EEC 118 Lecture #14: Low Power Circuits

Rajeevan Amirtharajah
University of California, Davis

Jeff Parkhurst
Intel Corporation

Announcements

- **HW7: Optional**
  - Issued later today
- **Lab 6: Memories**
  - Issued this evening, due last day of class
- **Quiz 4 Wednesday**

Outline

- **Review: Memory Basics**
- **Finish Memories: Rabaey 12.1-12.2 (Kang & Leblebici, 10.1-10.6)**
- **Low Power Circuits: Rabaey 5.5 (Kang & Leblebici, 11.1-11.3)**

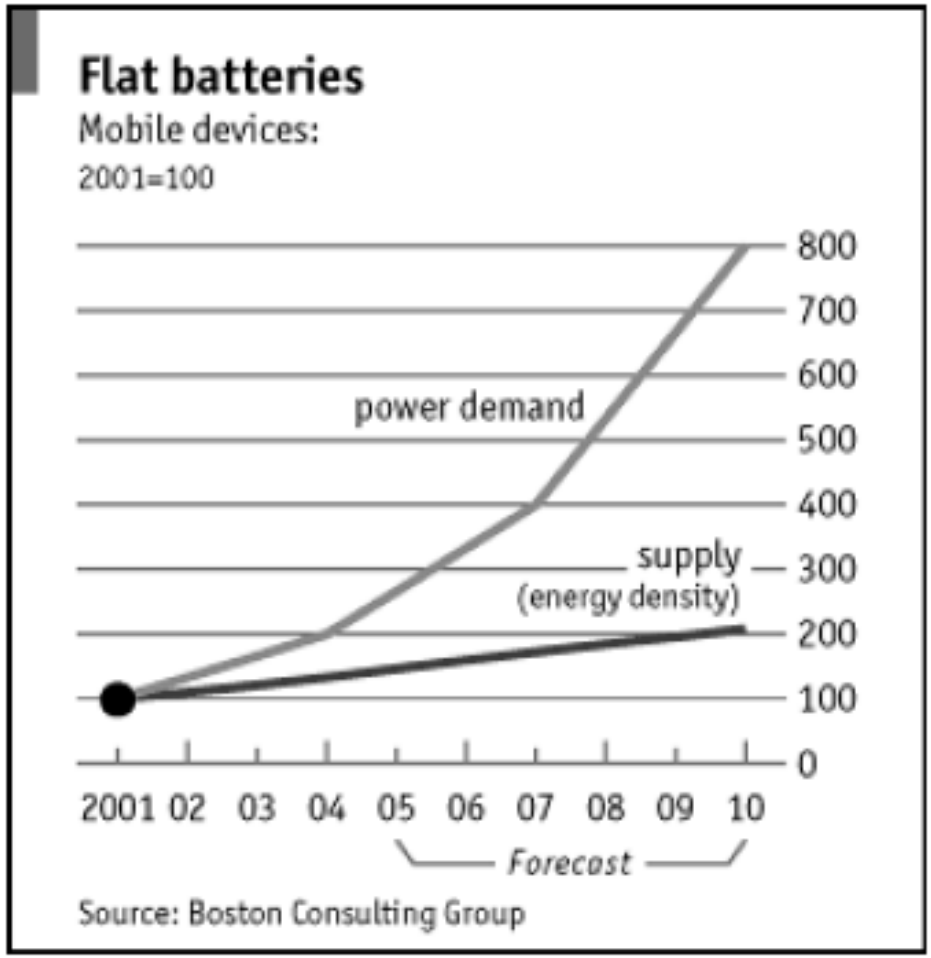
Why Power Matters

- **Packaging costs: many pins to get 10s of Amps into chip**
- **Power supply rail design: must get 10s of Amps through 1-10 μm^2 of on-chip wire area**
- **Chip and system cooling costs: large server farms might consume 1-10s MW**
- **Noise immunity and system reliability: high temperature bad for noise, devices degrade faster**
- **Battery life and weight (in portable systems)**
- **Environmental concerns**
  - Office equipment accounted for 5% of total US commercial energy usage in 1993

State-of-the-Art Processor Power

- **Reported at ISSCC 2004**
  - IBM POWER5: 130 nm SOI, 1.5 GHz at 1.3 V, incorporates 24 digital temperature sensors distributed over die for hot-spot throttling
  - Sun UltraSPARC: 130 nm CMOS, 1.2 GHz at 1.3 V, 23 W typical dissipation
  - IBM PowerPC 970: 130 nm SOI, 1.8 GHz at 1.45 V, 57 W typical dissipation
  - IBM PowerPC 970+: 90 nm SOI, 2.5 GHz at 1.3 V, 49 W typical dissipation
- **Careful design still keeping power below 100 W**
  - Montecito ISSCC 2005 (dual-core Itanium): 300 W down to 100 W

Recent Battery Scaling and Future Trends

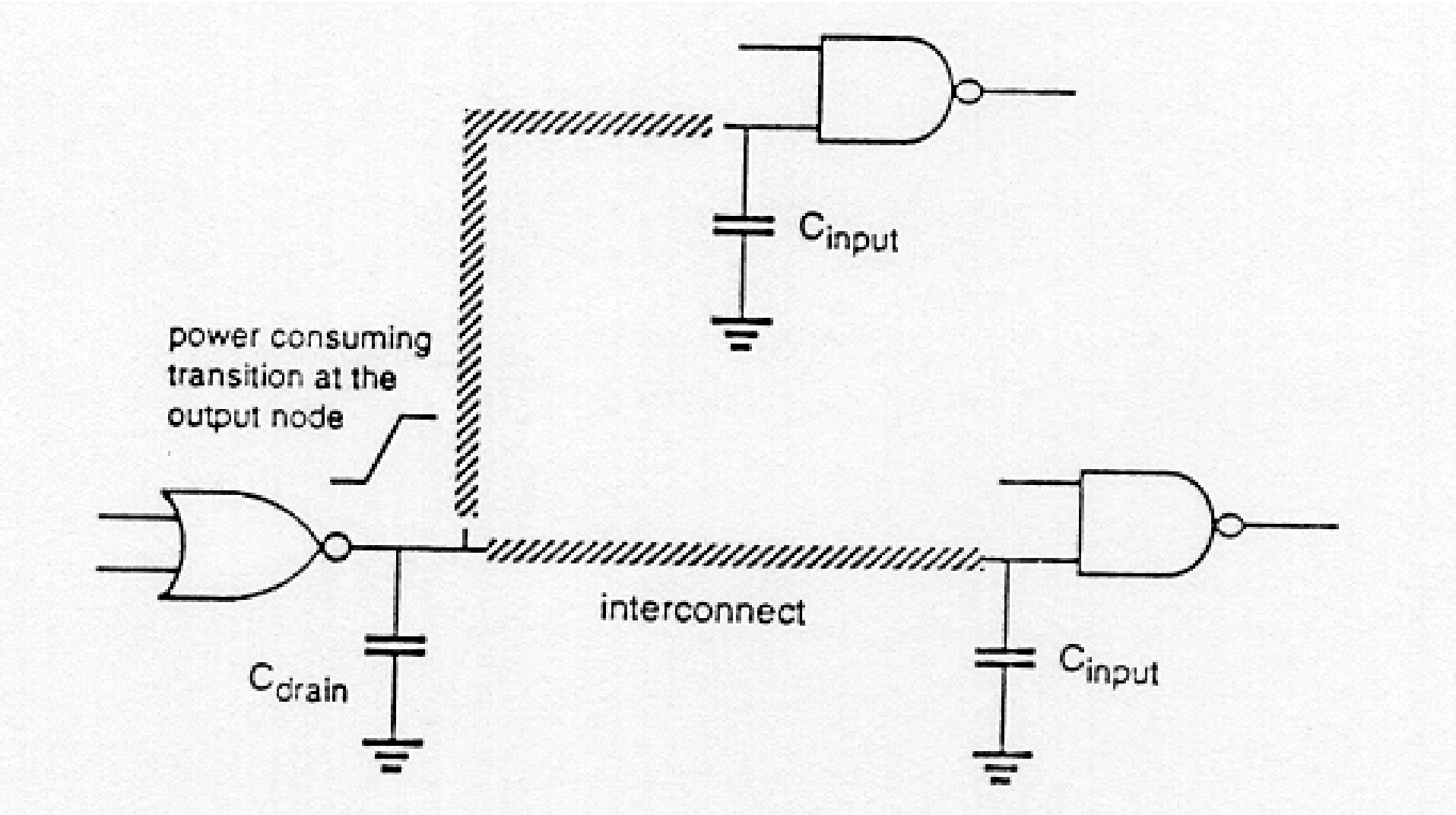


- **Battery energy density increasing 8% per year, demand increasing 24% per year (the Economist, January 6, 2005)**

