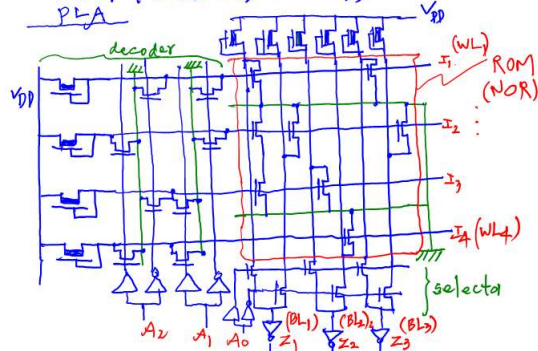


# EE 222 Lecture 13 A Feb. 19, 2019

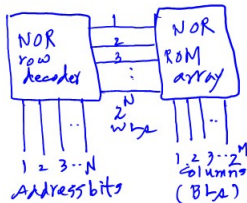
## ROM (Read Only Memory) - nMOS ROM PLA



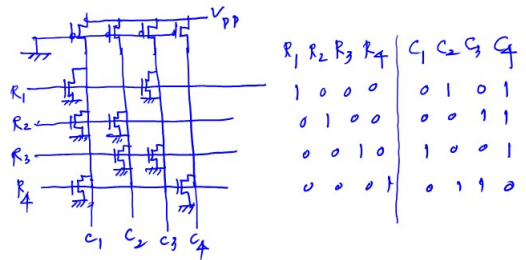
$$\begin{aligned} I_1 &= \overline{A_2 + A_1} \\ I_2 &= \overline{A_2 + A_1} \\ I_3 &= \overline{A_2 + A_1} \\ I_4 &= \overline{A_2 + A_1} \end{aligned} \left. \vphantom{\begin{aligned} I_1 \\ I_2 \\ I_3 \\ I_4 \end{aligned}} \right\} \text{NOR array}$$

$$\begin{aligned} Z_1 &= A_0 I_1 + I_2 + I_3 + \overline{A_0} I_4 \\ Z_2 &= A_0 I_3 + \overline{A_0} I_4 \\ Z_3 &= A_0 I_1 + \overline{A_0} I_2 \end{aligned} \left. \vphantom{\begin{aligned} Z_1 \\ Z_2 \\ Z_3 \end{aligned}} \right\} \text{NOR array gated by } A_0, \overline{A_0}$$

