Lemons are for Lemonade

Subramanian Iyer
(s.s.iyer@ucla.edu)
Center for Heterogeneous Integration and Performance Scaling
chips.ucla.edu

UCLA Samueli School of Engineering

CHIPS
CENTER FOR HETEROGENEOUS INTEGRATION AND PERFORMANCE SCALING

UCLA CHIPS

A UCLA Led partnership to develop Applications, Enablement and Core technologies and the eco-system required for continuing Moore's Law at the Package and System Integration levels and **develop our students & scholars to lead this effort**

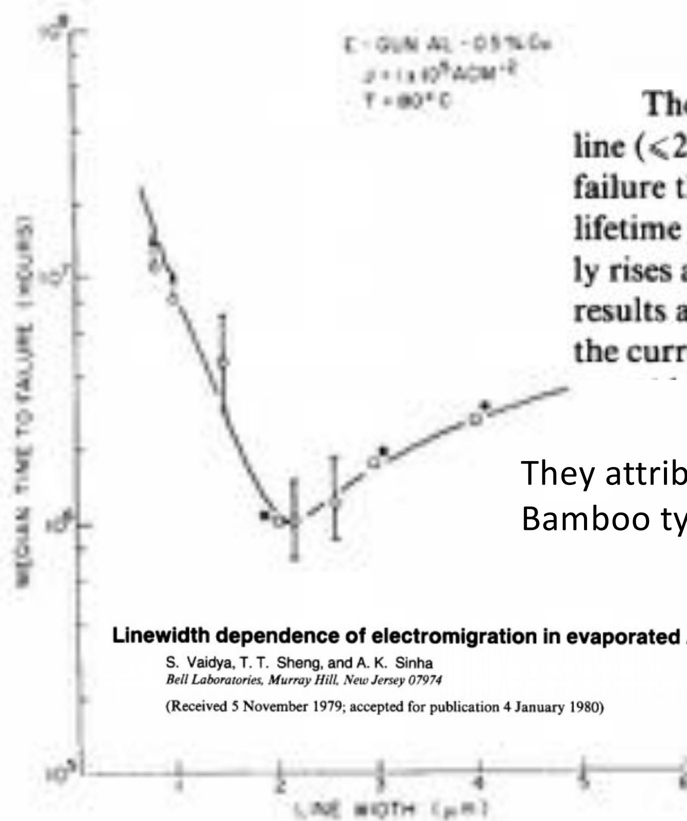
At the university, our main product is our Students.
Our research and development is a vehicle to educate and train our students

Our students learn by making mistakes

Two Examples – the power of mistakes

- Electromigration – how a ”bad” assignment impacts things in a good way decades later
- Salicide - the case of the leaky furnace leads to success in the nick of time
- Embedded DRAMS – how some problems just go away if you think hard enough (and wait long enough….)

Electromigration in fine Al-Cu* wires



Linewidth dependence of electromigration in evaporated Al-0.5%Cu

S. Vaidya, T. T. Sheng, and A. K. Sinha
Bell Laboratories, Murray Hill, New Jersey 07974

(Received 5 November 1979; accepted for publication 4 January 1980)

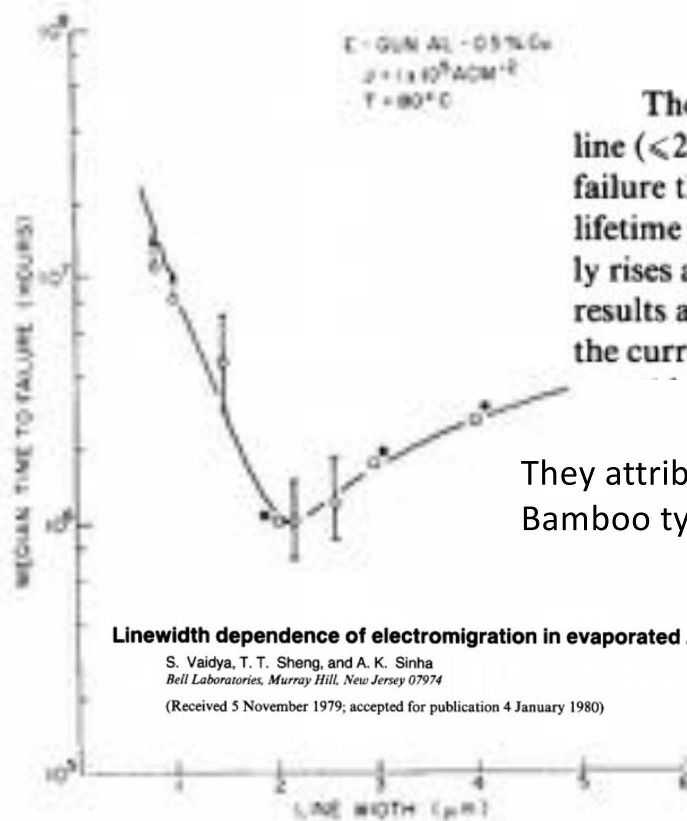
FIG. 1. Median time to failure as a function of linewidth for e-gun Al-0.5%Cu lines, at 80 °C and 1×10^6 A cm^{-2}.

The present work shows that *e*-gun-evaporated fine-line (<2 μm) Al is in fact is less prone to electromigration failure than the wider lines. With decreasing linewidth, the lifetime goes through a minimum at ~2 μm and then sharply rises again as the feature width is further reduced. The results are interpreted on the basis of the grain structure of the current-carrying stripes.

They attributed this "improvement" to finer lines having a Bamboo type of microstructure

*Al-Cu – that's another story

Electromigration in fine Al-Cu* wires



Linewidth dependence of electromigration in evaporated Al-0.5%Cu

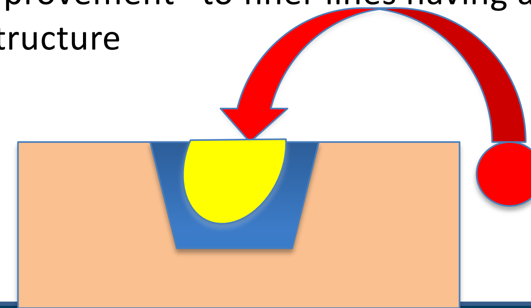
S. Vaidya, T. T. Sheng, and A. K. Sinha
Bell Laboratories, Murray Hill, New Jersey 07974

(Received 5 November 1979; accepted for publication 4 January 1980)

The present work shows that *e*-gun-evaporated fine-line (<2 μm) Al is in fact is less prone to electromigration failure than the wider lines. With decreasing linewidth, the lifetime goes through a minimum at ~2 μm and then sharply rises again as the feature width is further reduced. The results are interpreted on the basis of the grain structure of the current-carrying stripes.

FIG. 1. Median time to failure as a function of linewidth for e-gun Al-0.5%Cu lines, at 80 °C and 1 × 10⁶ A cm⁻².