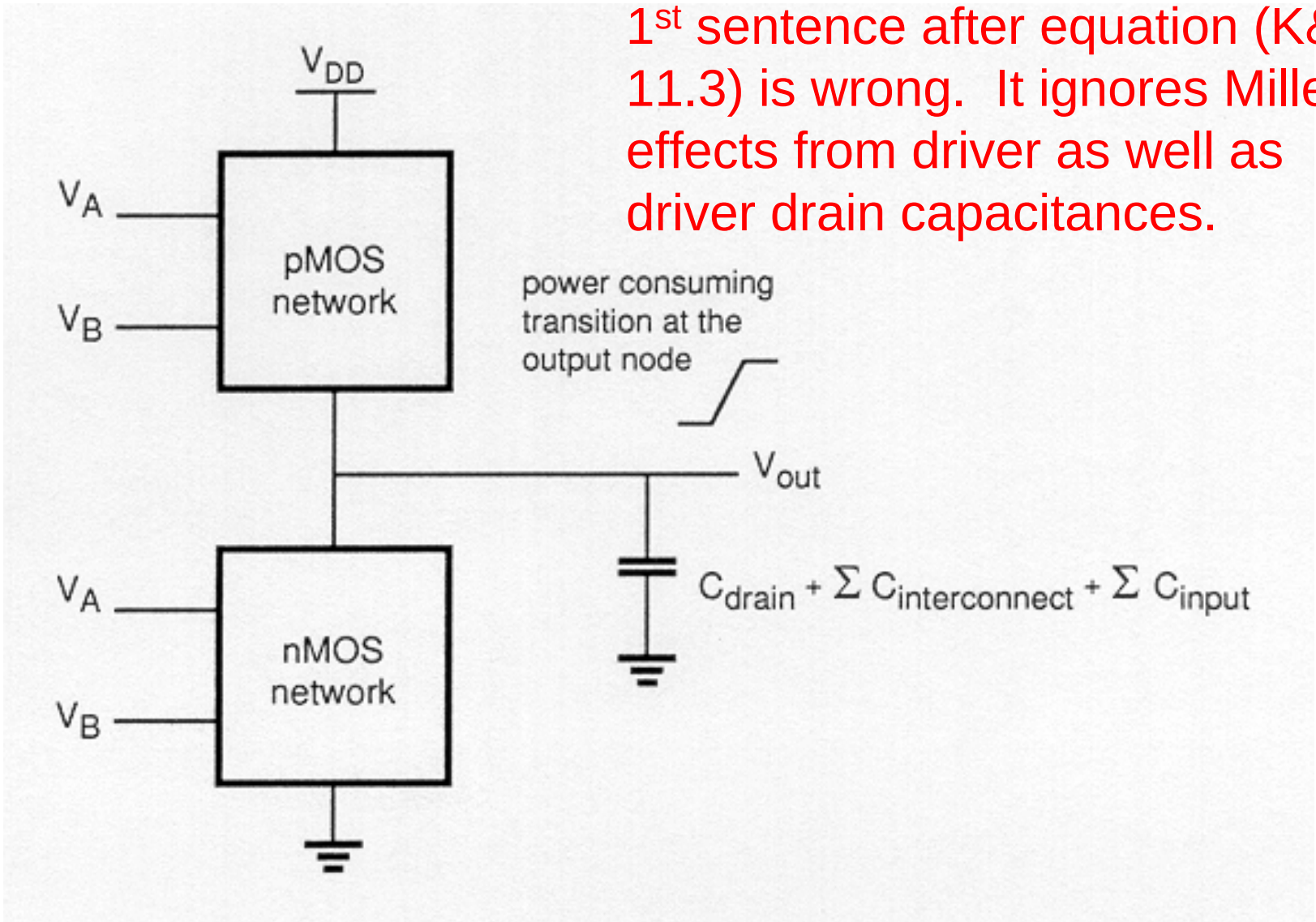
Overview of Dynamic Power Consumption

- **Dynamic (Switching) Power Dissipation**
  - Due to charging output node capacitance
    - Output node capacitance of driver
    - Total interconnect capacitance
    - Input node capacitance of receivers
- $\mathbf{P_{avg} = C_{Load} \times (V_{DD})^2 \times F_{clk}}$
  - Note power is a factor of
    - Supply voltage
    - Switching frequency
    - C_{load} (transistor sizing, interconnect width)
    - NOT dependent on rise/fall

Circuit Capacitances



Capacitance Analysis



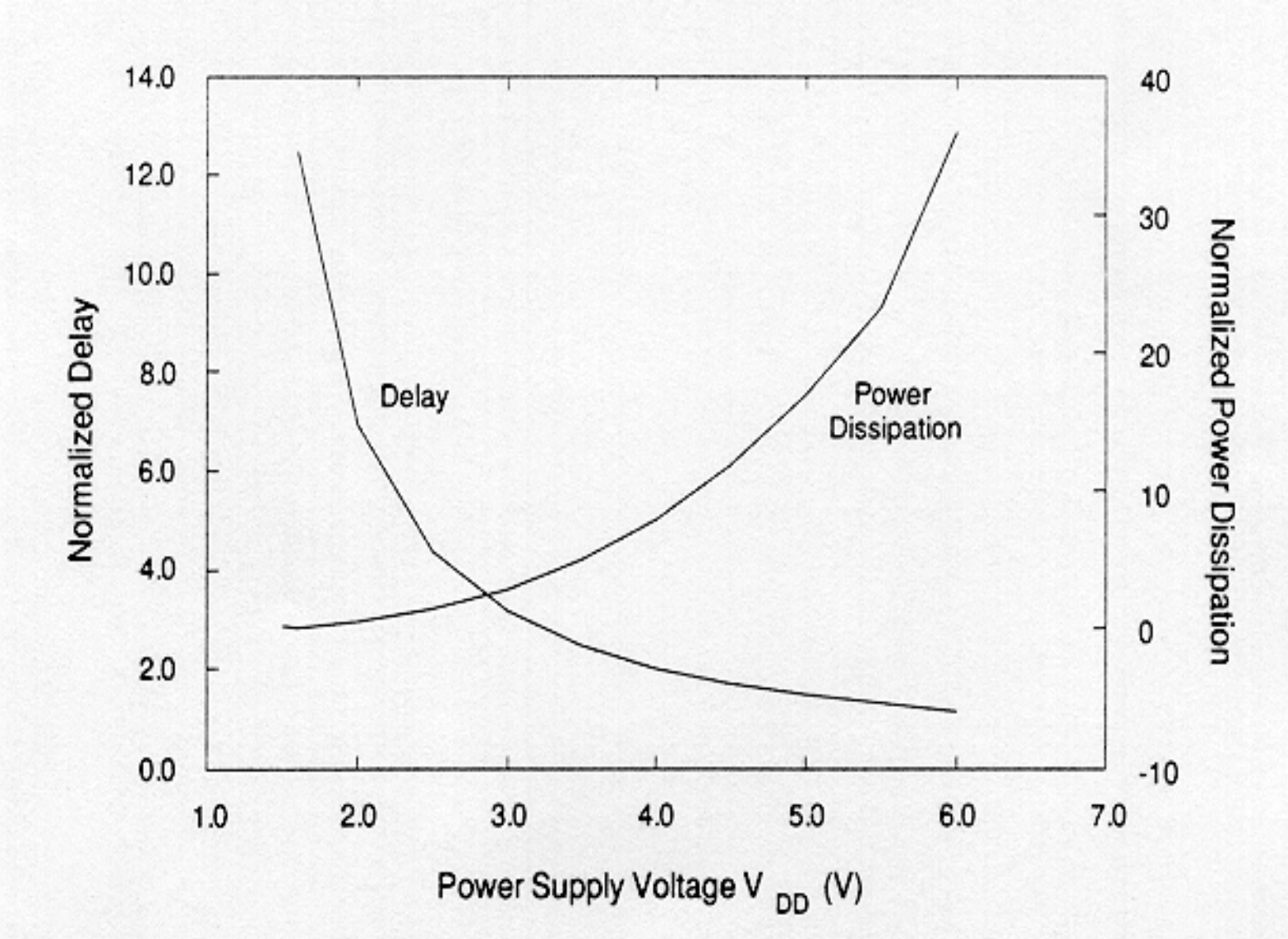
1st sentence after equation (K&L 11.3) is wrong. It ignores Miller effects from driver as well as driver drain capacitances.

