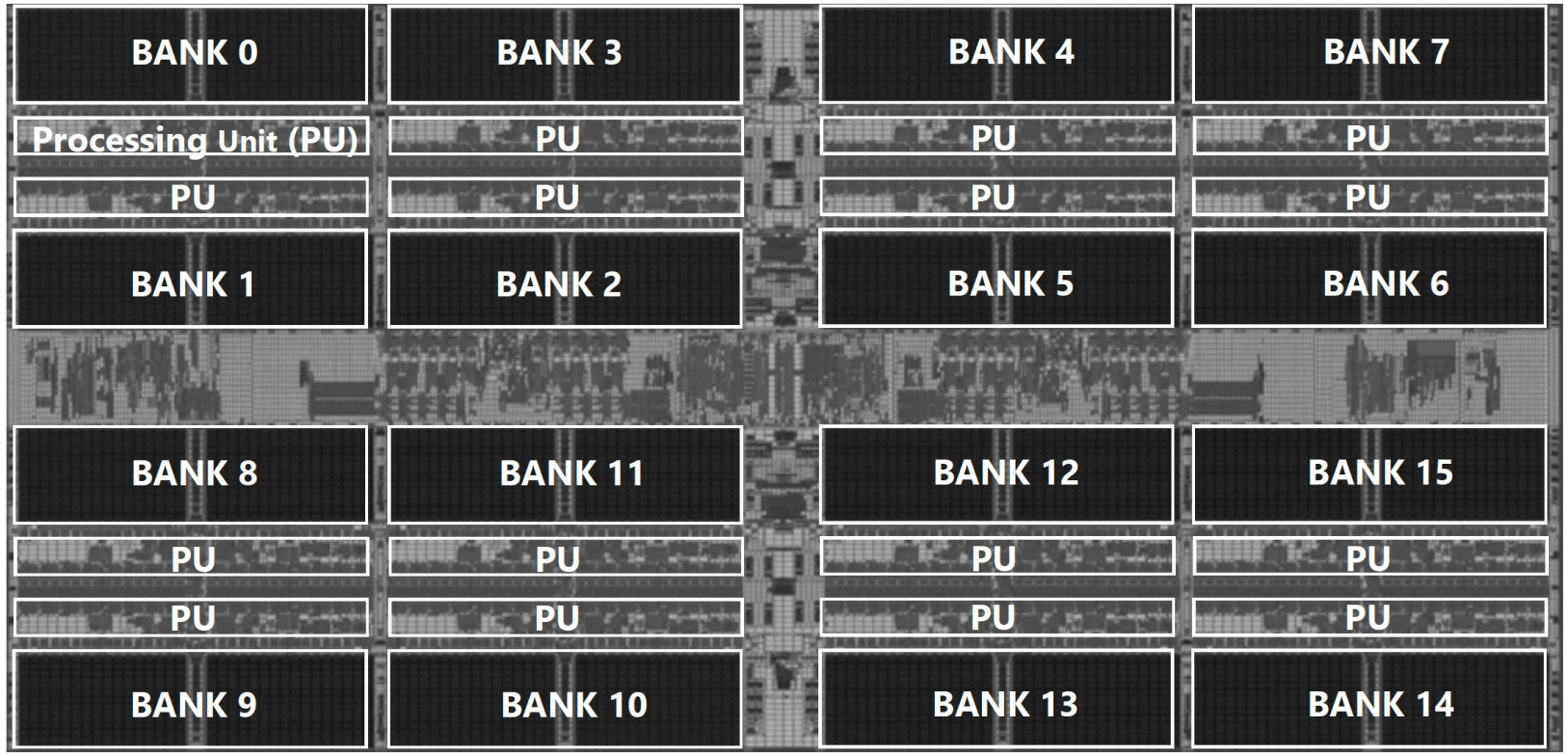
- NVIDIA **Hopper GPU**の例
  - Structured Sparsity
  - **2:4 Sparsity**
    - 4要素のうち２要素が非ゼロになるようにPruning
  - 非ゼロ要素とインデックスに圧縮
  - インデックスで入力Activationを選択して演算
  - **演算量が1/2**になり、実行性能が2倍に



出典：Bill Dally, "Hardware for Deep Learning," HOT Chips 2023 (Aug 2023)

Processing In-/Near-Memory（PIM/PNM）

- SK Hynixの最初のProcessing-In-Memory製品：GDDR6-AiM
- 演算性能（512GFLOPS, BF16）＝ メモリバンド幅（512GB/s）

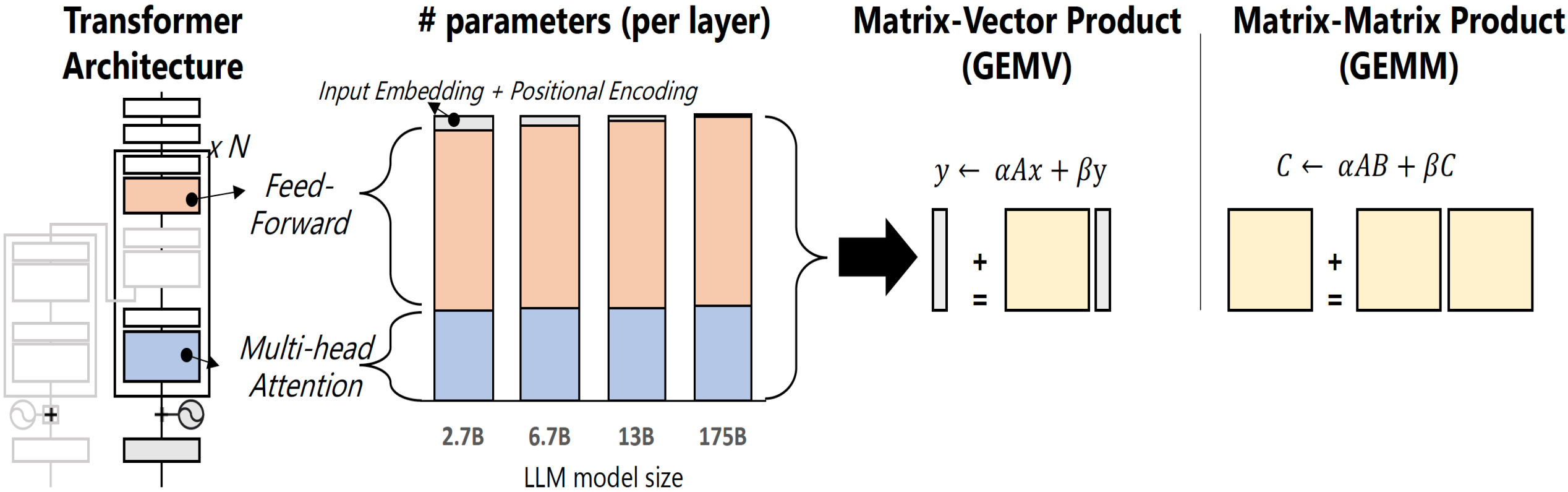


GDDR6-AiM* (per die)	
DRAM Type	GDDR6
Process Technology	1y
Memory Density	4Gb
Organization	X16
IO Data rate	16 Gbs/pin (@1.25V)
(External) Bandwidth**	32 GB/s
Operating Speed	1 GHz
Processing Unit (PU)	16 PU/die
Compute Throughput**	512 GFLOPS
Internal Bandwidth**	512 GB/s
Numeric Precision	BF16
Activation Function support***	Sigmoid, tanh, GELU, ReLU, Leaky ReLU, ...

