Lecture 16:
Energy & Power

Mark McDermott
Electrical and Computer Engineering
The University of Texas at Austin

Agenda

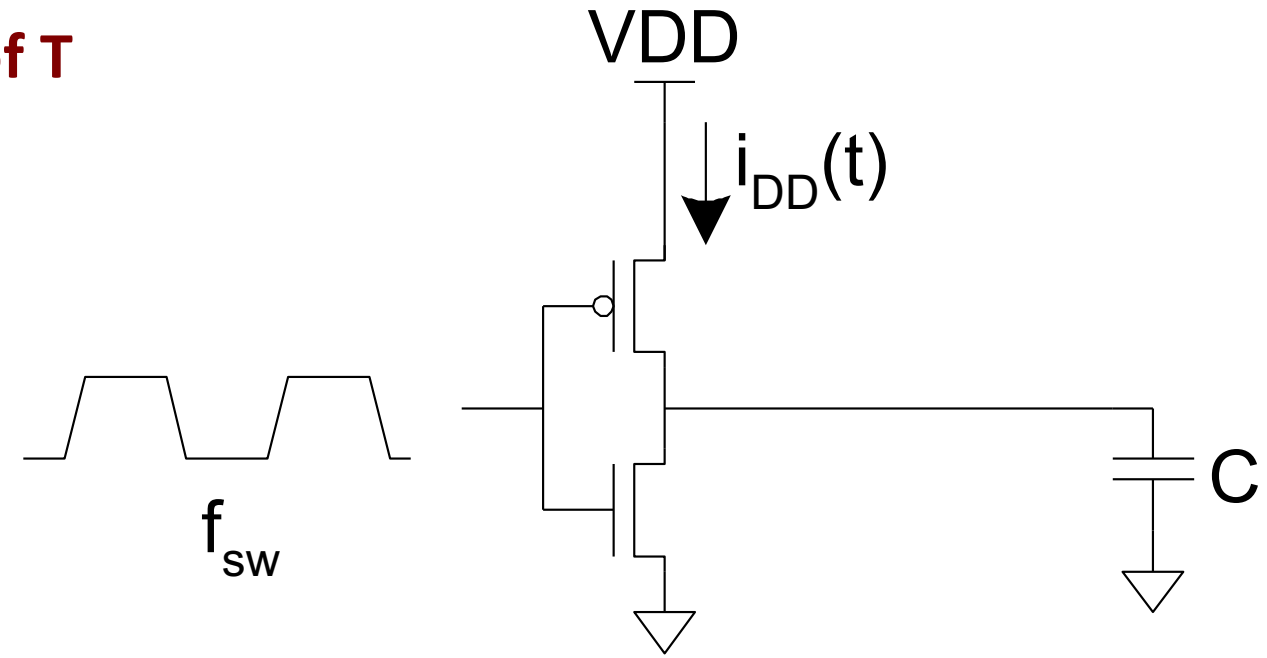
- Power and Energy
- Dynamic Power
- Static Power
- Low Power Design Techniques

Energy and Power

- **Energy is drawn from a voltage source**
- **Instantaneous Power:** $P(t) = i_{DD}(t)V_{DD}$
- **Energy:** $E = \int_0^T P(t)dt = \int_0^T i_{DD}(t)V_{DD}dt$
- **Average Power:** $P_{\text{avg}} = \frac{E}{T} = \frac{1}{T}\int_0^T i_{DD}(t)V_{DD}dt$

Dynamic Power

- Dynamic power required to charge and discharge load capacitances when transistors switch
- One cycle involves a rising and falling output
- On rising output, charge Q = CV_{DD} is required
- On falling output, charge is dumped to GND
- This repeats $T*f_{sw}$ times over an interval of T



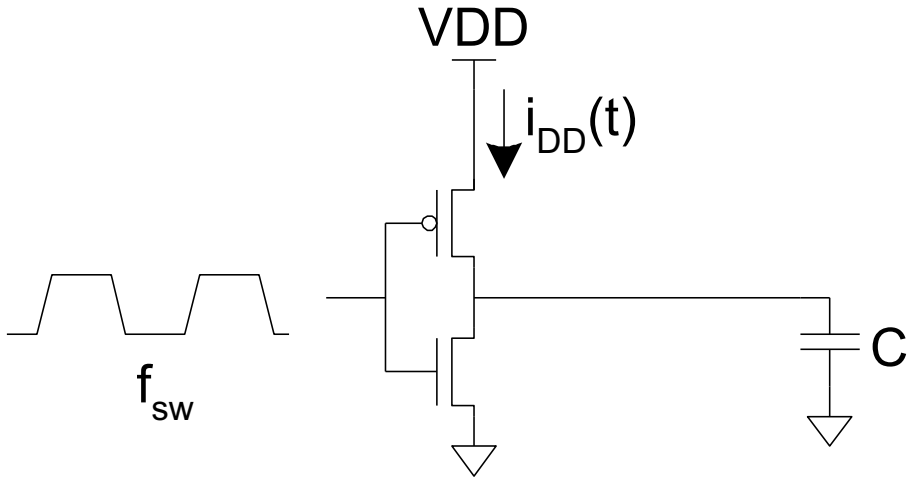
Dynamic Power (Cont.)

$$P_{\text{dynamic}} = \frac{1}{T}\int_0^T i_{DD}(t)V_{DD}dt$$

$$= \frac{V_{DD}}{T}\int_0^T i_{DD}(t)dt$$

$$= \frac{V_{DD}}{T}\left[Tf_{\text{sw}}CV_{DD}\right]$$

$$= CV_{DD}{}^2 f_{\text{sw}}$$



Activity Factor

- Suppose the system clock frequency = f
- Let $f_{sw} = \alpha f$, where α = activity factor
  - If the signal is a clock, $\alpha = 1$
  - If the signal switches once per cycle, $\alpha = ½$
  - Dynamic gates:
    - Switch either 0 or 2 times per cycle, $\alpha = 1$
  - Static gates:
    - Depends on the type of gate and logic network, but typically $\alpha = 0.1 – 0.2$
- Dynamic power: $P_{dyn} = \alpha * C * V_{dd} * \Delta V * freq$

