TOPICS IN COMPUTING WITH EMERGING TECHNOLOGIES

FALL 2020

PROF. IRIS BAHAR

SEPTEMBER 23, 2020

LECTURE 5: EMERGING MEMORY DESIGN

COMING UP FOR WEEK #4

- Paper discussions will start week #4
- Review the following papers:
  - Emerging NVM: A Survey on Architectural Integration and Research Challenges
  - Memory that never forgets: emerging nonvolatile memory and the implication for architecture design
- Papers are now available on Canvas (see under week #4)
  - You have been assigned to 1 or 4 groups (3 groups of 2, 1 group of 3)
  - Expect different team assignments weekly
- I will also assign discussion leaders for the week
  - For Monday, I will be the discussion leader
  - Jiwon will be the scribe
  - Anyone want to volunteer to be the scribe for Wednesday?
- Starting the 6th or 7th week of class, students will be able to choose papers to review

NMOS AND PMOS TRANSISTORS

Each transistor consists of a stack of a conducting gate, an insulating layer of silicon dioxide and a semiconductor substrate (body or bulk)

nMOS transistor

pMOS transistor

Body is typically grounded

Body is typically at supply voltage

THRESHOLD VOLTAGE CONCEPT

- Depletion region: area devoid of mobile carriers (holes)
- Inversion layer: n-channel region under oxide
- The value of $V_{GS}$ where strong inversion occurs is called the threshold voltage, $V_T$
What goes into charging/discharging the capacitance $C_L$?

\[ v(t) = V[1 - e^{-t/RC}] \]
\[ i(t) = C \frac{dv(t)}{dt} = \frac{V}{R} e^{-t/RC} \]

Charge on capacitor, $q(t) = C \cdot v(t)$

Current, $i(t) = dq(t)/dt = C \cdot dv(t)/dt$

Energy/transition = $C_L \cdot V_{DD}^2 \cdot \alpha$

$P_{dyn} = \text{Energy/transition} \cdot f = C_L \cdot V_{DD}^2 \cdot \alpha \cdot f$

$P_{dyn} = C_{\text{EFF}} \cdot V_{DD}^2 \cdot f$ \text{ where } $C_{\text{EFF}} = \alpha \cdot C_L$

Data dependent $\rightarrow$ a function of switching activity!
LOWERING DYNAMIC POWER

Capacitance: Function of fan-out, wire length, transistor sizes

Supply Voltage: Has been dropping with successive generations

Activity factor: How often, on average, do wires switch?

Clock frequency: Recent scaling back...

P_{dyn} = C_L V_{DD}^2 \alpha f_{ck}

LEAKAGE (STATIC) POWER DISSIPATION

- Reducing V_{DD} reduces dynamic power dissipation
- BUT… reduced V_{DD} also means reduced performance
- Improve performance by reducing V_{th}
- BUT… reduced V_{th} means increased leakage current

\[ I_{leakage} \approx I_0 e^{\frac{V_{GS}-V_{th}}{nVT}} \]

- I0 is a function of gate oxide, mobility, and size of device
- VT is the thermal voltage (26mV at T=300K)
- Even when V_{GS}=0V, leakage is non-zero. The closer V_{th} is to zero, the larger the leakage current at V_{GS} = 0V.

SUMMARY: LEAKAGE POWER

- Leakage power as a fraction of the total power increases as clock frequency drops.
  - Turning supply off in unused parts can save power.
- For a gate it is a small fraction of the total power; it can be significant for very large circuits.
- Scaling down features requires lowering the threshold voltage, which increases leakage power
  - roughly doubles with each shrinking.
CMOS PROCESS AT A GLANCE

- One full photolithography sequence per layer (mask)
- Built (roughly) from the bottom up
  5. metal 2
  4. metal 1
  2. polysilicon
  3. source and drain diffusions
  1. tubs (aka wells, active areas)