出典：Yongkee Kwon, "Memory-centric Computing with SK Hynix's Domain-Specific Memory," HOT Chips 2023 (Aug 2023)

Processing In-/Near-Memory（PIM/PNM）

- LLMのTransformer処理のFeed-ForwardとMulti-head Attentionにおいてレイヤー毎に2.7B～175Bのパラメータでgeneral matrix-vector product（GEMV）とgeneral matrix-matrix product（GEMM）を実行
- 特にGEMVがメモリアクセスネックになる

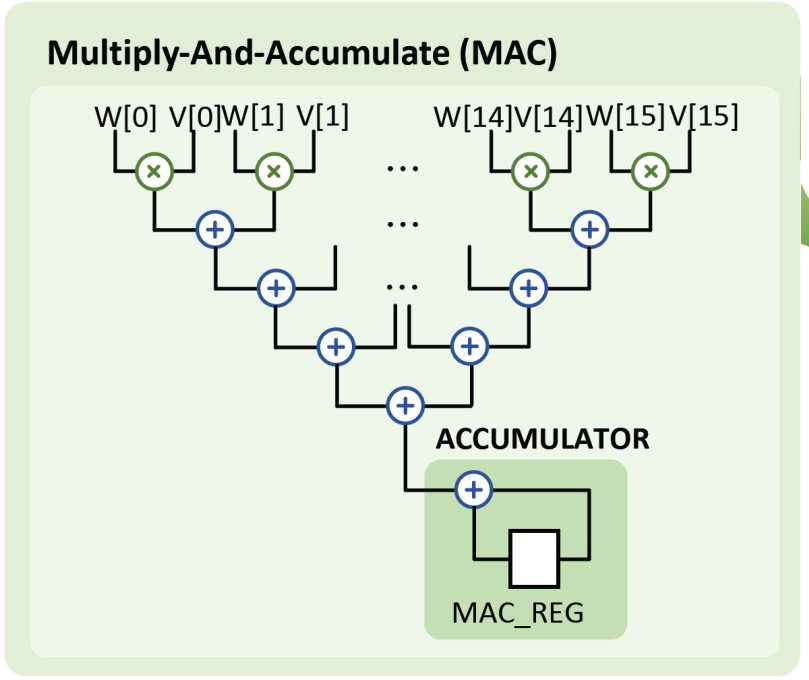


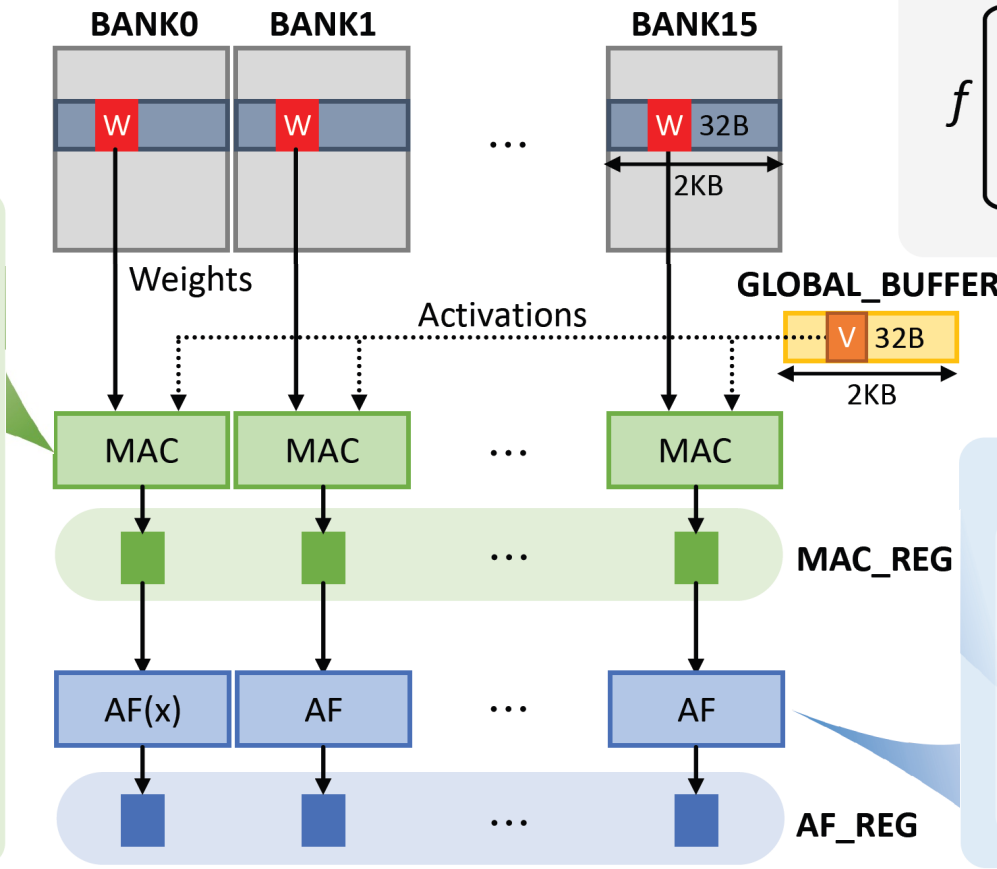
出典：Yongkee Kwon, "Memory-centric Computing with SK Hynix's Domain-Specific Memory," HOT Chips 2023 (Aug 2023)

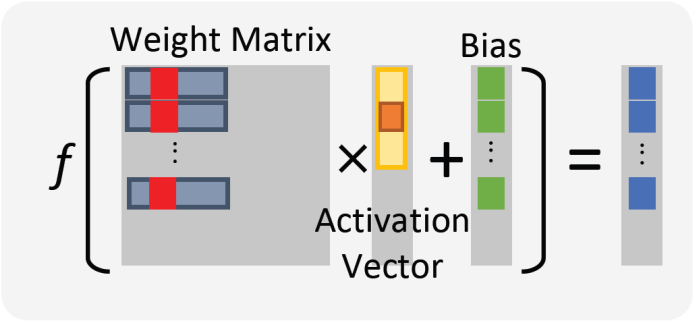
Processing In-/Near-Memory（PIM/PNM）

- GEMV（重み行列×ベクトル+バイアス）と活性化関数（AF）をIn-Memory処理
- 重み行列を16バンクから、ベクトルをGlobal Bufferから、それぞれ2B（BF16）16要素、計512B供給
- MAC REGをバイアスに初期化して４回廻せば、2kBのベクトルが求まり
  AFに通して、AF REGに格納

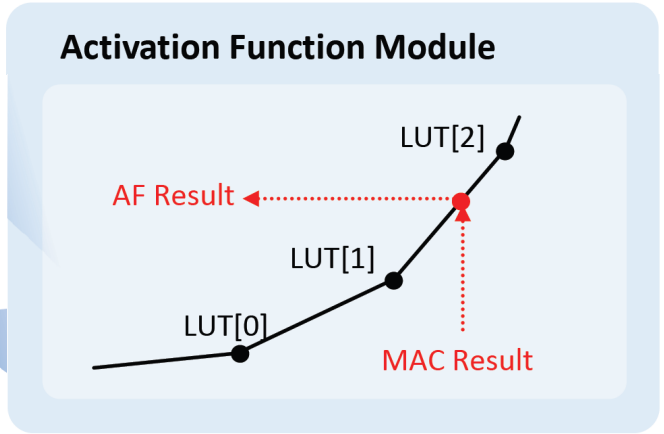
- MAC：32演算（16乗算, 16加算）×
  16バンク × 1GHz = 512GFLOPS