Activity Factor Estimation

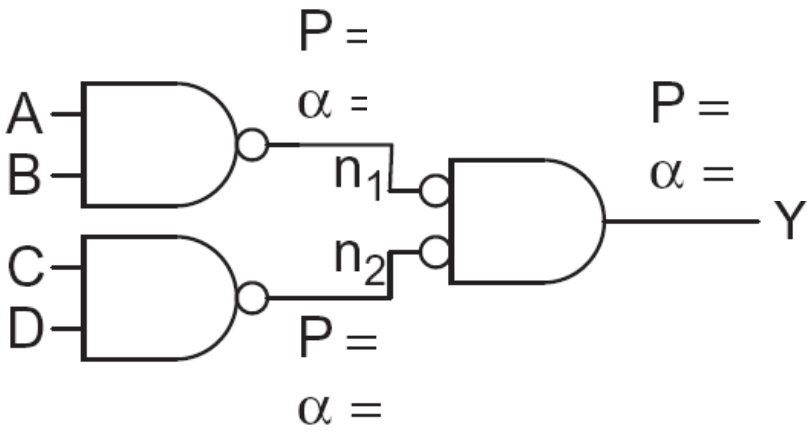
- Let P_i = Prob(node i = 1)
  - $\overline{P}_i = 1-P_i$
- $\alpha_i = \overline{P}_i * P_i$
- Completely random data has P = 0.5 and α = 0.25
- Data is often not completely random
  - e.g. upper bits of 64-bit words representing bank account balances are usually 0
- Data propagating through ANDs and ORs has lower activity factor
  - Depends on design, but typically $\alpha \approx 0.1$

Switching Probability

Gate	P_Y
AND2	$P_A P_B$
AND3	$P_A P_B P_C$
OR2	$1 - \overline{P}_A \overline{P}_B$
NAND2	$1 - P_A P_B$
NOR2	$\overline{P}_A \overline{P}_B$
XOR2	$P_A \overline{P}_B + \overline{P}_A P_B$

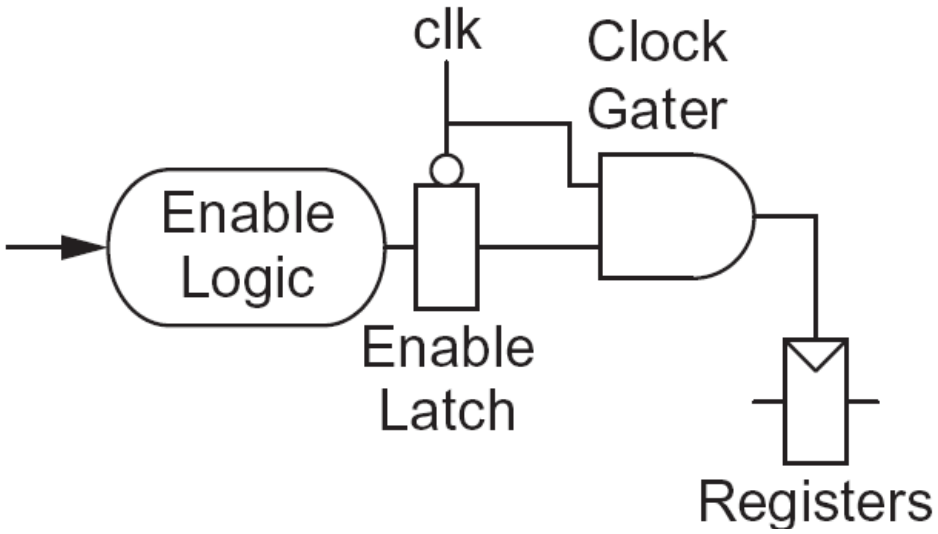
Switching Probability Example

- A 4-input AND is built out of two levels of gates
- Estimate the activity factor at each node if the inputs have P = 0.5



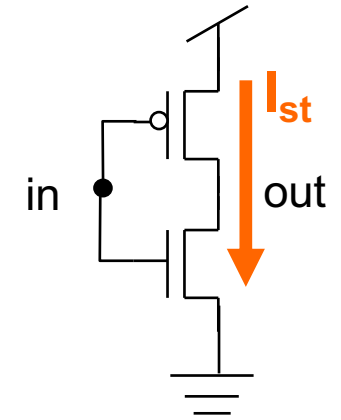
Clock Gating

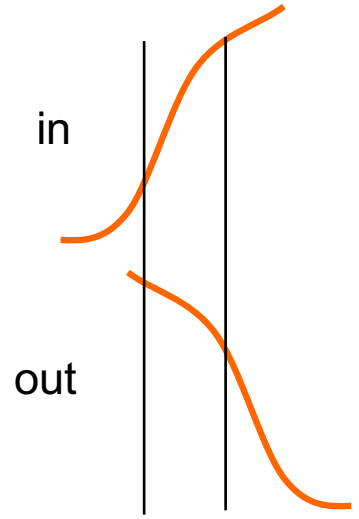
- **The best way to reduce the switching activity is to turn off the clock to registers in unused blocks**
  - **Saves clock activity (where α = 1)**
  - **Eliminates all switching activity in the block**
  - **Requires determining if block will be used**



Short Circuit (Shoot Through) Current

- When transistors switch, both nMOS and pMOS networks may be momentarily ON for a short period of time
- Leads to a blip of "short circuit" current.
- <8-15% of dynamic power if rise/fall times are comparable for input and output





Both Pmos & Nmos conducting

Dynamic Power Example

- 1 billion transistor chip
  - 50M logic transistors
    - Average width: 12 λ
    - Activity factor = 0.1
  - 950M memory transistors
    - Average width: 4 λ
    - Activity factor = 0.02
  - 1.0 V 65 nm process
  - C = 1 fF/μm (gate) + 0.8 fF/μm (diffusion)

- Estimate dynamic power consumption @ 1 GHz. Neglect wire capacitance and short-circuit current

Solution

$$C_{\text{logic}} = \left(50\times10^{6}\right)\left(12\lambda\right)\left(0.025\mu m/\lambda\right)\left(1.8\,fF/\mu m\right) = 27\text{ nF}$$

$$C_{\text{mem}} = \left(950\times10^{6}\right)\left(4\lambda\right)\left(0.025\mu m/\lambda\right)\left(1.8\,fF/\mu m\right) = 171\text{ nF}$$

$$P_{\text{dynamic}} = \left[0.1C_{\text{logic}} + 0.02C_{\text{mem}}\right]\left(1.0\right)^{2}\left(1.0\text{ GHz}\right) = 6.1\text{ W}$$

Static Power

- **Static power is consumed even when chip is quiescent.**
  - Ratioed circuits burn power in fight between ON transistors
  - Leakage draws power from nominally OFF devices

$$I_{ds} = I_{ds0} e^{\frac{V_{gs} - V_t}{n v_T}} \left[1 - e^{\frac{-V_{ds}}{v_T}} \right] \qquad V_t = V_{t0} - \eta V_{ds} + \gamma \left(\sqrt{\phi_s + V_{sb}} - \sqrt{\phi_s} \right)$$

- **Subthreshold leakage especially increasing in short channel devices (DIBL) and at high Temp –> 100-1000nA/u**
- **Subthreshold slope 85-110 mV/decade**
- **Cooling changes the slope….but can it be energy efficient?**

Leakage Power = $KVe^{(Vgs-Vt)q/nkT} (1 - e^{-Vds\ q/kT})$

Leakage Example

- Process has two threshold voltages and two oxide thicknesses
- Subthreshold leakage:
  - 20 nA/mm for low Vt
  - 0.02 nA/mm for high Vt
- Gate leakage:
  - 3 nA/mm for thin oxide
  - 0.002 nA/mm for thick oxide
- Memories use low-leakage transistors everywhere
- Gates use low-leakage transistors on 80% of logic

Leakage Example Cont.

