Device Challenges and Opportunities

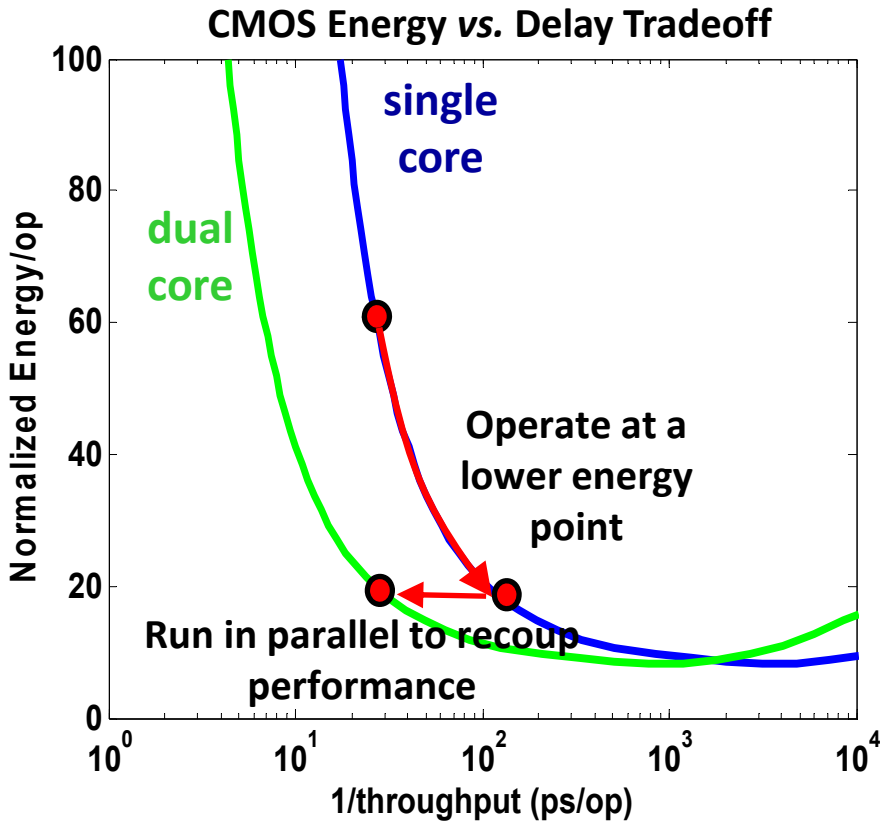
Prof. Tsu-Jae King Liu

Electrical Engineering and Computer Sciences Department
University of California at Berkeley

March 24, 2016

DOE Workshop on Energy Efficient Electronics

Improving CMOS Energy Efficiency

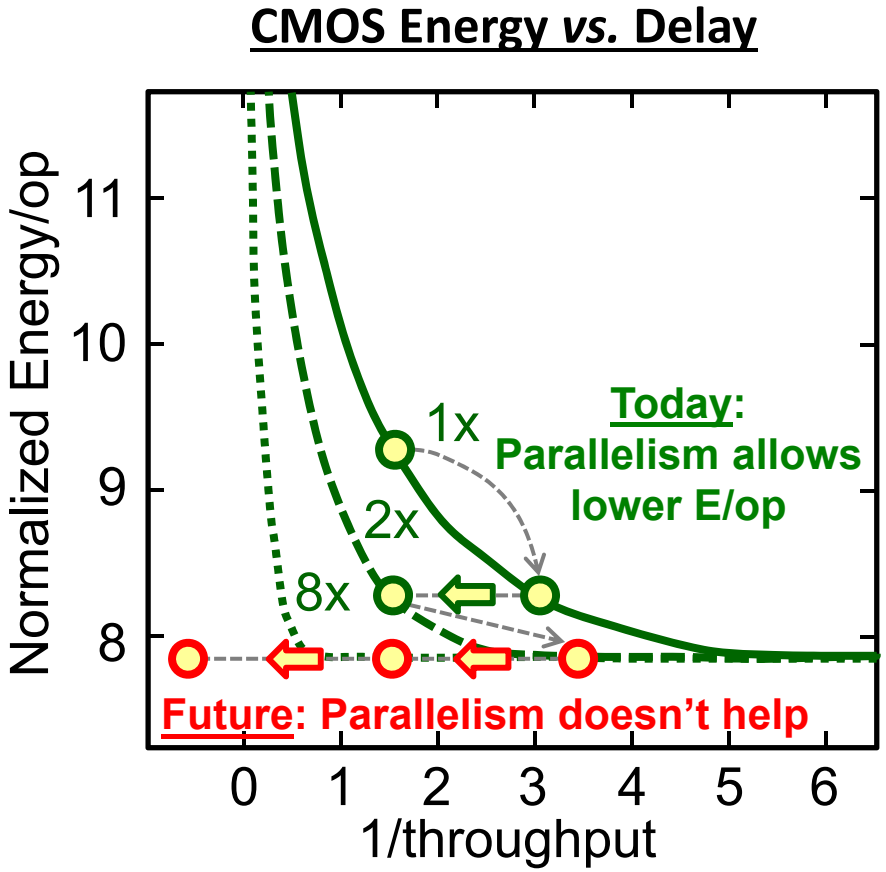


- To reduce power consumption, the chip operating voltage must be reduced – but this results in slower circuit operation.

→ Parallelism (multi-core processing) is used today to improve system throughput, within power constraints.

Game Over for CMOS

- **When each core operates at the minimum energy, increasing chip performance requires more power.**



Alternatives to the MOSFET as a logic switch will be needed!

- **The energy-delay tradeoff for a CMOS logic circuit can be understood by considering a cascade of inverters:**

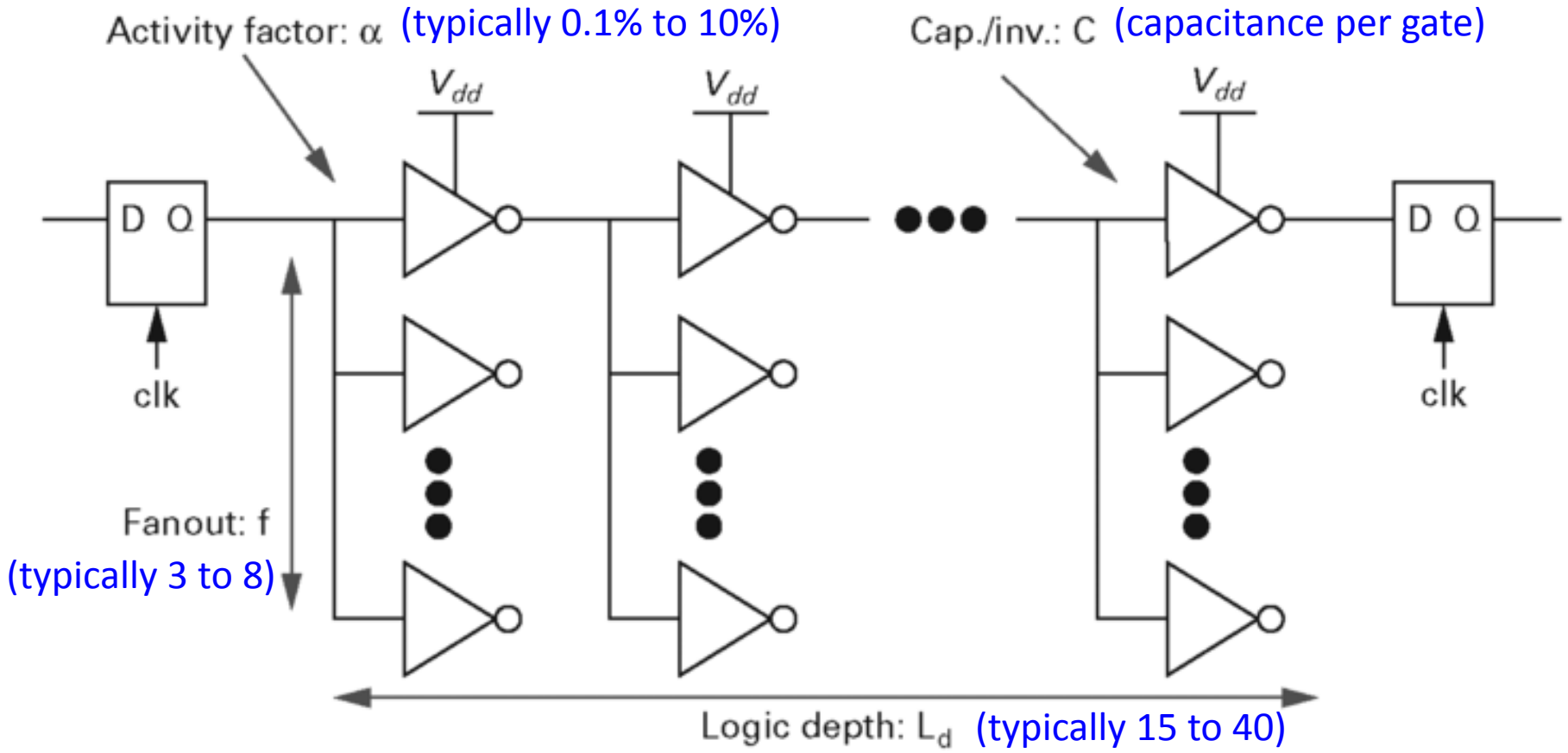
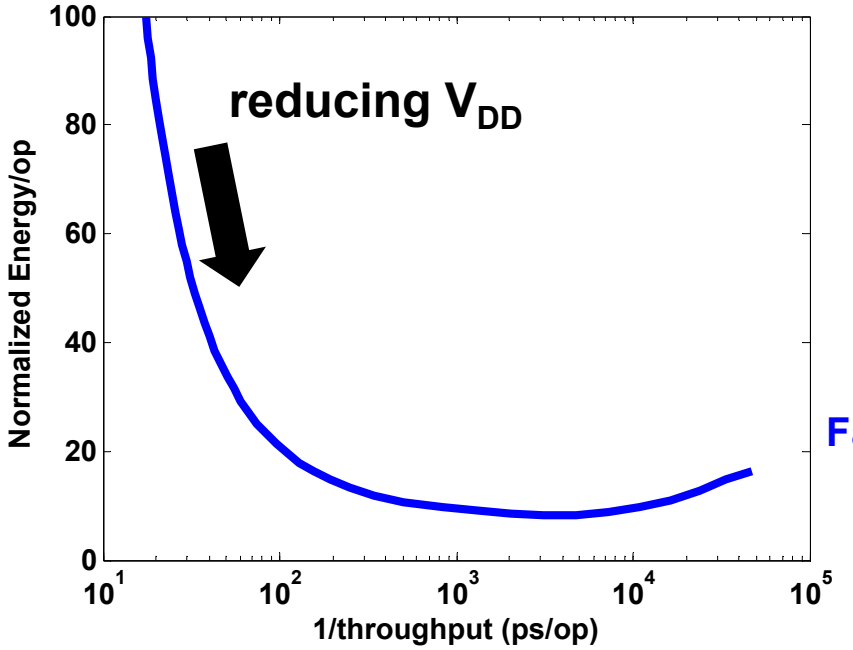


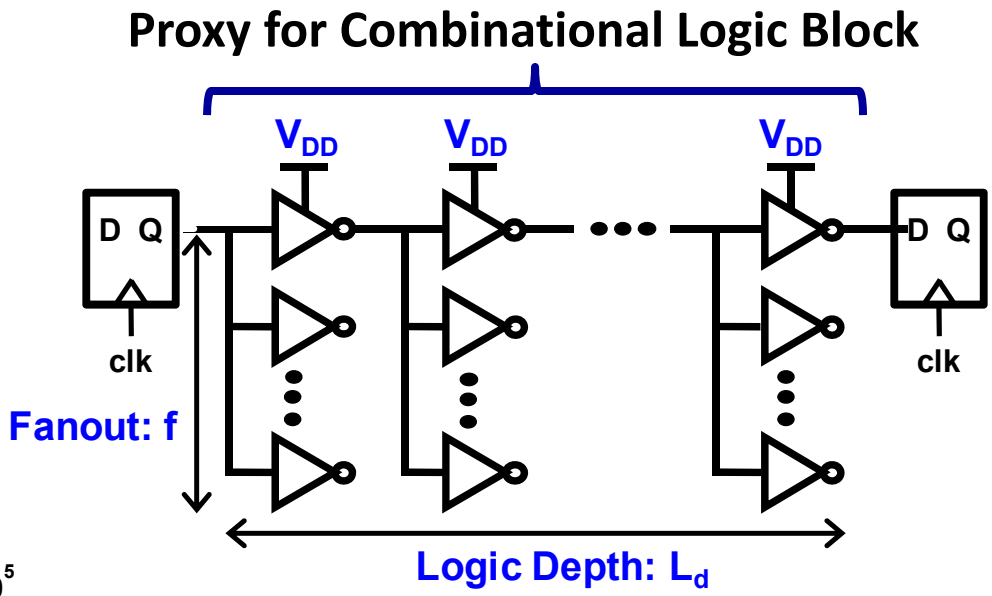
Fig. 1.2 Inverter-based model for combinational logic energy and performance.