Reducing Switching Power Consumption

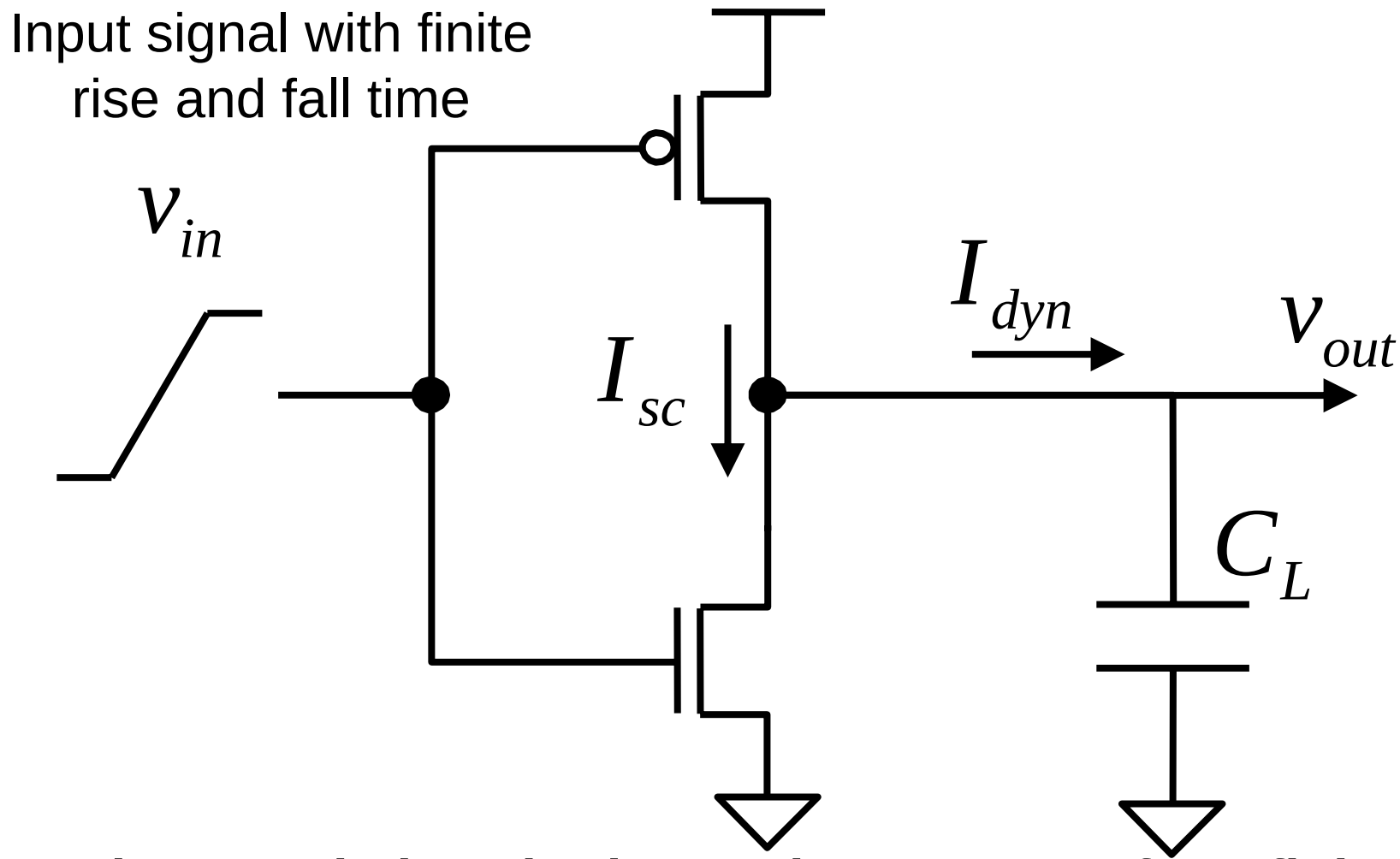
$$P_{avg} = C_{Load} \times (V_{DD})^2 \times F_{clk}$$

- **Reduce Power Supply voltage**
  - Process scaling accomplishes this due to reliability issues, but trend is slowing down
- **Reduce load capacitance**
  - Process scaling helps with this (approximately halves capacitance every node)
  - Proper sizing of transistors
- **Reduce activity factor (probability that capacitance is charged)**
  - Refer to Rabaey 6.2 (K&L 11.4)