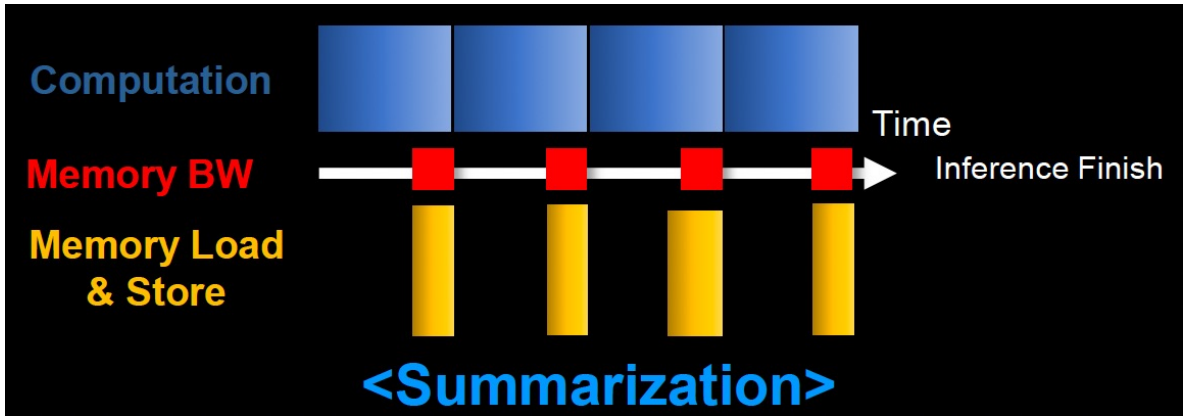
- AF：Sigmoid, tanh, GELU, ReLU, Leaky ReLUなど

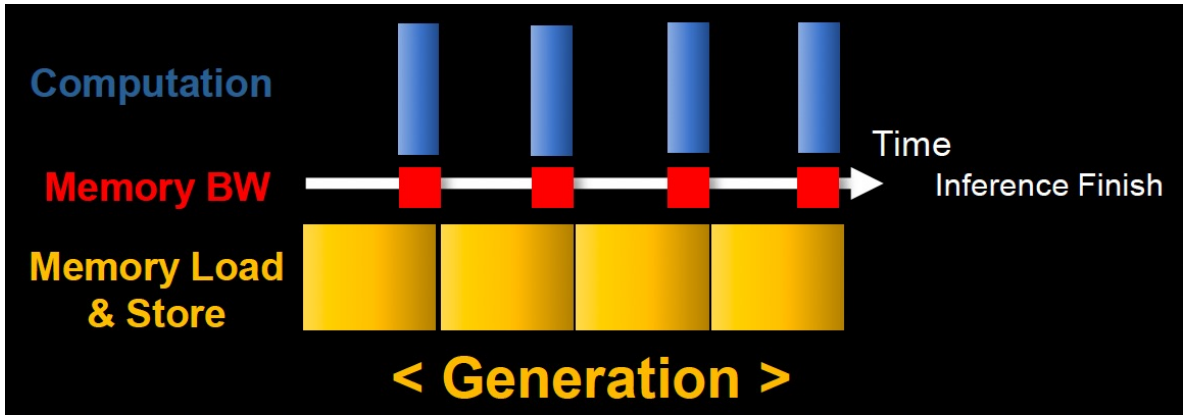


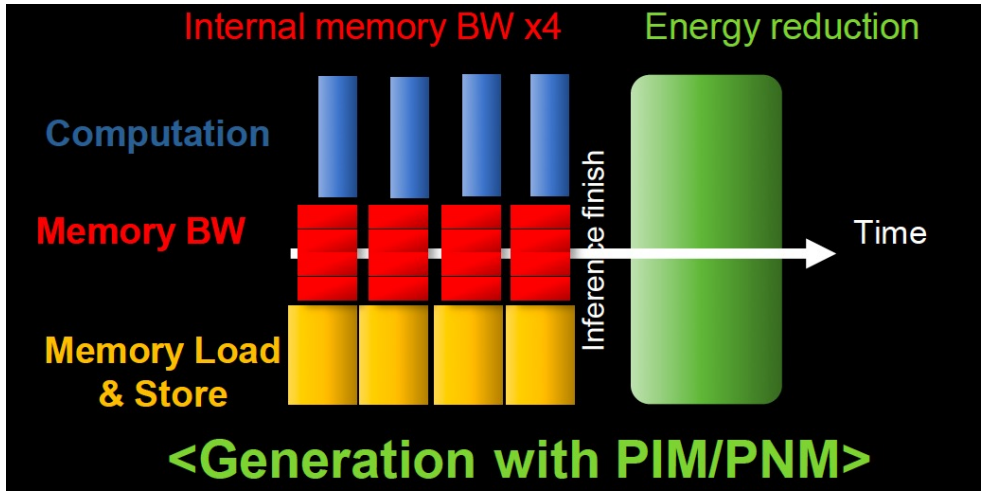
出典：Yongkee Kwon, "Memory-centric Computing with SK Hynix's Domain-Specific Memory," HOT Chips 2023 (Aug 2023)

Processing In-/Near-Memory（PIM/PNM）

- メモリー演算コア間の転送量削減により高効率化（転送時間>>演算時間の場合に有効）
- TrnsformerのSummarizationは演算ネック、Generationはメモリアクセスネック
- GenerationをPIM/PNMで高速化・エネルギー削減



