Addiction to Low Cost per Memory Bit

But at what cost?

Andy Walker
July 19th 2021
Contents

• Homage to the Field Effect Transistor (FET)
• Tunneling in Silicon Valley - Fairchild 1967-68
• The Golden Thread of Tunneling
• What’s happening in solid state memories?
  • SRAM, DRAM, NAND, Emerging Memories
• The “Creaking” Memory Hierarchy and the Rise of AI
• What is a “Negative Externality”?
• Examples in relation to solid state memories
  • Product Lifetimes – The Endurance Conundrum, Acquisition Cost and Total Cost
  • Compute insecurity – DRAM Rowhammer
  • Energy Use – The Problem with Memory
• Summary
• Prediction and a Call to Action
• Thank you
Homage to the Field Effect Transistor (FET)

SRAM: 6 FETs

DRAM: 1 FET + 1 Capacitor

NAND: FETs in a string

+ Tunneling

Tunneling in Silicon Valley - Fairchild 1967-68

With thanks to the Shustek Center at The Computer History Museum, Mountain View/Fremont, California
The Golden Thread of Tunneling
- From Fundamental Physics to Technological Innovation (2)

**NAND FLASH 1988**

**MULTI-BIT 1992**

**SYSTEM FLASH 1994**

**LIMIT of 2-D NAND FLASH ~2016**
The Golden Thread of Tunneling
- From Fundamental Physics to Technological Innovation (3)

**Thin Film Transistor (TFT) SONOS 2003**

3D TFT-SONOS Memory Cell for Ultra-High Density File Storage Applications

Abstract

For the first time, a scalable, low power, deep-submicron TFT-SONOS (Thin Film Transistor, Silion Oxide-Nitride-Oxide-Silicon) memory cell is described with characteristics rivaling those of single crystal devices (>10^6 cycles, -1.6V window after 10 years on cycled at 85°C) showing the promise of 3D integration and ultra small cell footprints. The ability to vertically stack device layers enables the current memory density record of ~200MByte/cm², set by 90nm NAND, to be surpassed.

(Keywords: TFT, SONOS, 3D, nonvolatile, memory)

Walker et al.

**TFT SONOS NAND 2006**

A Multi-Layer Stackable Thin-Film Transistor (TFT) NAND-Type Flash Memory

Abstract

A double-layer TFT NAND-type Flash memory is demonstrated, ushering into the era of three-dimensional (3D) Flash memory. A TFT device using bandgap- engineered SONOS (BE-SONOS) [1,2] with fully-depleted (FD) poly silicon (60 nm) channel and tri-gate P'-poly gate is integrated into a NAND array. Small devices (L/W=0.2/0.09 μm) with excellent performance and reliability properties are achieved. The bottom layer shows no sign of reliability degradation compared to the top layer, indicating the potential for further multi-layer stacking. The present work illustrates the feasibility of 3D Flash memory.

Lai et al.

**DG-TFT SONOS NAND 2008**

Sub-50nm DG-TFT-SONOS – The Ideal Flash Memory for Monolithic 3-D Integration

Abstract

A revolutionary 3-D stackable sub-50nm double-gate TFT SONOS technology is presented here with series strings of up to 64 cells consisting of the smallest silicon-based TFT's to date. Read- and program-pass disturbs have been extinguished. Excellent endurance and retention are shown. Monolithic 3-D integration and scalability are ensured through close to zero source/drain diffusion. Finally, comparisons with TANOS are given.

Walker

**The Rise of Monolithic 3-D Flash**
The Golden Thread of Tunneling
- From Fundamental Physics to Technological Innovation (4)

Bit Cost Scalable Technology with Punch and Plug Process for Ultra High Density Flash Memory

Tanaka et al.

Abstract
We propose Bit-Cost Scalable (BiCS) technology which realizes a multi-stacked memory array with a few constant critical lithography steps regardless of number of stacked layer to keep a continuous reduction of bit cost. In this technology, whole stack of electrode plate is punched through and plugged by another electrode material. SONOS type flash technology is successfully applied to achieve BiCS flash memory. Its cell array concept, fabrication process and characteristics of key features are presented.

Vertical Channel TFT SONOS BiCS NAND 2007

The Rise of Vertical Channel 3-D NAND Flash

Vertical Channel TFT SONOS TCAT NAND 2009

Jang et al.

Abstract
Vertical NAND flash memory cell array by TCAT (Terabit Cell Array Transistor) technology is proposed. Damascened metal gate SONOS type cell in the vertical NAND flash string is realized by a unique 'gate replacement' process. Also, conventional bulk erase operation of the cell is successfully demonstrated. All advantages of TCAT flash is achieved without any sacrifice of bit cost scalability.
The Golden Thread of Tunneling
- From Fundamental Physics to Technological Innovation (5)

Rise of 3-D NAND FLASH > 2013
In solid state memories, what’s happening is Scaling slowing down + Leaky.