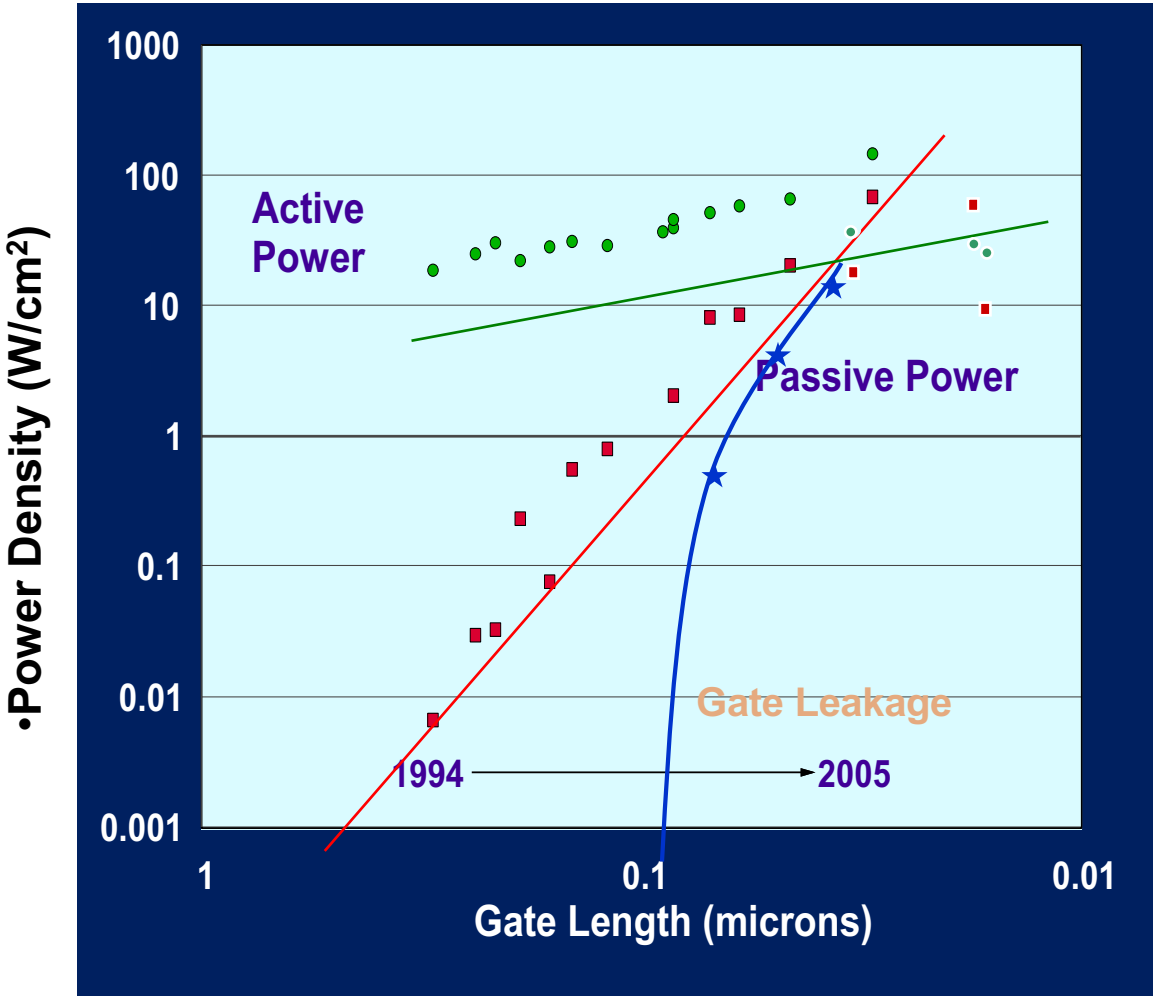
- **Estimate static power:**
  - **High leakage:** $\left(20\times10^{6}\right)\left(0.2\right)\left(12\lambda\right)\left(0.05\mu m/\lambda\right)=2.4\times10^{6}\,\mu m$
  - **Low leakage:** $\left(20\times10^{6}\right)\left(0.8\right)\left(12\lambda\right)\left(0.05\mu m/\lambda\right)+$
    $\left(180\times10^{6}\right)\left(4\lambda\right)\left(0.05\mu m/\lambda\right)=45.6\times10^{6}\,\mu m$

$$I_{static}=\left(2.4\times10^{6}\,\mu m\right)\left[\left(20nA/\mu m\right)/2+\left(3nA/\mu m\right)\right]+$$
$$\left(45.6\times10^{6}\,\mu m\right)\left[\left(0.02nA/\mu m\right)/2+\left(0.002nA/\mu m\right)\right]$$
$$=32mA$$
$$P_{static}=I_{static}V_{DD}=38mW$$