- One full photolithography sequence per layer (mask)
- Built (roughly) from the bottom up
  5. metal 2
  4. metal 1
  2. polysilicon
  3. source and drain diffusions
  1. tubs (aka wells, active areas)

SELF-ALIGNED GATES

1. Create thin oxide in the “active” regions, thick elsewhere
2. Deposit polysilicon
3. Etch thin oxide from active region (poly acts as a mask for the diffusion)
4. Implant dopant

LAYOUT AND CROSS SECTION OF A CMOS INVERTER

I skipped a lot of steps!
THE INSIDE OF AN INTEGRATED CIRCUIT

BEYOND BULK CMOS PROCESSES
SILICON-ON-INSULATOR & FINFET TRANSISTORS

BULK NMOS AND PMOS TRANSISTORS

Each transistor consists of a stack of a conducting gate, an insulating layer of silicon dioxide and a semiconductor substrate (body or bulk)

nMOS transistor

Body is typically grounded

pMOS transistor

Body is typically at supply voltage

SILICON-ON-INSULATOR (SOI)

Bulk Transistor

Fully Depleted SOI (FDSoI)

Thin Si layer is expensive. Cost can be reduced using Partially Depleted SOI

[Intel]
TRI-GATE TRANSISTOR (FINFET)

- Bulk Transistor
- Fully Depleted Tri-Gate Transistor

FinFET width quantization
\[ W_{\text{FIN}} = 2 \times h_{\text{FIN}} + t_{\text{FIN}} \]
\[ W = N_{\text{FIN}} \times W_{\text{FIN}} \]

FINFET LAYOUT – INVX3

- Drawn \( L_g = 21\text{nm} \)
- Effective \( L_g = 7\text{nm} \)

EMERGING NON-VOLATILE MEMORIES

- DRAM and SRAM are charge-based memories
- Require energy to keep the stored value
- Both technologies struggle to keep up with today’s memory requirements
- Can we do better? Look at emerging memory technologies (ReRAM, PCM, STT-RAM, ….)
EMERGING NVM DEVICE OPTIONS

- Emerging non-volatile memories (NVMs) often involve new mechanisms and/or materials
  - Ferroelectric dielectrics
  - Ferromagnetic metals
  - Carbon materials
- Switching mechanisms often extend beyond classical electronics
  - Quantum mechanical phenomenon
  - Ionic reactions
  - Phase transitions
  - Molecular reconfiguration
- Scalability is a main driver

PHASE CHANGE MEMORY (PCM)

- Based on chalcogenide materials
  - Sulphide, selenide, telluride
- The active region can be switched between amorphous (high resistance state) and crystalline (low resistance)
- Most develop of eNVMs
- <100ns switching speed
- 10⁹ cycle endurance

MAGNETIC TUNNEL JUNCTION (MTJ)

- MTJ: two ferromagnetic layers with a barrier layer in between
- Free bottom layer can switch between parallel and anti-parallel magnetization (LRS and HRS respectively)
- Many commercial implementations already available (STT-RAM)

SPIN-TRANSFER TORQUE MAGNETIC RAM (STT-RAM)

Similar to DRAM, but storage device is connected to sense line
- Read: small voltage applied between sense and bit lines and small current is sensed. → non-destructive, non-volatile
- Write: large current needed to change magnetic orientation of free layer → slow, high-power operation
FERROELECTRIC-FET (FEFET)

- For a FeFET, a ferroelectric layer is used as the gate dielectric.
- Change in ferroelectric polarization modulates FET channel conductance.
- Dielectric remembers its state, even without power.
- Low switching power, but can have high gate leakage and low endurance ($10^4$-$10^5$ cycles).

NON-IDEALITIES

- Retention
  - Non-volatility requires data retention for 10+ years
  - Tested by applying several read cycles at high temperature (85°C).
- Endurance
  - After several write cycles the memory window (high-resistance to low-resistance ratio) degrades.
  - The device is eventually stuck in the LRS or HRS.
- Variability
  - Process variability
  - Resistance variability
  - Cycle-to-cycle variation.