They attributed this “improvement” to finer lines having a Bamboo type of microstructure



*Al-Cu – that’s another story

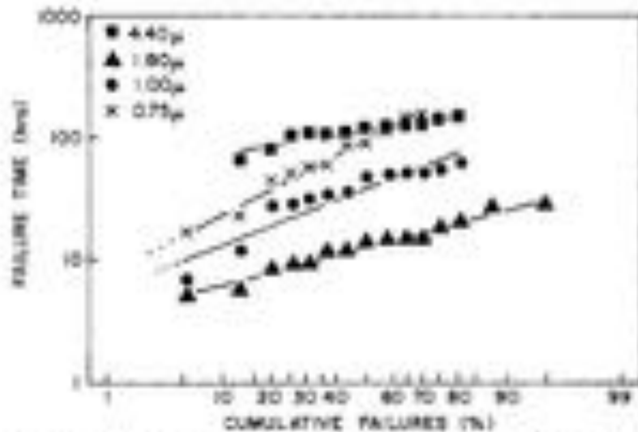


Fig. 1. Failure density plotted with time for a few different linewidths on a log-normal scale.

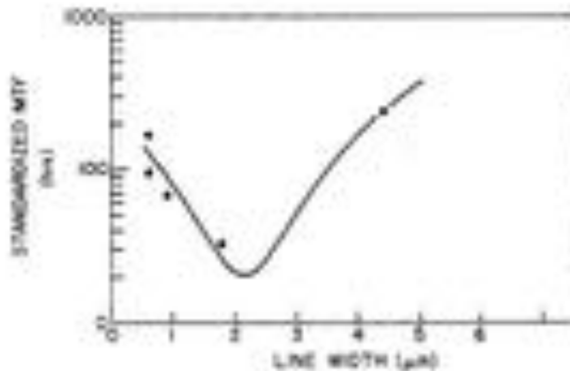


Fig. 2. Standardized mean time to fail, t_{50} plotted against line width for 0.5-μm-thick Al-4 percent Cu lines. Results are standardized to 250°C and 1×10^{6} A/cm^{-2}. Actual tests were done close to those conditions.

We were able to reproduce the results as far as mean time to fail !!

But

Iyer, S.S.; Chung-Yu Ting, "Electromigration lifetime studies of submicrometer-linewidth Al-Cu conductors," in Electron Devices, IEEE Transactions on , vol.31, no.10, pp.1468-1471, Oct 1984

We don’t care when half our sample fails – we worry about the first one failing !

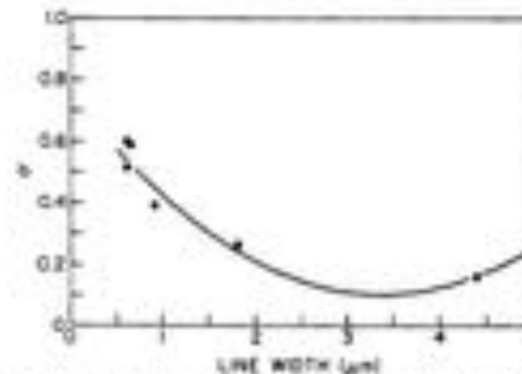


Fig. 3. Variation of sigma of the assumed log-normal distribution for different linewidths.

Lesson:
Controlling material variation is perhaps more important than coming up with better median /nominal properties

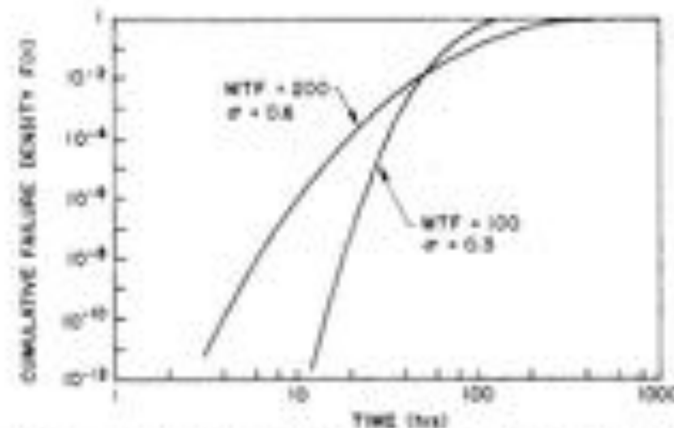


Fig. 4. Cumulative failure density for a log-normal distribution plotted as a function of time for different parameters. Note the influence of σ particularly at early times on the cumulative failure density.

This led to the so-called Sandwich metallurgy

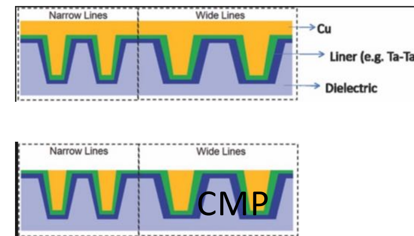
This led to the Extrusion monitor

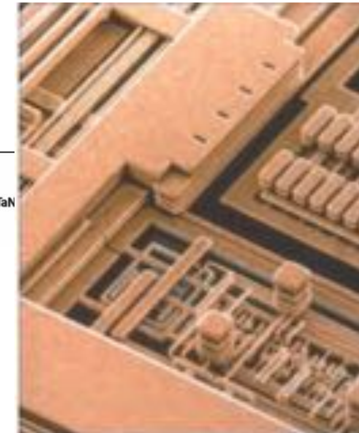
But the new failure mode was by hillock formation: shorts to neighboring wires
And passivation cracking

Two more big events

- The transition from Bipolar chips to CMOS chips
  - The power and currents dropped significantly
  - Electromigration was less of a problem
  - Went back to Al-Cu

- But Wire delay was mounting
  - Made the transition to copper wires
  - Dual damascene with TaN liners
  - Electromigration went away (for a while at least)

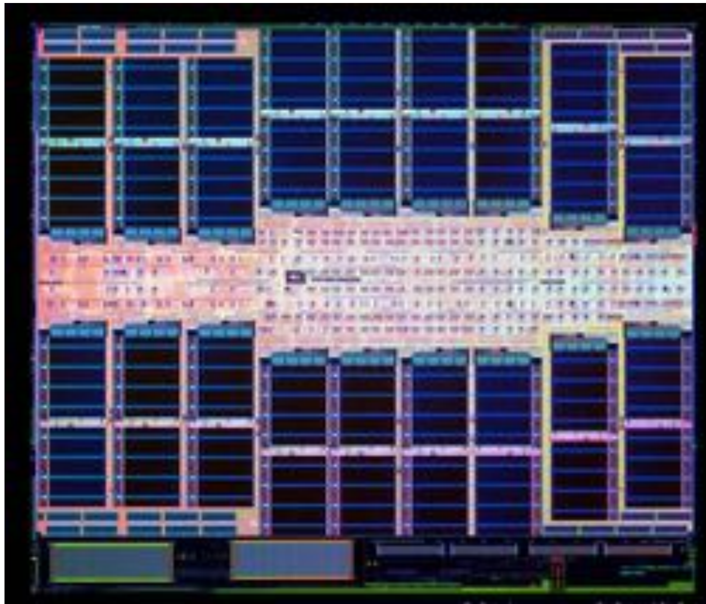




UCLA Samueli School of Engineering
CHIPS
CENTER FOR HETEROGENEOUS INTEGRATION AND PERFORMANCE SCALING