Presentation to SAFARI Research Group ETH Zurich
What’s happening in solid state memories? - DRAM

- Scaling slowing down + Increasing Rowhammer susceptibility


What’s happening in solid state memories? - NAND

- Reaching diminishing returns from cell/string stacking
- String currents a major challenge – new materials and epitaxial growth needed
- Electrical bits per cell exhausted – SLC, MLC, TLC, QLC…. at expense of endurance
- Active under memory array trick done

What's happening in solid state memories? – Emerging Memories

- Endless forms most beautiful and most wonderful have been and are being evolved
  - STT-MRAM, ReRAM, FeRAM, PCM
- Less well understood switching mechanisms compared to charge storage
- New materials
- Highly conservative manufacturing culture as a barrier to entry
- What specific problems in the “memory hierarchy” are being solved?

---

(1) The Origin of Species, Charles Darwin, 1859

(2) S.-Y. Wu, “A New Ferroelectric Memory Device, Metal-Ferroelectric-Semiconductor Transistor”, IEEE T-ED, August 1974

The “Creaking” Memory Hierarchy and the Rise of AI

- SRAM, DRAM “main memory” and NAND storage all separate chips
- Interchip data movements highly energy intensive
- Data “explosion” and AI drive volume of interchip data movements and dramatic rise in energy use
- Training a single AI model can emit as much CO₂ as five cars in their lifetimes\(^{(1)}\)
- AI data centers to consume > **10% of world energy capacity by 2025**\(^{(2)}\)


Negative Externalities

- A negative externality (NE) is a cost that is suffered by a third party as a consequence of an economic transaction. In a transaction, the producer and consumer are the first and second parties, and third parties include any individual, organisation, property owner, or resource that is indirectly affected. Externalities are also referred to as spillover effects, and a negative externality is also referred to as an ‘external cost’\(^{(1)}\)

\(^{(1)}\) Economics Online (https://www.economicsonline.co.uk/Market_failures/Externalities.html)
NE 1 – Memory Endurance, Limited SSD Lifetime and Total Cost of Ownership

**Cost metric for Consumer applications: $/GB**
- GB = SSD Capacity in GigaBytes
- Initial hardware acquisition cost
- \( GB \propto (N_{\text{Layers}} \times \text{Bits/Cell}) \) in 3-D NAND
- \( N_{\text{Layers}} \) sub-linearly increasing with time
- \( \text{Bits/Cell} \) almost exhausted (SLC=1, MLC=2, TLC=3, QLC=4 .....

**Cost metric for Enterprise applications**(1) $/PBW
- PBW = Maximum PetaBytes Written to SSD after which failure occurs
- Data centers and corporations
- Includes life of SSD and therefore **Cell Endurance**
- \( PBW = (\text{SLC SSD Capacity}) \times (\text{Bits/Cell}) \times (\text{Cell Endurance}) \times (1/WAF) \)
- Capacity in PB, WAF = Write Amplification Factor (between 1.5 and 4)
- 3-D NAND cell endurance very strong function of bits/cell(2) and limited by physics
- Main factor now in 3-D NAND to increase PBW – increase \( N_{\text{Layers}} \)
- Segmentation of market into read intensive and write intensive

---

(2) [https://www.micron.com/products/nand-flash/qlc-nand](https://www.micron.com/products/nand-flash/qlc-nand)
NE 1 – Limited Endurance and the Consequences

- Pass disturb reduction requires thick tunnel oxide
- Samsung “discovered” this between 2003 and 2006
  - 2.5nm (2003), 4nm (2005), 5nm (2006)
- Thick oxide tunneling causes severe P/E cycling damage
- Fundamentally Limited endurance inherent to NAND (compared to Classic SONOS\(^{(1)}\))
- Built-in product obsolescence, shortened replacement cycles, long term sales
- Environmental impact in resources and energy use

\[^{(1)}\] A. J. Walker et al., “3D TFT-SONOS memory cell for ultra-high density file storage applications,” VLSI Symp., 2003
\[^{(2)}\] A.J. Walker “Sub-50-nm Dual-Gate Thin-Film Transistors for Monolithic 3-D Flash”, IEEE T-ED, Nov. 2009
\[^{(3)}\] https://www.youtube.com/watch?v=zbwIhsBy808

Presentation to SAFARI Research Group ETH Zurich

July 19 2021
NE 1 – Limited Endurance – What Can be Done? – Tunnel Engineering

• New 3D Flash device architectures can allow thin tunnel oxides
  - 3-D Dual-Gate SONOS allows millions of endurance cycles

• STT-MRAM endurance boosted by design techniques
  - Up to 6 orders of magnitude gain allowing potential SRAM/DRAM replacement

(1) A.J. Walker “Sub-50-nm Dual-Gate Thin-Film Transistors for Monolithic 3-D Flash”, IEEE T-ED, Nov. 2009
NE 2 – DRAM Rowhammer (RH) and Compute Insecurity

- Repeated access of a DRAM row (wordline) causes bit-flips in nearby rows
- First public unmasking in 2014\(^1\)
- Google showed how to use RH to hack into a laptop in 2015\(^2\)
- Newer node DRAM products are becoming more susceptible
- All mitigations have been worked around except for the recent one (so far)
- Effect is due to a bipolar effect (electron injection and capture) and a MOS effect (crosstalk)\(^3\)

---


NE 2 – DRAM Rowhammer (RH) and Compute Insecurity – What Can be Done?

