☆ 委員会報告☆

ニア・メモリ・ コンピューティング

論説委員会

近年メモリ技術が進展し、「CPU 演算回路のごく近傍に大容量のメモリを配置する」、あるいは「メモリを演算回路として使用する」などのニア・メモリ・コンピューティングが現実のものになってきている。特にエッジ側で動かす AI の分野では、高速処理と低消費電力が必須であることから、この分野に関して非常に大きな関心が集まっている。そこで本報では、ニア・メモリ・コンピューティングの現状をまとめ、この分野における我が国のステータスを論説する。

ニア・メモリ・コンピューティングの必要性

通常 CPU と主メモリは個別のチップであるが、両チップ間を往来するチップ外信号に関しては、チップ内信号と比べて信号伝搬速度が極端に遅くなるだけではなく、信号伝搬に伴う電力消費も桁違いに大きくなる。信号伝搬媒体である信号当たりの平均配線長が 5~6 桁違うため、配線に起因する寄生容量値と寄生抵抗値がチップ内外で極端に異なることにその理由がある。1940 年代にプログラム内蔵型の計算機が誕生したが、当初から現在まで言われ続けてきたいわゆる「von Neumann bottle-neck (メモリ・アクセスがシステム性能上のネック)」である。スマホなどユーザの手元で大量の情報処理を行うようになった現在、このボトル・ネックを何としても排除しなければならない状況になったのだ。対応策は明瞭で、「演算回路とプログラム及びデータが格納されているメモリを極限まで近傍に配置する」ことである。これを実現したのがニア・メモリ・コンピューティングであるが、どのような実現方法があるのか、次節で分類してみよう。

ニア・メモリ・コンピューティングの形態

大きく分類すると図1のように、「チップ・レベルで近傍に配置する方法」と「シリコン内にモノリシック状態で近傍に配置する方法」とがある。以下合計5種類ある形態を順に説明する。(A)はHBM(High Bandwidth Memory)と呼ばれるもので、Encore 第101号p29でも述べたようにNVIDIAやAMDなどAI関連チップでは広く採用されている形態だ。Si インターポーザあるいは、最近ではSiブリッジを使い、TSV(Through Silicon Via)を用いて積層されたDRAM群とCPU/GPUチップを接続している。TESALAはこのHBM構造で、10^{12}バイト / 秒のメモリ・バンド幅を実現している。ここで、メモリ・バンド幅とは(バイト数/メモリ・バンク)×(メモリ・バンク数) ×(クロック周波数)のことである。この広帯域幅によって8ビット整数演算で1兆回 / W の高速・低消費電力チップを実現しており、Level 3 以上の自動運転用車載チップへの要求性能値(Encore 第102号p26参照)をクリアしたものとなっている。(B)は既存のCPUチップとメモリ・チップを積層したもの。HBMと同様にTSV を用いてチップ間の信号伝搬を実現している。(C)はシリコンの層として演算回路層とメモリ層を積層した形態、(D)はCPUチップ内に大容量のメモリを搭載した形態、(E)はメモリを演算機能として使用するCIM (Computing in Memory)と呼ばれる形態である。

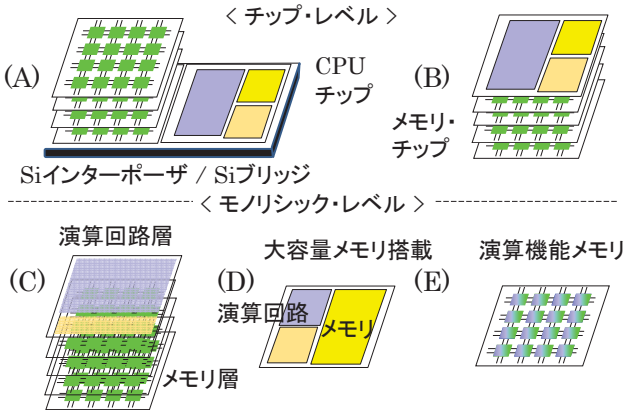

図1 ニア・メモリ・コンピューティングの形態

モノリシック・メモリ・コンピューティングの実例

ここで、モノリシック・メモリ・コンピューティング形態(C)、(D)、(E)の実例をそれぞれ挙げることにする。

図2 はISSCC2018でStanfordとUC Berkeleyから報告された形態(C)の「文章自動認識用AIチップ」だ。CMOS-Trを使った論理回路の上空にエンベデッド・メモリとしてReRAM(Resistive RAM) を積層し、最上部にはゲートとしてカーボン・ナノ・チューブを使った論理回路を積層している。「チップ・レベルで積層する形態(B)と比べて、4.5倍高速で、消費電力が1/8になった」としている。

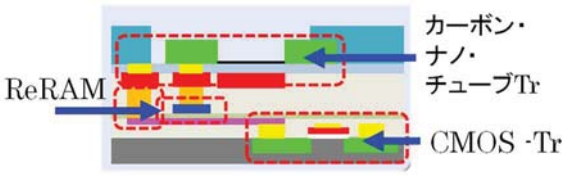

図2 エンベデッドReRAMを内蔵するAIチップ

図3は2019年に東北大学から報告された形態(D)のCPUである。ARMをプロセッサ・コアとし、ユーザ論理はFPGA(Field Programmable Gate Array)で構成している。プロセッサのキャッシュとFPGAの構成データ用メモリとしてSTT-MRAM (Spin Transfer Torque MRAM)を用いてい