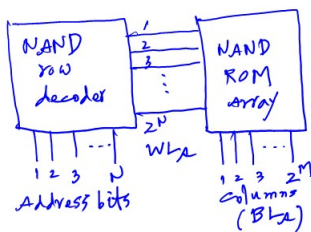
## NOR decoder - NOR ROM Array



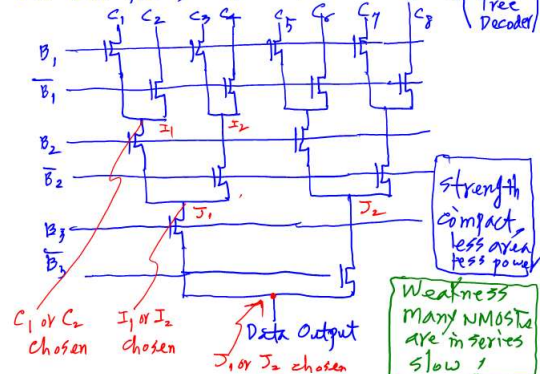
## 4 x 4 NOR-based ROM



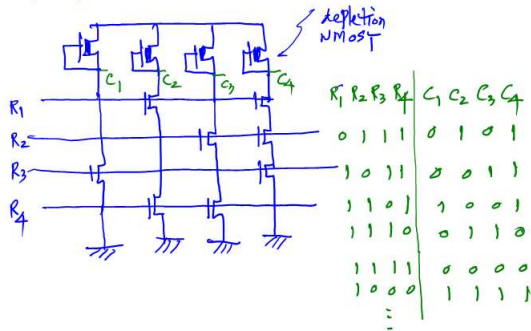
## NAND-row decoder - NAND ROM Array



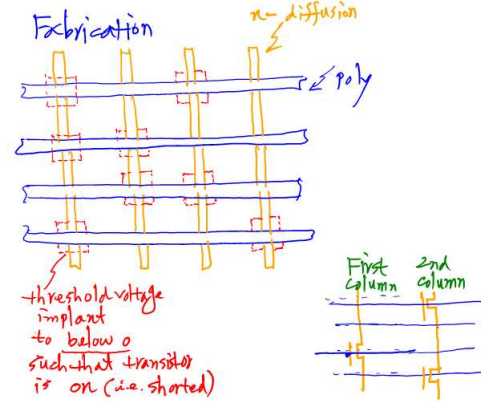
## An Example of Column Decoder Circuit (Binary Tree Decoder)



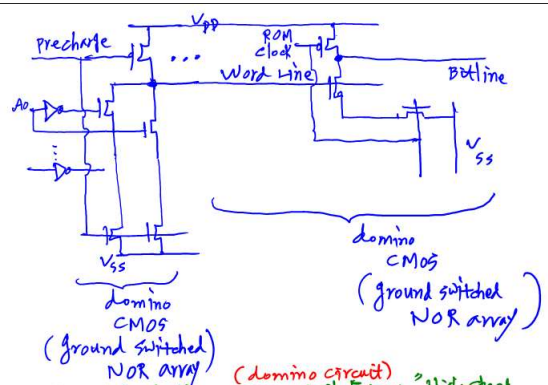
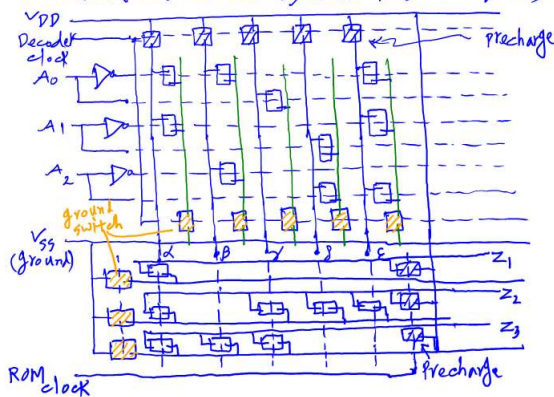
### 4x4 NAND-based ROM



### Fabrication



### A Structure of CMOS Dynamic PLA (Layout)



(domino circuit)  
Ref: R.H. Krambeck, C.M. Lee and M.F. Law, "High-Speed Compact Circuits with CMOS", IEEE J of Solid-State Circuits, 5(1) (1970)

$$\tau_{HL} = \frac{\beta (C_{wire} + N C_{drain}) V_{DD}}{I_{drive}} < \tau_{spec}$$

For  $\tau_{spec} = 0.25 \text{ ns}$  (WL/BL delay)

$$C_{drain} = 0.01 \text{ pF}$$

$$V_{dd} = 1.2 \text{ V}$$

$\beta = 2$  (empirical constant)

$$I_{d1} = 0.6 \mu\text{A}, C_{wire} < N C_{drain}$$

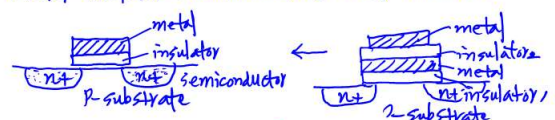
$$N < \frac{0.25 \times 10^{-9} \times 0.6 \times 10^{-6}}{2 (0.01 \times 10^{-15}) \times 1.2} \frac{0.15 \times 10^{-15}}{2.4 \times 10^{-15} \times 10^3} \approx 62 \text{ (number of nMOS's per WL/BL)}$$

### CMOS reference

F.M. Wanlass and C.T. Sah, "Nanowatt Logic using Field-Effect Metal-Oxide Semiconductor Triodes", IEEE Solid-State Circuits Conf. Philadelphia, PA (1963) (First CMOS paper)