E. Alon, Ch.1, *CMOS and Beyond: Logic Switches for Terascale Integrated Circuits*, Cambridge University Press, 2015.

- **The clock frequency of the microprocessor is limited by the delay of the combinational logic between the clocked registers.**
  L_d stages, only 1 actively switching at a time (the other stages are static)

CMOS Energy per Operation





Active Energy Passive Energy

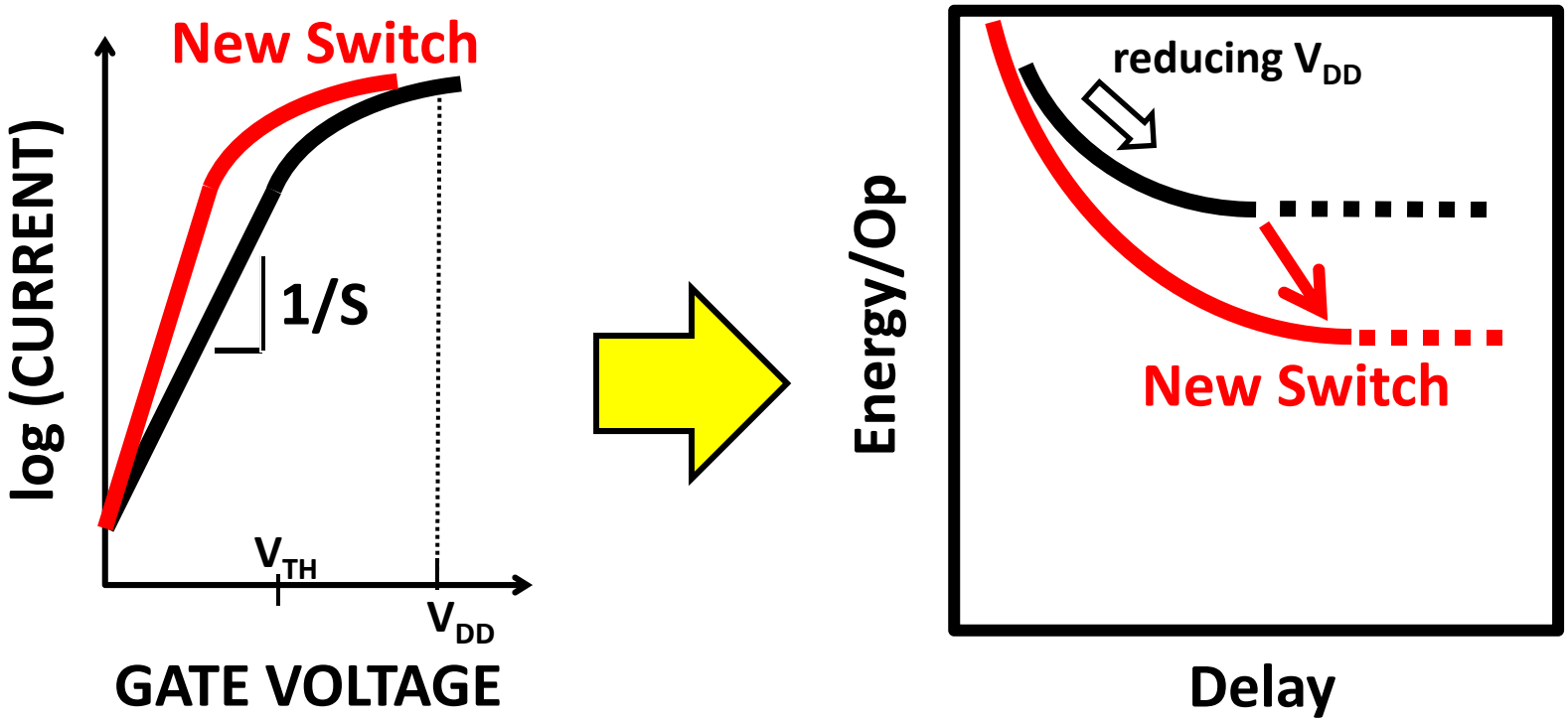
$$E_{total} = \alpha L_d f C V_{DD}^2 + L_d f I_{OFF} V_{DD} t_{delay}$$

$$t_{delay} = L_d f C V_{DD} / (2 I_{ON})$$

- Active Energy: consumed during gate switching (to charge/discharge voltages)
- Passive Energy: consumed when gates are static,due to MOSFET OFF-state leakage

New Logic Switch Requirement

$$E_{total} = aL_d fCV_{DD}^2 [\ 1 + (L_d f/2\alpha) / (I_{ON}/I_{OFF})\ ]$$

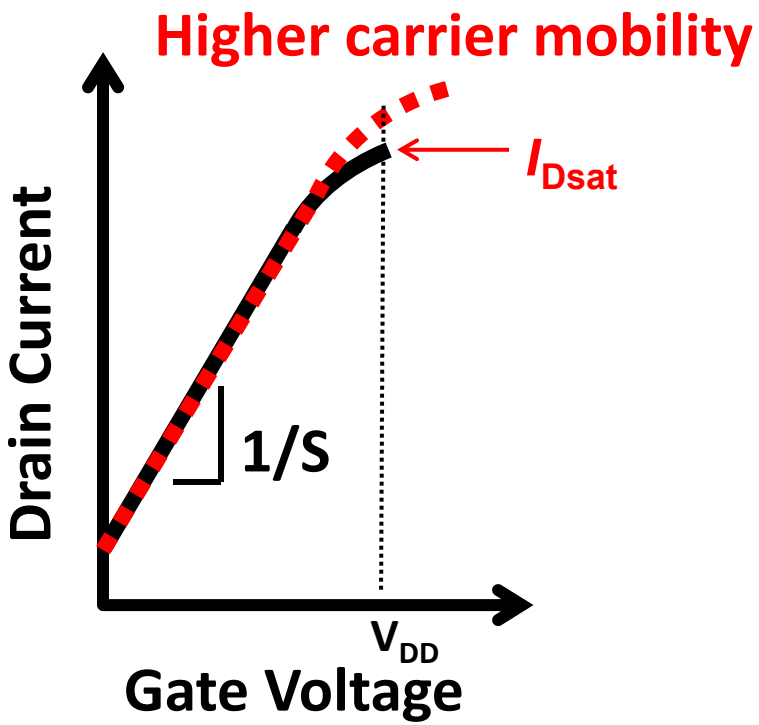


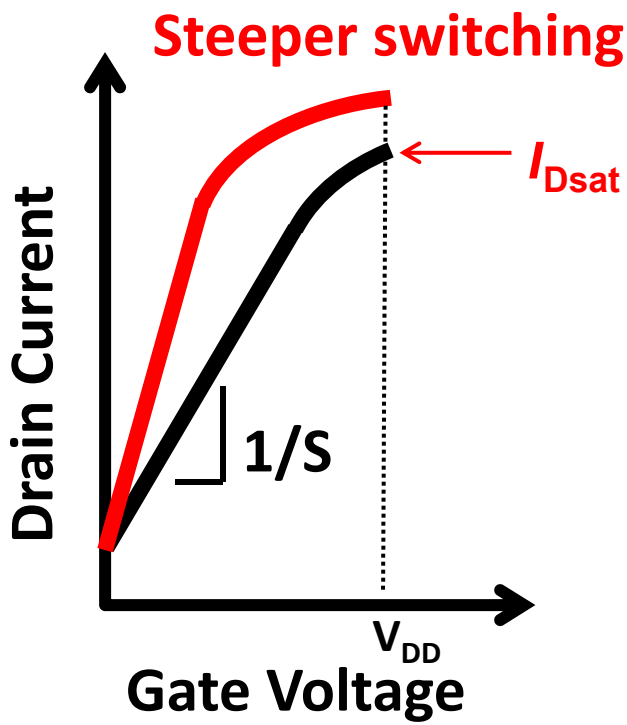
- Higher I_{ON}/I_{OFF} → lower Energy/op

→ Much steeper switching behavior is needed!

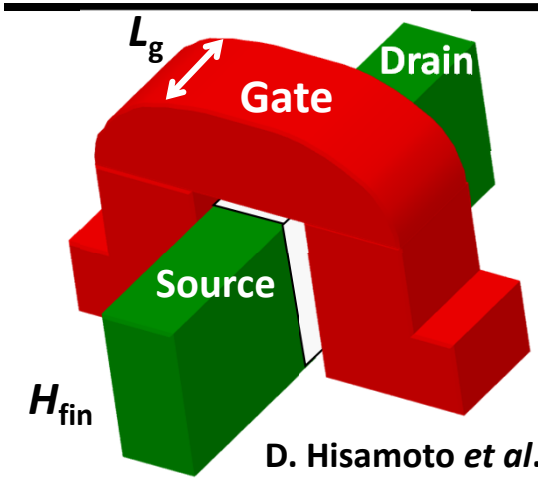
Approaches to Facilitate Voltage Scaling

- To operate with lower V_{DD} without sacrificing circuit performance (*i.e.* maintaining high ON-state current) for a given I_{OFF} specification, the MOSFET ON/OFF current ratio must be improved:

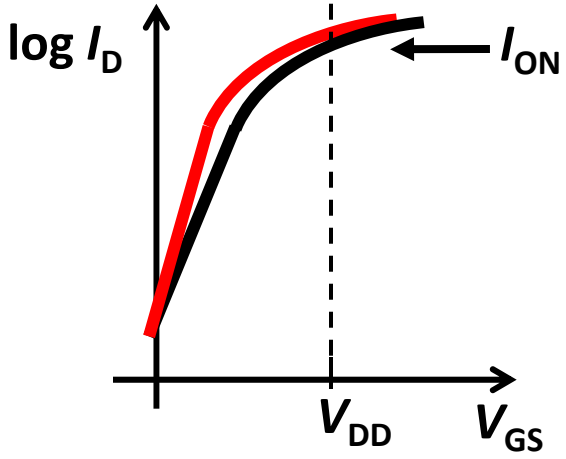




3-D Transistor ("Tri-gate" or "FinFET")



D. Hisamoto *et al.* (UC Berkeley), *IEDM* 1998

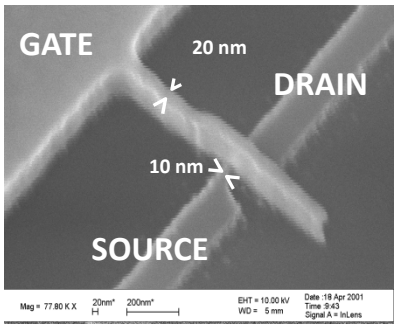


- Superior gate control

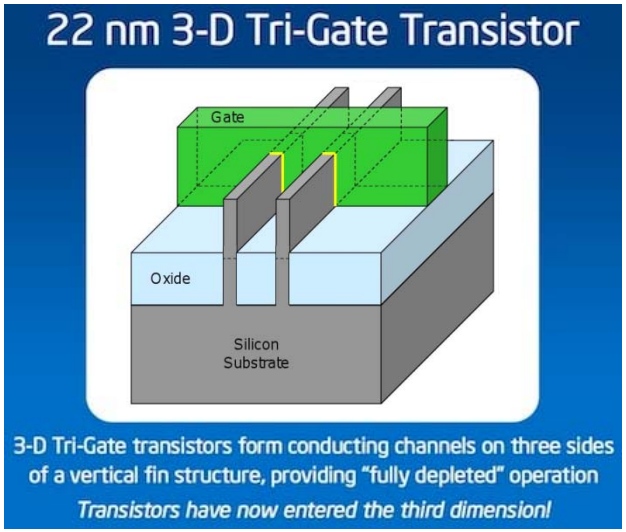
→ steeper switching

→ Lower V_{DD} for target I_{ON}

15nm L_g FinFET



Y.-K. Choi *et al.*, (UC Berkeley) *IEDM* 2001



Intel Corp., May 2011

- Multiple fins can be connected in parallel to achieve higher drive current.