- **If no low leakage devices, P_{static} = 749 mW**

Passive Power Continues to Increase

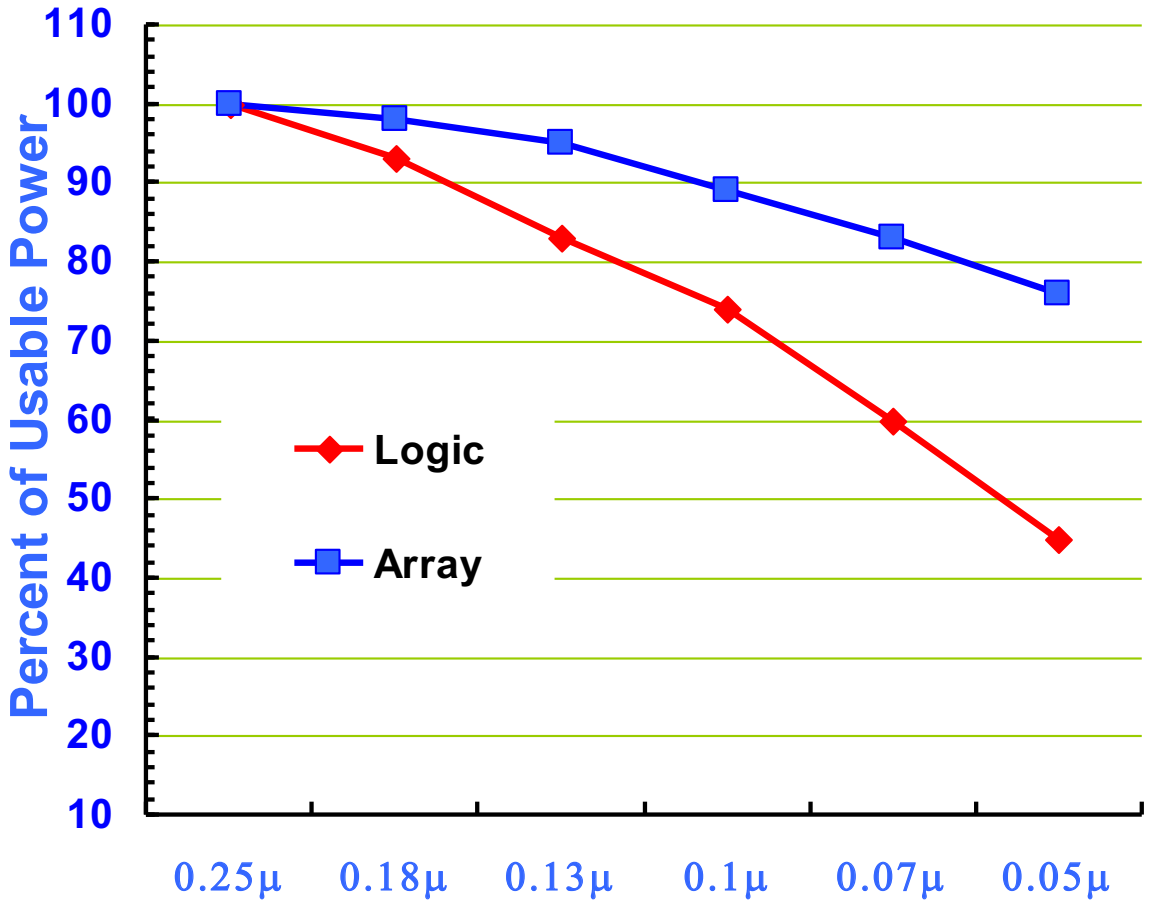
- **Leakage is the price we pay for the increasing device performance**



Src: Nowka, et al

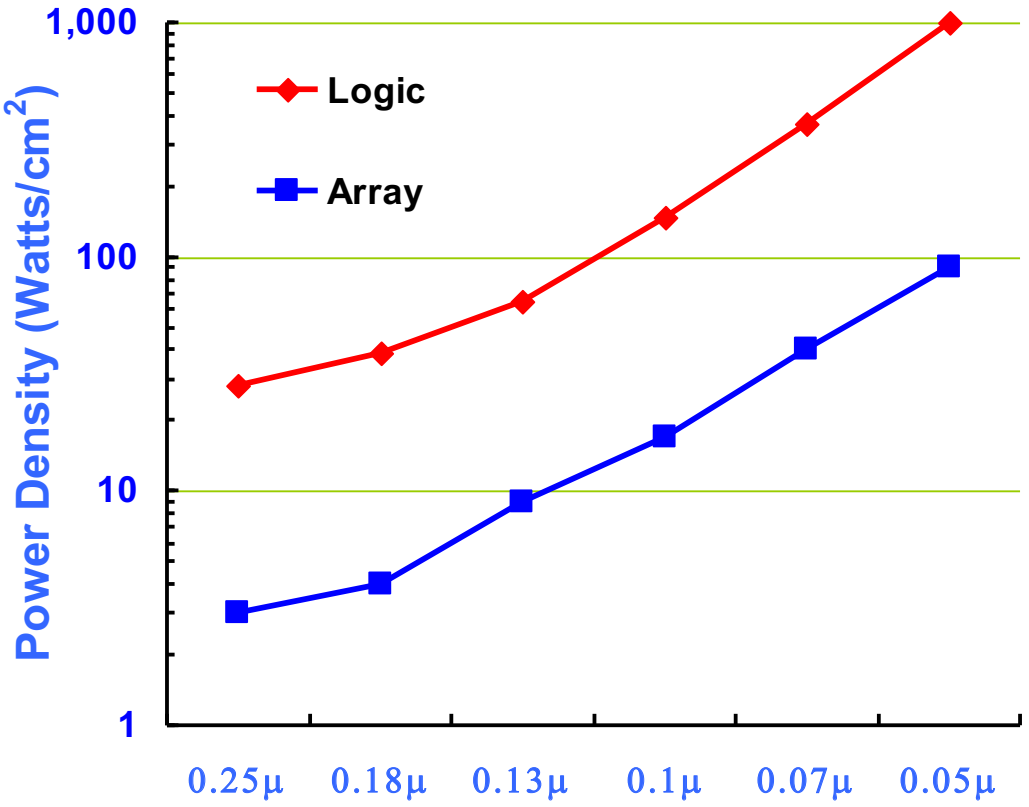
Usable power is decreasing

- Usable power is diminishing due to various transistor effects.
- Logic devices are degrading faster than array devices.