Delay and Power versus Supply Voltage

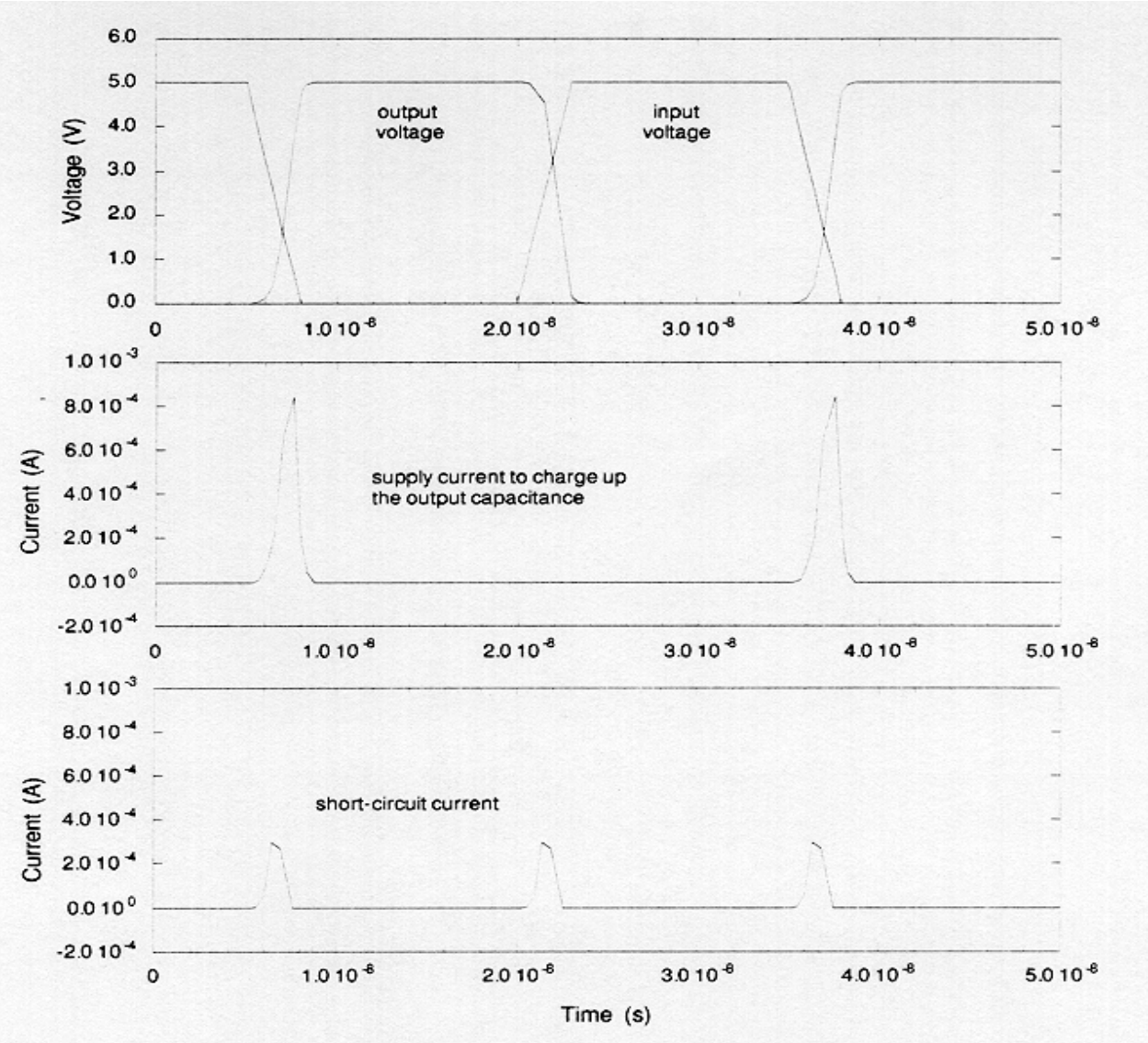


CMOS Inverter Short Circuit Current

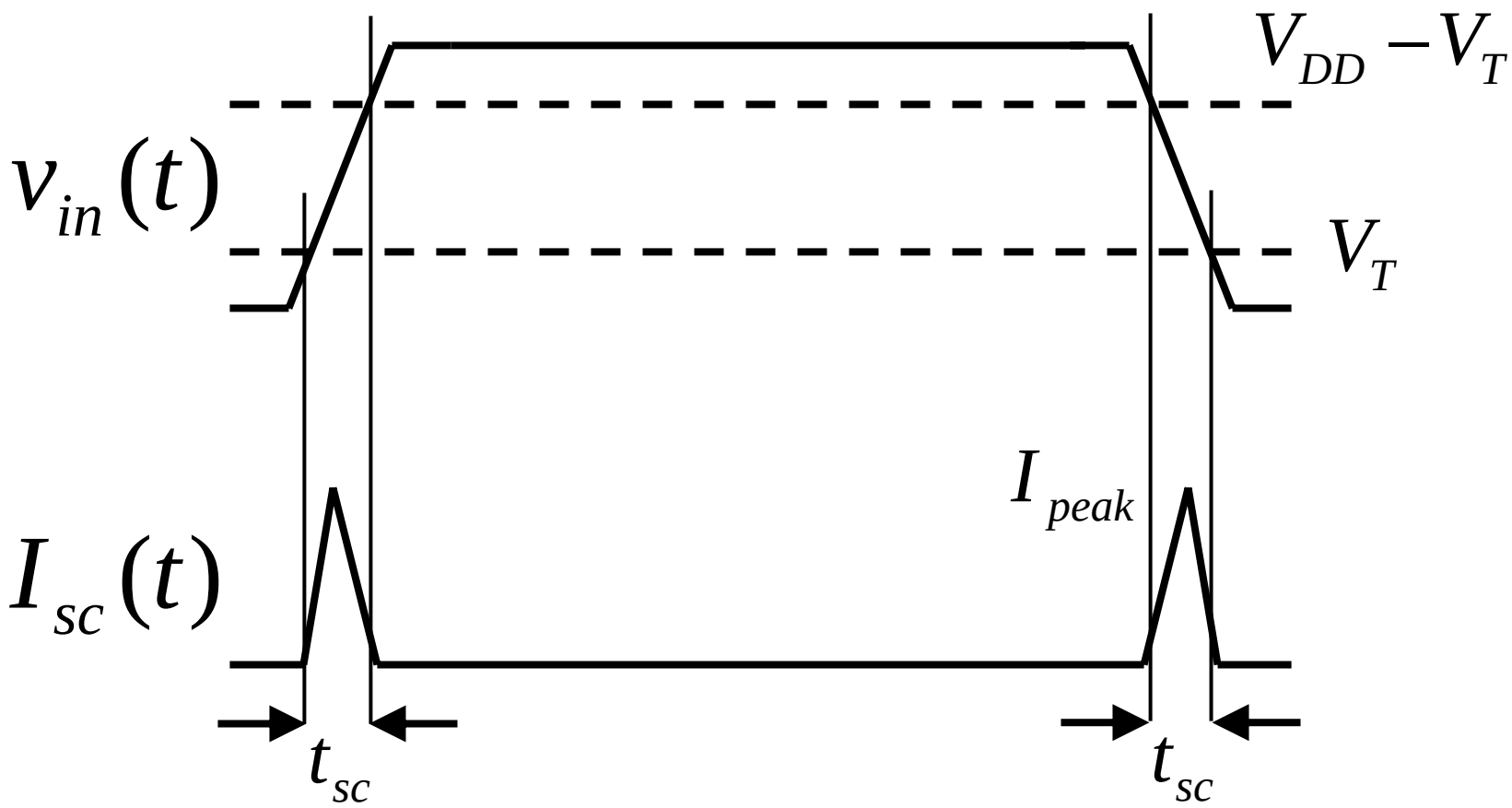


- **As input switches, both transistors are on for a finite amount of time: current travels from Vdd directly to Gnd**