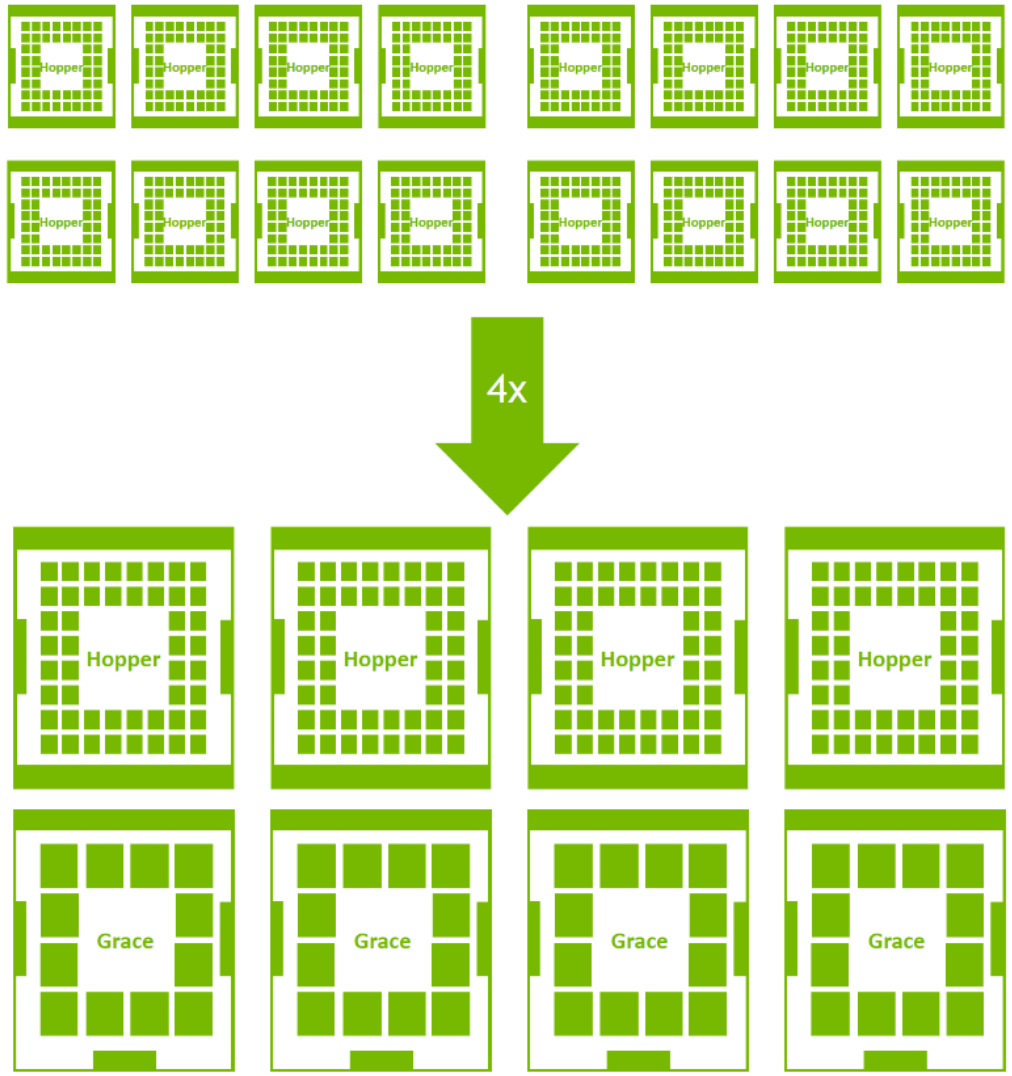




出典：Jin Hyun Kim, "Samsung AI-cluster system with HBM-PIM and CXL-based Processing-near-Memory for transformer-based LLMs," HOT Chips 2023 (Aug 2023)

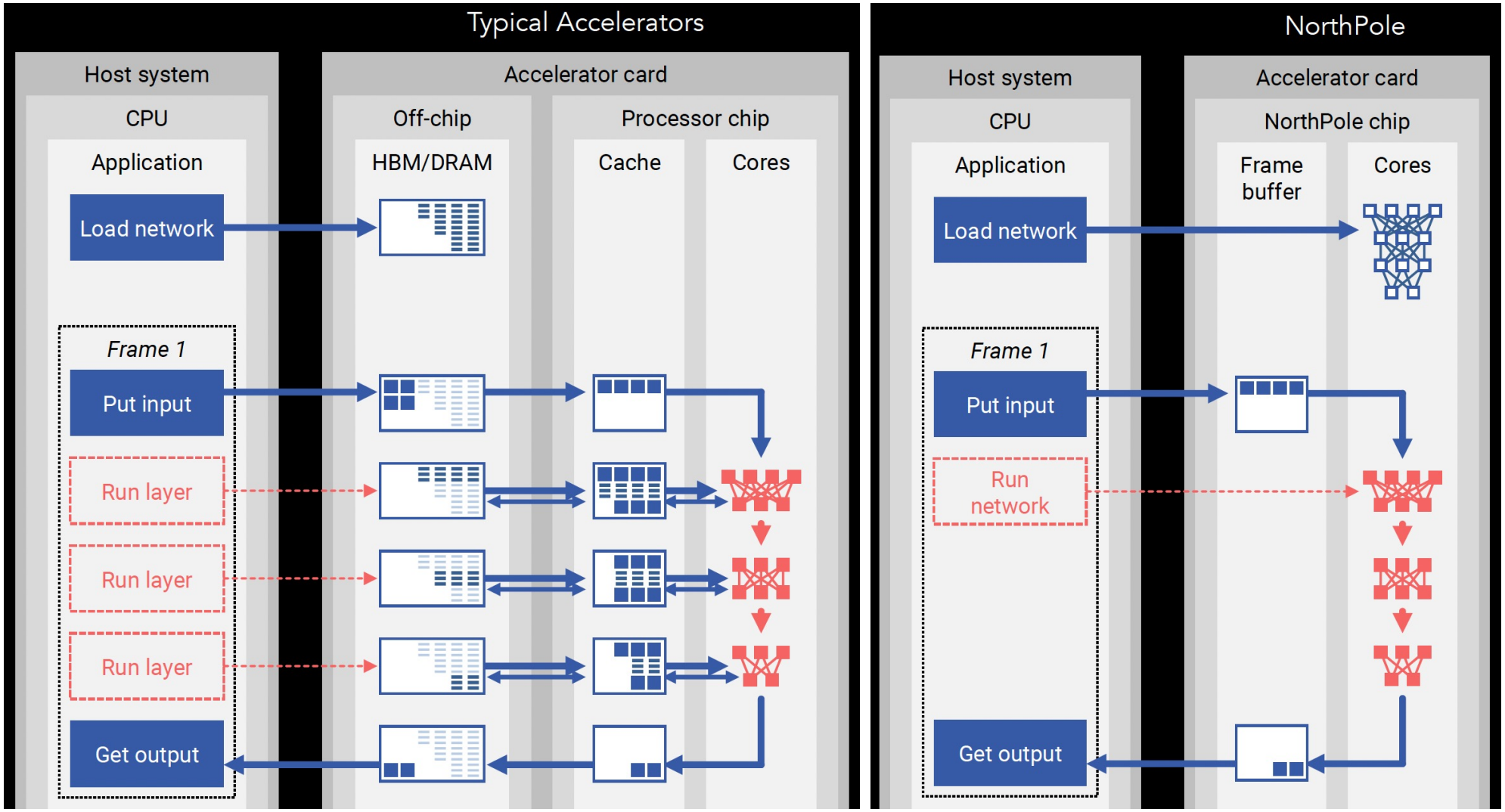
汎用プロセッサとAIアクセラレータの密結合

- NVIDIA **Grace Hopper Superchip**
- 自然言語処理の**GPT-3**学習の例では**2.5TB**のメモリが必要
- Hopperのみでは**16チップ**必要なのに対して**4ペア**でOK
- Hopper GPU：
  - テンソルコア：ピーク性能1PFLOPS（INT8, FP8）
  - メモリインタフェースはHBM3、3TB/s、80GB
- Grace CPU：
  - NVIDIA初のサーバー用CPU、Arm v9.0コアを72個集積
  - メモリはLPDDR5x (32-channel), 546GB/s, 512GB
  - Hopper GPUと**NVLink-C2C**で高速接続（**900GB/s**）
- 大規模データ供給は**Grace**のような汎用ハイエンドプロセッサが得意だが、効率が悪い
- アクセラレータが**大規模データ対応**することが望ましい



出典：Jonathon Evans, "NVIDIA's Grace CPU," HOT Chips 2022 (Aug 2022)