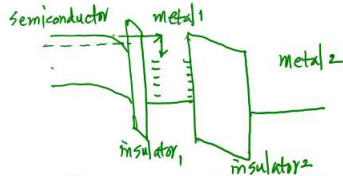
### Nonvolatile Semiconductor Memory

#### Metal-Insulator-Semiconductor (MIS) Structure

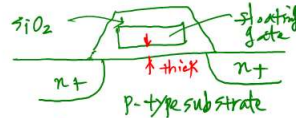


By D. Kahng and S.M. Sze  
Bell System Technical Journal (1967)

### Operation Principle of metal-insulator-metal-insulator-semiconductor nonvolatile memory

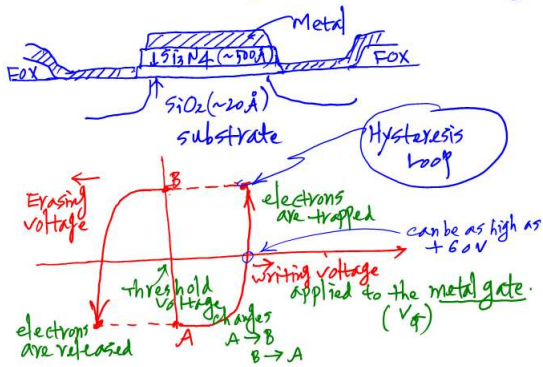


Electrons are injected into the floating gate (metal 1) by tunneling through insulator 1, and are stored semi-permanently.

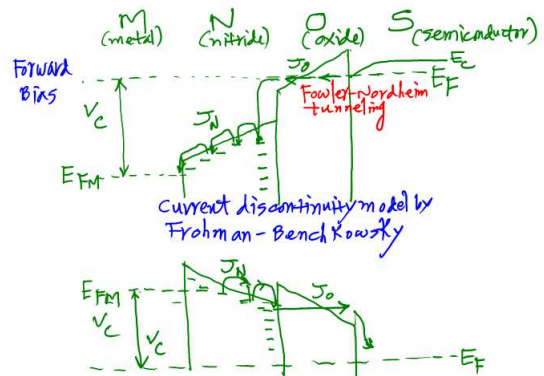


electrons are injected into the 'floating gate' by crossing the oxide ( $\text{SiO}_2$ ) potential barrier and are stored in the floating gate. Thus, named FAMOS (Floating-gate Avalanche injection MOS)

### MNOS structure (Metal Nitride Oxide Semicon)



### Energy Band Diagram



$$J_N = C_N E_N^2 \exp(-E_N/E_F)$$

$$J_{ox} = C_{ox} E_{ox}^2 \frac{\pi k T / E_{ox}}{\sin(\pi k T / E_{ox})} \exp(-E_{ox}/E_{ox})$$

$E_N$  = electric field in the nitride layer  
 $E_{ox}$  = " " oxide layer

$k$  = Boltzmann's constant

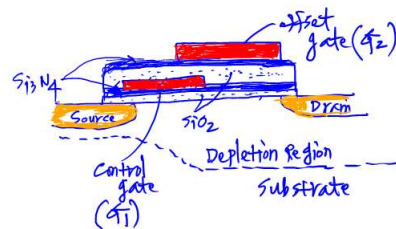
$T$  = temp in  $^{\circ}\text{K}$

$J_N$  = nitride layer current density

$J_{ox}$  (or  $J_{ox}$ ) = oxide layer "  $V_G = E_{ox} t_{ox} + E_N t_N$

e.g.  $C_{ox} = 10^{-5} \text{ A/V}^2$   $C_N = 3.5 \times 10^{-10} \text{ A/V}^2$   
 $E_N = 2.54 \times 10^8 \text{ V/cm}$   $E_{ox} = 1.2 \times 10^8 \text{ V/cm}$   
 $t_{ox} = 50 \text{ Å}$   $t_N = 1000 \text{ Å}$