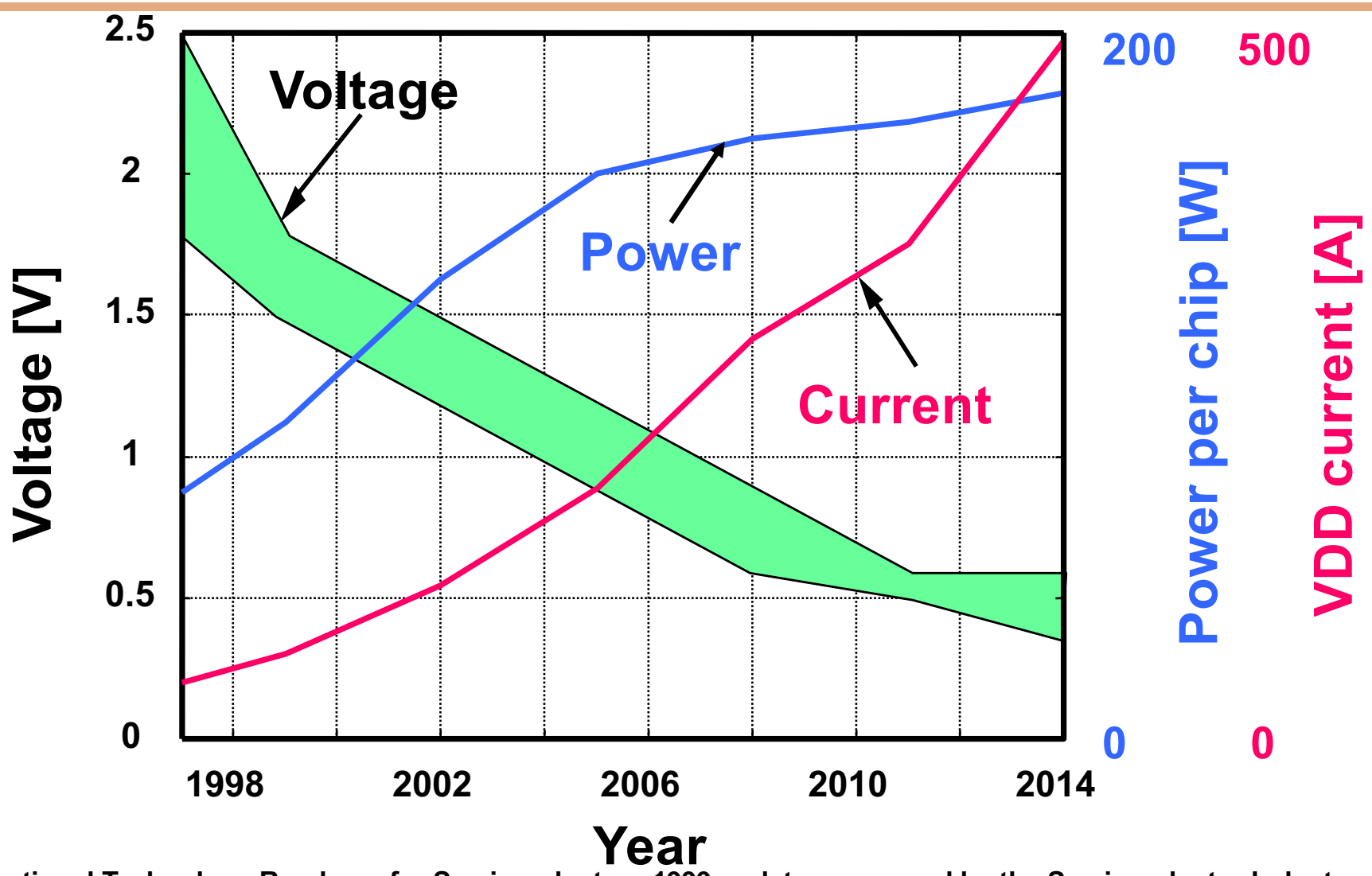


Power density is increasing

- Next generation process will provide more transistors than can be used (@ fixed power).
- Use of array devices may provide better POWER-PERF optimization.

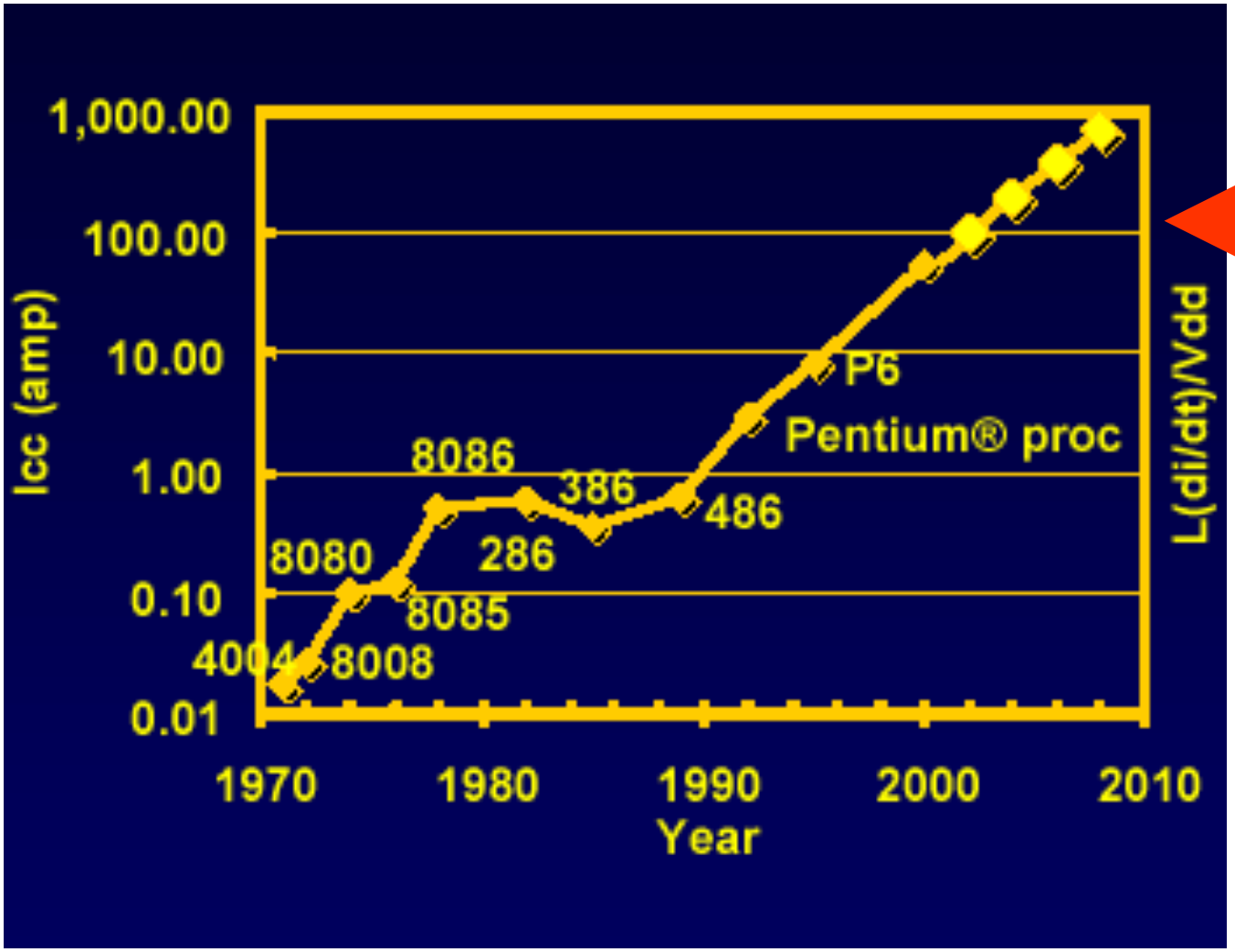


VDD, Power and Current Trend



International Technology Roadmap for Semiconductors 1999 update sponsored by the Semiconductor Industry Association in cooperation with European Electronic Component Association (EECA) , Electronic Industries Association of Japan (EIAJ), Korea Semiconductor Industry Association (KSIA), and Taiwan Semiconductor Industry Association (TSIA) (* Taken from Sakurai's ISSCC 2001 presentation)

Current Delivery Problem (Circa 2005)



Arc Welder

Source: Shekhar Borkar, Intel

Let's talk about Low Power Design Techniques

Low Power Design

- **Reduce dynamic power**
  - α: clock gating, sleep mode
  - C: small transistors (esp. on clock), short wires
  - V_{DD}: lowest suitable voltage
  - f: lowest suitable frequency

- **Reduce static power**
  - Selectively use ratioed circuits
  - Selectively use low V_t devices
  - Leakage reduction
    - Stacked devices, body bias, low temperature

Functional Clock Gating

$$P = \tfrac{1}{2} C_{sw} V_{dd} \Delta V f + I_{st} V_{dd} + I_{static} V_{dd}$$

- 25-50% of power consumption due to driving latches (Bose, Martinozi, Brooks 2001 50%)
- Utilization of most latches is low (~10-35%)
- Gate off unused latches and associated logic:
  - Unit level clock gating – turn off clocks to FPU, MMX, Shifter, L/S unit, … at clk buffer or splitter
  - Functional clock gating – turn off clocks to individual latch banks – forwarding latch, shift-amount register, overflow logic & latches, …qualify (AND) clock to latch
- Asynch is the most aggressive gating – but is it efficient?