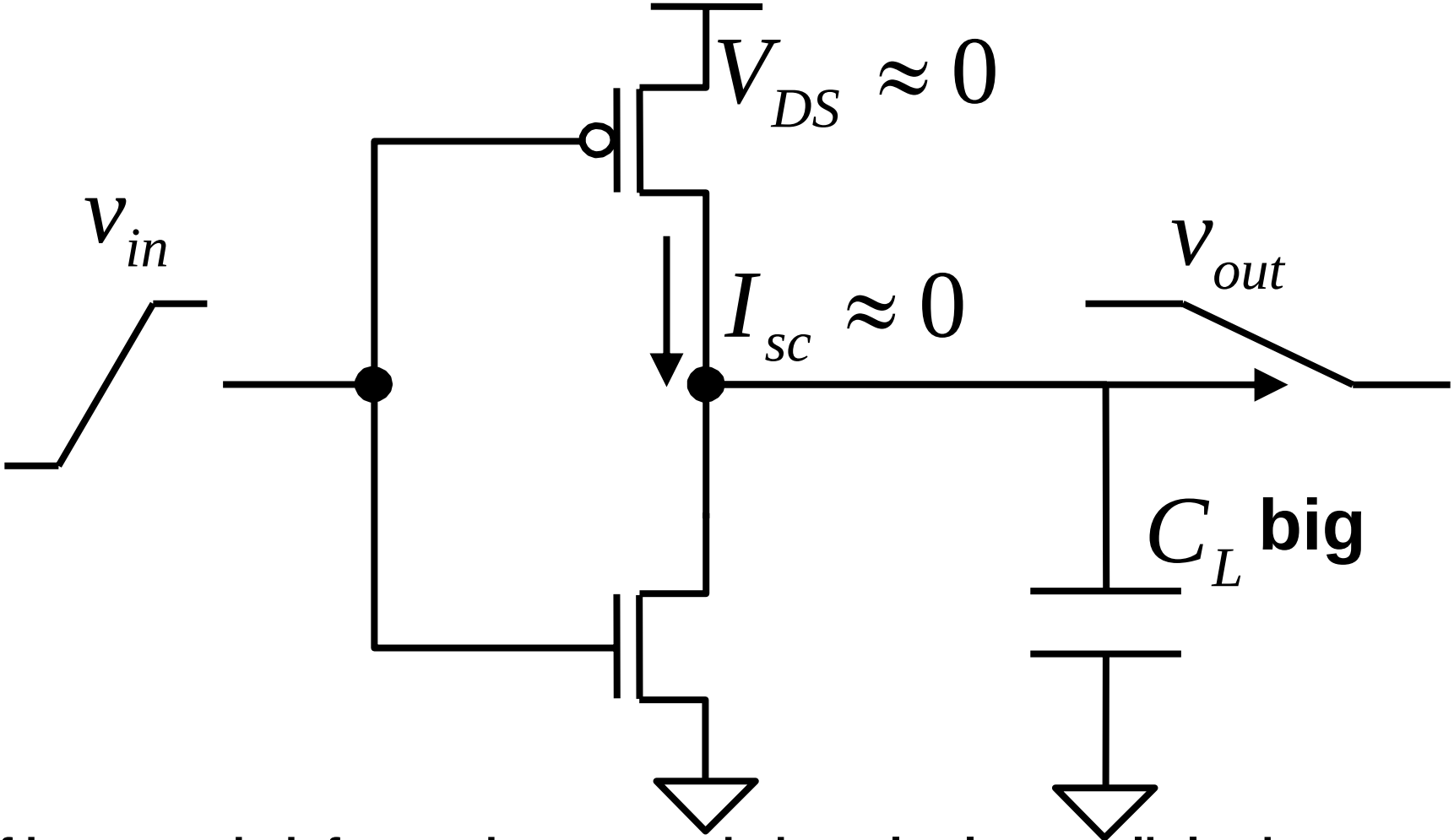
Short Circuit Power Dissipation



Short Circuit Current Triangle Approx.