推論プロセッサの例：IBM NorthPole



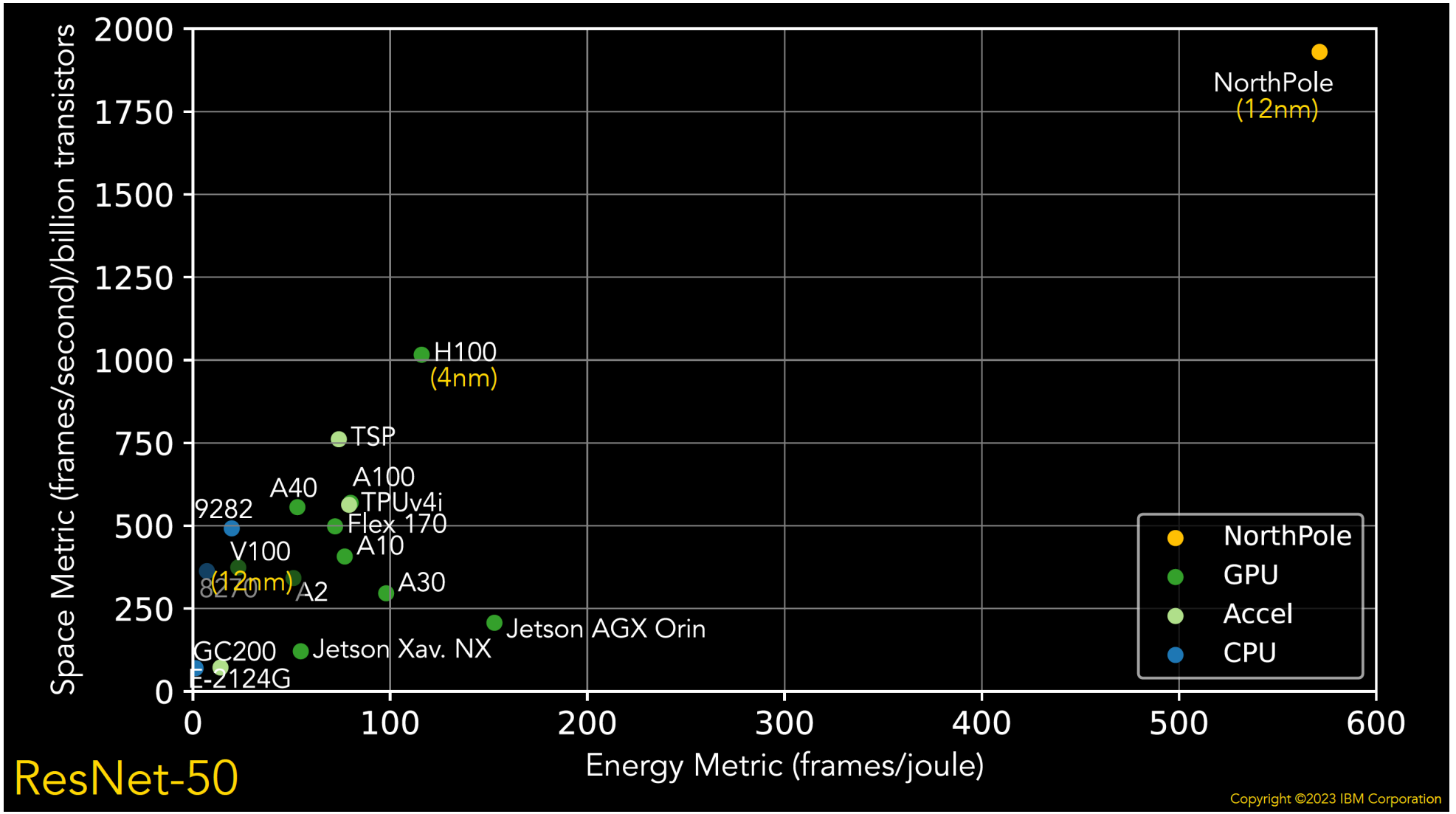
- 小規模アプリに特化して高効率化
- アクセラレータに学習結果を全てロードして保持

- 外部メモリ不要
- レイヤー毎の命令不要
- 実行中のデータ転送不要

（キャッシュ ⇄ コア
キャッシュ ⇄
外部メモリ）

出典：Dharmendra Modha, "IBM NorthPole Neural Inference Machine," HOT Chips 2023 (Aug 2023)

推論プロセッサの例：IBM NorthPole

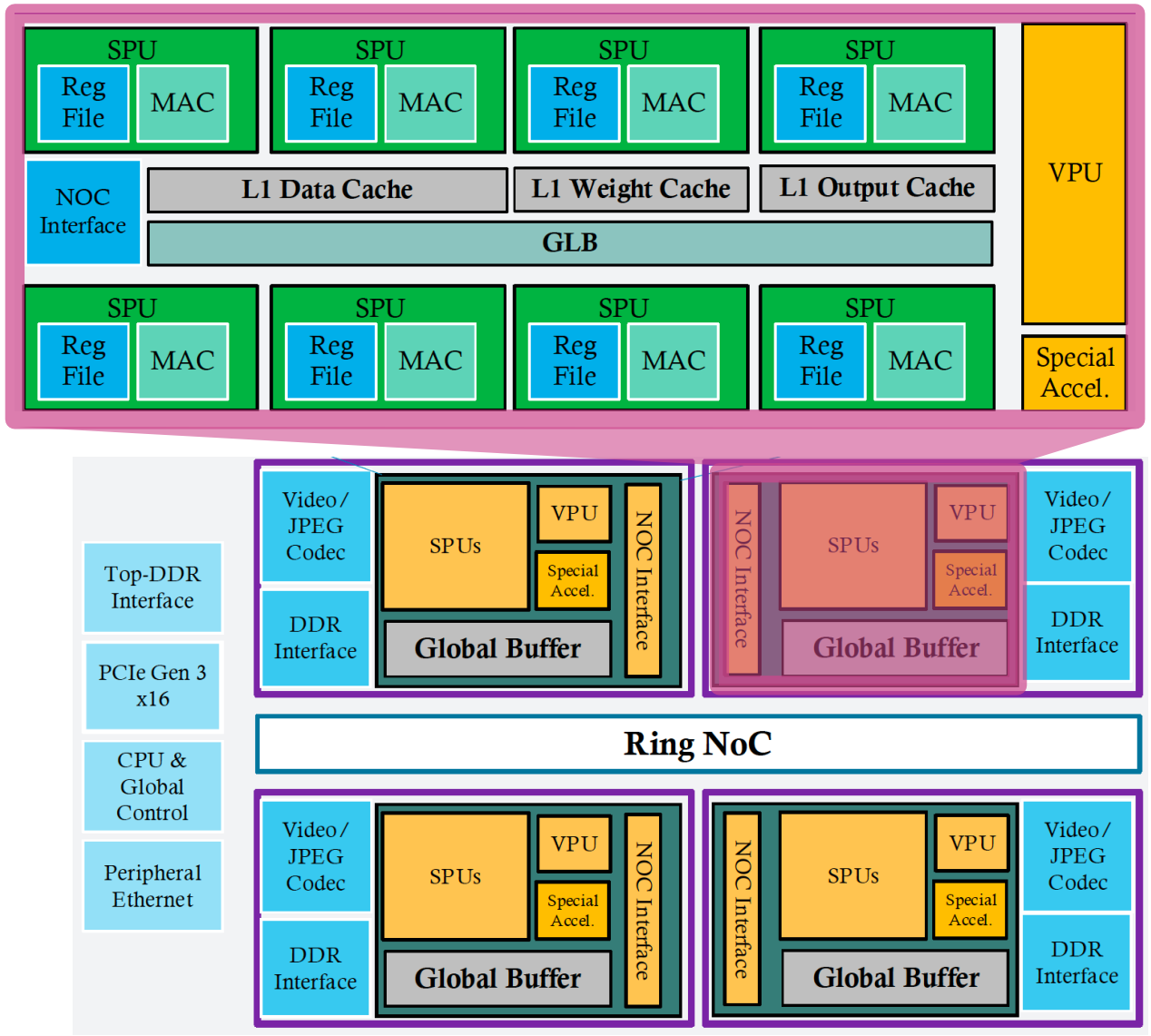


- トランジスタ効率、エネルギー効率で他を圧倒
- 大規模データを扱うGPT、LLMには不向き

出典：Dharmendra Modha, "IBM NorthPole Neural Inference Machine," HOT Chips 2023 (Aug 2023)