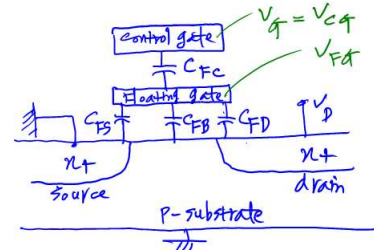
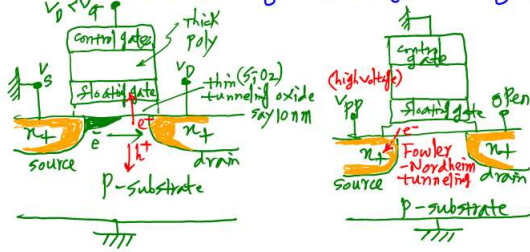
### Stacked Gate Tetraode Proposed by H. G. Dill & T. N. Tombs (1969)





## Flash Memory

Memory cell is a transistor with a floating gate whose threshold voltage can be programmed (changed) repeatedly by applying an electric field (through  $V_g$  voltage) to its gate.



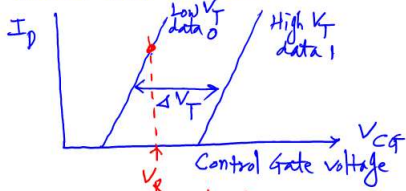
$$C_{total} = C_{fc} + C_{fs} + C_{fb} + C_{fd}$$

$$V_{fg} = \frac{Q_{fg}}{C_{fg}} + \frac{C_{fc}}{C_{total}} V_{cg} + \frac{C_{fd}}{C_{total}} V_d$$

$Q_{fg}$  = charge stored in the floating gate

$$V_T(Cg) = \frac{C_{total}}{C_{fc}} V_T(Fg) - \frac{Q_{fg}}{C_{fc}} - \frac{C_{fd}}{C_{fc}} V_d$$

$$\Delta V_T(Cg) = - \frac{\Delta Q_{fg}}{C_{fc}}$$



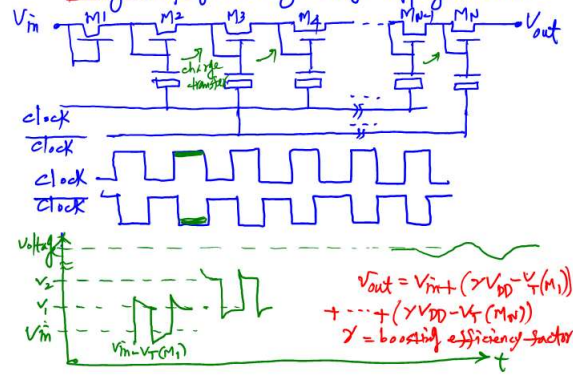
NOR Flash Memory

	BL1	BL2	operation
Signal	open	open	0V 1V
WL1	open	open	0V 0V
WL2	12V	0V	0V 0V
Source line	0	0	12V 5V

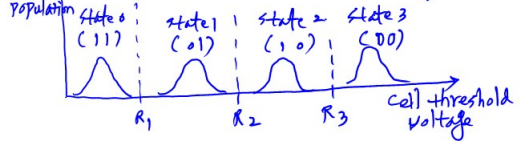
NOR array - faster but many contacts (area ↑)  
\* NAND array \* slower but compact (area ↓)

	BL1	BL2	operation
Signal	open	open	0V 1V
WL1	0	10V	5V
WL2	0	10V	5V
WL3	0	20V	0V
WL4	0	10V	5V
WL5	0	10V	5V
WL6	0	10V	5V
Source line	20V	0V	0V
Select line 2	open	0V	5V
P-wall	20V	0V	0V
N-tub	20V	0V	0V