L3 caches in servers

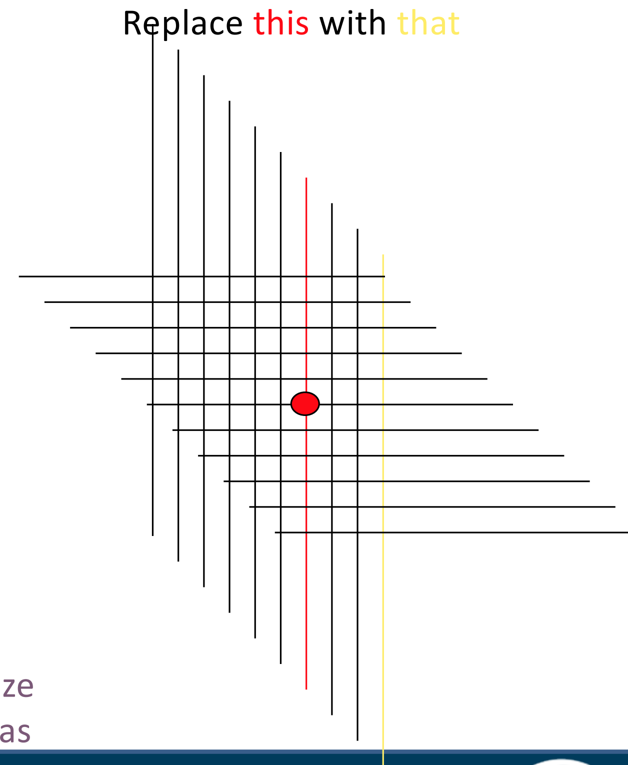
Yielding Large Memory Chips
Requires redundancy



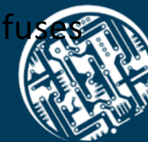
L3 cache used in P5
344 Mb, 430M transistors
Largest ASIC made at IBM

Note: a chip this size can have as many as 32K fuses

Replace this with that



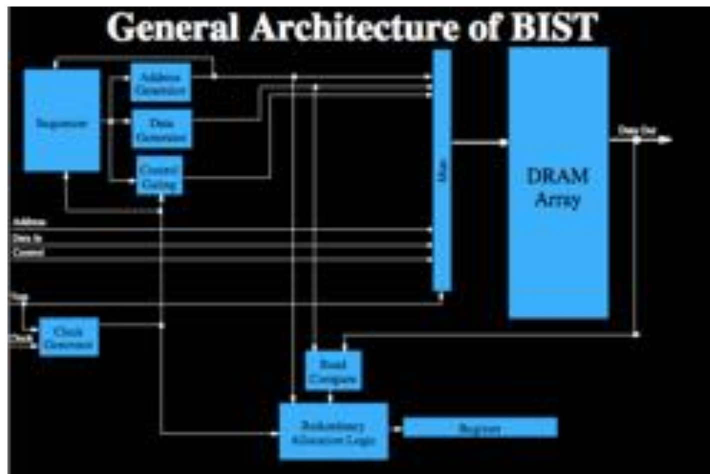
Redundancy invoked using fuses



Built-in Self Test and repair

Key Enablers: BIST & Redundancy

- Test at High speed
- Very large bandwidth but very few pin-outs
- Solution : Since you have access to a high speed logic technology why not build the tester on-chip
- Next step : repair faulty chips!



- On chip test engine
- Hard & Soft Patterns
- Allocates Redundancy
- Tests redundant elements
- Generates Fuse String

Laser Fuses

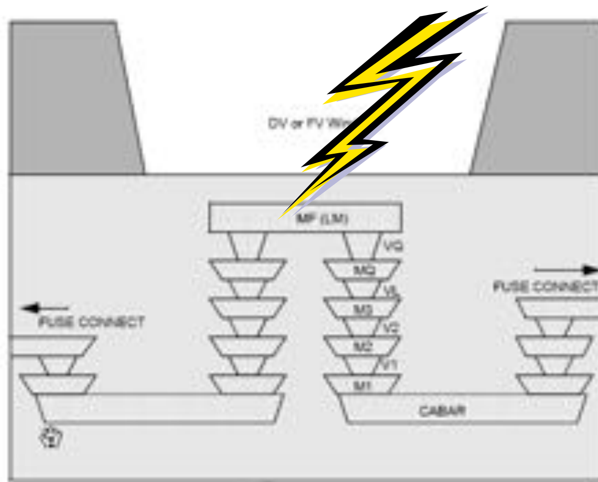


Figure 27: Fuse Cross-Section example - LM

Test
Determine repairs
Move wafer to laser fuser
Move repair data to laser fuse
Blow laser fuses
Take wafer back to tester
Test again to verify
Hope nothing breaks again ever!

- Do not scale & occupy too much space
- Block wiring and C4s
- Need to be exposed
- Can be blown only at wafer level
- Need precise mechanical alignment
- Require a complex laser fuser
- Require multiple wafer handling and data manipulation

Net: Laser Fuses are a pain!!

Copper Fuses were even more painful as they corroded as well

You could heat up a wire till it breaks

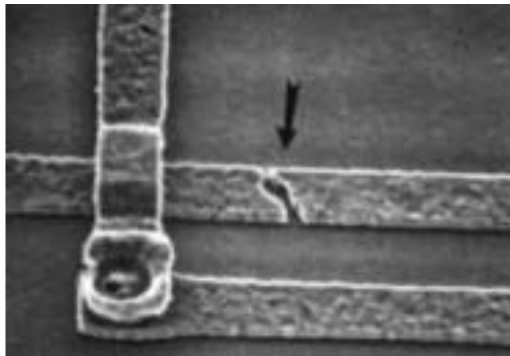
But seriously do you want this on your chip ?