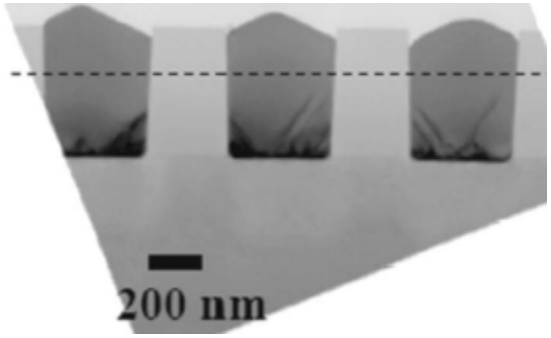
Advanced Channel Materials

- **High-mobility semiconductor materials potentially can provide for improved performance:**
  - **Ge for PMOS**
  - **(In)GaAs for NMOS**

	Si	Ge	GaAs
Electron mobility (cm^2/Vs)	1500	3900	8500
Hole mobility (cm^2/Vs)	450	1900	400
Lattice constant (Å)	5.431	5.646	5.653
Band gap (eV)	1.12	0.66	1.424
Dielectric constant	12	16	13

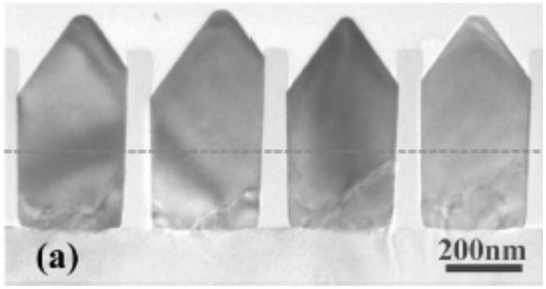
- **Selective epitaxial growth directly on Si is facilitated by the use of a corrugated substrate:**

Ge on Si



J.-S. Park *et al.*, *Appl. Phys. Lett.* 90 052113, 2007

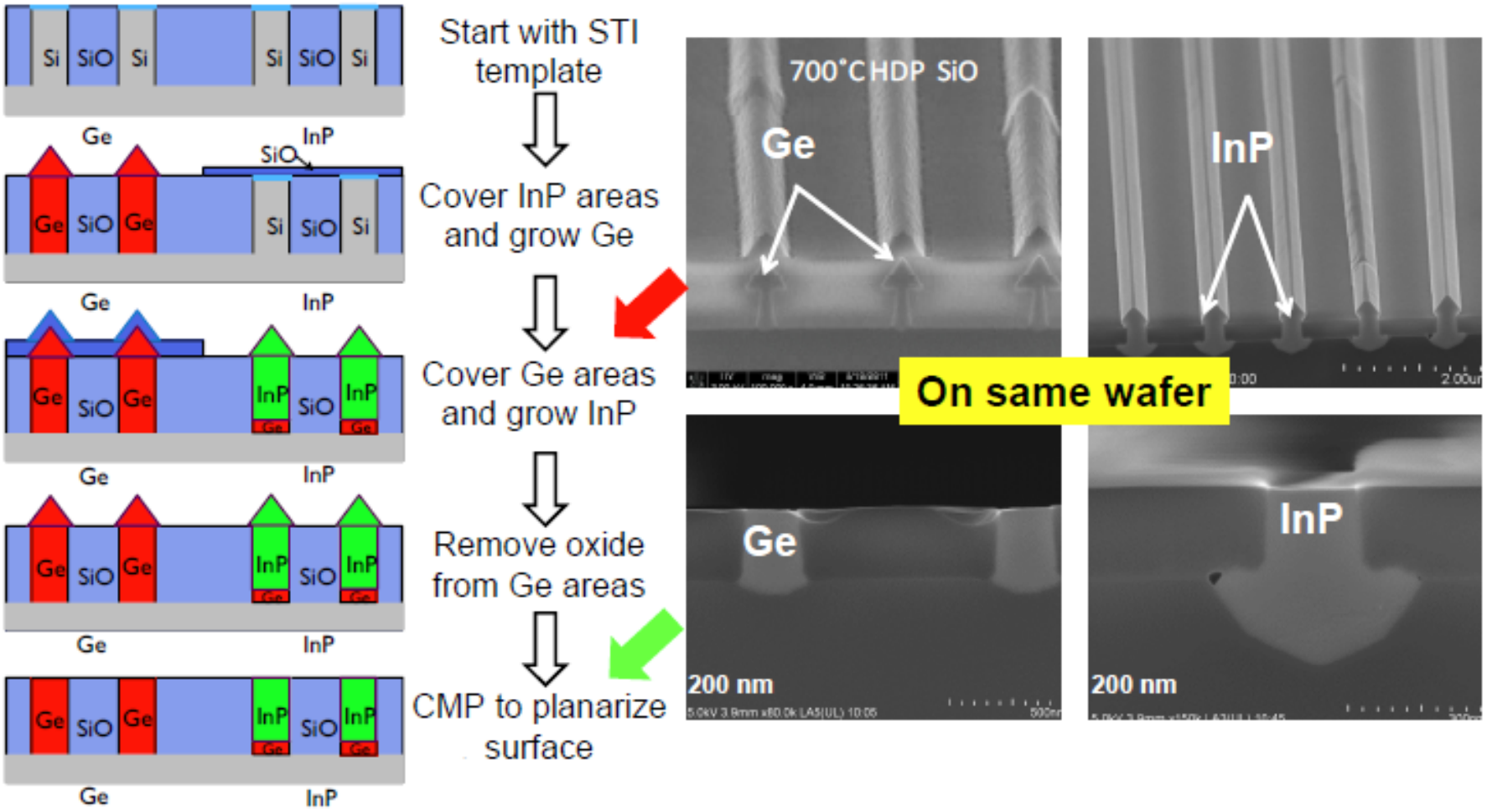
GaAs on Si



J. Z. Li *et al.*, *Appl. Phys. Lett.* 91 021114, 2007

Heterogeneous CMOS Integration

M. Heyns (IMEC), *EuroNanoForum* 2013

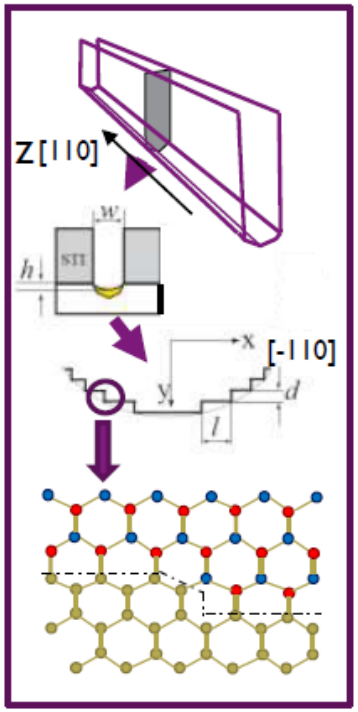


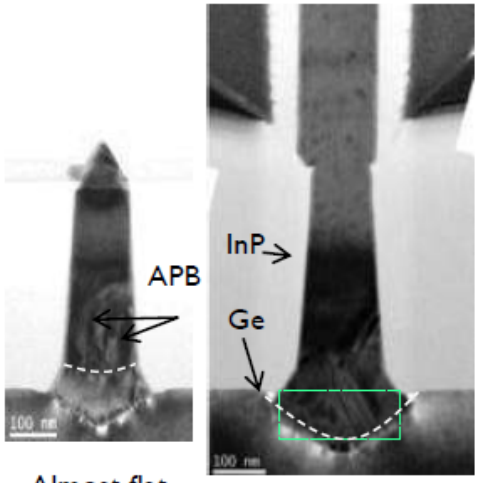
- Demonstration of CMOS Ge/InP virtual substrate by ART (Aspect Ratio Trapping)

Remaining Issues with ART

N. Waldron (IMEC), *ISTDM* 2012

"Perpendicular" view





Almost flat Ge surface

w/h=2.9

Efficient defect necking effect

Effective double step formation on the "rounded-Ge" surface

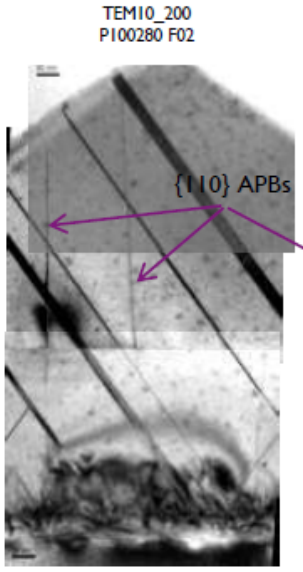
- APB observed only with an almost flat Ge surface

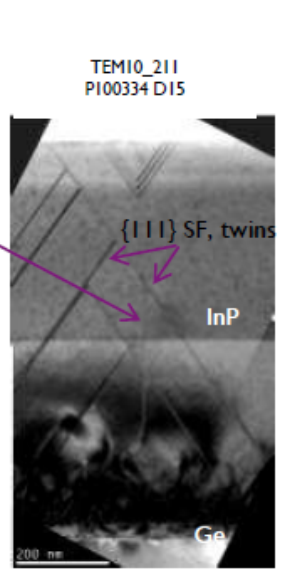
"Parallel" view

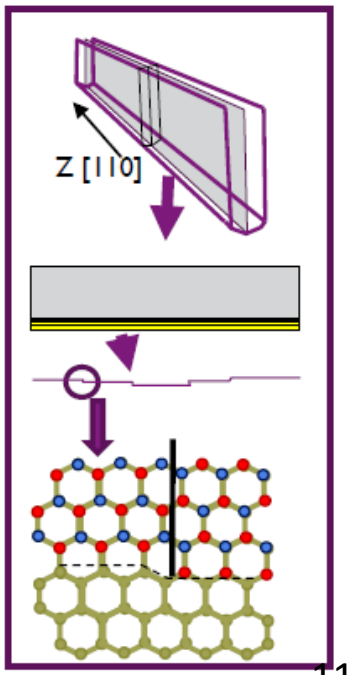
High defect density in parallel view

- twins/Stacking Faults/APBs

APBs originate from single steps along [110]?