- Mitigation involving refresh engineering - containment
- Shielding of both injected electrons and crosstalk effects \(^{(1,2)}\)
- 3-D DRAM \(^{(3)}\)

---

![Diagram of DRAM structure](image)

---


(3) A. Belmonte et al., "Capacitor-less, long-retention (>400s) DRAM cell paving the way towards low-power and high-density monolithic 3D DRAM," in IEDM Tech. Dig., Dec. 2020

NE 3 - Energy Use and The Problem with Memory

• Fetching/storing data in solid state memory uses >/~ 60% of system energy\(^{(1)}\)
  - On-chip SRAM access ~ \(10X\) energy of CPU data manipulation
  - Off-chip DRAM access ~ \(1000X\) energy of CPU data manipulation

• On-chip SRAM has fundamental leakage – wasted energy

---

\(^{(1)}\) A. Boroumand et al., “Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks”, ASPLOS’18, March 2018


\(^{(3)}\) Simulations/estimates from 7nm transistor data in IEDM 2017
Physics of Charging and Discharging Capacitors

**Charging**

- Total energy stored in capacitor = \( \frac{1}{2} CV_b^2 \)
- Total energy converted into heat = \( \frac{1}{2} CV_b^2 \)

**Discharging**

- Total energy stored in capacitor = \( \frac{1}{2} CV_0^2 \)
- Total energy converted into heat = \( \frac{1}{2} CV_0^2 \)

• During charging half** of electrical energy converted into heat and half** stored on capacitor
• During discharging all** of stored capacitor energy converted into heat
• Actual energy transfer and conversion wonderfully more complex\(^{(2,3)}\)

\(^{(1)}\) http://hyperphysics.phy-astr.gsu.edu


** About right since does not count EM radiation
Relevance to Integrated Circuits, Systems and AI

• Any IC and system is an electrical power supply and a network of capacitors and resistors
• Data movements require charging and discharging of wires

\[ \text{Wires are capacitors with } C \propto L_{\text{wire}} \]

\[ \text{Energy conversion into heat } \propto L_{\text{wire}} \]

• Long wires between data in on-chip SRAM and processor
• Longest wires between data in off-chip DRAM and processor
• Most energy conversion into heat takes place in data transactions with memory

• AI needs extremely intensive store and recall between processor and memory
• Memory requirements pose a huge challenge in energy efficiency for deep learning models
AI and the Colossal Energy Demand

• Energy efficiency key constraint for AI proliferation
  - Training a single AI model can emit as much CO₂ as five cars in their lifetimes\(^{(1)}\)
  - AI data centers to consume > 10\% of world energy capacity by 2025\(^{(1)}\)

NE 3 - What Can We Do About This for AI?

• Domain Specific Architectures for AI accelerator chips
• Algorithms that minimize data flow interactions with off-chip DRAM
• Package solutions that minimize inter-chip impedances (capacitances)
• In-memory compute
• Near memory compute
• Principle of locality in time and space – cache structure and control
• Data compression to minimize weight populations
• Reduced precision arithmetic
• Maximize stand-alone main memory single chip capacity
• Minimize static energy loss
• Maximize on-chip memory capacity
• Fault tolerance and voltage manipulation
Summary

• The cost per memory bit is becoming a weakened metric
  • Directly associated with acquisition cost
  • Running cost and negative externalities not built in
• Negative externalities becoming more important as scaling proceeds
  • Short product lifecycles due to limited endurance
    • Wasted energy and resources
  • Compute insecurities affecting Clouds, laptops and mobile phones are one mitigation work-around away
  • Dramatic energy use increases driven by the data deluge and AI
    • Can green energy keep up with the demand?
• Industry consolidation has led to uniformity in memory offerings
• Well nigh impossible to disrupt the main players
• Unless.......
Prediction and a Call to Action

• Challenges are opportunities!
• Energy and resource considerations will change the definition of cost per memory bit to include a mix of process cost, die area and the relevant negative externalities associated with the markets
• The memory hierarchy will open up from its “closed shop” position
• There will be room for technology innovation from non-tier 1 players
• Who will be the next Tesla of solid state memories?

• Thank you!
Speaker biography

• Andy Walker has been working in silicon technology since 1985. After a BSc(Hons) in physics from Dundee University in Scotland (where he first met the Field Effect Transistor and tunneling) he joined Philips Research Laboratories (the “Nat Lab”) in Eindhoven, The Netherlands. His PhD from the Technical University of Eindhoven arose from his research work at Philips encompassing MOS transistor physics and technology, nonvolatile memories, dopant diffusion, shallow junctions, and inversion layer mobility. In 1994 he came to Silicon Valley and worked at various companies including Cypress, Matrix and Spin Memory. He also founded Schiltron Corporation to develop new forms of monolithic memories. He has been fortunate in being able to work in many interesting areas of silicon with some of the best engineers and scientists in the industry.