Nowka EE382M

Glitch suppression

$$P = \tfrac{1}{2} C_{sw} V_{dd} \Delta V f + I_{st} V_{dd} + I_{static} V_{dd}$$

- Glitches can represent a sizeable portion of active power, (up to 30% for some circuits in some studies)
- Three basic mechanisms for avoidance:
  - Use non-glitching logic, e.g. domino
  - Add redundant logic to avoid glitching hazards
    - Increases cap, testability problems
  - Adjust delays in the design to avoid
    - Shouldn't timing tools do this already if it is possible?

Nowka EE382M

Voltage minimization

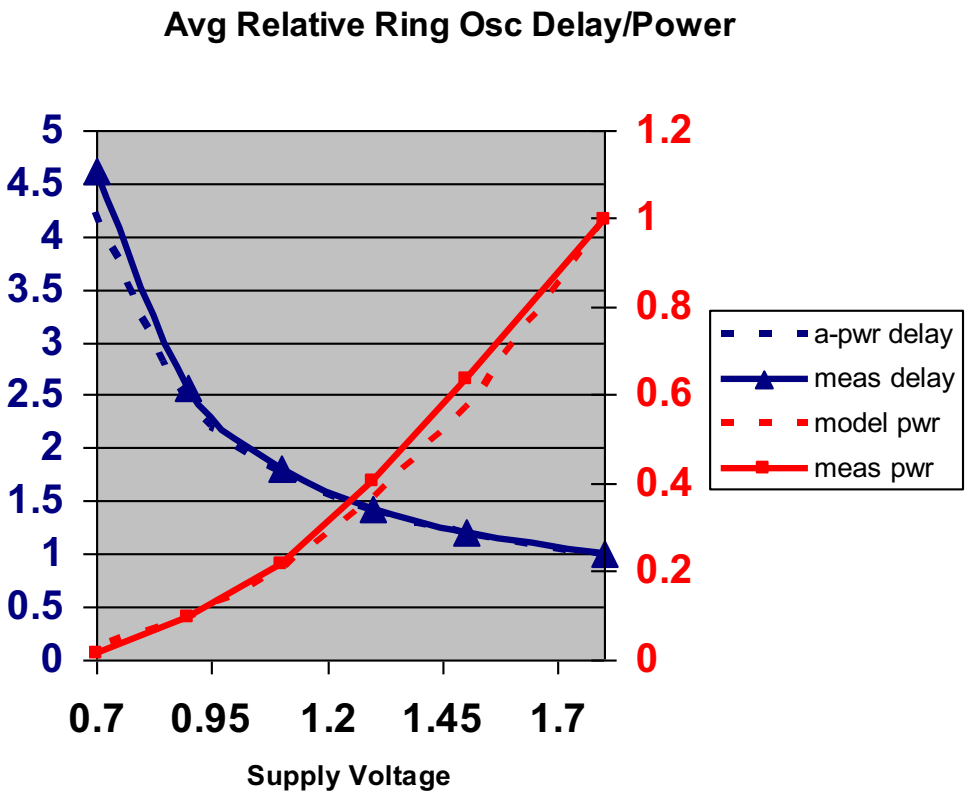
$$P = \tfrac{1}{2} C_{sw} V_{dd} \Delta V f + I_{st} V_{dd} + I_{static} V_{dd}$$

- Lowering voltage swing, DV, lowers power
  - Low swing logic efforts have not been very successful (unless you consider array voltage sensing)
  - Low swing busses have been quite successful
- Lowering supply, Vdd and DV, (voltage scaling) is most promising:
  - Frequency ~V, Power ~V^3

Nowka EE382M

Voltage Scaling Reduces Active Power

- **Voltage Scaling Benefits**
  - Can be used widely over entire chip
  - Complementary CMOS scales well over a wide voltage range => Can optimize power/performance (MIPS/mW) over a wide range
- **Voltage Scaling Challenges**
  - Custom CPUs, Analog, PLLs, and I/O drivers don't voltage scale easily
  - Sensitivity to supply voltage varies circuit to circuit – esp SRAM, buffers, NAND4
  - Thresholds tend to be too high at low supply



After Carpenter, Microprocessor forum, '01

Nowka EE382M

Frequency minimization

$$P = \tfrac{1}{2} C_{sw} V_{dd} \Delta V f + I_{st} V_{dd} + I_{static} V_{dd}$$

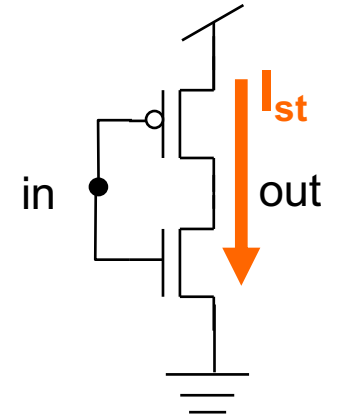
- Lowering frequency lowers power linearly
  - DOES NOT improve energy efficiency, just slows down energy consumption
  - Important for avoiding thermal problems

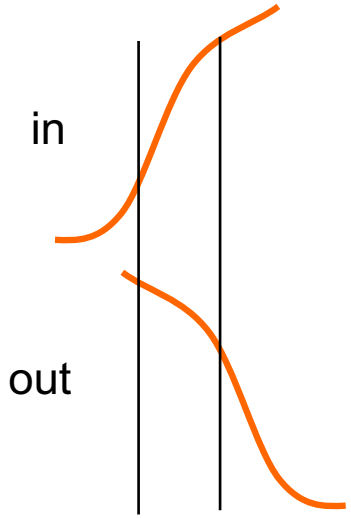
Nowka EE382M

Shoot-through minimization

$$P = \tfrac{1}{2} C_{sw} V_{dd} \Delta V f + I_{st} V_{dd} + I_{static} V_{dd}$$

- For most designs, shoot-thru represents 8-15% of active power.
- Avoidance and minimization:
  - Lower supply voltage
  - Domino?
  - Avoid slow input slews
  - Careful of level-shifters in multiple voltage domain designs





Both Pfet &
Nfet conducting

Nowka EE382M

Standby-Power Reduction Techniques

- **Standby power can be reduced through:**
  - Frequency scaling for circuits that are still toggling
  - Voltage-scaling
  - Power gating
  - Vdd/Vt selection