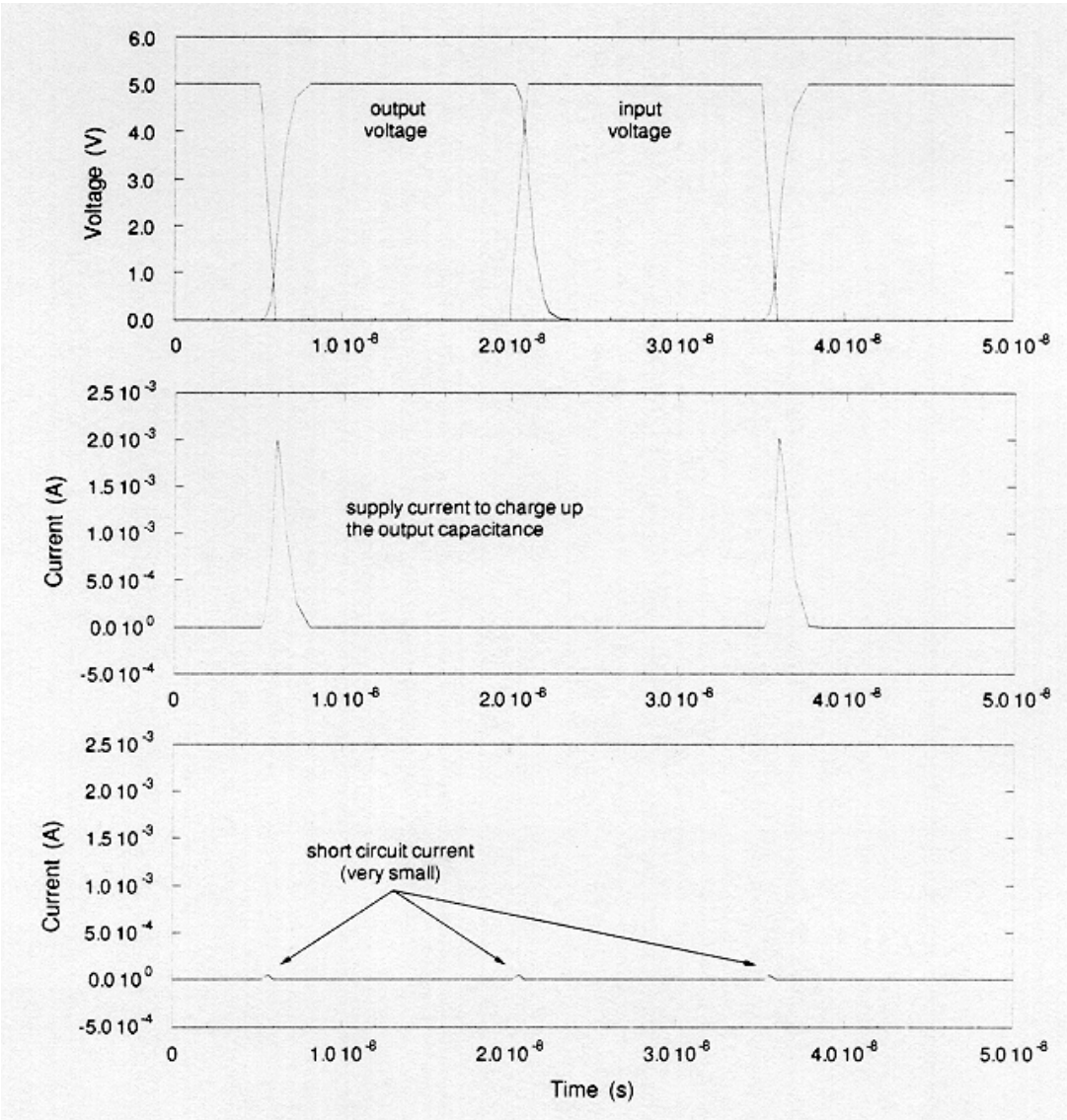


Short Circuit Current With Large Load

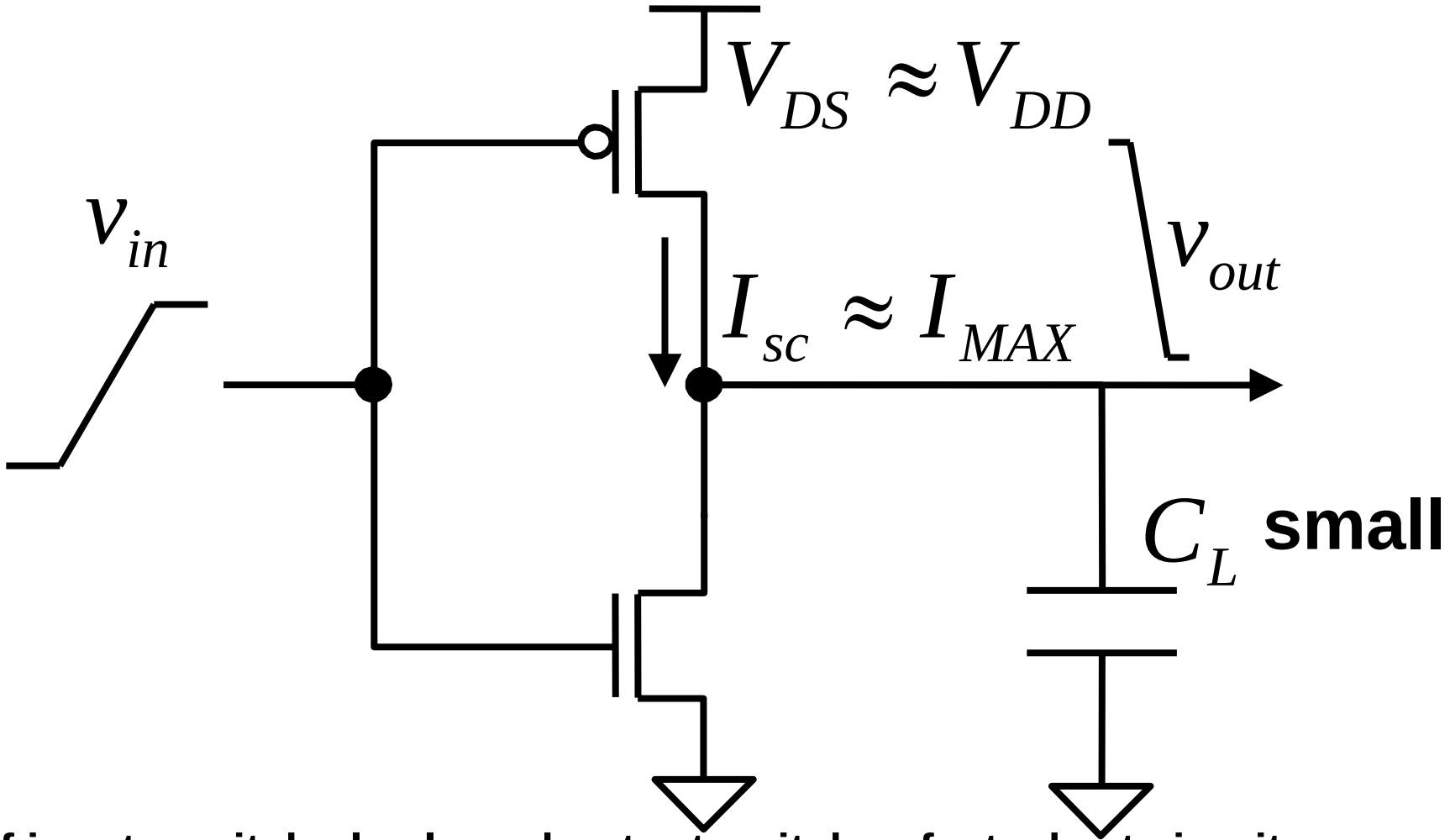


- **If inputs switch fast and output switches slowly, very little short circuit current results**
  - Translates to slower propagation delays which might not be tolerable

Short Circuit Power Dissipation



Short Circuit Current With Small Load



- **If inputs switch slowly and output switches fast, short circuit current maximized since $V_{DS}=V_{DD}$ for most of input transition**

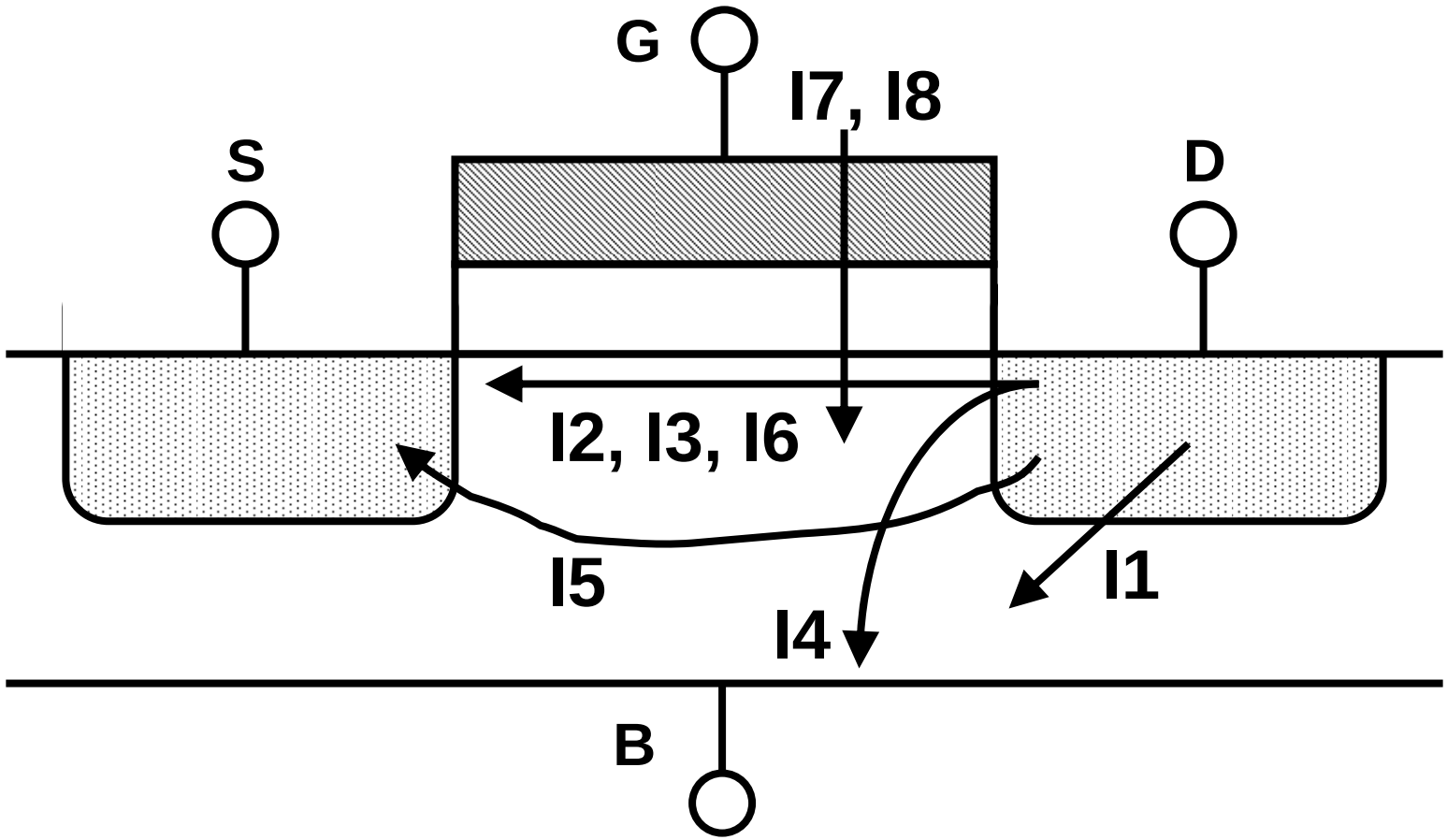
Minimizing Short Circuit Power

- **Peak current determined by MOSFET saturation current, so directly proportional to device sizes**
- **Peak current also strong function of ratio between input and output slopes as shown in previous 2 slides**
- **For individual gate, minimize short circuit current by making output rise/fall time much bigger than input rise/fall time**
  - Slows down circuit
  - Increases short circuit current in fanout gates
- **Compromise: match input and output rise/fall times**

Some Final Words on Short Circuit Power

- **When input and output rise/fall times are equalized, most power is associated with dynamic power**
  - <10% devoted to short circuit currents
- **Can eliminate short circuit dissipation entirely by very aggressive voltage scaling**
  - Need $V_{DD} < V_{Tn} + |V_{Tp}|$
  - Both devices can't be on simultaneously
- **Short circuit power becoming less important in deep submicron**
  - Threshold voltages not scaling as fast as supply voltages