推論プロセッサの例：Moffett Antoum

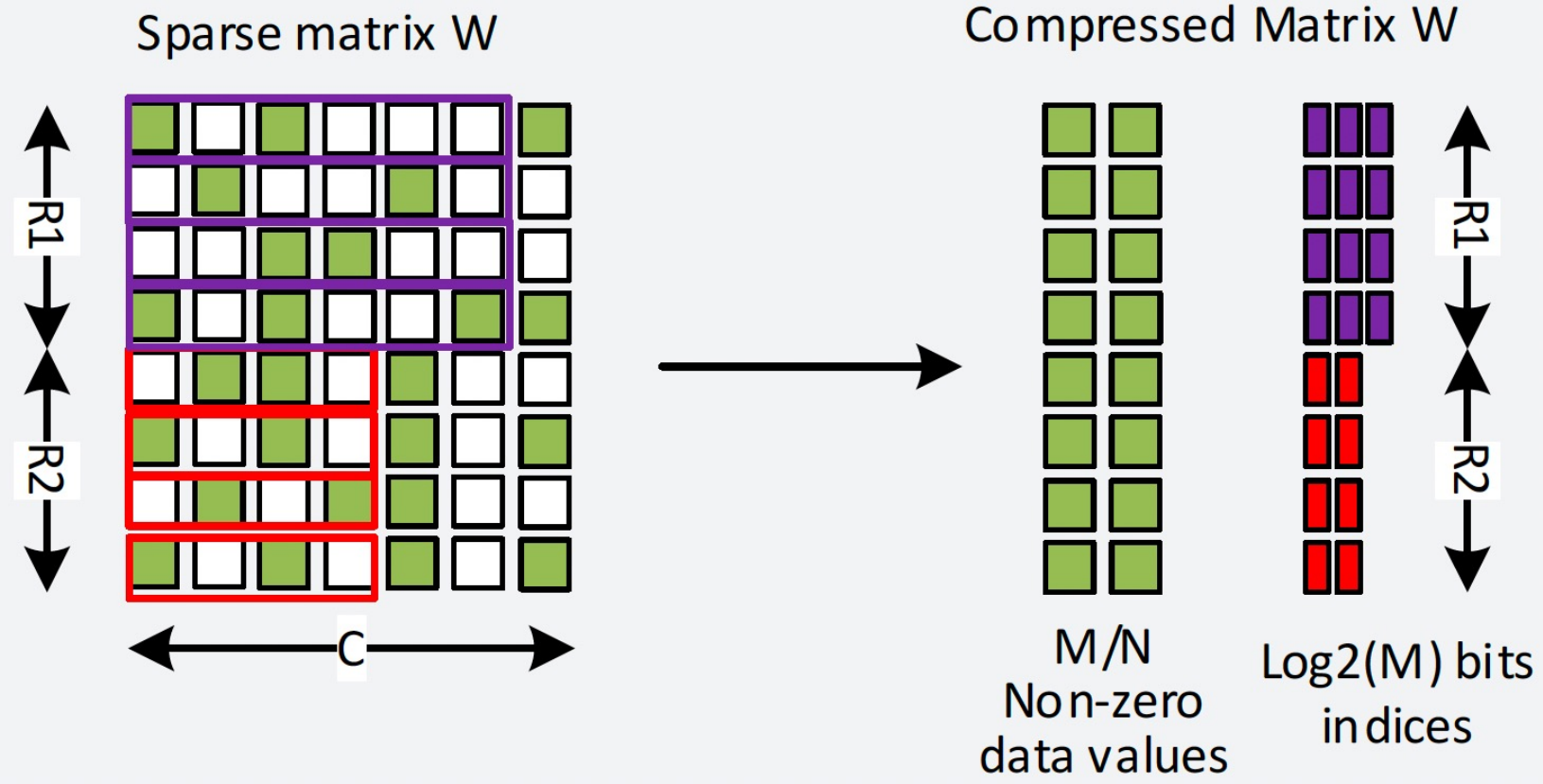
- NNCore Subsystem
  - 8x Sparse Processing Units (SPUs)
  - 1x CPU、1x VPU (ベクトル長：512 Byte)
  - Special Accelerators (活性化関数など)
  - 階層メモリ：Reg File, L1, Global Buffer
- Full SoC
  - 4x NNCore Subsystems
  - 16x PCIe Gen3
  - 5x LPDDR4x (84GB/s, 20GB)
  - 4x ARM CPU (Linux)
  - 800MHz, 70W (TDP), 340mm^2@12nm
- ピーク性能
  - 29.5TOPS（INT8）
  - 14.7（SPU）+ 3.7（VPU）TFLOPS（BF16）



出典：Zhibin Xiao, "Moffett Antoum: A Deep-Sparse AI Inference System-on-Chip for Vision and Large Language Models," HOT Chips 2023 (Aug 2023)