Poly Si Fuse Programmed by rupture

Fine Print: Explosion here has been somewhat exaggerated

The Kinder Gentler Fuse

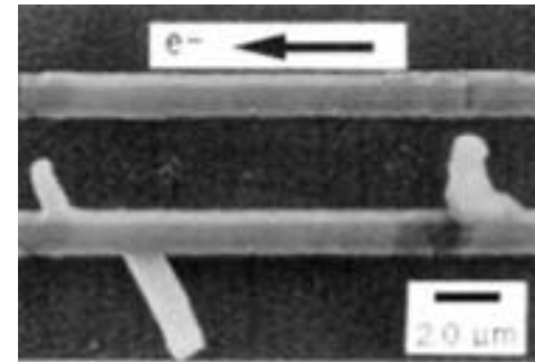
We need to induce an electrical open without needing material to disappear.



$$\vec{F} = N \frac{D(T)}{kT} Z^* q \vec{E}$$

$$\vec{F} \neq \mathbf{0}$$

$$= f(T, J^n)$$

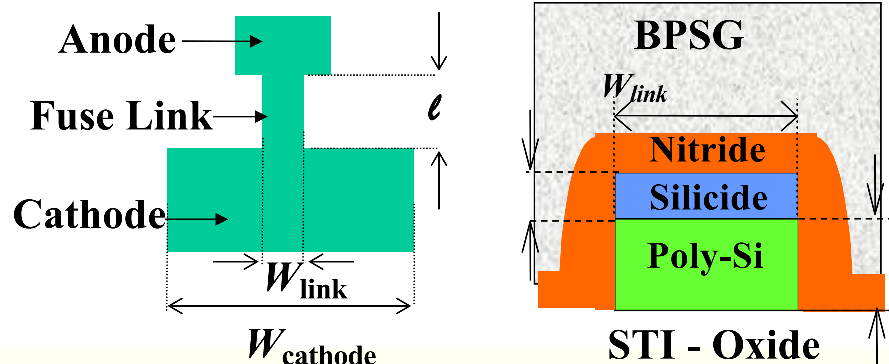


Can we employ electromigration of metal lines ?

Modern interconnects are electromigration resistant

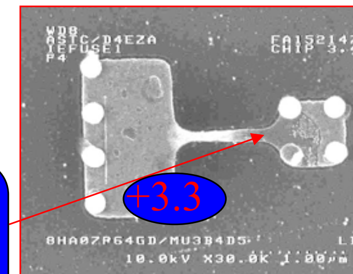
Need to control the electromigration and complete in a reasonable time e.g. 200 µs – Cu line not an option

eFUSE - physical layout

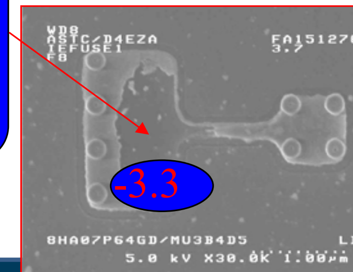


Key Features

- Geometry
- Thermal environment
- Current Characteristics

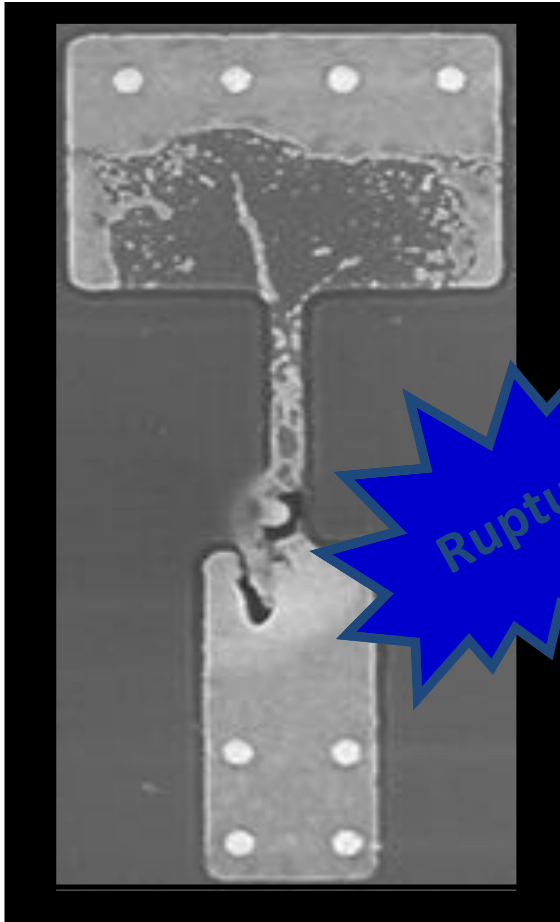


Silicide removal always occurs at cathode



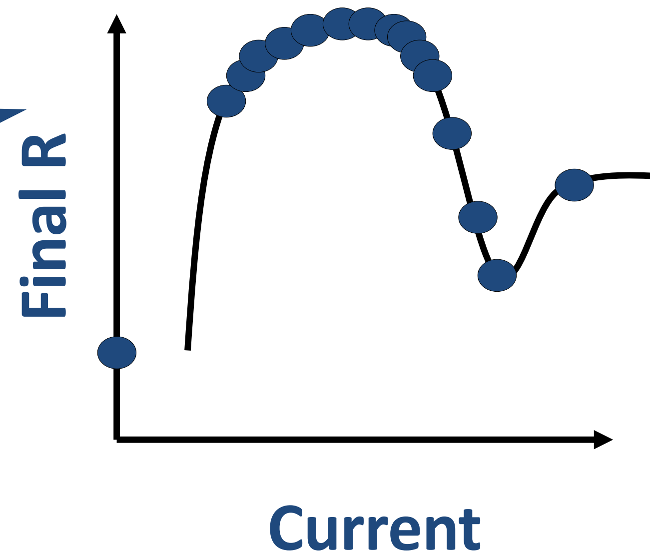
Mechanism:

- Current driven through silicide
- Temperature rises & gradient set up
- Silicide electromigrates but current is sustained as the Poly Si is hot intrinsic and conductive
- Electromigration of silicide is forced to completion
- Current turned off, everything cools and link is high resistance