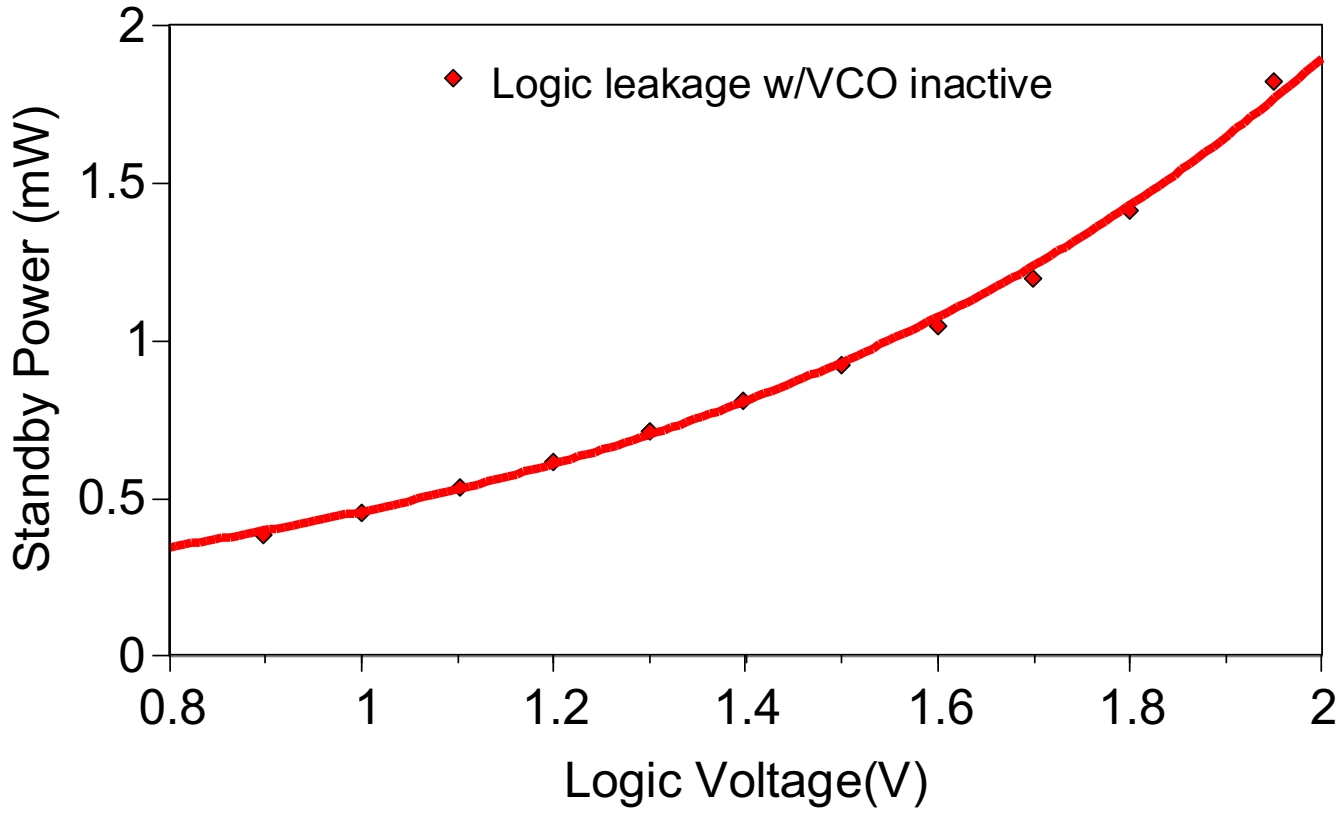
Capacitance minimization

- **Only the devices (device width) used in the design leakage calc!**
  - **Runs counter to the complexity-for-IPC trend**
  - **Runs counter to the SOC trend**
  - **Transistors are not free -- Even though they are not switched they still leak**

Nowka EE382M

Voltage Scaling Standby Reduction

Decreasing the supply voltage significantly improves standby power



Subthreshold dominated technology

Nowka EE382M

Supply/Power Gating

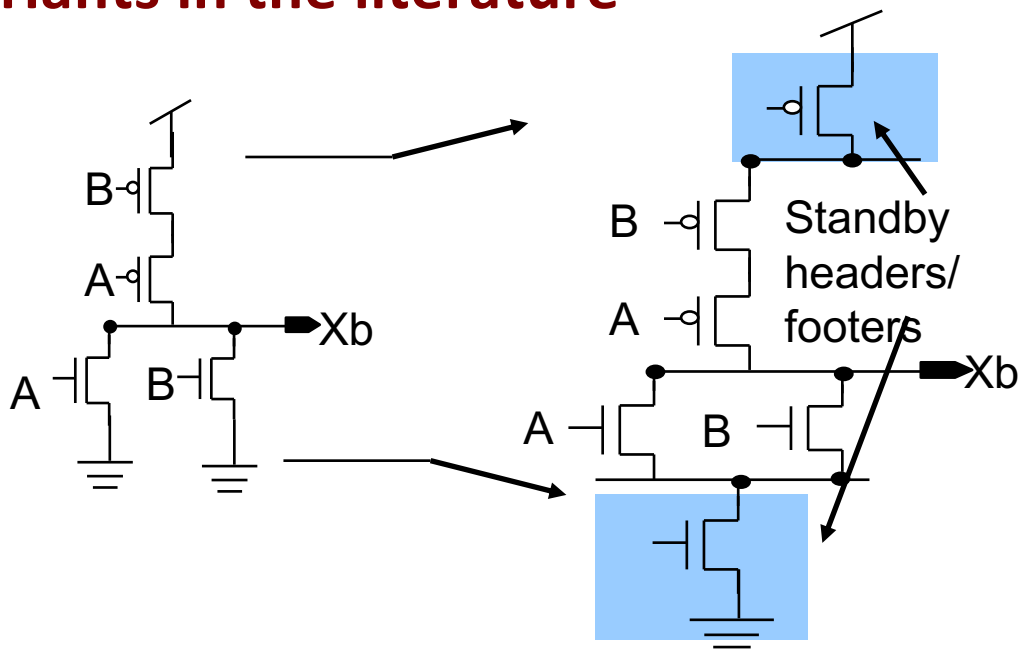
- **Especially for energy constrained (e.g. battery powered systems).**

  Two levels of gating:

  - **"Standby, freeze, sleep, deep-sleep, doze, nap, hibernate": lower or turn off power supply to system to avoid power consumption when inactive**
    - **Control difficulties, hidden-state, entry/exit, "instant-on" or user-visible.**
  - **Unit level power gating – turn off inactive units while system is active**
    - **Eg. MTCMOS**
    - **Distribution, entry/exit control & glitching, state-loss…**