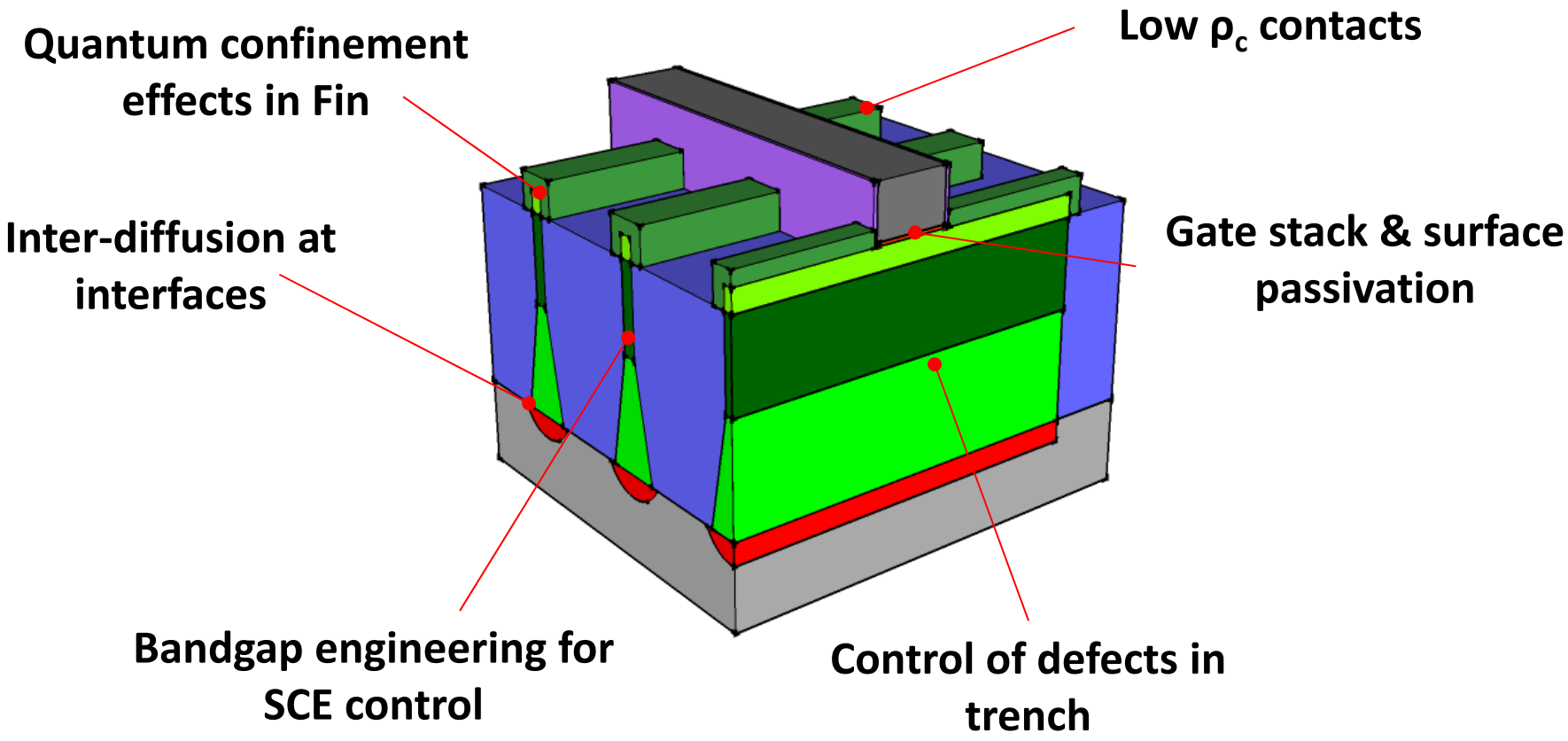






Challenges for FinFET Architecture

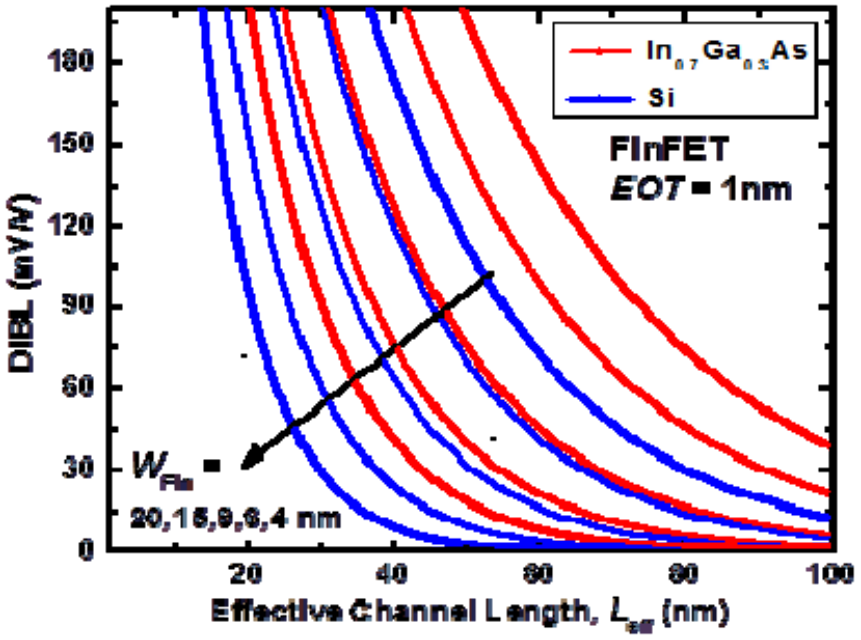
N. Waldron (IMEC), *ISTDM* 2012



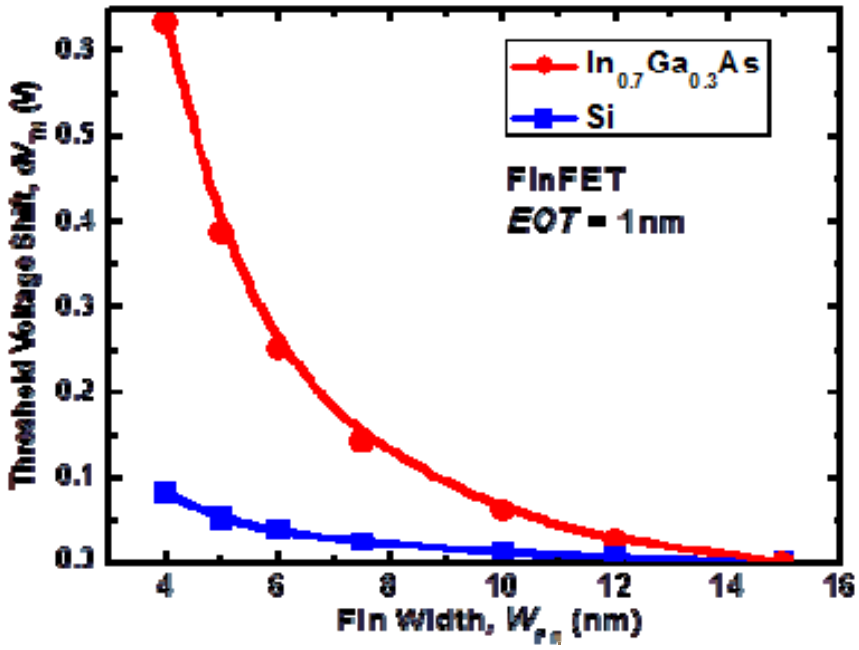
Si *vs*. $In_{0.7}Ga_{0.3}As$ FinFETs

N. Xu (UC Berkeley), unpublished

Drain-Induced Barrier Lowering



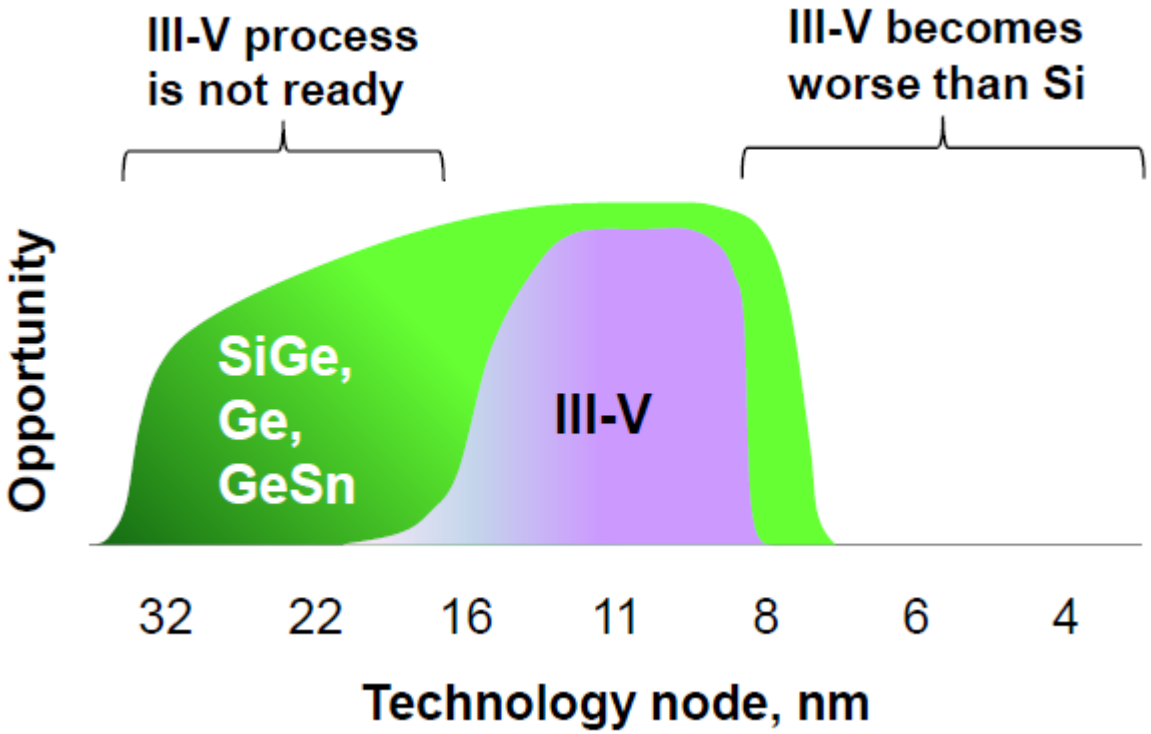
Threshold Voltage Increase



- Narrower fin width is required for InGaAs FinFET *vs*. Si FinFET
- V_{TH} is more sensitive to W_{fin} for InGaAs FinFETs

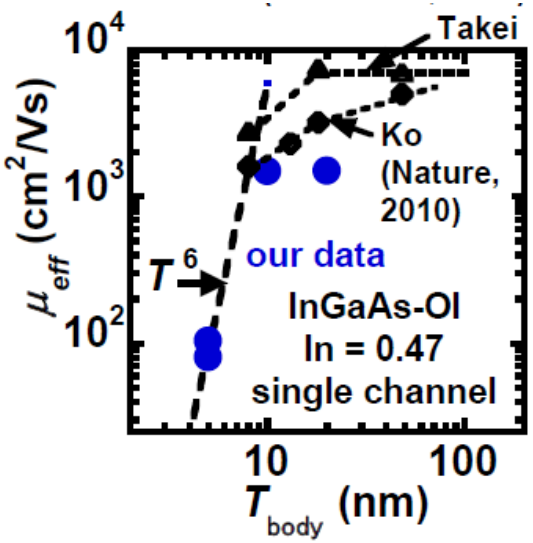
Outlook for III-V MOSFETs

courtesy V. Moroz (Synopsys, Inc.)



Lighter m^* → larger T_{inv}

→μ degrades at larger T_{body}



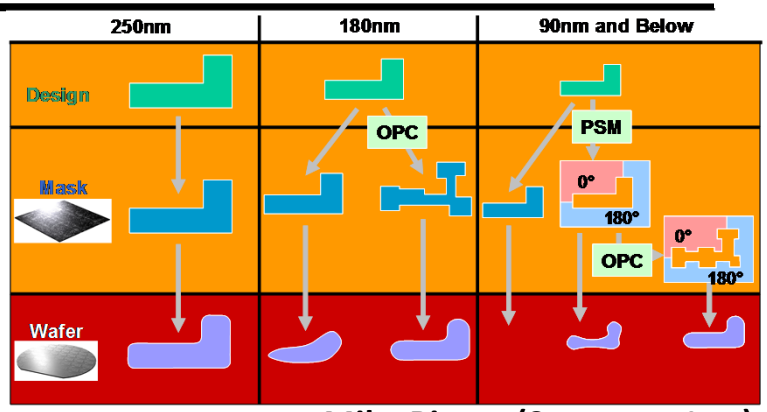
S. Takagi *et al.*, *ECS Transactions* 53, pp. 107-122, 2013

- Any new technology should last for at least 2 technology nodes
- $Si_{1-x}Ge_x$ channel is easier to manufacture

→ III-V channel materials have a narrow window of opportunity (?)

Sources of Variability

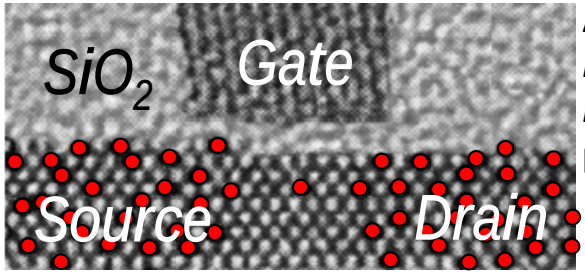
- **Sub-wavelength lithography:**
  - **Resolution enhancement techniques are costly and increase process sensitivity**



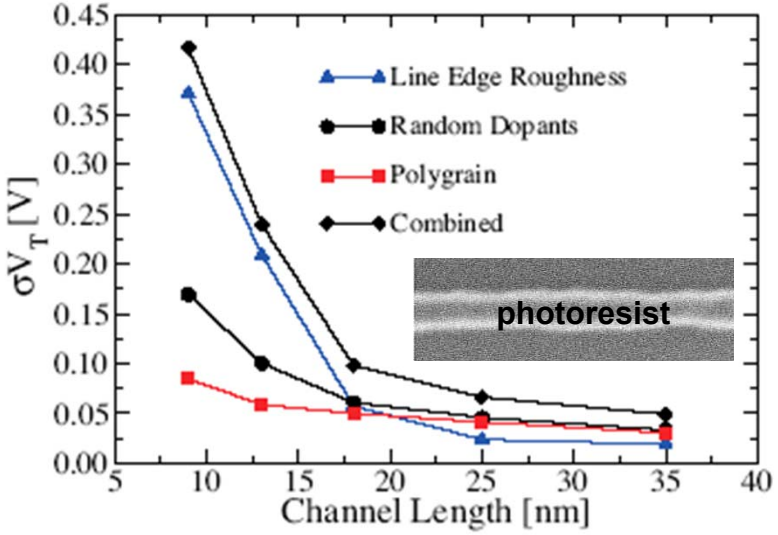
courtesy Mike Rieger (Synopsys, Inc.)

- **Layout-dependent transistor performance:**
  - **Process-induced stress is dependent on layout**

- **Random dopant fluctuations (RDF):**
  - **Atomistic effects become significant in nanoscale FETs**

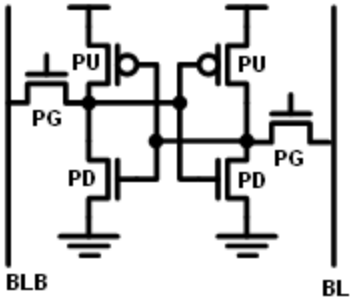


A. Brown *et al.*, *IEEE Trans. Nanotechnology*, p. 195, 2002



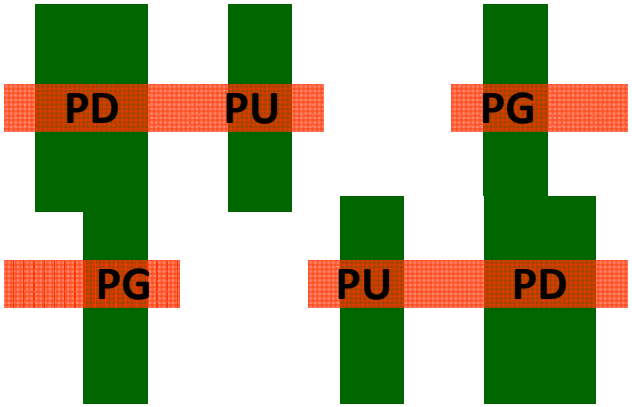
A. Asenov, *Symp. VLSI Tech. Dig.*, p. 86, 2007