Charge Pumping → High voltage  $V_{pp}$  generation



Multi-level-cell Threshold Voltage Distribution in Flash (4 values) - 2bits/cell



## subthreshold operation

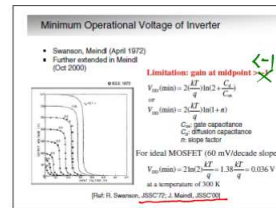
### Opportunities for Ultra-Low Voltage

- Number of applications emerging that do not need high performance, only extremely low power dissipation
- Examples:
  - Standby operation for mobile components
  - Implanted electronics and artificial senses
  - Smart objects, fabrics, and e-textiles
- Need power levels below 1 mW (even  $\mu\text{W}$  in certain cases)

### Slide 11.4

Although keeping the power density constant is one motivation for the continued search to lower the EOP, another, maybe even more important, reason is the exciting applications that only become feasible at very low energy/power levels. Consider, for instance, the digital wrist-watch. The concept, though straightforward, only became attractive once the power dissipation

[R<sub>1</sub>]



### Slide 11.5

The question of the minimum operational voltage of a CMOS inverter was addressed in a landmark paper [Swanson72] in the early 1970s. Published even before CMOS integrated circuits came in vogue, the paper for an inverter to be regenerative and to have two distinct steady-state operation points (a "1" and a "0"), it is essential that the absolute value of the gain of the gate in the transition region be larger than 1. Solving for those conditions leads to an expression for  $V_{min}$  equal to  $2kT/q \ln(1+n)$ , where  $n$  is the slope factor of the transistors. One important observation is that  $V_{min}$  is proportional to the operational temperature  $T$ . Cooling down a CMOS circuit to temperatures close to absolute zero (e.g., liquid Helium), makes operation at mV levels possible. (Unfortunately, the energy going into the cooling more than offsets the gains in operational energy.) Also, the closer the MOS transistor operating in sub-threshold mode gets to the ideal bipolar transistor behavior, the lower the minimum voltage. At room temperature, an ideal CMOS inverter (with a slope factor of 1) could marginally operate at as low as 56 mV.

$$\frac{\partial V_o}{\partial V_{in}} = g_{m,p} / g_{m,n} > 1$$

$$I_{DS} = I_S e^{\frac{V_{GS}-V_{th}}{n V_T}} \left(1 - e^{-\frac{V_{DS}}{V_T}}\right) = I_0 e^{\frac{V_{GS}}{n V_T}} \left(1 - e^{-\frac{V_{DS}}{V_T}}\right)$$

where  $I_0 = I_S e^{-\frac{V_{th}}{n V_T}}$

### Sub-threshold Modeling of CMOS Inverter

- From Chapter 2:

$$I_{DS} = I_S e^{\frac{V_{GS}-V_{th}}{n V_T}} \left(1 - e^{-\frac{V_{DS}}{V_T}}\right) = I_0 e^{\frac{V_{GS}}{n V_T}} \left(1 - e^{-\frac{V_{DS}}{V_T}}\right)$$

where

$$I_0 = I_S e^{-\frac{V_{th}}{n V_T}}$$

(DIBL can be ignored at low voltages)  
of clarity. For low values of  $V_{DS}$ , the DIBL effect can be ignored.

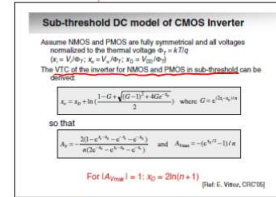
### Slide 11.6

Given the importance of this expression, a quick derivation is worth undertaking. We assume that at these low operational voltages, the transistors operate only in the sub-threshold regime, which is often also called the weak-inversion mode. The current-voltage relationship for a MOS transistor in sub-threshold mode was presented in Chapter 2, and is repeated here for the sake

$$x_i = V_i / \phi_T \quad x_o = V_o / \phi_T \quad x_D = V_{DD} / \phi_T$$

Thermal voltage

$$x_o = x_D + \ln \left( \frac{1 - e^{-x_D} + \sqrt{(1 - e^{-x_D})^2 + 4 e^{-x_D}}}{2} \right), \quad e = \frac{x_D x_o}{n}$$



