3D Stackable Circuits and Memory

KARL-MAGNUS PERSSON AND LARS-ERIK WERNERSSON
Talk Outline

Introduction

• Energy efficient computing with big data - background and motivation
• A promising candidate - ReRAM

Considerations for ReRAM Integration

• 3D – methods and possibilities
• Scaling – materials, trends, and challenges

Ongoing Research Efforts

• Stanford – monolithic 3D integration of circuit and memory
• Lund – vertically integrated nanowire selectors and RRAM
The Era of Big Data and Recognition

Machine Learning Hardware Implications

• Iterative re-programming of memory
• Performance limited by read/write of none-volatile-memory (NVM)

Hardware Challenges

• Component improvement stagnated - Moore’s law has halted
• NVM technologies - 10,000x slower than computing
• Separate compute and memory circuitry infer large inefficiencies

Possible solutions

• Creative systems → co-integrated circuits and memory in 3D, introducing new materials
• True neuromorphic hardware → synaptic networks using computational units
NAND Flash and Inherent Limitations

- **The upside**
  - 3D integrated, 128 layers in the near future
  - Minimal feature size down to 5 nm

- **The downside**
  - Read in ns but write in ms
  - Further scaling not possible due to tunneling

![RC delay diagram](https://www.storagenewsletter.com)

Bit Cost Scalable NAND Flash

https://www.storagenewsletter.com
3D ReRAM – A Promising Candidate

Resistive switching
- Mobile ions form a filament
- Low voltage programming (1-2 V)

Speed
- 10 ns read/write

Scaling
- $6F^2$/layer
- $F < 5$ nm

3D ReRAM
- 128 layers
- 64 Tb per chip

Stanford is a leading institution in the field of ReRAM under the supervision of Prof. H.-S. Philip Wong
RRAM Mechanics

H.-S. Philip Wong et al - Proceedings of the IEEE 2012
RRAM Mechanics

H.-S. Philip Wong et al - Proceedings of the IEEE 2012
Overview of different ReRAM Technologies

- **RRAM (oxRRAM)**
  - Anode filament, oxygen vacancies form conductive path
  - High endurance, $1^{12}$ cycles at device level
  - 3D compatible

- **CBRAM**
  - Cathode filament, bridging with metal ion movement
  - Similar structure to RRAM
  - Endurance questionable (finite number of switches)

- **PCRAM**
  - Phase change memory, a flash heating switches dielectric film between amorphous and crystalline state
  - Endurance questionable (finite number of switches)

- **STT-MRAM**
  - Spin-transfer-torque magnetic RAM
  - Changing orientation of spin changes the conductivity
  - Advanced material stack, 3D compatibility unlikely
Overview of different ReRAM Technologies

- **RRAM (oxRRAM)**
  - Anode filament, oxygen vacancies form conductive path
  - High endurance, 112 cycles at device level
  - 3D compatible

- **CBRAM**
  - Cathode filament, bridging with metal ion movement
  - Similar structure to RRAM
  - Endurance questionable (finite number of switches)

- **PCRAM**
  - Phase change memory, a flash heating switches dielectric film between amorphous and crystalline state
  - Endurance questionable (finite number of switches)

- **STT-MRAM**
  - Spin-transfer-torque magnetic RAM
  - Changing orientation of spin changes the conductivity
  - Advanced material stack, 3D compatibility unlikely

Stanford Memory Trends, H.-S. P. Wong et al
https://nano.stanford.edu/stanford-memory-trends
The Memory Hierarchy

- **Core 1**
  - Registers [R]
  - L1 Instruction Cache (SRAM) [L1I]
  - L1 Data Cache (SRAM) [L1D]
  - L2 Cache (SRAM) [L2]
  - L3 Cache (eDRAM) [L3]

- **Core N**
  - Peripheral controller

### Memory Components

- **Main Memory (DRAM)**
  - CPU cycles: 1 - 100
  - Speed: 0.1’s ns
  - Size (bytes): 100 K M G T P

- **Secondary Storage (Flash, Disk)**
  - CPU cycles: >100
  - Speed: 10’s -100’s ns
  - Size (bytes): 10^6 - 10^8
  - Time: >10^9

- **Tertiary Storage (Disk, RAID)**
  - CPU cycles: >100
  - Speed: 0.1-10’s ms
  - Size (bytes): >10^9