6-T SRAM Cell

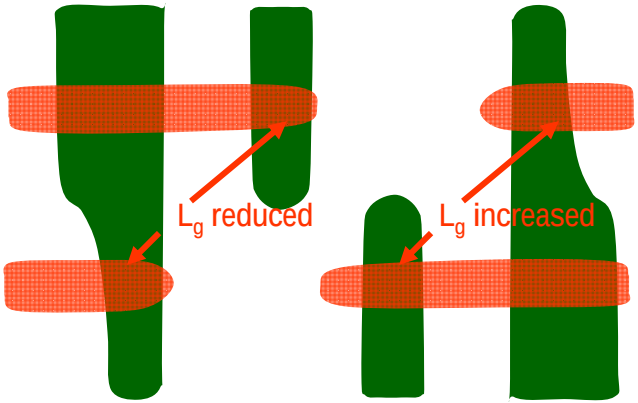


Impact of Misalignment

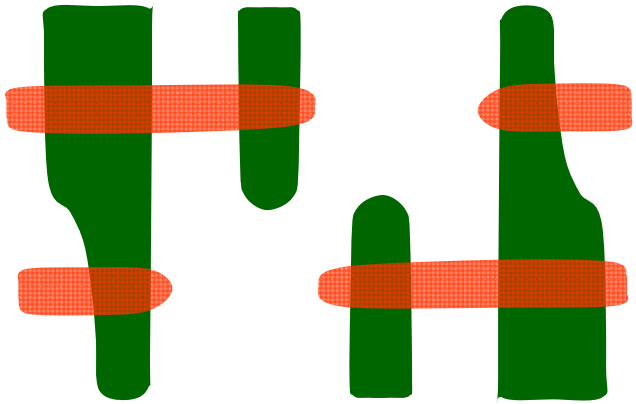
Desired layout
(6-T SRAM cell)



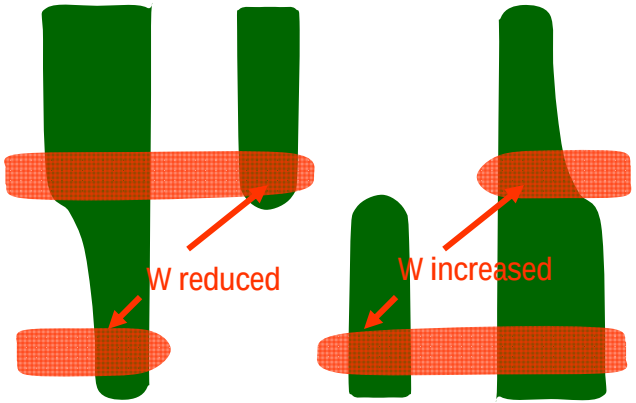
Actual layout w/ lateral misalignment
(gate length variations)



Actual layout
(corner rounding)

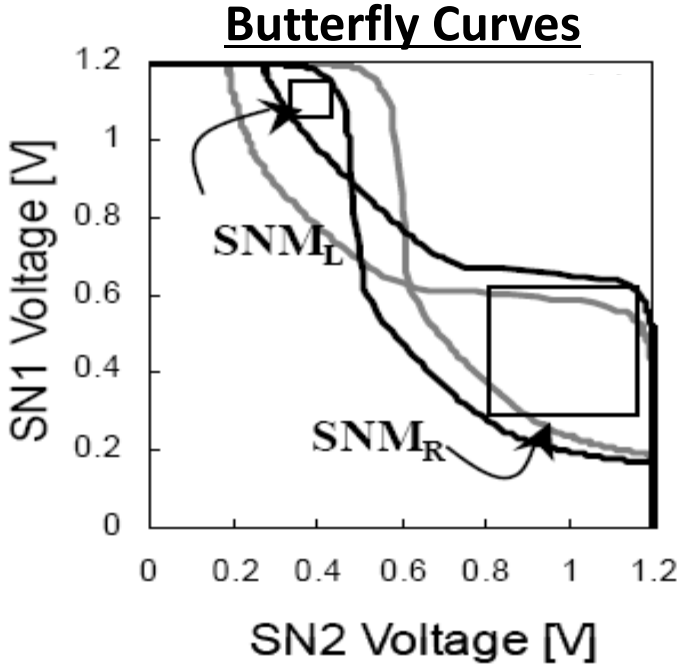


Actual layout w/ vertical misalignment
(channel width variations due to active jogs)



Impact of Variability on SRAM

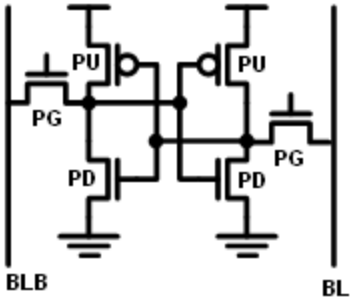
- V_{TH} mismatch results in reduced static noise margin.
  → lowers cell yield, and limits V_{DD} scaling



Y. Tsukamoto *et al.*, *Proc. IEEE/ACM ICCAD*, p. 398, 2005

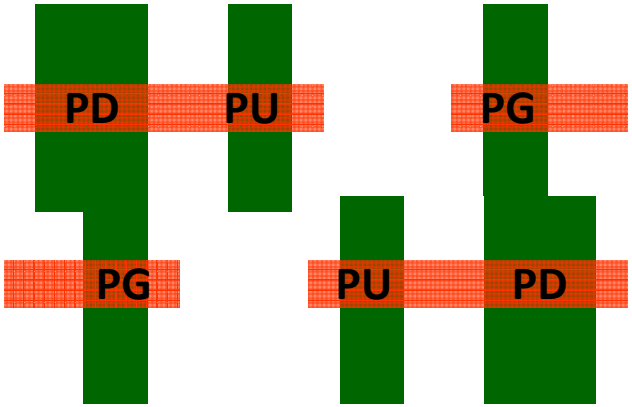
→ Immunity to short-channel effects (SCE) and narrow-width effects as well as RDF effects is needed to achieve high SRAM cell yield.

6-T SRAM Cell

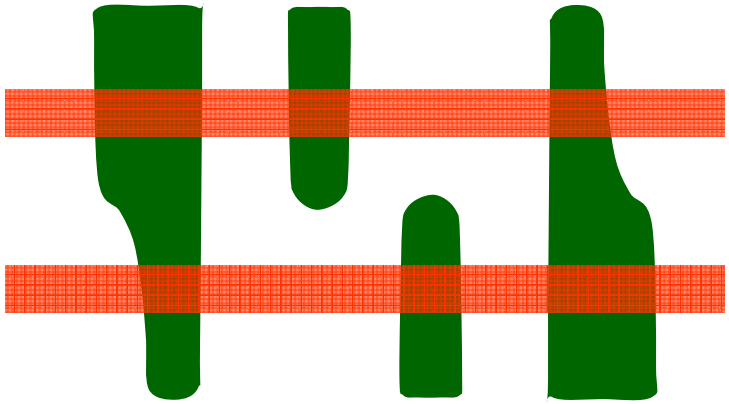


Double Patterning of Gate

Desired layout
(6-T SRAM cell)

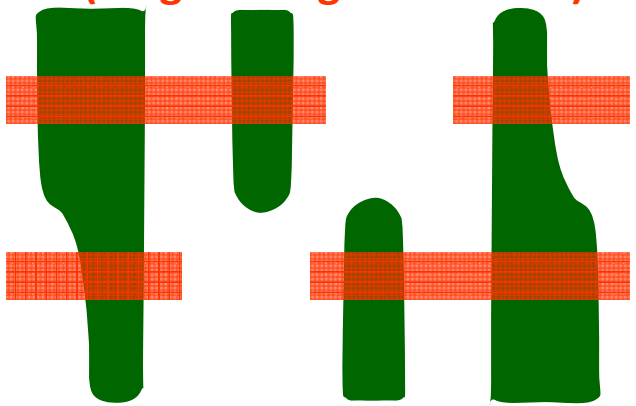


Actual layout after 1st gate patterning



Actual layout after active patterning

Actual layout after 2nd gate patterning
(no gate length variation)

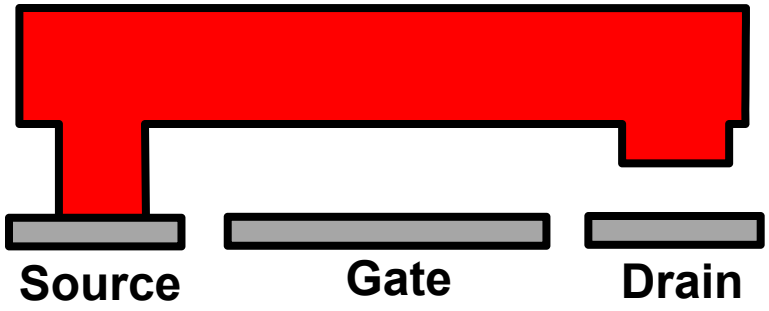


Future Device Requirements

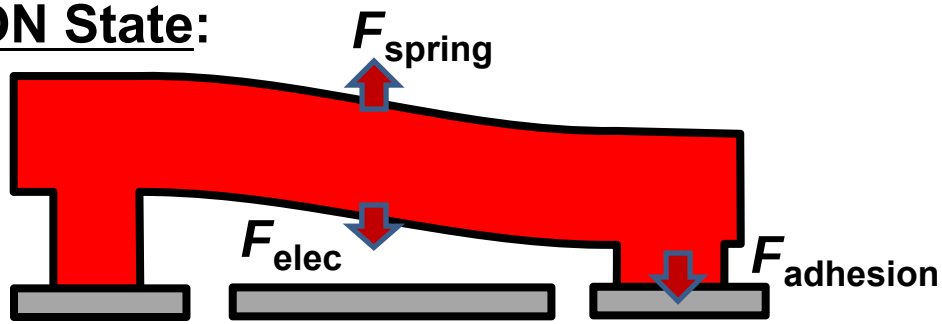
- **Low operating voltage** → **Low active power**
- **Robust to variations** → **Low cost**
- **Zero leakage** → **Zero standby power**

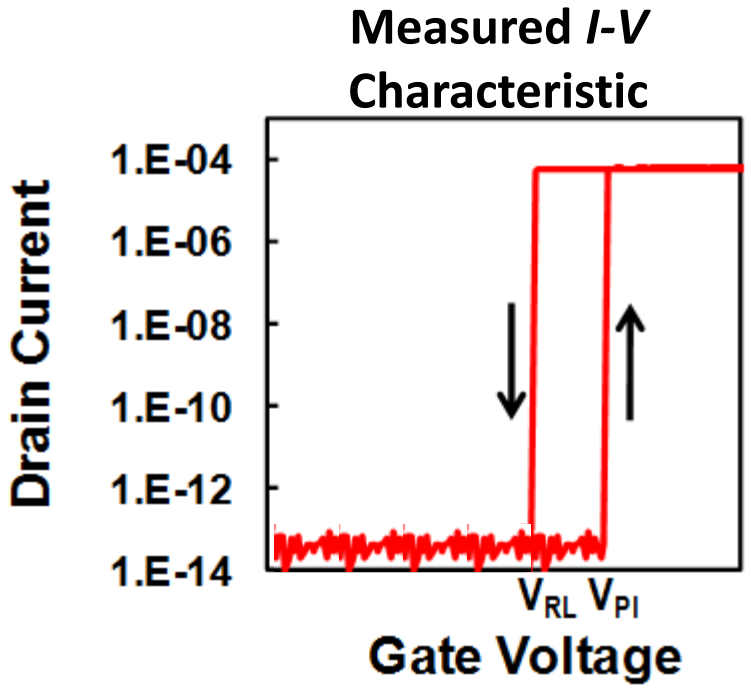
Micro-Electro-Mechanical Switch

OFF State:



ON State:

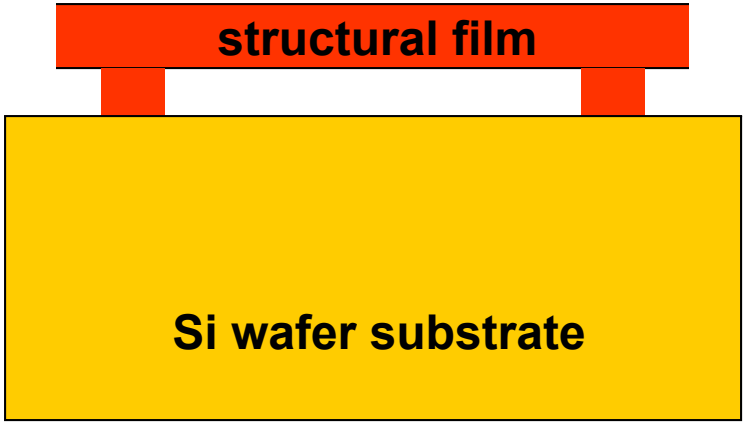




- **Zero OFF-state current (I_{OFF}); abrupt switching**
  - **Turns on by electrostatic force (F_{elec}) when $|V_{GS}| \geq V_{PI}$**
  - **Turns off by spring restoring force (F_{spring}) when $|V_{GS}| \leq V_{RL}$**

Surface Micromachining Process

Cross-sectional View



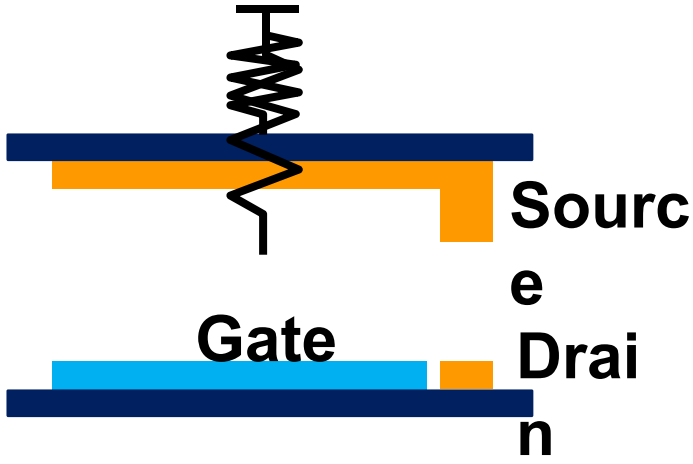
- **Mechanical structures can be made using conventional microfabrication techniques**
- **Structures are freed by selective removal of sacrificial layer(s)**

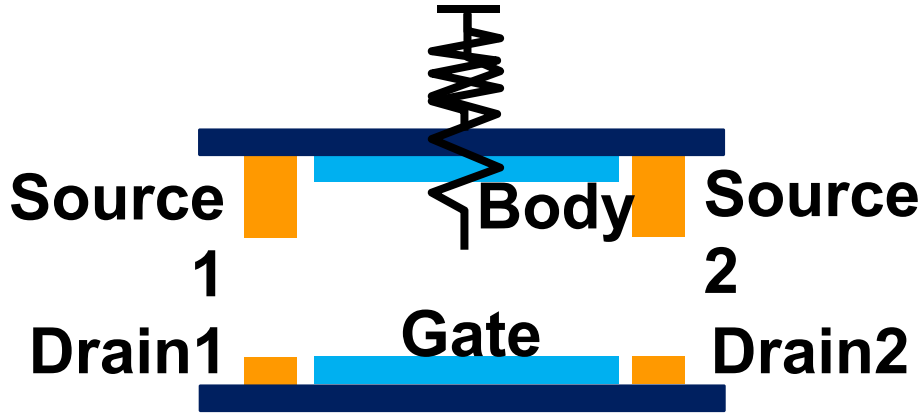
Relay Design for Digital Logic

3-Terminal design
Source as reference

vs.