推論プロセッサの例：Moffett Antoum

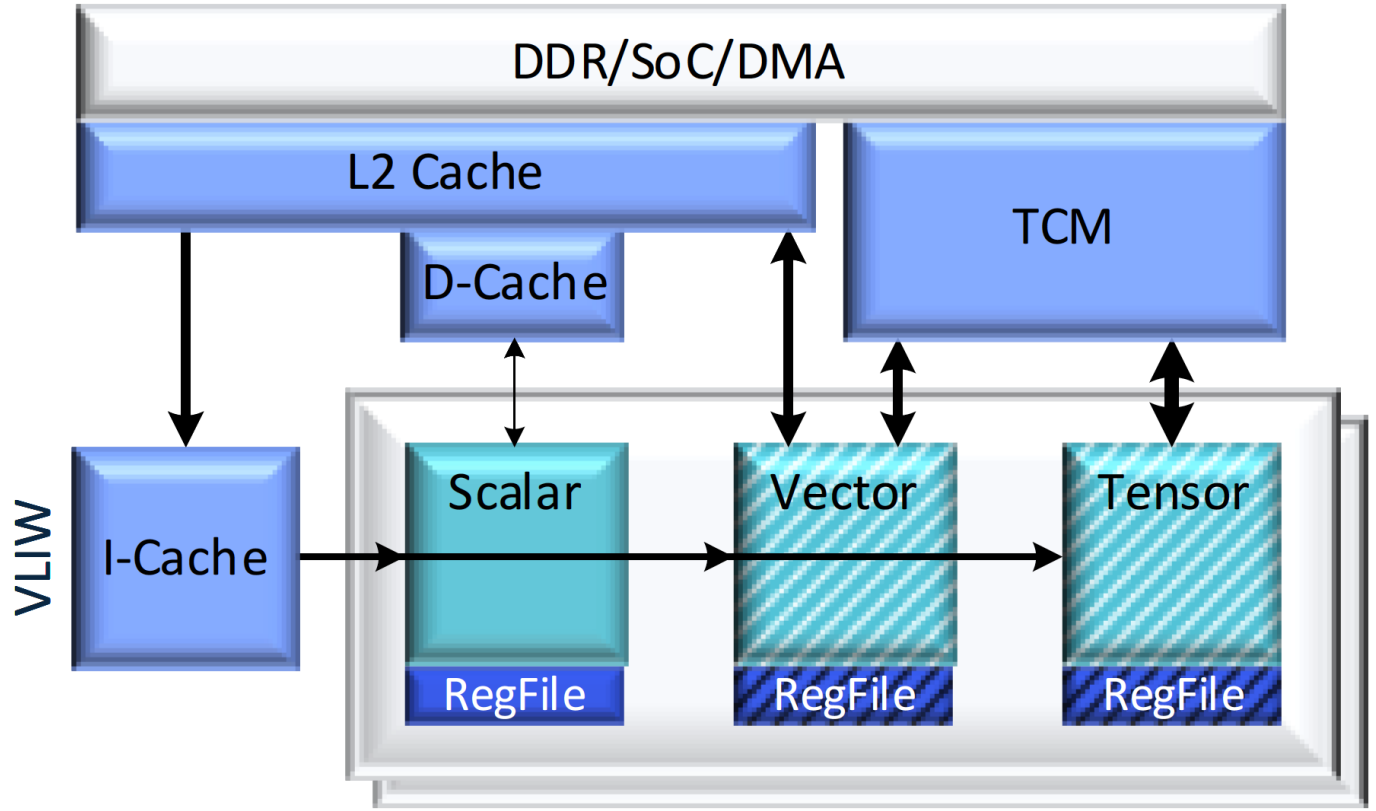
- スパース性活用
  - Hopper GPUと同様のStructured Sparsity
  - **1/N Sparsity** (M要素のうちM/N要素が非ゼロ、M：16,32,64、N：1 to 32)、**演算量が1/N**
  - 畳込み、逆畳込み、行列演算に対応



出典：Zhibin Xiao, "Moffett Antoum: A Deep-Sparse AI Inference System-on-Chip for Vision and Large Language Models," HOT Chips 2023 (Aug 2023)

推論プロセッサの例：Qualcomm Hexagon NPU

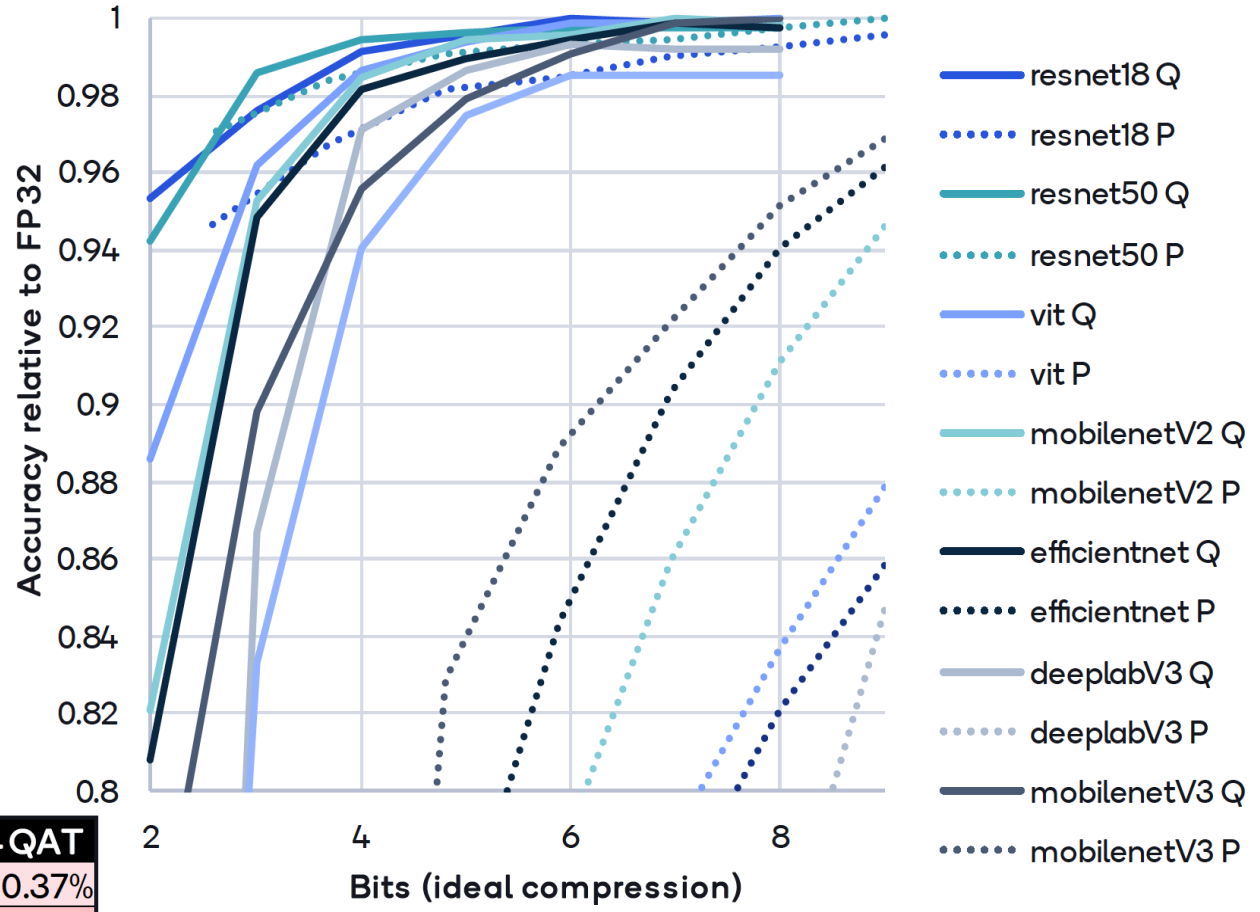
- Scalar, Vector, Tensorの３命令セット、それぞれ専用ハードウェア（演算器、レジスタ等）
- Vector：VLIW (演算4、ロード、ストア)、128Byte SIMD、Scatter/Gather、 L2直接アクセス
- Tensor：2/3/4次元対応、INT4/8/16、FP16、混合精度、行列乗算、畳込み、活性化関数など
- TCM：Vector/Tensor用Tightly-Coupled Memory、大容量、ソフトウェアプリフェッチ、Scatter/Gather



出典：Eric Mahurin, "Qualcomm® Hexagon™ NPU," HOT Chips 2023 (Aug 2023)

推論プロセッサの例：Qualcomm Hexagon NPU

- 量子化を積極活用（右図）
  - 様々なアプリでPruningと比較
  - 同じ圧縮率では量子化の方がAccuracyが高い
  - Pruningもサポートしているが
- 量子化方式
  - PTQ (Post Training Quantization)
  - QAT (Quantization Aware Training)
- Accuracy比較（下表）
  - CNN, RNN系モデルはINT8 QATが高い
  - Transformer系はFP8E4 QATが高い



