Advanced Technologies for Brain-Inspired Computing

Fabien CLERMIDY*

*CEA-Leti, Grenoble, France
**CEA-List, Paris, France
# INRIA, Paris, France
Context: dark silicon...

- Dramatic changes in the area of computing architectures
- Power has become the main limiting factor for the scalability of microprocessors
  - Multi-cores architectures
  - Simultaneous
    - Heterogenous
    - Need for versatile hardware accelerators
Dramatic changes in the area of computing architectures

- Robustness issues

Robust accelerators

(Borkar 2005)
for new applications

- Today: scientific computing, cryptography, ...
- Tomorrow: data recognition, mining, synthesis
A potential answer: Neural networks

- Neural Networks are good candidates
  - They provide Robustness

- Good application scope
  - Signal processing
  - Optimization

- Approximation
  - Classification
  - Clustering

- Prediction

But classically limited by 2D hardware solutions
- Limited interconnections, costly long range ones
- But the HW situation has changed!
• Context
• Exploiting resistive memories
• Analogue neurons and 3D-TSV
• Towards monolithic 3D...
• Conclusion
New memory technologies

Volatile Memory
- SRAM
- DRAM
- Charge Trap
- FLASH
- SONOS
- FRAM

Polarization change

Non-Volatile Memory
- Resistance Change = memristors
  - PRAM
  - MRAM
  - RRAM

Charge-based Programming & Reading

Phase dependent Resistance changes
Magneto Resistance changes
Interface or bulk Resistance changes

Current-based Programming & Reading
<table>
<thead>
<tr>
<th></th>
<th>DRAM</th>
<th>FLASH</th>
<th>MRAM</th>
<th>PRAM</th>
<th>RRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compatibility with CMOS</td>
<td>- -</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>++</td>
</tr>
<tr>
<td>Scalability</td>
<td>32nm</td>
<td>15nm</td>
<td>20-30nm</td>
<td>10-20 nm</td>
<td>10 nm</td>
</tr>
<tr>
<td>Density</td>
<td>6-8f²</td>
<td>4f²</td>
<td>35-40f²</td>
<td>4-6f²</td>
<td>4-6f²</td>
</tr>
<tr>
<td>Maturity</td>
<td>++</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>- -</td>
</tr>
<tr>
<td>Byte @</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Writing cycle</td>
<td>50ns</td>
<td>0.1 ms</td>
<td>20ns</td>
<td>100 ns</td>
<td>150 ns</td>
</tr>
<tr>
<td>Endurance</td>
<td>1,00E+16</td>
<td>1,00E+05</td>
<td>1,00E+15</td>
<td>1,00E+09</td>
<td>→1,00E+06</td>
</tr>
<tr>
<td>Plus</td>
<td>Perf.</td>
<td>NV</td>
<td>Perf. + NV + Endurance</td>
<td>Perf. + NV+ density</td>
<td>Perf. + NV+ density</td>
</tr>
<tr>
<td>Minus</td>
<td>Scalability</td>
<td>Endurance</td>
<td>Perf. - -</td>
<td>Costly tech. Delay wrt CMOS nodes</td>
<td>consumptio n</td>
</tr>
</tbody>
</table>
Memristor as a Synapse

- Concept introduced by Chua
  - “Memristor-The missing circuit element” - 1971
- Strukov & Snider developed it
  - “The missing memristor found” - Nature 2008
  - “Spike-timing-dependent learning in memristive nanodevices” – Nanoarch 2008
- Concept “memristor = synapse weight”
  - So easy ?

\[ s = u_1 + u_2 \]
\[ w = 1 \]
Crossbar of Synapses

- Require “gradually” programming memristors
  - Difficulty is to obtain equivalent potentiation / depression phases

Phase-Change Memories

Material with 2 stable phases
- Chalcogenide alloys: GST, GeTe, ...
- Hysteresis cycle between 2 states

Transition by Joule heating

- Polycrystal: High conductivity
- Amorphous: Low conductivity

Reversible Phase Change

$T_x$ $\Rightarrow$ Polycrystal
$t_{\text{QUENCH}} > 100 \text{ ns}$

$T_{\text{melt}}$ $\Rightarrow$ Amorphous
$t_{\text{QUENCH}} < 10 \text{ ns}$
A solution: 2-PCM Synapse

Symmetric Long-Term-Potentiation and Long-Term depression behavior by using pulsed crystallization.
• Context
• Exploiting resistive memories
• **Analogue neurons and 3D-TSV**
• Towards monolithic 3D…
• Conclusion
**Analogue neurons**

- Compact design (5x gain vs digital)
- Low power consumption (20x gain vs digital)
- Ability to interface sensors directly with the processing part
- Computational efficiency

**Drawback:**
Large capacitance required (0.5-1pF)  
⇒ MIM  
⇒ Large area (not scalable !)


Leaky Integrate and Fire Neuron in CMOS 65nm accept spikes in 1kHz-1 Mhz range
Through Silicon Vias

- Used for crossing circuits in 3D stacking
- Present large “parasitic” capacitance

Demonsrated feasibility

A. Joubert et al., “Capacitance of TSVs in 3-D stacked chips a problem? Not for neuromorphic systems”, DAC’12
3D integration using TSV

(a) Standard 2D neuromorphic architecture

(b) 3D stacking with standard 2D neurons

(c) et (d) 3D stacking with TSV-based neurons

Chaabouni et. al. (2010)
• Context
• Exploiting resistive memories
• Analogue neurons and 3D-TSV
• Towards monolithic 3D...
• Conclusion
• 3D technologies

• **Parallel Integration**
  Dies fabricated separately then vertically stacked, such as:
  - Through-Silicon-Via (TSV)
  - Copper-to-Copper (Cu-Cu) contact
  - High-Density TSVs

• **Sequential Integration**
  Second transistor layer is fabricated directly on the top of first one. (Cold process)
  - 3D Monolithic Integration (3DMI)
Integration is monolithic using a “cold” process

- Allows 100 nm 3D vias pitch

=> Open perspectives for real 3D implementation

P. Batude et al., "GeOI and SOI 3D Monolithic Cell integrations for High Density Applications" VLSIT 2009
Using 3D TSV allows a “cube” of neurons
- Drastically reduce interconnects
- Example shown here is 3x more efficient in 3D

184 neurons with 1181 connections
Conclusion

- Neuromorphic architectures are appealing
  - Solve dark silicon issues
  - Adapted to modern applications

- And can take benefit of advances technologies
  - Synapse => Memristors = RRAM, PCM
  - Neurons => Analogue neurons + high-capacitance TSV
  - Interconnections => high-density 3D

- The path is open for the next neuromorphic revolution!
Questions ?