4- or 6-terminal design
Body as reference





Switching is dependent on source voltage

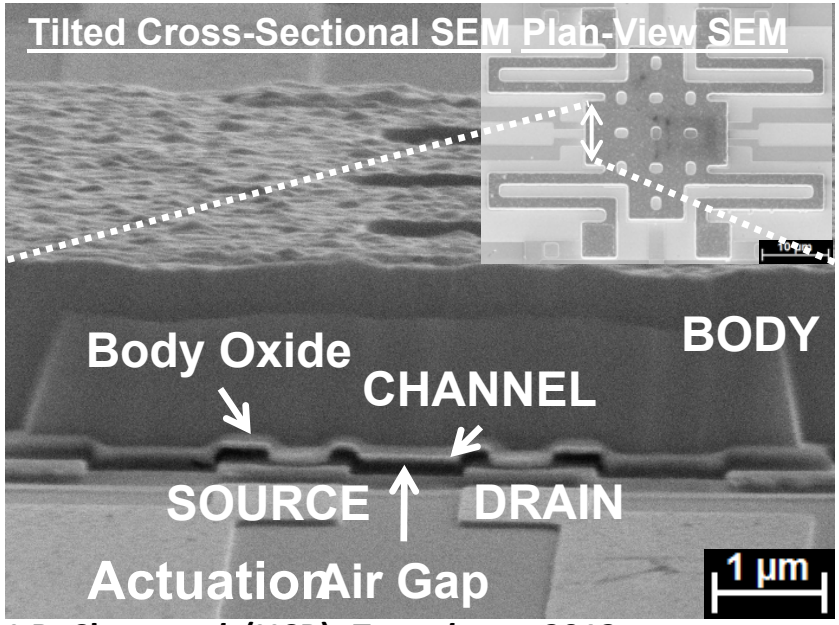
(Source is common to input and output)

Switching is independent of source voltage

Can act as "pull-up" or "pull-down" switch

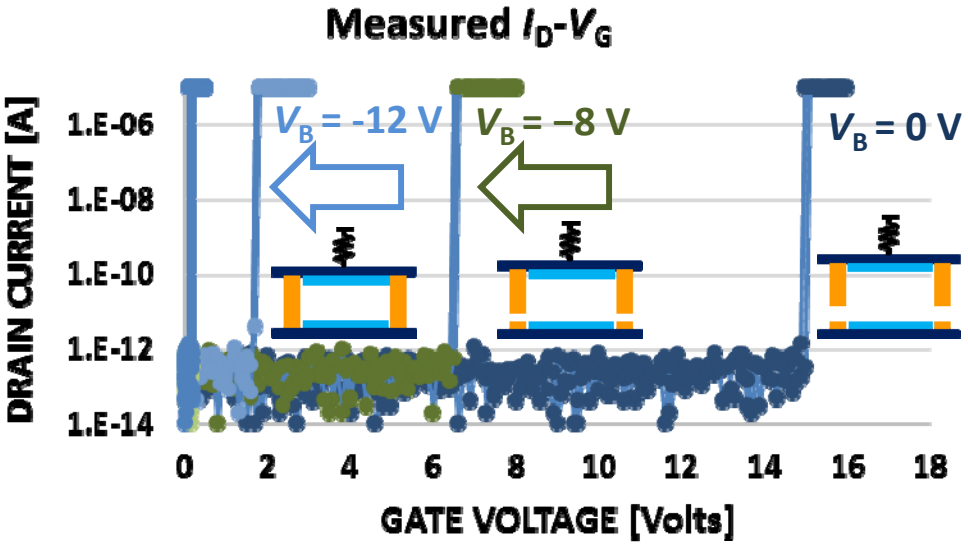
→ more suitable for pass-gate logic

Logic Relay Structure & Operation

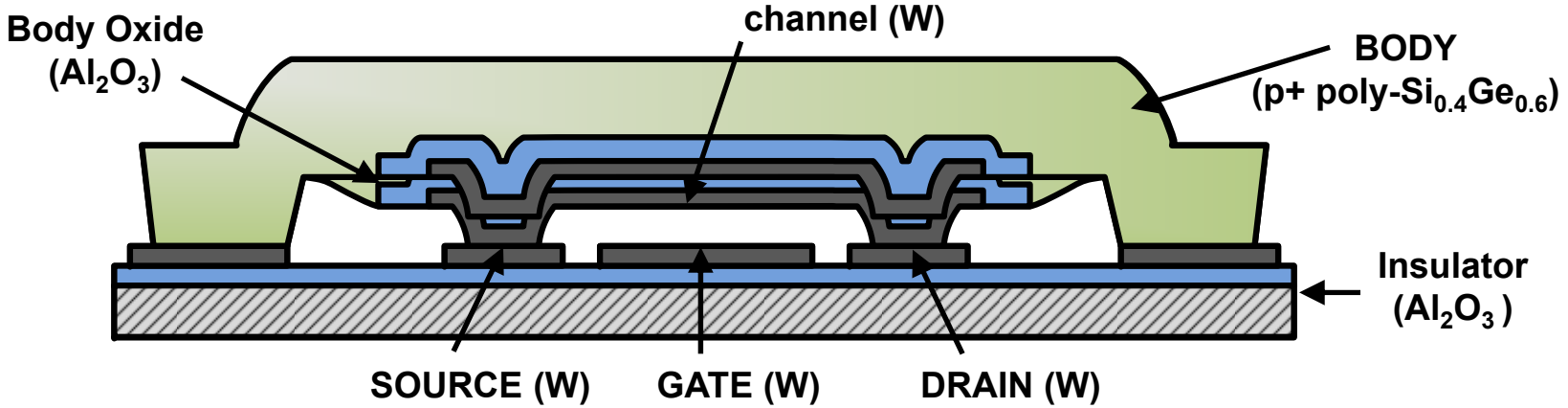


I-R. Chen *et al*. (UCB), *Transducers* 2013

➢ **Body biasing to enable low V_{DD}**

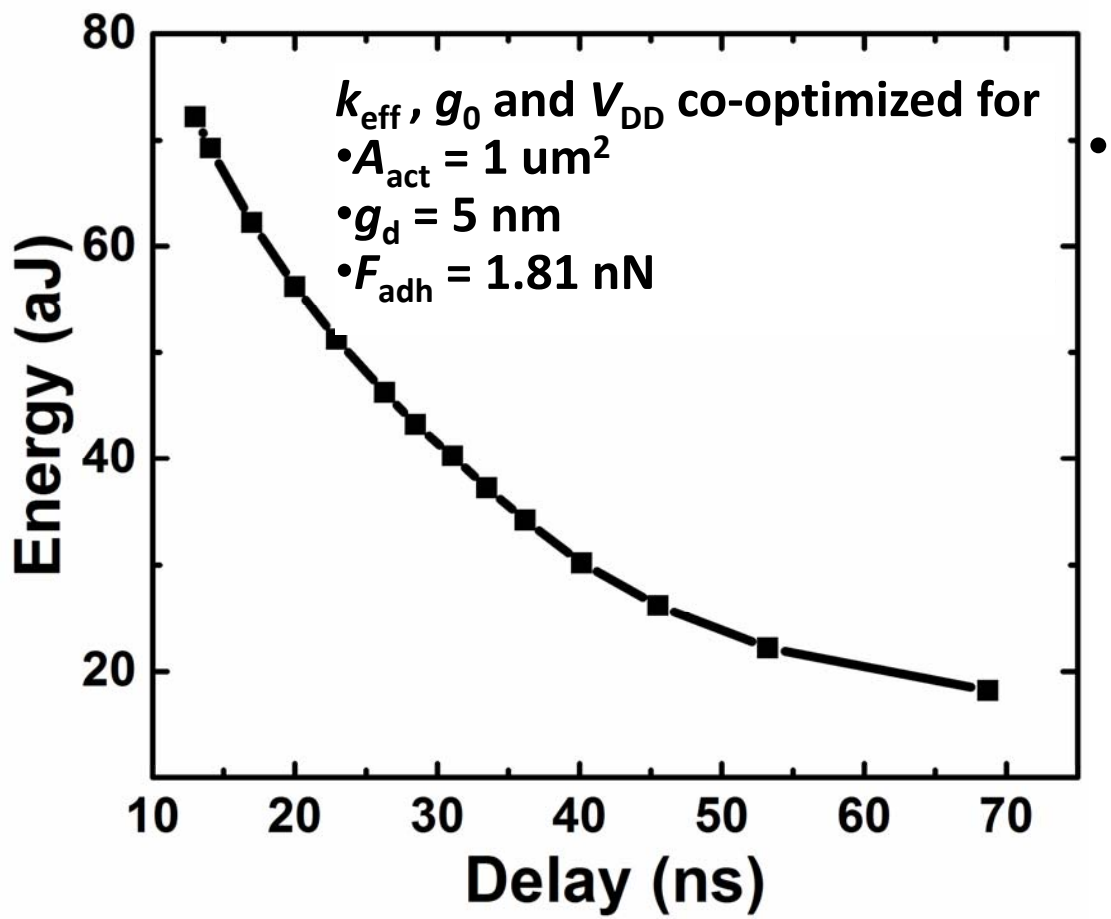


Schematic Cross-section



NEM Relay Switching Energy-Delay

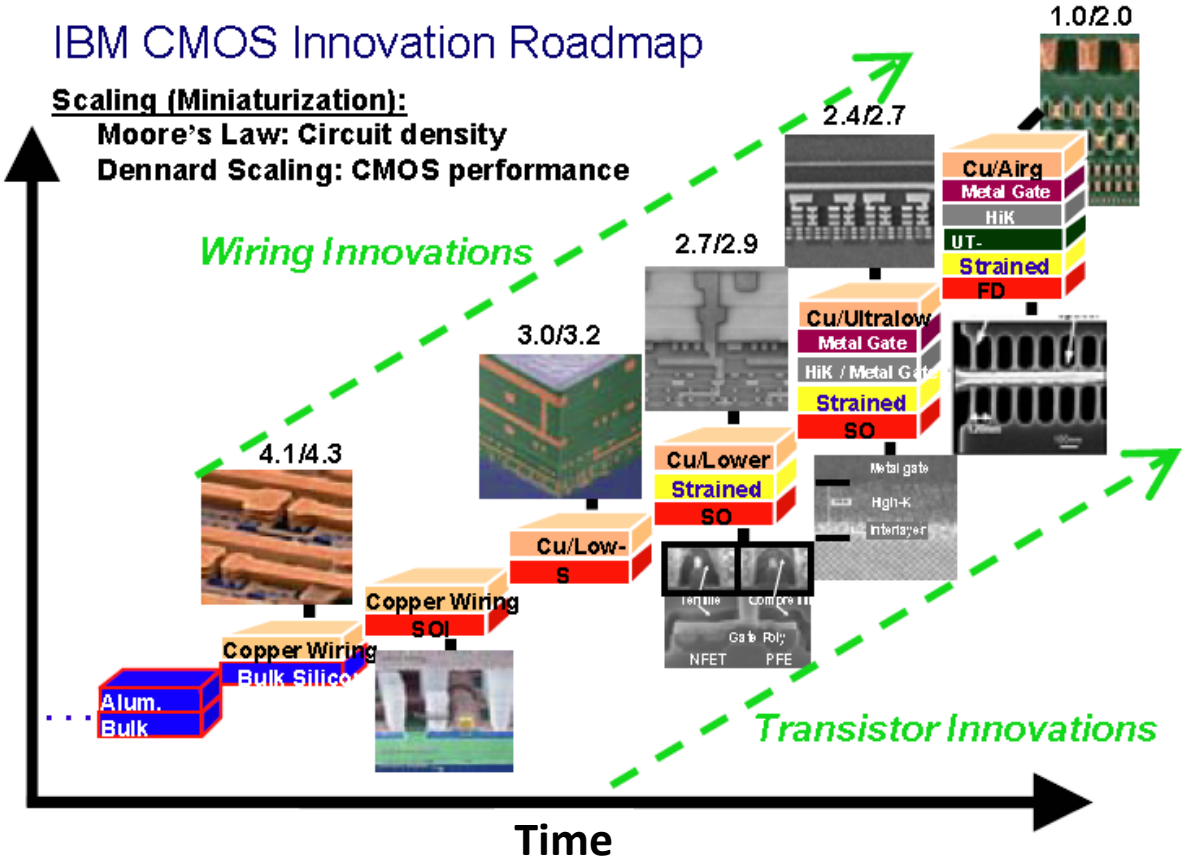
C. Qian *et al.*, 2015 International Electron Devices Meeting (Paper 18.1)