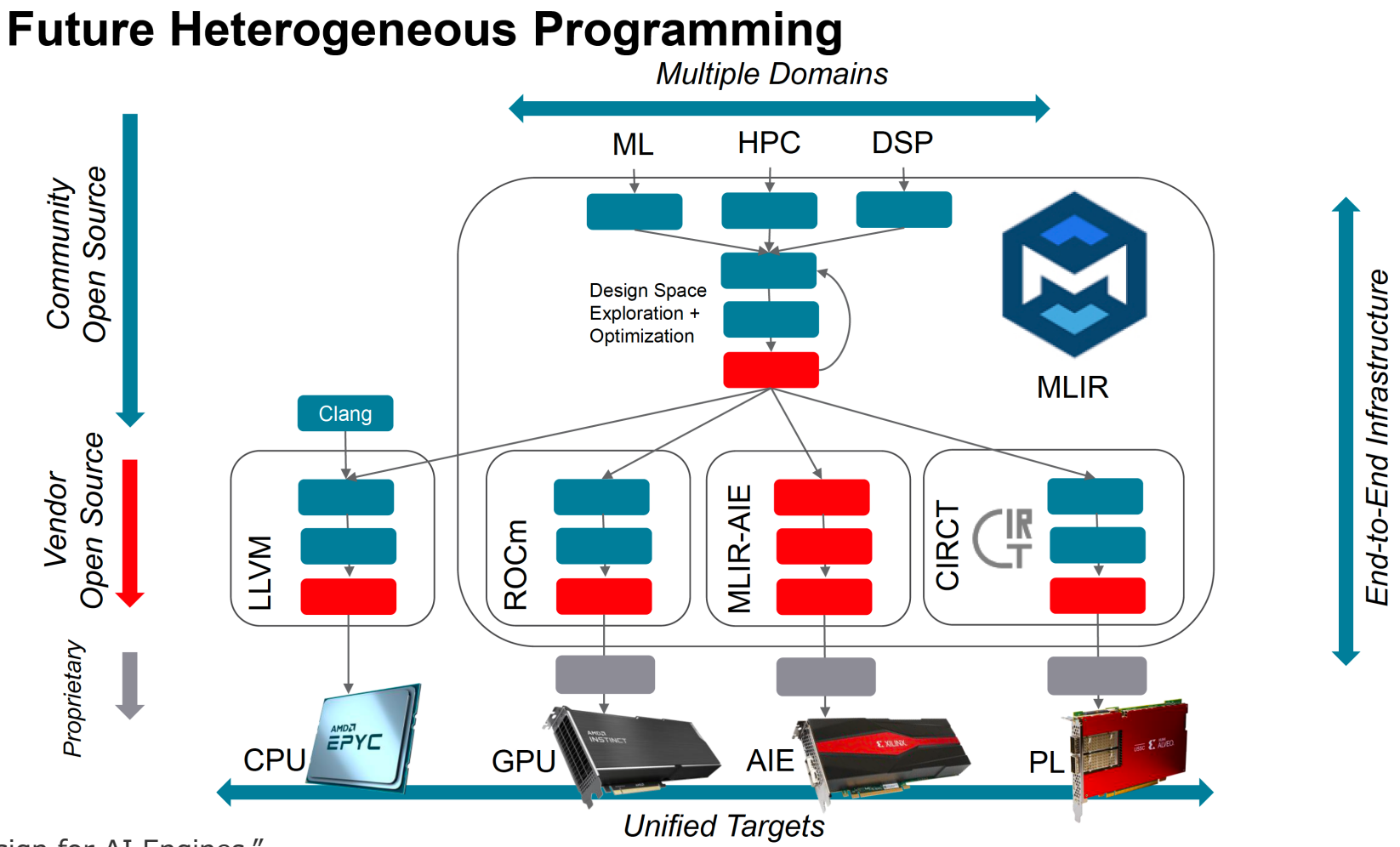
Model	FP32	INT8 PTQ	FP8E4 PTQ	A8W4 QAT	INT8 QAT	FP8E4 QAT
ResNet18	69.72%	-0.08%	-1.15%	0.29%	0.71%	-0.37%
MobileNetV2	71.70%	-0.76%	-5.65%	-0.53%	0.12%	-0.81%
HRNet	81.05%	-0.12%	-0.28%		0.22%	0.01%
DeeplabV3	72.91%	-1.67%	-34.98%	0.10%	1.08%	0.31%
SalsaNext	55.80%	-1.58%	-0.68%		-0.80%	-0.60%
BERT(GLUE)	83.06%	-12.03%	-0.26%	-0.42%	0.20%	0.85%

A8W4: Activation 8b, Weight 4b
FP8E4: 指数部が4bのFP8

出典：Eric Mahurin, "Qualcomm® Hexagon™ NPU," HOT Chips 2023 (Aug 2023)

AMDのMLIR活用例

- ML, HPC, DSPなど複数ドメインの様々な言語のプログラムをMLIRを使って変換・最適化を繰り返して、様々なターゲットハードウェアのコードにコンパイル
- ハードウェア非依存の部分はコミュニティオープンソースを活用
- 依存部分はベンダーオープンソースを提供
- 新アーキテクチャや複数アーキテクチャのサポートが容易に



出典： Jack Lo, et al., "Leveraging MLIR to Design for AI Engines," Tutorial T2, FPGA 2023 (Feb 2023)

アーキテクチャ・研究開発の変化

Trend	Issue	from	to
Open Innovation	R&D	Closed	**Open**
	ISA, Architecture	Proprietary	**Open**
	Interface	Proprietary	**Open Standard**
	Competitive design	Full custom	**Expertised part**
Multi-many cores (Distributed Processing)	Processing style	Master/Slave	**Distributed**
	Operation start	Code driven	**Data driven**
	Data/Code transfer	Pull	**Push**
Time to Space	Optimizations	Order	**Place**
	Process switch	Interrupt	**Sleep/Wake**
Finer process	Major cost	Processing	**Transfer**
	The Same data	Copy	**Regenerate**
Power limitation	Activity	Full utilization	**Dark Silicon**
	Core micro-arch.	Complex	**Simple**

まとめ

- **プロセッサ研究開発の潮流**
  - AI応用の急速な発展により**新アーキテクチャ**のプロセッサが多数の機関から発表・製品化
  - **MLIR**の標準化によりコンパイラの新アーキテクチャや複数アーキテクチャサポートが容易に
  - RISC-Vの普及によりアーキテクチャの**オープン化**が進展
  - サーバー系を中心に**Chiplet**化が進展、**UCIe**が標準化、**CXL**が台頭
- **組込みマルチコア**
  - ADAS、自動運転、ロボットなどの進化により、組込み系にも**AI応用**を中心に**高性能化**要求
  - **AI推論に適したレスポンス重視の高効率な組込みマルチコア**を研究開発すべき
- **AI向けプロセッサ技術**
  - 演算量削減：高並列**低ビット**演算、**混合精度**演算、**スパース性**活用
  - 転送量削減：Processing **Near-/In-Memory**、メモリ階層の最適化、内蔵メモリの大容量化
- **推論プロセッサの例**（HOT Chips 2023より）
  - IBM: North Pole、Moffet AI: Moffett Antoum、Qualcomm: Hexagon NPU
- **アーキテクチャ・研究開発の変化**

ご清聴
ありがとうございました