Leakage Currents in Deep Submicron

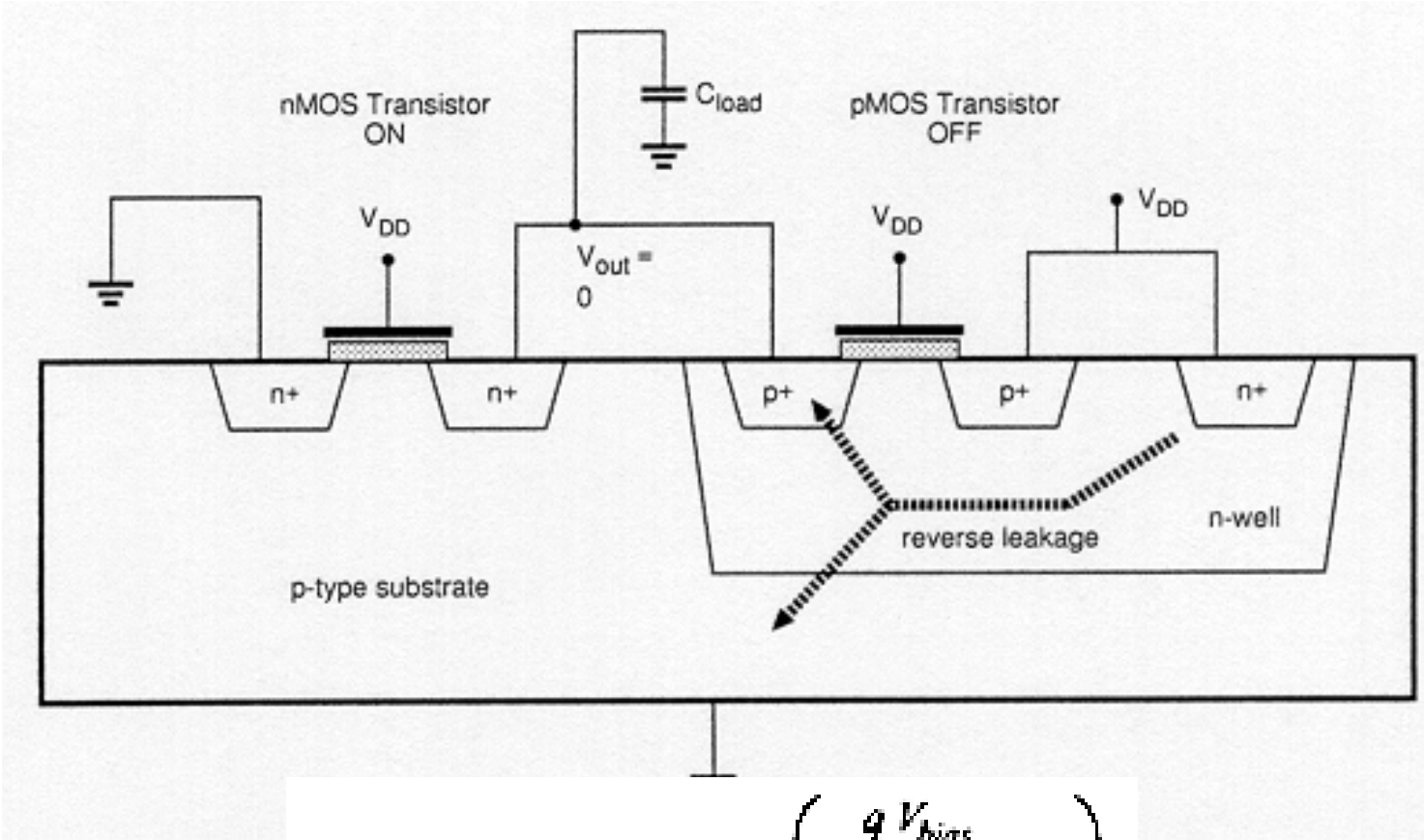


- **Many physical mechanisms produce static currents in deep submicron**

Transistor Leakage Mechanisms

1. pn Reverse Bias Current (I1)
2. Subthreshold (Weak Inversion) (I2)
3. Drain Induced Barrier Lowering (I3)
4. Gate Induced Drain Leakage (I4)
5. Punchthrough (I5)
6. Narrow Width Effect (I6)
7. Gate Oxide Tunneling (I7)
8. Hot Carrier Injection (I8)