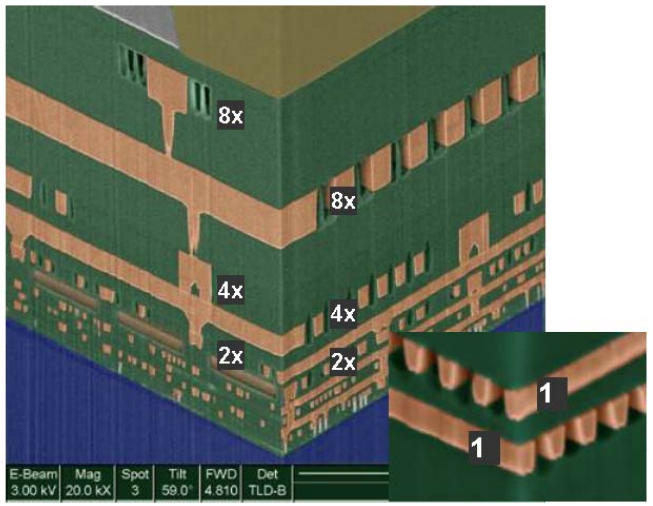


- **In comparison, LSTP CMOS is projected to have switching energy > 100 aJ at the 5nm node**

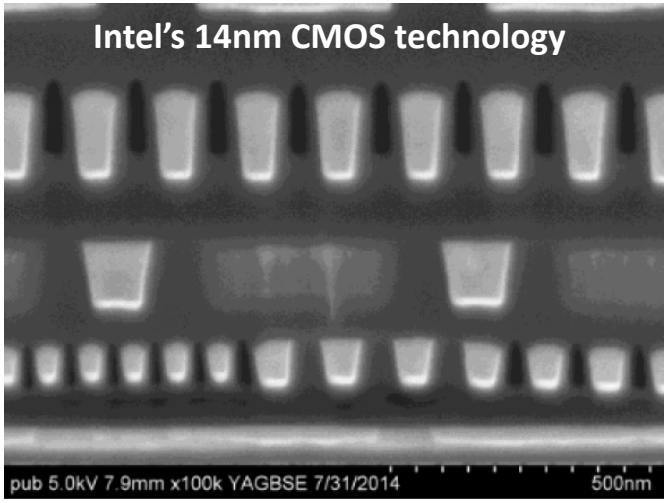
[W. Cao *et al.*, *IEEE-TED* Nov 2015]

IC Technology Advancement





D. C. Edelstein, *214th ECS Meeting*, Abstract #2073, 2008



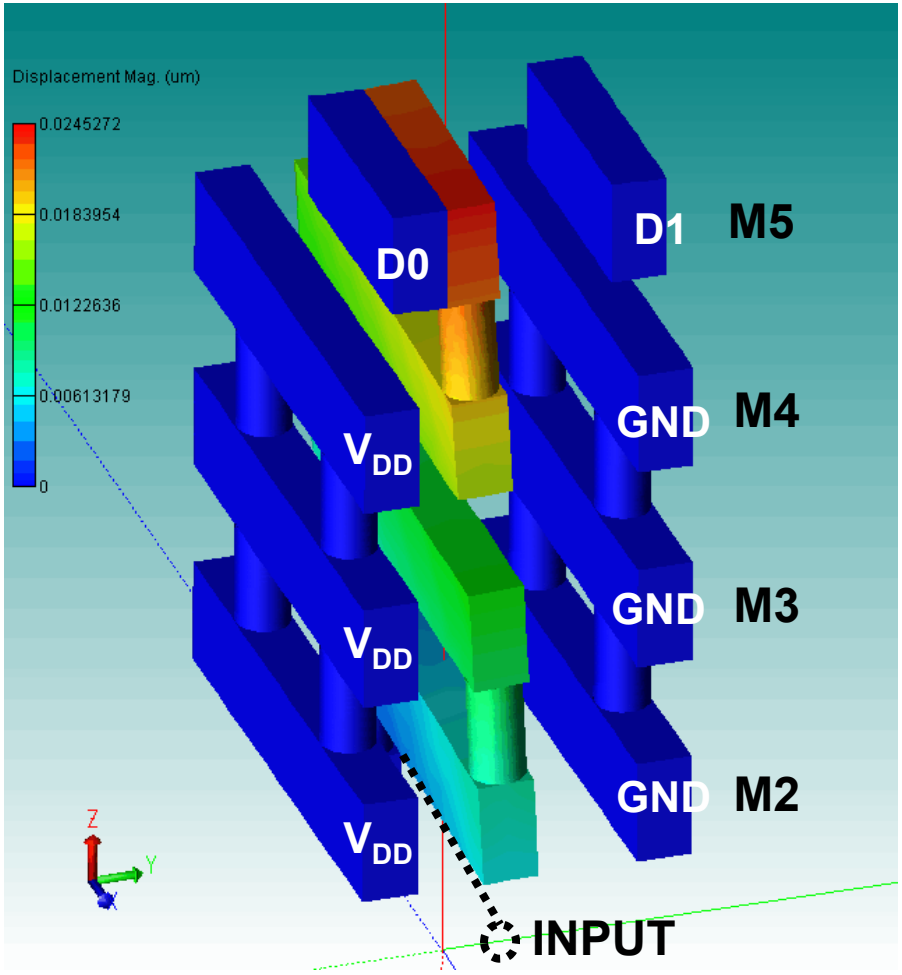
S. Natarajan *et al.* (Intel), *IEDM* 2014

- **Advanced back-end-of-line (BEOL) processes have air-gapped interconnects**

→ **can be adapted for fabrication of compact NEMS!**

BEOL NEM Switch

N. Xu *et al*. (UC Berkeley), 2014 IEEE International Electron Devices Meeting



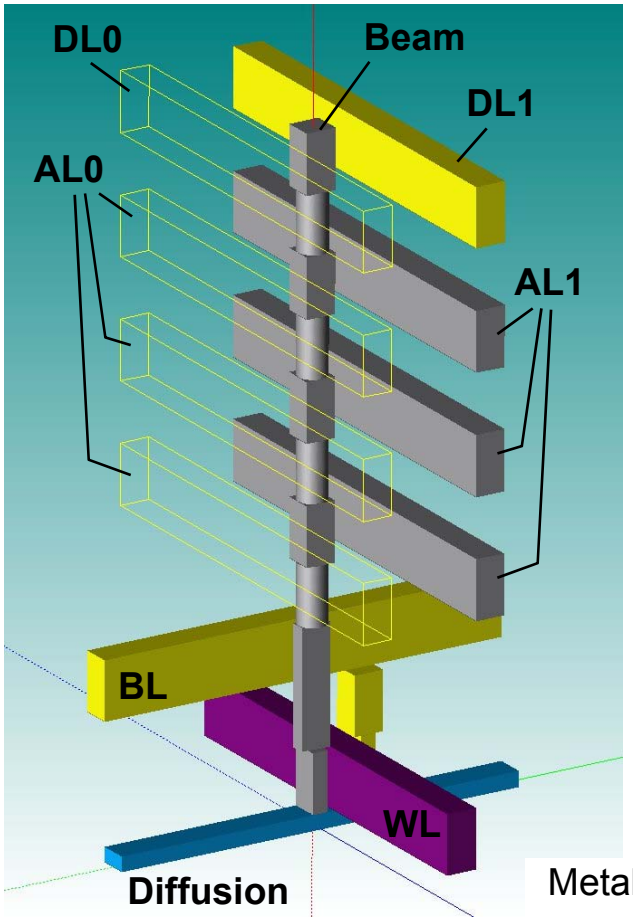
courtesy of Dr. Kimihiko Kato (UC Berkeley)

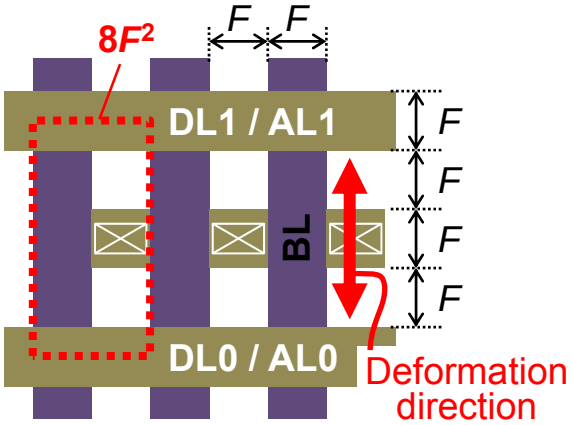
- **A relay can be implemented using multiple metal layers**
  - **Vias can be used for electrical connection and as torsional elements for lower k_{eff}**
- **Actuation electrodes on opposite sides of movable electrode structure**

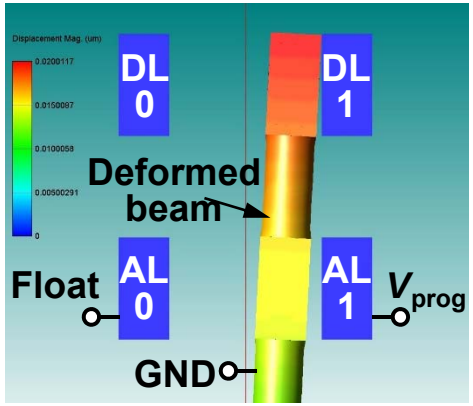
  → 2 stable states (contacting D0 or D1)
- **Low-voltage (<1 V) operation can be achieved with small footprint (< 0.1 μm²).**

Non-Volatile NEMory Cell Structure

K. Kato *et al.*, *IEEE Electron Device Letters*, Vol. 37, pp. 31-34, 2016







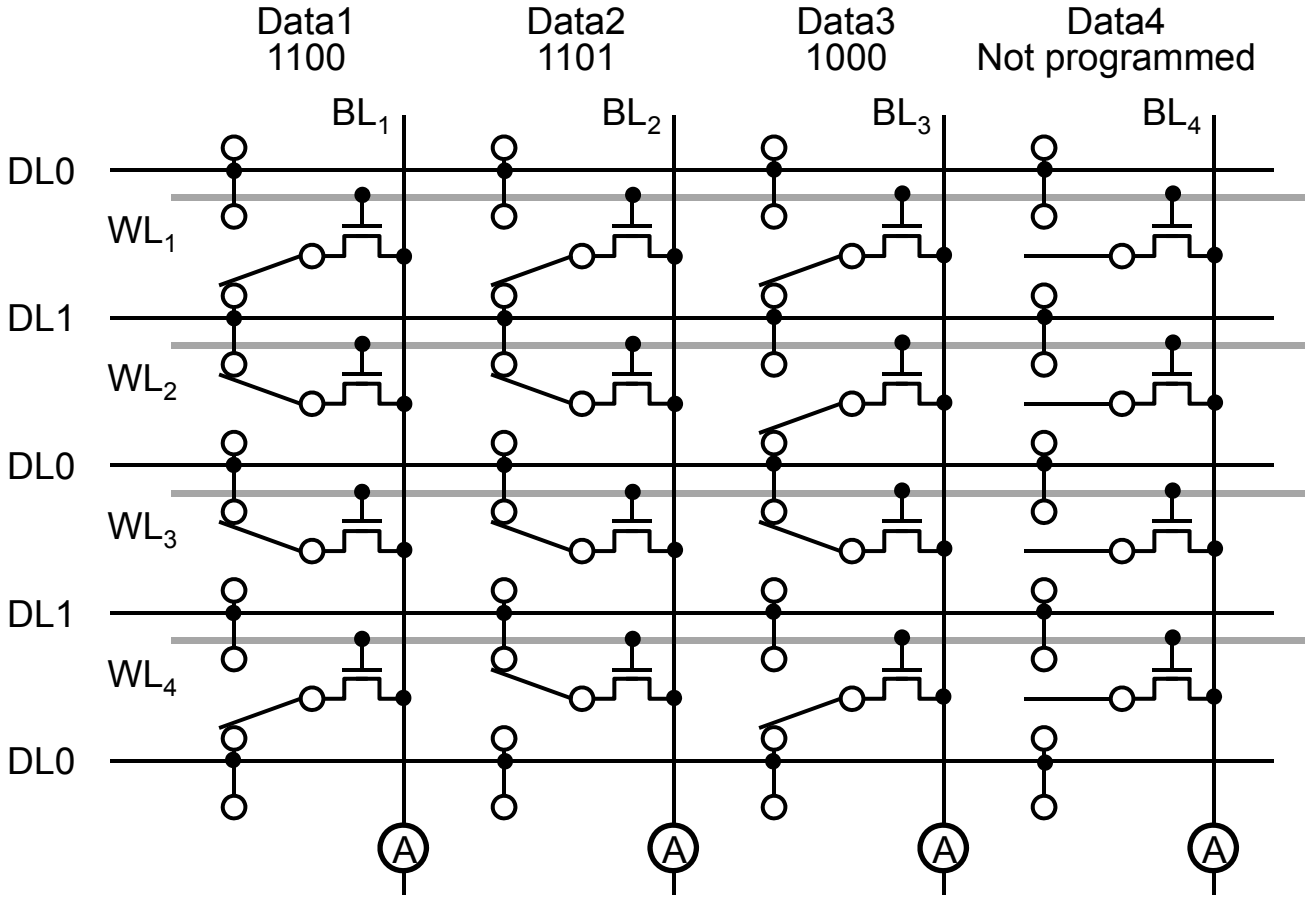
Metal: Al (Young's modulus: 70 GPa)
Gap: Air