### Slide 11.7

The (static) voltage transfer characteristic (VTC) of the inverter is derived by equating the current through the NMOS and PMOS transistors. The derivation is substantially simplified if we assume that two devices have exactly the same strength when operating in sub-threshold. Also, normalizing all voltages with respect to the thermal voltage  $\phi_T$  leads to more elegant expressions. Setting the

$$k_p = k_n$$

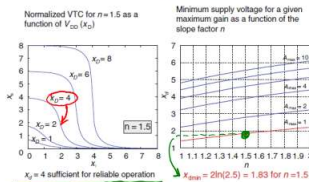
where

$$A_V = - \frac{2(1 - e^{-x_D}) - x_D - x_o}{n(2e^{-x_D} - e^{-x_D} - e^{-x_o})}$$

$$|A_V| = 1 \Rightarrow x_D = 2 \ln(n+1) \Rightarrow V_{DD} = 2 \ln(n+1) \frac{kT}{q}$$

### Results from Analytical Model

#### Sub-threshold Inverter



thermal voltage leads to reasonable noise margins (assuming  $n = 1.5$ ). This is approximately equal to 100 mV.

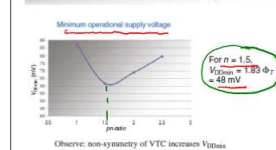
$$X_D = V_{DD} / \phi_T$$

$$X_D = 4 \approx 100 \text{ mV}$$

$$\phi_T = \frac{kT}{q} \text{ (thermal voltage)}$$

$$= 26 \text{ mV at } T = 300 \text{ K (room temp.)}$$

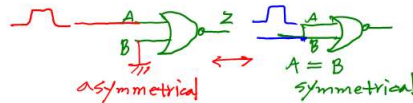
### Confirmed by Simulation (at 90 nm)



### Slide 11.9


Simulations (for a 90 nm technology) confirm these results. When plotting the minimum supply voltage as a function of the PMOS/NMOS ratio, a minimum can be observed when the inverter is completely symmetrical, that is when the PMOS and NMOS transistors have identical drive strengths. Any deviation from the symmetry causes  $V_{min}$  to rise. This implies that transistor

sizing will play a role in the design of minimum-voltage circuits. Also worth noticing is that the simulated minimum voltage of 60 mV is slightly higher than the theoretical value of 48 mV. This is mostly owing to the definition of "operational" point. At 48 mV, the inverter is only marginally functional. In the simulation, we assume a small margin of approximately 25%.



## Minimum Energy per Operation

Predicted by von Neumann  $\sqrt{fT(E)}$



J. von Neumann  
(Theory of Self-Organizing Automata, 1966)

- Moving one electron over  $V_{DD}/2$**

  - $E_{min} = qV_{DD}/2 = q \cdot 250mV/2 = 4 fT(E)$
  - Also called the Von Neumann 2-electron energy bound
  - At room temp  $k_B T \approx 25 mV$ ,  $E_{min} = 90 pJ$  per operation
- Minimum sized CMOS inverter at 90 nm operating at 1V**

  - $E = C_{VDD} \cdot V_{DD}^2 \approx 1.8 \cdot 10^{-4} J$ , or 4 orders of magnitude larger!

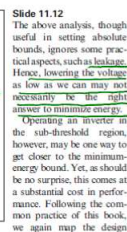
How close can one get?

(Prof. J. Van Neumann, 1966)

How close can one get?

[Ref.: J. Von Neumann, 1956]

- This expression for the minimum energy for a digital operation was already predicted much earlier by John von Neumann (as reported in [von Neumann, 1966]). Landauer later established that this is only the case for "logically irreversible" operations in a physical computer that dissipate energy by generating a corresponding amount of entropy for each bit of information that then gets irreversibly erased. This bound indeed does not hold for reversible computers (if such could be built) [Landauer, 1961].