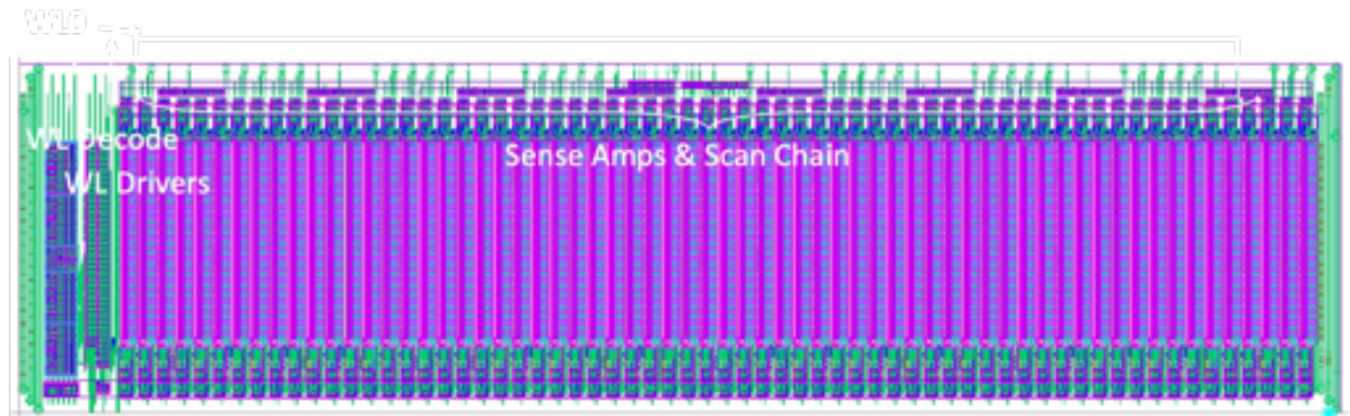
Poly Si/silicide eFUSE

Intact Fuse

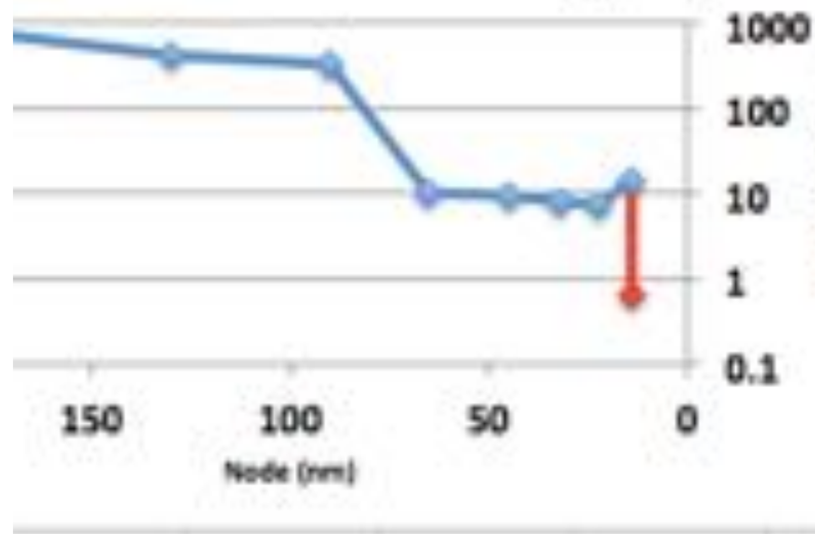


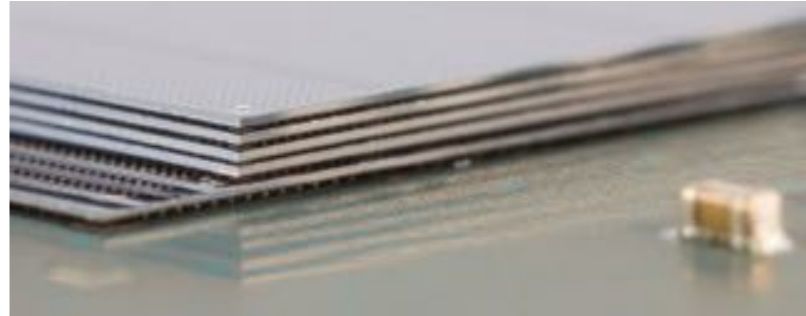
Further developments

- Optimizing chips for power and performance
- More autonomic functions
- Supply Chain management
- Chip identifiers for authentication
- RFID tags
- Logic compatible 2D OTP



Scaling eFUSE

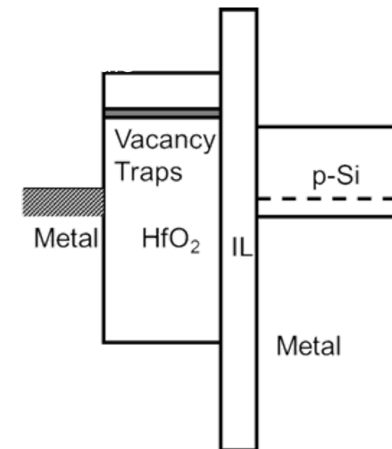




- After some dramatic scaling, scaling eFUSE scaling saturated
- In "14" nm it reverse scaled
- At the same time there has been another big transition to Hi K metal gate CMOS – No more poly-Si gate
- And 3D stacked memory made even more demands on redundancy
- And yet another opportunity to leverage this new material

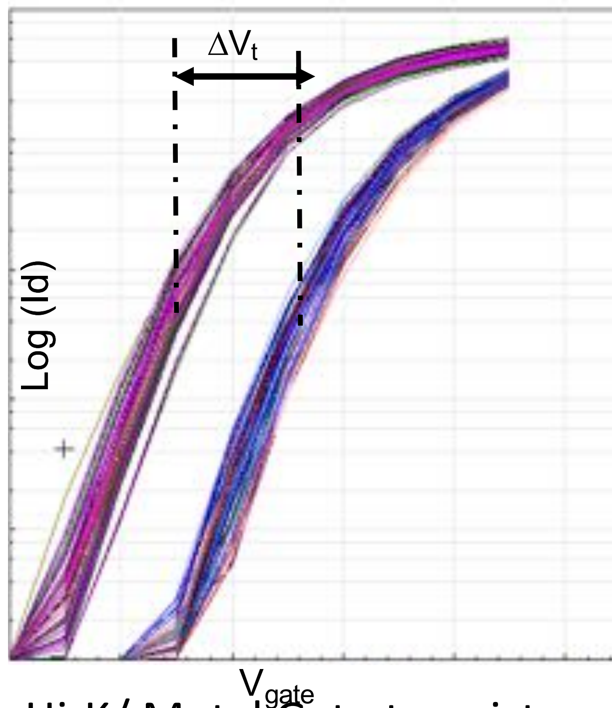
Hafnium Oxide vacancy

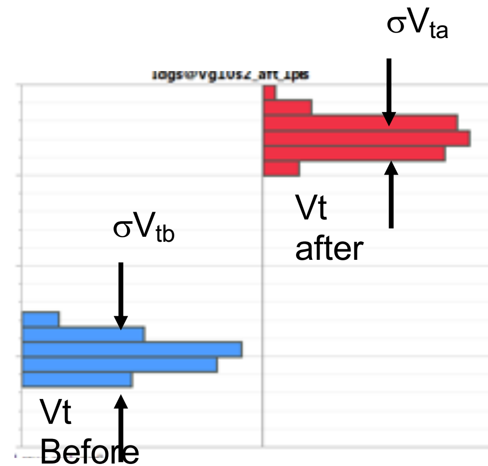
- Hafnium Oxide is the gate dielectric of choice for advanced CMOS (High K)
- Propensity to form oxygen vacancies
- Low formation energy
- Traps carriers resulting in threshold shifts
  - Similar to SONOS not Floating Gate
  - But CMOS compatible!!