Year	2015	2017	2019	2021	2023	2025
Half pitch, *F* (nm)	25.3	20.1	15.9	12.6	10.0	7.09

In-Memory Computing

K. Kato *et al.*, *IEEE Electron Device Letters*, Vol. 37, pp. 31-34, 2016

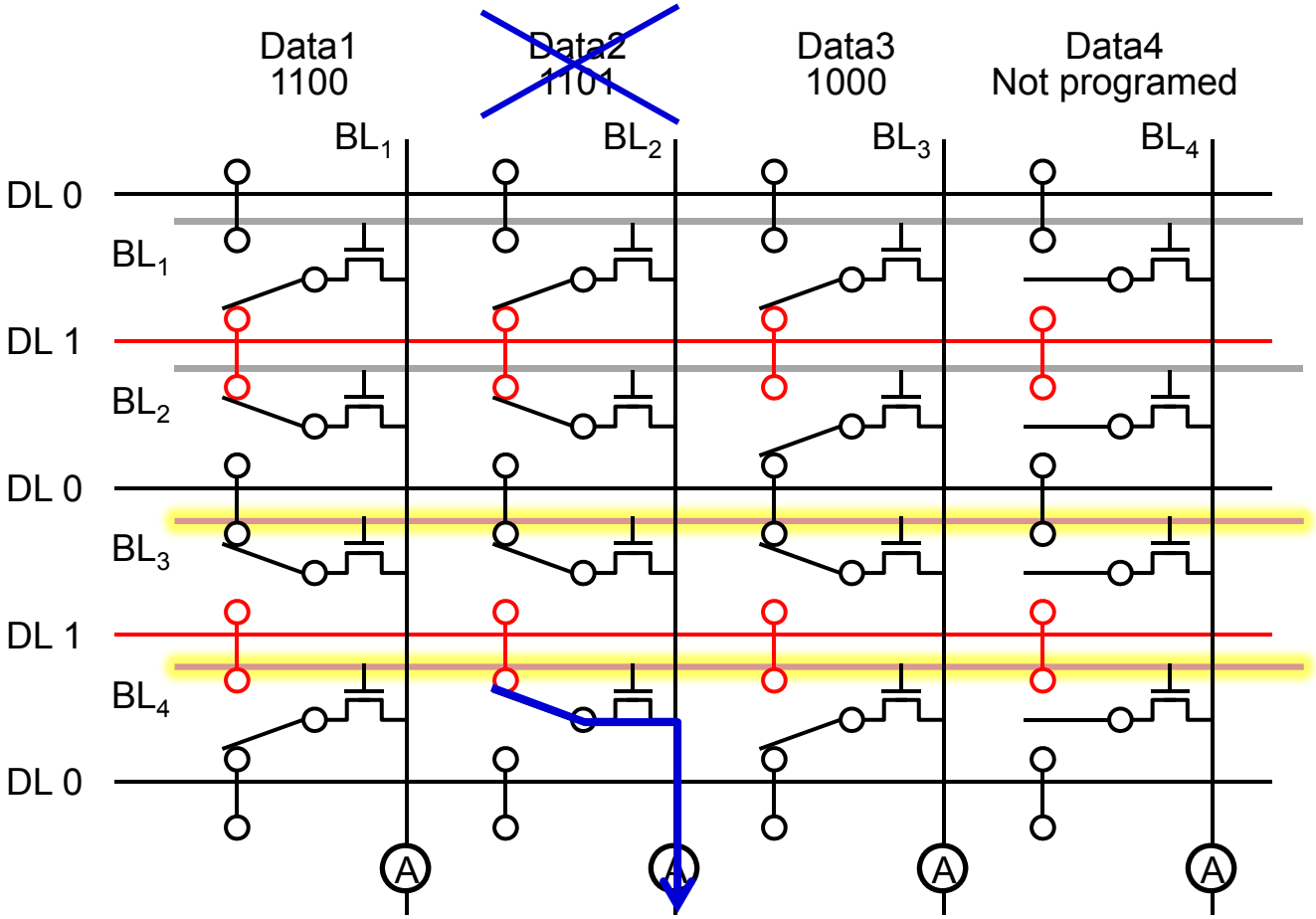
- **NV-NEMory cell array for memory-based super-parallel data searching**



Data Search Step 1: Match "0"

K. Kato *et al.*, *IEEE Electron Device Letters*, Vol. 37, pp. 31-34, 2016

Reference Data: 1100



a) Activate DL1 lines

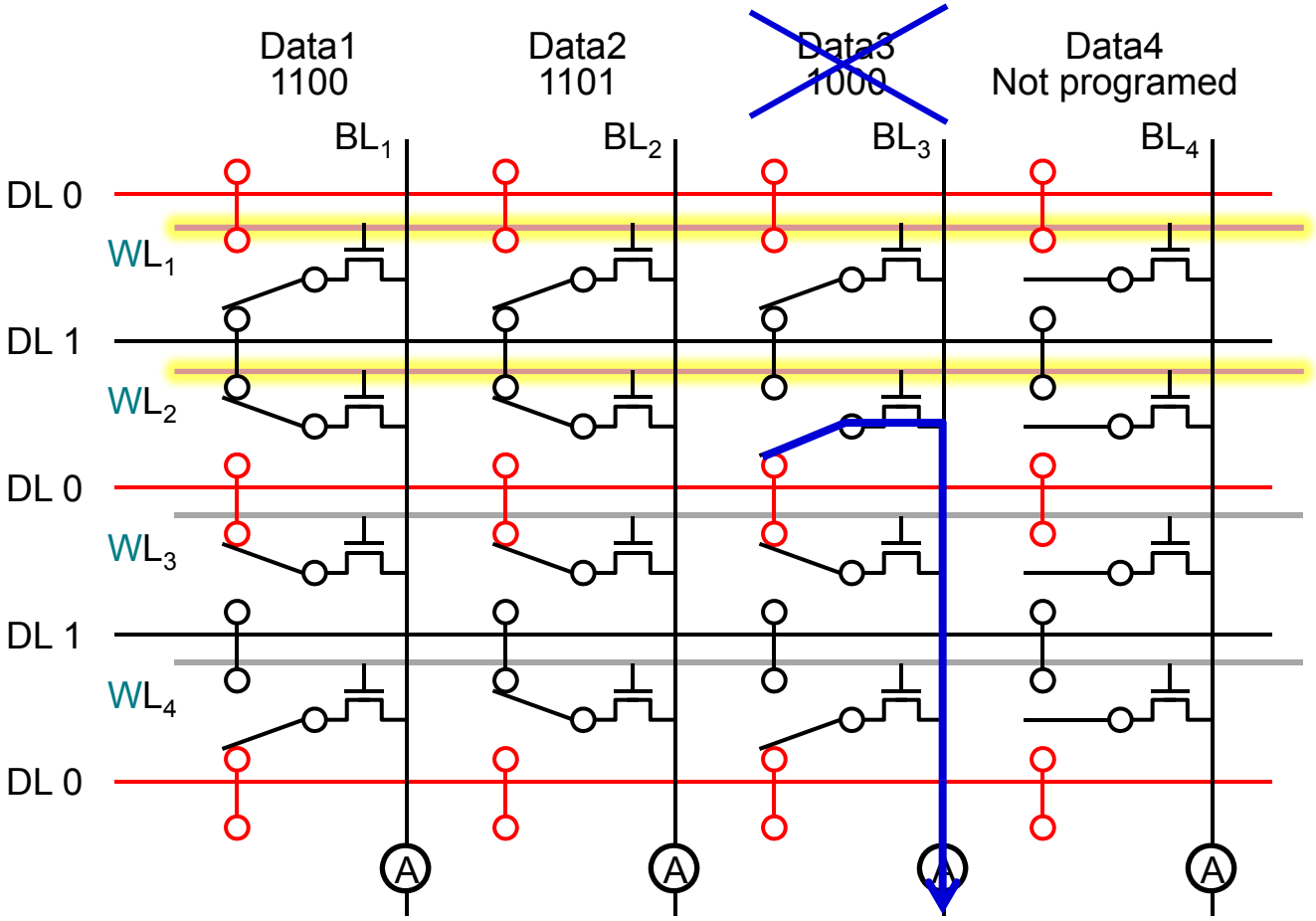
b) Access rows for 0 bits in reference

c) Zero BL current → Matched 0s

Data Search Step 2: Match "1"

K. Kato *et al.*, *IEEE Electron Device Letters*, Vol. 37, pp. 31-34, 2016

Reference Data: 1100



a) **Activate DL0 lines**

b) **Access rows for 1 bits in reference**

c) **Zero BL current → Matched 1s**

Energy and Delay for Data Search

K. Kato *et al.*, *IEEE Electron Device Letters*, Vol. 37, pp. 31-34, 2016

256 × 256 NV-NEMory Array

	Energy			Delay
Cells involved:	1 column × 1 row	1 column × 256 rows	256 columns × 256 rows	
Program (V_{prog} = 2.5 V)	15 fJ	2.0 pJ	N/A	< 10 ns
Match "0" or Match "1"	N/A	N/A	1.2 pJ	< 0.2 ns

- The location of a data string can be found in <0.5 ns with less than 2.5 pJ.

→ For a die size of 42 mm^2 (same as DDR4 DRAM) at F = 20 nm and cell density of 65% (similar to DRAM), a NV-NEMory chip would have the capacity 8 Gb and would consume only 300 nJ to find a match on the whole chip.

→ In comparison, it would take CPU+DRAM ~90 mJ, 80 ms for the same task.

Relatively fast read speed & low power consumption make NV-NEMory technology well-suited for real-time data searching applications!

Summary

- **Challenges:**
  - **CMOS technology has a fundamental limit in energy efficiency, due to non-zero transistor OFF-state current.**

  →New logic switch designs are needed to overcome this limit!
    - **Steeply switching with zero I_{OFF}**
    - **Robust to process-induced variations**
- **Opportunities:**
  - **2-D semiconductor materials, negative capacitance FETs, …**
    - **Semiconductor device designs which do not utilize doping**
  - **Nanomanufacturing innovations to lower cost per function**
  - **Collaboration across domains of expertise to co-optimize device technology, circuit/system architecture, algorithms**
    - **Examples: Reconfigurable specializers, communication-avoiding and write-avoiding algorithms**

Cell-Level Comparison of Emerging NVM Technologies

K. Kato *et al.*, *IEEE Electron Device Letters*, Vol. 37, pp. 31-34, 2016

	NAND Flash	PCM	Redox RRAM	STT-MRAM	NV-NEMory	Stand alone DRAM
Cell area	$2.5F^2$	$6F^2$	$5\text{-}8F^2$	$20\text{-}40F^2$	$8F^2$	$6F^2$
Program voltage	18-20 V	3 V	0.5 V	1.8 V	~ 2 V	1.5 V
Program time	> 10 µs	50 ns	5 ns	100 ns	< 10 ns	< 10 ns
Program current	n/a	100 µA	0.4 µA	100 µA	zero	n/a
Program energy	> 1fJ	2 pJ	1 fJ	4 pJ	~ 50 aJ	2 fJ
Read voltage	0.1-0.5 V	3 V	0.2 V	0.5 V	< 0.1 V	1.5 V
Read time	15-50 µs	60 ns	10 ns	10-20 ns	< 0.1 ns	< 10 ns
Endurance	$10^4\text{-}10^5$	10^{15}	10^{16}	2×10^{12} @10ns 2×10^{6} @10ms	$>10^{16}$	N/A

- **NV-NEMory technology offers much lower programming energy per bit and fast read access time as compared with other NVM technologies.**