Forging the Way in AI Architecture with ReRAM Based Computational Memory

Hagop Nazarian
VP of Engineering, Cofounder
Crossbar Technology Inc.
AI Architectural Evolution

Embedded ReRAM with NN Processors
- Computational ReRAM array
- Monolithic solution at advanced CMOS nodes
- Interface bus size defined by AI System architects – not by memory or GPU manufacturers
- Large parallelism directly coupled to the processor
  - Yields huge bandwidth increase
  - Drastically reduces overall system energy consumption
ReRAM Bridging The Generational Gap

- eFlash is ~4 generations behind advanced Logic !!
- Flash technology does not scale with CMOS Logic process
- ReRAM is the non-Volatile memory choice for the advanced nodes in major foundries
- ReRAM Development already in Progress at least in 3 major foundries
What does an AI system Do?

AI systems can think and learn

• Training
• Inferring
• Classifying information
• Evaluating
• Use Low Energy and Low Latency
• Do all above in real-time at the edge
How does an AI system operate?

• To Train
  • Learns from the massive data
  • Establish relationships and trends
  • Interpolation and extrapolation models to be used for optimum solutions
  • Store them

• To Infer
  • Deduce estimated solutions or trends from observations and the trained model
  • Adapts to the environment and optimize the system
  • Store the new scheme

• To Classify information
  • Compares information, establishes relations, and stores them

• To Evaluate
  • Calculates, compares and finds the best given match conditions
AI Operations

- Add/Subtract/Multiply/Divide
- Comparison => Classification
- Randomization => to speed up searches
- Best fit, Best match
- Matrix operations – Convolution – Sparse
- Energy efficiency of the AI system
Example: Convolution

\[ G[i, j] = \sum_{u=-k}^{k} \sum_{v=-k}^{k} H[u, v] F[i - u, j - v] \]

This is called a **convolution** operation:

\[ G = H * F \]

- F is the image Matrix
- H is the Kernel Matrix
- G is the output
Mean Filter (Blurring)

1 1 1
1 1 1
1 1 1

Sharpening Filter

-1 -1 -1
-1  9  -1
-1 -1 -1

\[ H \ast F = G \]
Edge Detection

Source: Matlabtricks.com
Edge Detection Calculation

Original image data Matrix $F$

Edge detected image data $G = H \ast F$

$$\begin{array}{ccc}
0 & 1 & 0 \\
1 & -4 & 1 \\
0 & 1 & 0 \\
\end{array}$$

$$\begin{array}{ccc}
0 & 1 & 0 \\
1 & -4 & 1 \\
0 & 1 & 0 \\
\end{array}$$

$$\begin{array}{cccccccc}
70 & 26 & 92 & 8 & 68 & 7 & 38 & 67 & 59 \\
95 & 81 & 100 & 37 & 70 & 91 & 46 & 92 & 90 & 78 \\
60 & 72 & 68 & 69 & 61 & 72 & 23 & 43 & 23 & 79 \\
\end{array}$$

$$\begin{array}{cccccccc}
-35 & -6 & -25 & -19 & 3 & 14 & -37 & 13 & -1 & -7 \\
19 & -9 & 24 & 24 & -10 & -24 & 11 & 3 & 12 & 6 \\
-19 & 21 & -22 & -19 & 18 & 2 & 6 & -3 & -28 & -17 \\
-7 & -14 & 13 & -14 & 3 & 16 & -9 & 4 & 24 & 15 \\
5 & 3 & -11 & -1 & 16 & -24 & 21 & -16 & -18 & -23 \\
4 & 5 & -1 & -6 & -5 & -10 & -8 & 19 & -19 & -17 \\
-15 & 7 & 26 & -25 & 30 & -8 & 19 & 3 & 1 & 1 \\
-19 & 4 & -23 & 20 & -9 & -12 & 4 & -19 & -11 & -10 \\
-8 & -9 & -4 & -12 & -4 & -11 & 1 & 6 & 11 & -24 \\
\end{array}$$

$$\begin{array}{ccc}
70*(0)+26*(1)+7*(0)+95*(1)+81*(-4)+100*(1)+60*(0)+72*(1)+68*(0) \end{array} \ast (1/9) = -4$$
Parallelisms in Computation

Sequential Computation
1700 operations
900 multiplications and 800 additions

Parallel Computation
30 operations
900 multiplications and 800 additions
• ReRAM computational array architectures provides magnitudes orders of performance & energy improvement
## ReRAM Computational Memory Arrays & IPs

<table>
<thead>
<tr>
<th></th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly Parallel Memory</td>
<td>Classification</td>
</tr>
<tr>
<td>Computational Arrays</td>
<td>Matrix Operations, Multiply Accumulate, Sparse Matrix</td>
</tr>
<tr>
<td>Comparison/Evaluation</td>
<td>Matching, Best Fit, statistics</td>
</tr>
<tr>
<td>Configurable logic</td>
<td>Configuration bits, FPGA,</td>
</tr>
<tr>
<td>Power management</td>
<td>Memory Shadowing</td>
</tr>
<tr>
<td>Embedded memory, standalone memory, &amp; OTP</td>
<td>Embedded at Code/Data Memory at advanced nodes</td>
</tr>
</tbody>
</table>
Highly Parallel Memory
Simultaneous Processing with Deterministic Performance

- Parallel comparison against all identities
- If no match, new identity created (learning)
- Classification performed in one cycle independent of number of identities

Classification of 100,000s Identities stored in ReRAM in one iteration

Non-Volatile, Instant-on

Image Capture
Facial Detection (HOG)
Preprocessing
Neural Network (FaceNet Inception Model)

Representation (128 dimensions)
Computational ReRAM Array
Convolution - MAC
ReRAM Computational Array Properties - MAC

- High Bandwidth Multiply Accumulate operations (MAC) are performed with crossbar ReRAM array architectures
- Many MAC operations are simultaneously calculated
- Multiple Word lines are activated simultaneously
- Linear equations are solved with low latency, and low energy consumption

For example:

\[ V_d = \frac{(V_1 \cdot Gm1) + (V_2 \cdot Gm2) + (V_3 \cdot Gm3)}{G_0 \cdot (1 + Gm1 + Gm2 + Gm3)} \]

where \( G = \frac{1}{R} \)
Computational ReRAM Array
Comparison – Evaluation – Best Fit
- Large vector comparison/detection (i.e. > 512 bits) is performed in few nanoseconds
- Vector evaluation performed within the memory array
- Providing major system energy savings and reduced latency
Configurable Logic – Power management
RRAM for FPGA Configuration Bits, NVRAM, State Retainer

**Configuration Bit**

- Instant On
- Eliminates external non-volatile memory
- Security

**NVRAM**

**State Retainer**

- Stores data at power down
- Recalls at power up
- Power saving
ReRAM based Computational Arrays and IPs

- ReRAM Technology provides AI architects:
  - Breakthrough computational ReRAM memory arrays with
    - High computation bandwidth and high parallelism
    - Low energy
    - Low latency
  - Freedom to architect
  - Monolithic integration with advanced CMOS & FPGAs
Everyone says AI is the key, but we know the key to AI is ReRAM
Don’t be left behind

Rethink Artificial intelligence with ReRAM