One interesting by-product of operating in the sub-threshold region is that the equations are quite simple and are exponentials (as used to be the case for bipolar transistors). Under the earlier assumptions of symmetry, an expression of the inverter delay is readily derived. Observe again that a reduction in supply voltage has an exponential effect on the delay!

$$\frac{\tau}{\tau_0} = \frac{\cancel{\frac{C}{\epsilon_0}} \sqrt{\frac{\mu_0}{\epsilon_0}}}{\cancel{\frac{1}{\epsilon_0}} e^{\frac{\mu_0 R^2}{\epsilon_0}}} \cdot \frac{\cancel{\epsilon_0}}{\cancel{\frac{1}{\epsilon_0}}} = \frac{X_D}{e^{X_D/n}} = \boxed{X_D e^{-X_D/n}} \text{ normalize delay}$$

Year of Production	2013	2015	2017	2019	2021	2023	2025	2028
Feature Pitch: "Node Name" Label	140/140	110	90	70	55	45	35	28
Gate Length (nm)	40	32	25	20	16	13	10	7
Finch Pitch (2D) (nm)	18	15	13	11	9	8	6	5
Gate Pitch (nm)	28	24	20	16	14	12	10	7
Finch Fin Half-pitch (nm) (nm)	30	24	19	15	12	9	7.5	5.5
Finch Fin Width (nm) (nm)	7.8	7.2	5.8	4.4	3.1	2.7	2.4	2.0
Finch Fin Width (nm) (nm)	3.006	2.803	2.261	1.705	1.205	0.905	0.705	0.505
Finch Fin Width (nm) (nm)	0.248	0.157	0.099	0.062	0.039	0.025	0.018	0.009
Finch Fin Width (nm) (nm)	4.030/40	3.730/35	3.150/34	2.610/34	2.100/34	1.650/34	1.400/34	1.200/34
Finch Fin Width (nm) (nm)	1000	800	600	450	350	250	180	120
Finch Fin Width (nm) (nm)	16-32	16-32	16-32	32-64	64-128	128-256	256-512	512-1024
Finch Fin Width (nm) (nm)	46mm	54mm	45mm	30mm	28mm	27mm	25mm	22mm
Finch Fin Width (nm) (nm)	40	80	80	160	320	320	320	320
Finch Fin Width (nm) (nm)	2018							
Finch Fin Width (nm) (nm)	0.80	0.83	0.80	0.77	0.76	0.71	0.68	0.64
Finch Fin Width (nm) (nm)	1.10	1.03	1.75	1.87	2.10	2.20	2.52	2.57
Finch Fin Width (nm) (nm)	5.00	5.50	6.44	6.96	7.73	8.14	8.8	9.6
Finch Fin Width (nm) (nm)	10	12	14	16	18	20	22	24
Finch Fin Width (nm) (nm)	28	22	18	14	11	9	7	5
Finch Fin Width (nm) (nm)	2017							
Finch Fin Width (nm) (nm)	20	17	14	12	10	8	7	5
Finch Fin Width (nm) (nm)	23	19	16	13	11	9	8	6

\*\* Note: from the PIDS working group data; however, the calibration of Vdd, GLph, and I/CV is ongoing for improved targets in 2014 ITRS work

Flash memory has become a new FEOL technology driver for critical dimension scaling, materials and processing (lithography, etching, etc.) technology, ahead of DRAM and logic. Continued Flash density improvements in the near term rely on the thickness scaling of the tunnel oxide and the integrate dielectric. To guarantee the charge retention and endurance requirements, the introduction of high- $k$  materials will be necessary. Cost effective implementation of 3-D NAND flash beyond 256 Gb with MLC and acceptable reliability performance remains a difficult challenge. New challenges also include the inception into mainstream manufacturing of new memory types and storage concepts such as magnetic RAM (MRAM), phase-change memory (PCM), resistive RAM (ReRAM) and ferroelectric RAM (FeRAM).