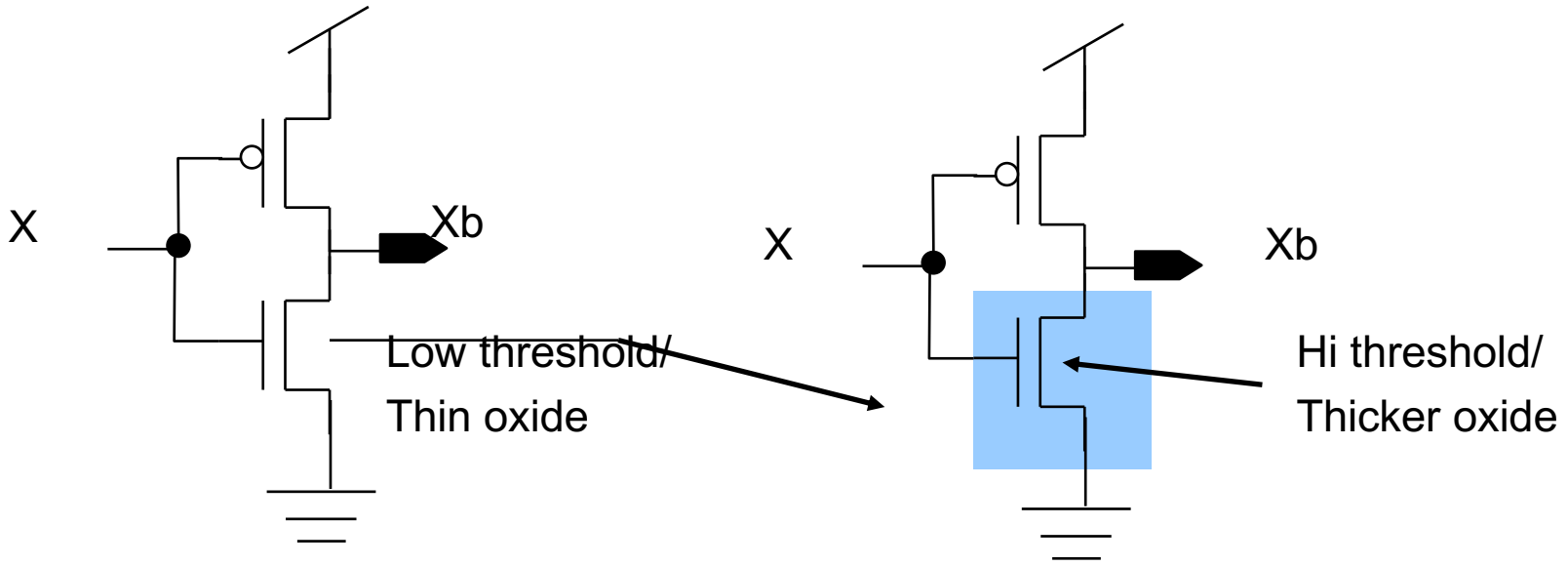
MTCMOS

- Use header and/or footer switches to disconnect supplies when inactive.
- For performance, low-Vt for logic devices.
- 10-100x leakage improvement, ~5% perf overhead
- Loss of state when disconnected from supplies
- Large number of variants in the literature



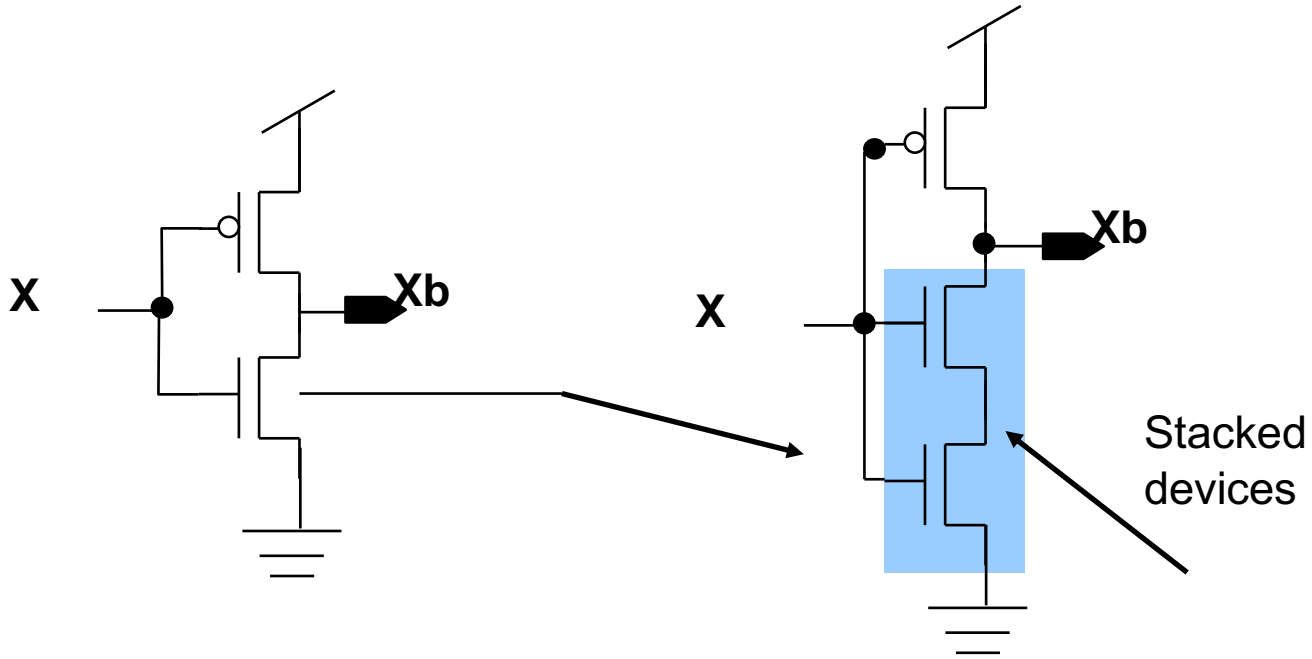
Vt / Tox selection

- Low Vt devices on critical paths, rest high Vt
- 70-180mV higher Vt, 10-100x lower leakage, 5-20% slower
- Small fraction of devices low-Vt (1-5%)
- Thick oxide reduces gate leakage by orders of magnitude



Device Stacking

- Decreases subthreshold leakage
- Improvement beyond use of long channel device
- 2-5x improvement in subthreshold leakage
- 15-35% performance penalty



Summary

- Technology scaling hasn't solved the power/energy problems.
- So what to do? We've shown that,
- Do less and/or do more in parallel at low V_{dd}. For the circuit designer this implies:
  - Support low V_{dd}
  - Support power-down modes
  - Choosing the right mix of Vt,
  - Sizing devices appropriately
  - Choosing right V_{DD} per block (voltage islands)

Nowka EE382M