HiK/Metal gate
Band diagram

Clear distinction between low and High states





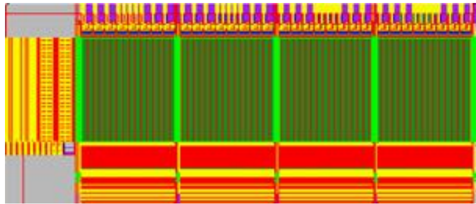
Can it be used as a replacement for eFUSE with the advantage of limited rewritability ?

$$\Delta V_t > \{\sigma V_{ta}, \sigma V_{tb}\}$$

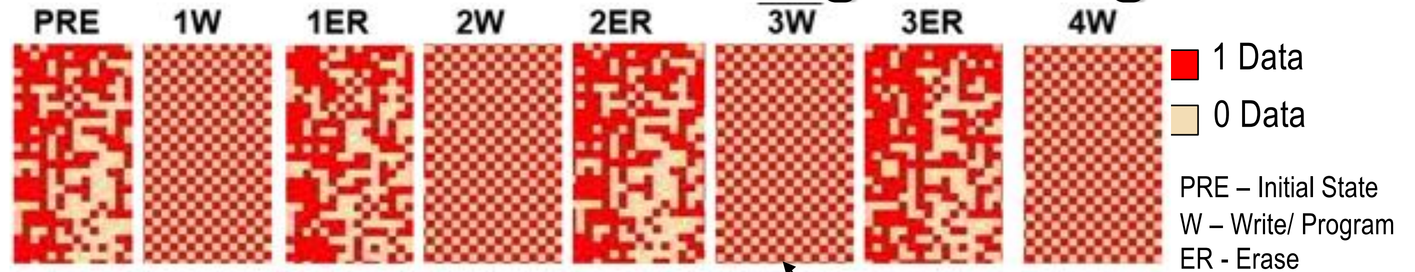
- Hi-K/ Metal Gate transistors have inherent fluctuations
- But, shift is larger than the variability due to other sources -> stable memory operations

Kothandaraman et al (2015)

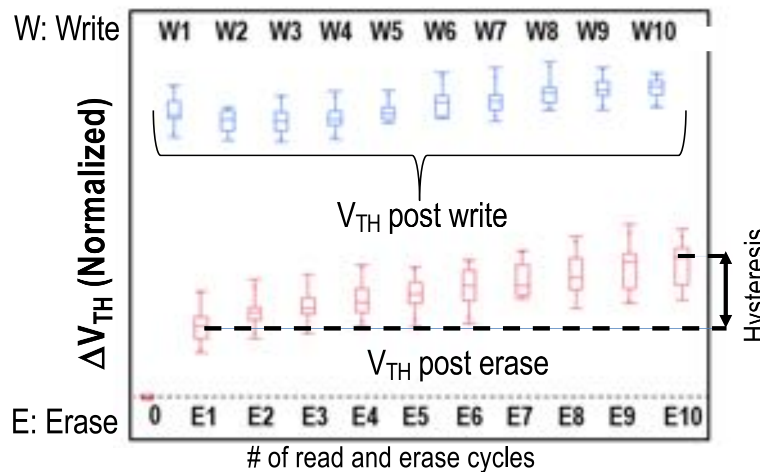
Hardware Results – Multi-Time Programming



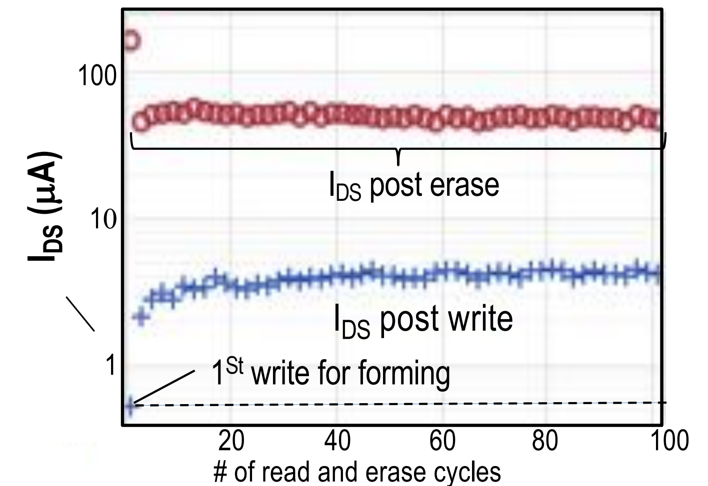
Chip image of 64Kb array



a: Multiple Write (4X) with OWP



b: ΔV_{TH} Change in Multi-Write (32nm)

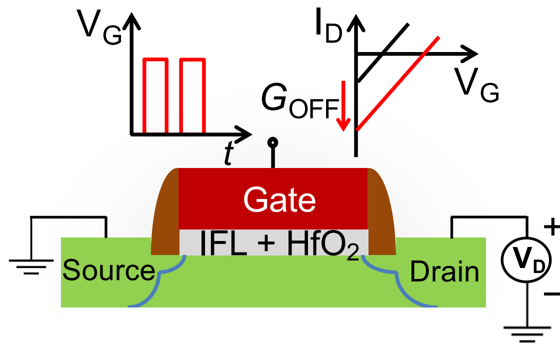


c: Current Change in Multi-Write (22nm)

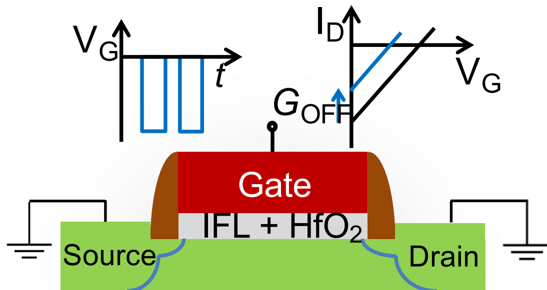
V. Janakiraman et al
VLSI Symp 2016

Use of CTT as analog memory

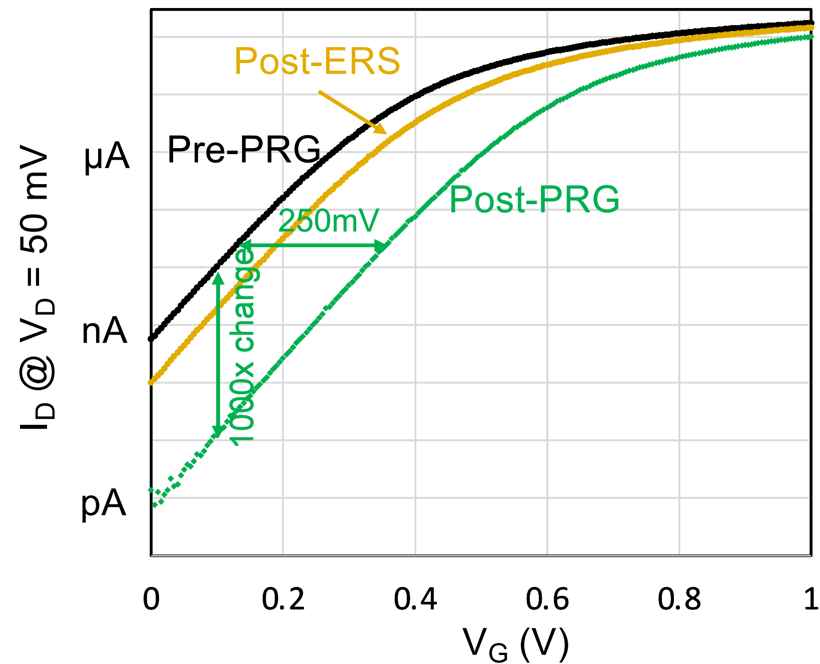
Increasing threshold voltage:



Reducing threshold voltage:

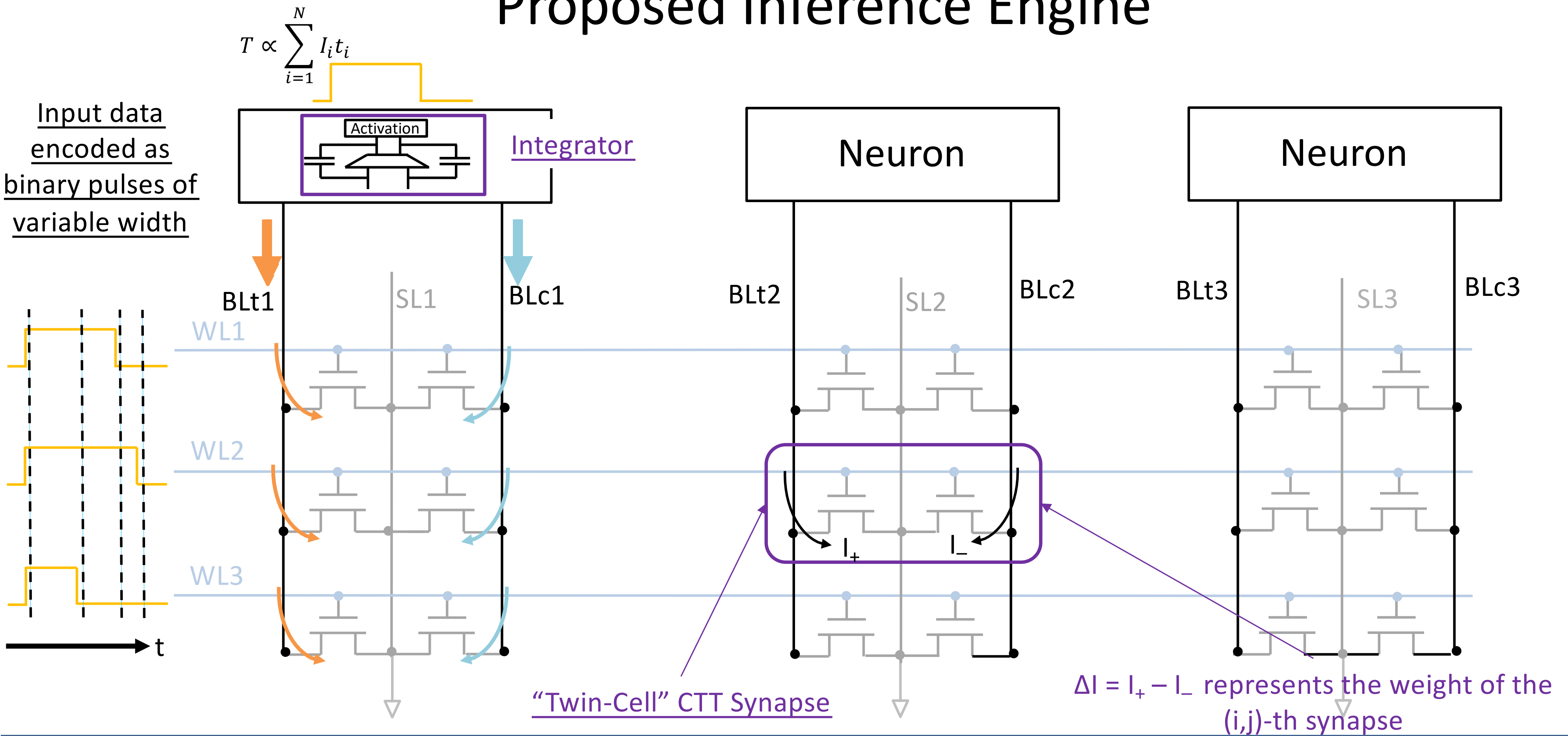


The CTT can be programmed and erased



Gu et al (2017)

Proposed Inference Engine



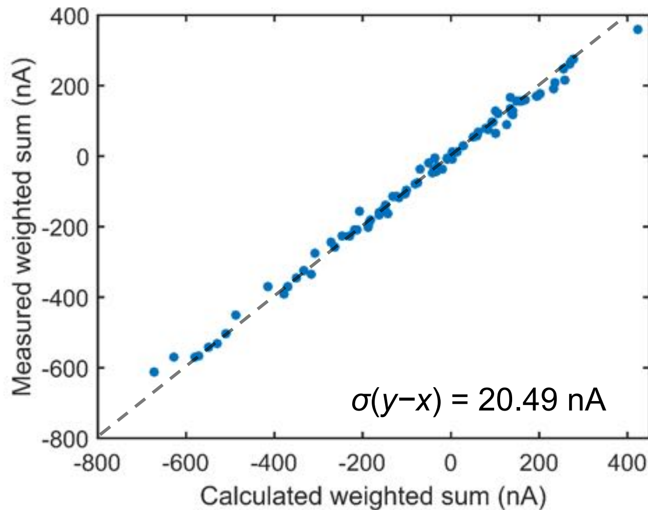
Stability of the Dot Product Engine

- Measurement time contributes to the difference between measured and calculated weighted sum