る。プロセッサ・コア、FPGA、そしてSTT-MRAMを数ミリ角のシリコン上に搭載しているため、「200MHzで50μW」と驚異的な性能を実現している。「GHzオーダのクロックでも、プロセッサと外付けDARM組み合わせの消費電力と比べれば、桁違いに小さなものになる」との報告だ。

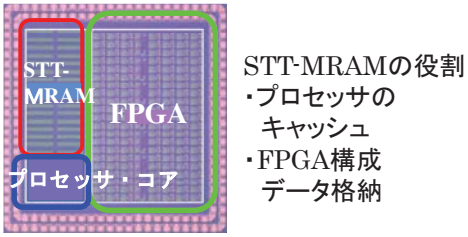

図 3　大容量 STT-MRAM を搭載する CPU チップ

図 4 は ISSCC2018 で National Tsing Hua Univ.から報告された形態 (E)の「手書き文字自動認識用 AI チップ」だ。複数のアドレスを同時に指定すると、メモリ・セルに記憶された複数の情報がデータ線上に電流として同時に流れるが、電流は加算されることから、センス・アンプに入る電流値はメモリ・セルに記憶された内容に対応して、様々な値となる。この振る舞いを巧妙に使い、メモリを演算回路として使うアイデアだ。メモリとして ReRAM を用いている。また通常 8 ビットの整数で neural net の重みを記憶するところ、"+1"と"－1"の 2 ビットに重み情報を圧縮した BNN(Binary Neural Net)を採用している。手書き文字自動認識システムのベンチ・マーク MNIST において 95%以上の認識率を実現し、2 ビット整数演算で 53 兆回 /W の高速・低消費電力を実現したチップだ。IBM は SRAM を用いた同様のチップ「TrueNorth」を 2016 年に発表しているが、それよりも一段と優れた性能を発揮している。いずれにせよ「ここ数年の ISSCC において、CIM を利用した AI 推論用チップが毎年報告されている」ことからも分かるように注目の技術だ。

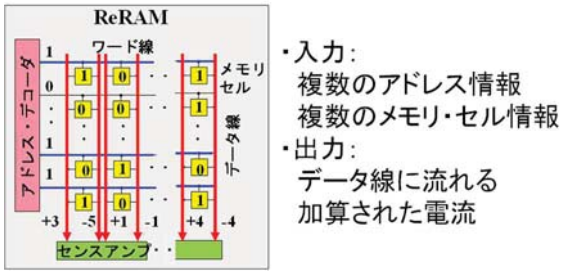

図 4　ReRAM を用いた CIM

モノリシック・メモリ・コンピューティングを実現する鍵

モノリシック・メモリ・コンピューティングとしての例を挙げたが、いずれも「不揮発性次世代メモリ」だから実現可能になったものだ。ReRAMとSTT-MRAMは次の特徴がある。

(1) 1ビットを記憶するSRAMのメモリ・セル・サイズ70F² (Fは配線ピッチの1/2の寸法) に対して「ReRAM は 4F² 、STT-MRAMは8F²」と1桁小型
(2) メモリ単体としての読み・書き時間はDRAM並み
(3) メモリ単体としての読み・書き消費電力は30μW

メモリ・セル・サイズが小さいことから、大容量エンベデッド・メモリが可能になる。しかも、メモリ単体としての高速処理と配線遅延の低減が相乗効果を発揮し、同様にメモリ単体としての低消費電力と配線に起因する消費電力の低減が相乗効果を発揮して、圧倒的な性能を実現している。

我が国のステータス

次世代メモリの中でも STT-MRAM に代表される強磁性体メモリの研究分野では、我が国の大学が大いに活躍している。 この分野に関しては、日本学術振興会が主催する OPERA プロジェクト及び JSPS プロジェクト, 内閣府が主催する SIP プロジェクト、そして NEDO が主催するプロジェクト、など産官学共同プロジェクトが進行中であり、研究レベルだが、優れた結果が数多く報告されている。

しかし過去を振り返ると、学術的に優れた成果が出ても、ビジネスに結びつかなかった例が山ほどある。今は米国・ヨーロッパのみならず、中国・台湾や韓国における実用化へのスピード感が凄い。例えば、「エンベデッド・メモリとしの STT-MRAM」に関しては、すでに TSMC、UMC、Intel、Samsung などの海外メーカが開発に乗り出している。特に TSMC は ISSCC2020 において「従来の論理マスクに加えて 2～5 枚の追加マスクで 32 ビット・エンベデッド STT-MRAM を開発」と発表しているが、製品化に向けてのこのスピード感が必要なのである。

「ISSCC の論文採択件数が半導体技術レベルを反映している」と言えるが、2020 年の採択件数は、米国が 71 件、韓国が 35 件、中国（香港、マカオを含む）が 23 件、台湾が 22 件、日本が 12 件となっている。米国が減少傾向にある中、韓国は前年から 10 件の増加、中国も 5 件増やしている。「ビジネスだけではなく、技術レベルでも他国の後塵を拝し始めている」との感じが否めない。

まとめ

ニア・メモリ・コンピューティングの現状を概観し、不揮発性次世代メモリとの関連を述べた。過去のEncoreで何度となく指摘してきたが、この分野においても、危機感とスピード感を持って研究・開発そしてビジネスに臨まない限り、またもや外国勢にしてやられる結果となるのだろう。

ご意見を論説委員会 ronsetsu@ssis.or.jp までお寄せ下さい。
論説委員：鈴木五郎(委員長)　渡壁弥一郎(副委員長)　井入正博
川端章夫　長尾繁雄　　吉岡信行　　野中敏夫(アドバイザー)