### Interconnect

- L2 Cache
- L3 Cache (eDRAM)
- Peripheral controller

---

Research Trend I – Stacking Circuits and Mem

- Performance predictions of ReRAM with large scale circuit simulations using calibrated models
- Study shows benchmarks of a contemporary Intel Xeon Phi system VS a system with CNT-cores and STT-MRAM + 3D RRAM
- Major part of the improved performance in memory intense applications like PageRank comes from the new memory technology
- Proposed system shows up to 1000x gains in combined power and speed

Conventional CPU is idle 97% of the time!!
Implications for Neuromorphic Computing

IBM TrueNorth
- SyNAPSE - DARPA funded initiative to simulate the brain
- Dedicated neuromorphic hardware
- 4096 computational units (1 unit pictured)
- Memory occupies about 40% of the chip area
- Conventional memory (SRAM)

RRAM
- 3D RRAM with 128 layers
- 64 Tb per chip

SRAM $\sim 120-140 \, F^2 \rightarrow 1T1R \, RRAM \, 20x \, smaller!!$
3D RRAM - Architectural Concepts

HRRAM (horizontal)
- Most simplistic
- Superior performance due to low RC interconnects
- Diode selector limits feasible number of layers
- Least cost efficient
- Intel Xpoint

VRRAM type I
- Interconnect resistance limited performance
- More energy efficient than VRAM type II

VRRAM type II
- Interconnect capacitance limited performance
- Litho-free stacking
- Bit performance improves with # layers
- Most scalable/cost efficient
Vertical RRAM – Lithography Free

Litho-free formation of a stair-case structure

Tanaka et al – VLSI symposium 2007
Research Trend II – In Memory Computations

- 3D RRAM allows for hyper-dimensional vectors with computations directly in the memory
- NOR and NAND operations can be accomplished with propagation of specific pulse trains
- It is not determined if in-memory computations will be a viable way forward, could be task specific

H. Li et al – VLSI 2016
• RRAM switching is ideally independent of device area as only one filament forms.

• Area dependence is instead partly coupled to self-capacitance, and a reduction in parasitic current discharge.

• However, the probability to form a filament increases with area.

• To reduce increased forming voltage and spread in the distribution, surface roughness and material quality at the interfaces are of crucial importance.
**RRAM - Oxide Thickness Scaling**

- Scaling the dielectric necessary to reduce minimum feature size
- Surface roughness affect the spread of the performance distribution
- Etched out vertical pillar have a smoother surface, 3D thus reduces the feature size

Zhao et al – IEDM 2014
Considerations on Scaled 3D Arrays

- Large arrays require MOSFET selectors to reduce leakage.
- Vertical geometry allows for more aggressive thickness scaling as it reduces roughness.
- Simulations show metal plane thickness will limit array size due to resistance of the vias, sub 6-nm metal is highly resistive.
- Graphene and other 2D materials way become a viable way forward for large scaled arrays.

Joon Sohn – IEDM 2014
Stanford Research: CNTs with 3D RRAM

Pillar connecting each CNFET to 4 RRAM layers

20 nm
1T4R 64-bit Monolithic Memory Cell

- 16 CNT MOSFETs and 4 layers of RRAM
- ~18 lithography steps + countless of fabrication procedures
TiN RRAM Layer 1 and 2 – Form and Reset
Lund Research: RRAM with NWFETs
Initial 2D RRAM Tests

- Not established which materials are most beneficial
- Currently investigating different materials and combinations
- Voltage envelope is a great concern (<1 V) for low-power MOSFETs

ITO/Al₂O₃/HfO₂/Ti-stack
Lund Research: RRAM with NWFETs

• Integration on vertical MOSFETS only demonstrated on individual Si-pillars

• III-Vs offer low-power operation

• Lund has demonstrated record high T-FET performance using nanowires

• A unique approach would be to combine T-FETs and low-power RRAM technology for ultra-low-power operation
Conclusions

• Contemporary NVM technologies hugely limiting memory intense applications

• 3D stacking circuits and ReRAM could potentially improve efficiency for data intense computing by 1000x

• MOSFET selector needed for large arrays and 3D integration

• New approach with 3D integration of RRAM directly on top of vertical MOSFETs, no array demonstration to date

• Implementing TFETs selectors, RRAM would benefit in the same way as for CMOS logic, enabling larger, faster, and more energy efficient circuits
This work has been funded by the Wallenberg Foundation