MRAM: Memory for the Edge… And Beyond

Jeff Lewis
SVP Business Development
Intelligence Is Moving to the Edge

Edge AI Required When:

- Latency Unacceptable: ADAS, Robotics, Security, Industrial Process
- Communication Power: Local Processing more efficient than transmission
- Privacy Concerns: GDPR, HIPAA, Surveillance
- Localized Training: Local surveillance, Google Federated Model, other emerging NN architectures

Devices must get more sophisticated to prevent Fraud, Hacking, Mischief - Next step is Personalization -- Voice ➔ MY Voice
Edge AI: Smarter = Power Hungry

- Complex local computation
  - E.g., Facial recognition, ADAS, Surveillance

- SRAM size/power grows as computation grows
  - Or add external DRAM…

- Power consumption huge challenge
  - Limits model sophistication
  - Large battery
  - Heat effects – wearables, cameras

---

Accuracy of ImageNet classification versus neural net complexity – showing reducing complexity (and power) by 10x reduces accuracy by 15-20%  
(source Google MobileNet V2)
Thanks for the Memories

• Edge AI Is All About (RAM) Memory

• “Never Enough Memory”…
  • Most AI Chip Energy and Die Area consumed by Memory
  • Han, ISCA 2016, many others

• … And it better be On-chip
  • Google study*: 40%-60% of total mobile system energy consumed in DRAM ➔ Chip data transfers

*Boroumoud, Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks, © 2018
App-Targeted MRAM for SoC’s

SRAM Replacement
- LLC, AI, DDI, many others
- 10-15ns R/W
- >$10^{13}$ cycles
- Days-months retention

High Speed & Endurance NVM
- IoT, Edge AI
- 25-50ns R/W
- >$10^{11}$ cycles
- >10 years retention

eNVM
- eFlash replacement
- 25ns Rd / 50-500ns Wrt
- $10^{6-8}$ cycles
- >10 years retention
MRAM Replaces SRAM

✓ **SIZE**: 1T-bitcell – MRAM block 70%-80% smaller than SRAM

✓ **LEAKAGE**: No bitcell leakage

✓ **BEOL**: Simple Integration

✓ **NO S.E.U.**: Rad hard bitcells

✓ **PERSISTENT**: Data retained

* R. de Werdt et al., IEDM 1987; R.D.J. Verhaar et al., IEDM 1990; S.H. Kang et al., IEDM17
<table>
<thead>
<tr>
<th>Challenge</th>
<th>MRAM Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lots of On-Chip RAM</td>
<td>MRAM ¼ Size of SRAM ➔ 4x More Memory in Same Footprint</td>
</tr>
<tr>
<td>Lowest Possible Cost</td>
<td>Size vs. SRAM</td>
</tr>
<tr>
<td></td>
<td>Low-cost BEOL Adder vs. Flash</td>
</tr>
<tr>
<td></td>
<td>Execute-in-Place Merged NVM + RAM</td>
</tr>
<tr>
<td>Lowest Possible Power</td>
<td>MRAM Low Write Energy</td>
</tr>
<tr>
<td></td>
<td>MRAM Persistence – No Leakage vs. SRAM</td>
</tr>
<tr>
<td></td>
<td>Easy + Efficient Sleep Management</td>
</tr>
<tr>
<td>Many Updates</td>
<td>MRAM 100x – 10,000x Lower Write Power than Flash, and Much Higher Endurance</td>
</tr>
</tbody>
</table>

5/14/2019
MRAM Unified Execute-In-Place Memory

Benefits:

- **Cost**: Much smaller die area
- **Cost**: MRAM wafer cost lower than flash wafer cost
- **Power**: Eliminates memory data transfer latency + power
- **Lifetime**: Enables frequent data updates, data logging, other write-intensive NVM

Merged NVM + SRAM for up to ~40MHz Operation
MRAM Maximizes Sleep Cycles

More Sleep = Longer Battery Life

- Going in / out of sleep burns energy
  - Mostly storing / reloading SRAM
- Only sleep when:
  \[ \text{EnergySavings}_{\text{Sleep}} > \text{EnergyCost}_{\text{LoadStore}} \]
  - Limits sleep to few, long periods

Using MRAM instead of SRAM:
- Simply power memory off!
- Eliminates store/reload energy cost
- Enables frequent “Micro-Naps”
  - save significant additional energy
Edge AI – The Model is Never “Complete”

Normal Processing Device

- Code loaded to part at test
- Zero to a few code updates during lifetime

Flash write power + performance: *No Problem*
Edge AI – The Model is Never “Complete”

**Normal Processing Device**
- Code loaded to part at test
- Zero to a few code updates during lifetime
- Flash write power + performance: No Problem

**“Smart” Edge Device**
- (maybe) Code loaded to part at test
- Very frequent NN model updates from Cloud
- Anomalous data sent to Cloud
- Flash write power + performance: BIG Problem – MRAM 100-1,000X Better

Local training model updates – e.g., Google Federated Model scheme
MRAM: For The Edge… and Beyond!

• Transformative memory ➔ Enables applications other memories can’t

• Will be as significant as SRAM and other memories

Will finally be able to answer the question: – Is this a Rabbit… or a Duck?