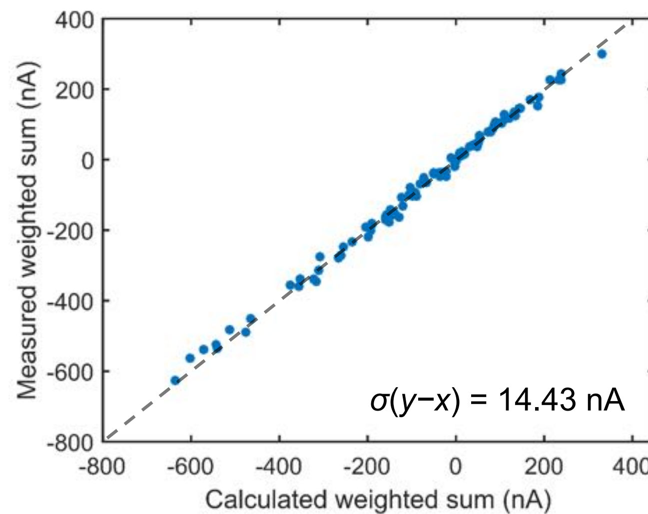
Input • Weight matrix = Output

measured from t_1 to t_2 measured from t_3 to t_4

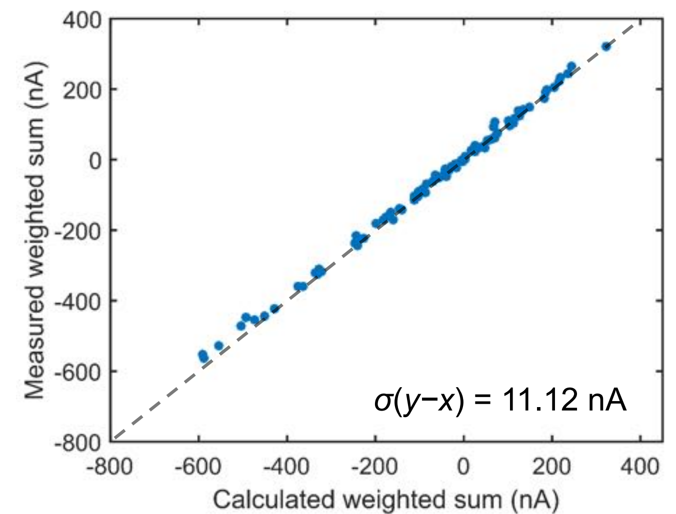
Immediately after programming



6 hours later



18 hours later

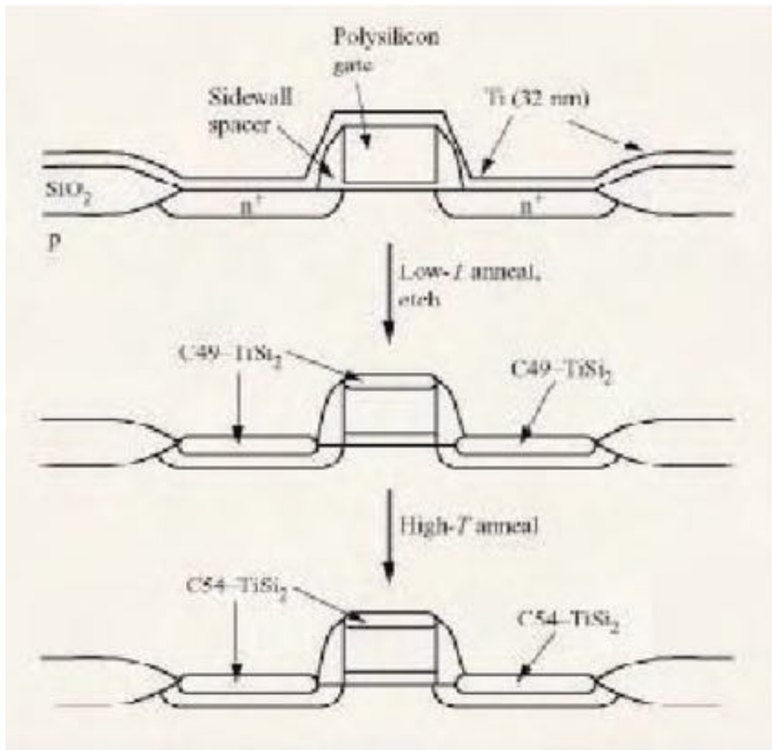


Recap

- Reliability issues are not necessarily bad news

- When used judiciously, they can be leveraged to solve other reliability and yield problems

Salicide – the case of the leaky furnace



- Bake off between different metal candidate: Ti vs Pt, Ni, Co
- Ti was preferred because of its higher temperature stability but it required extreme purity of the annealing gas
- But it formed by Silicon diffusion through the formed silicide to the metal silicide interface
- The others formed by metal diffusion through the silicide into the silicon

Why is this important ? -Bridging

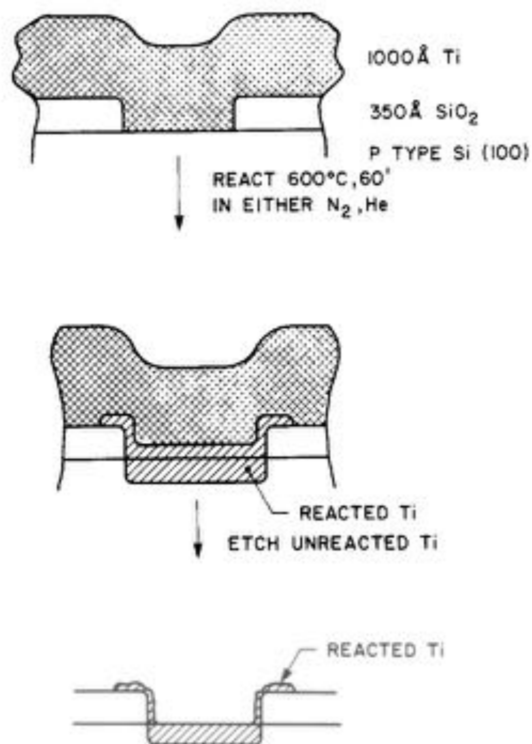
The leaky furnace

- Every now and then a leak would develop in the furnace
  - Usually because we let it cool too fast
- In those cases, the resistance of the silicide would be very high

But there would be no bridging

So, what was happening ?

So, what was happening ?



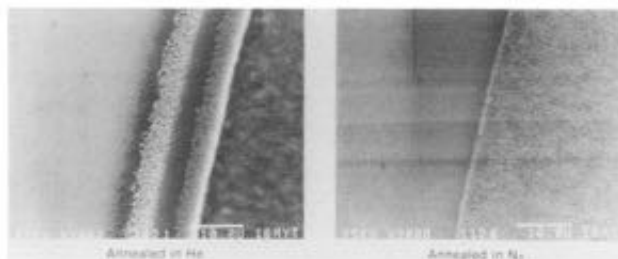


Fig. 12. Lateral formation of $TiSi_2$. a: Structure. b: SEM micrographs for 4000Å Ti on a 1000Å thick SiO_2 window after 8h at 600°C and Ti etch. The N_2 annealing arrests lateral migration completely, as compared to He annealing. Note that the N_2 annealed surface exhibits a rough surface morphology.

The solution was to stop annealing in pure He and anneal in Forming gas

Titanium Salicide wins round one!

- Titanium Salicide had a ling run – all the way to 0.25 μm in Logic and Memory and well below 90nm in BiCMOS
- There were many process enhancements that were needed to scale it to these dimensions
  - But I decided to quit when I was ahead

Building embedded DRAM

- The challenges of Building a DRAM in logic technology
  - Logic is leaky compared to DRAM
  - But logic transistors are significantly higher performance
- The world was divided:
  - The DRAM guys: you can never make a DRAM in Logic technology (that's why we have DRAM technology)
  - The Logic guys: DRAM technology cannot be used for Logic – it sucks
- The challenge was retention time

But was it ?

The parable of the two Alaskan Hikers

UCLA Samueli
School of Engineering

CHIPS
CENTER FOR HETEROGENEOUS INTEGRATION
AND PERFORMANCE SCALING

In cache applications date gets stale withing a few clock cycles

So, retention just needs to be a few clock cycles :
100’s of nano Secs vs 100’s of milli Secs

And the refresh rate is completely controllable by the system unlike in a commodity DRAM

To Conclude

I would like to thank my management and colleagues during my long career at IBM for not just tolerating me but encouraging me to make mistakes and learn

Never forget what you learned – it can come in useful later

Be observant: experiments that go wrong offer useful clues – even if you did not plan the mistakes

Sometimes roadblocks are not