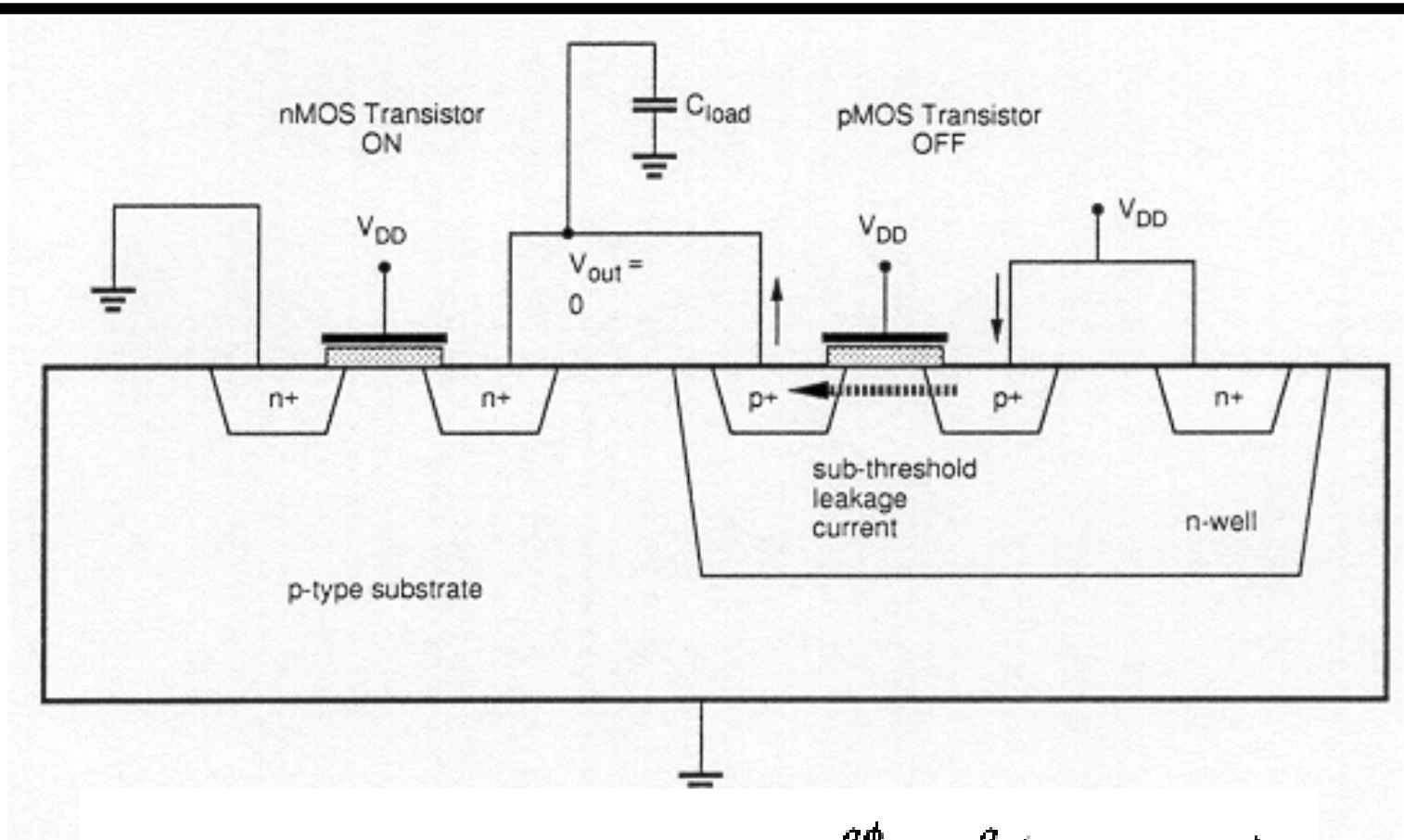
Reverse Diode Leakage Current



$$I_{reverse} = A \cdot J_S \left(e^{\frac{q V_{bias}}{kT}} - 1 \right)$$

Reverse leakage current paths in a CMOS inverter

Subthreshold Leakage Current



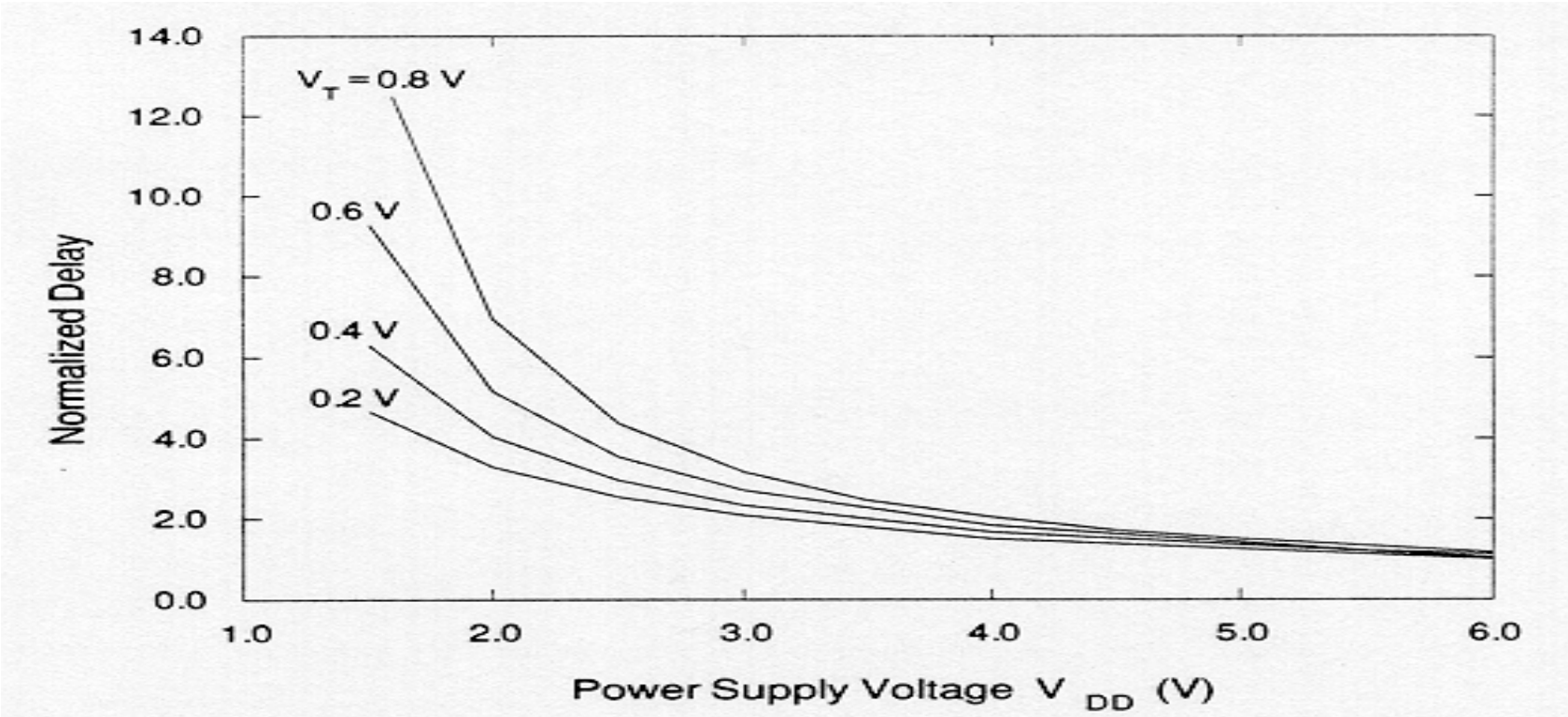
$$I_D(subthreshold) \cong \frac{qD_n W x_c n_0}{L_B} \cdot e^{\frac{q\phi_r}{kT}} \cdot e^{\frac{q}{kT}(A \cdot V_{GS} + B V_{DS})}$$

Subthreshold leakage current path in CMOS inverter

Leakage Power

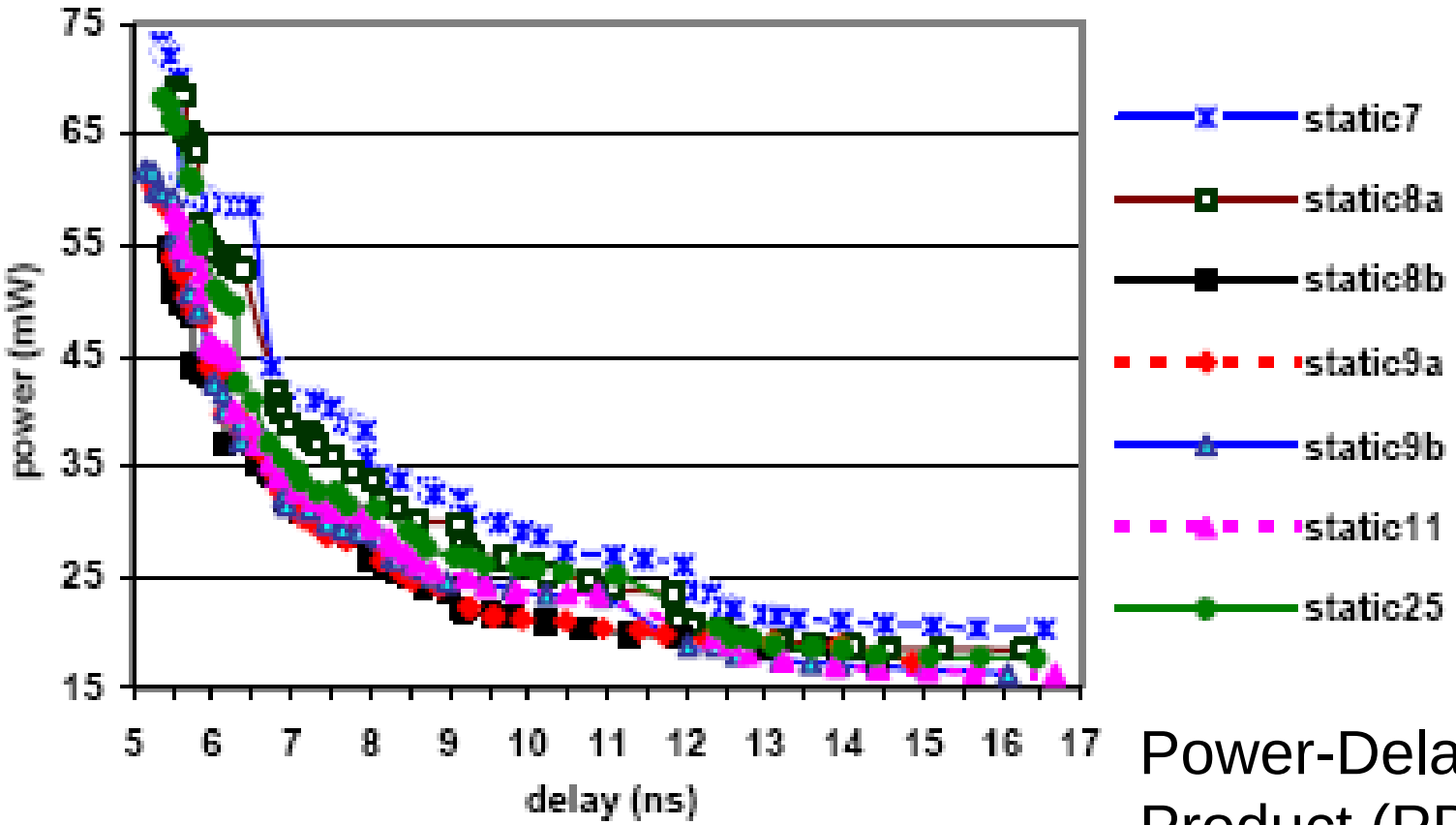
- **Reverse bias diode leakage current**
  - Diode between well and substrate reverse biased
  - Reverse saturation current Is drains power from Vdd
- **Sub-threshold leakage current**
  - Due to channel being in weak inversion instead of being completely off
  - Noise on ground line can contribute to sub-threshold leakage (negative noise voltage yields positive V_{GS})
  - Avoid low Vt transistors to minimize leakage (limit to <10% of total transistor count)
  - Will dominate total power consumption if scaling trend continues

Reducing Power by Voltage Scaling



- **Plot of Normalized delay vs Power supply for different Vt**
  - Increasing power supply voltage decreases delay
  - Decreasing Vt for a given Vdd also decreased delay(up to a point)
    - Note it is important to linearly scale Vt with Vdd when process scaling to meet delay specs, but subthreshold leakage increases as we scale
  - Use Multiple Threshold transistor solution in your design (if allowed)

Figure of Merit: Power Delay Product



W/L decreasing →

Power-Delay Product (PDP) tries to balance power and delay tradeoff

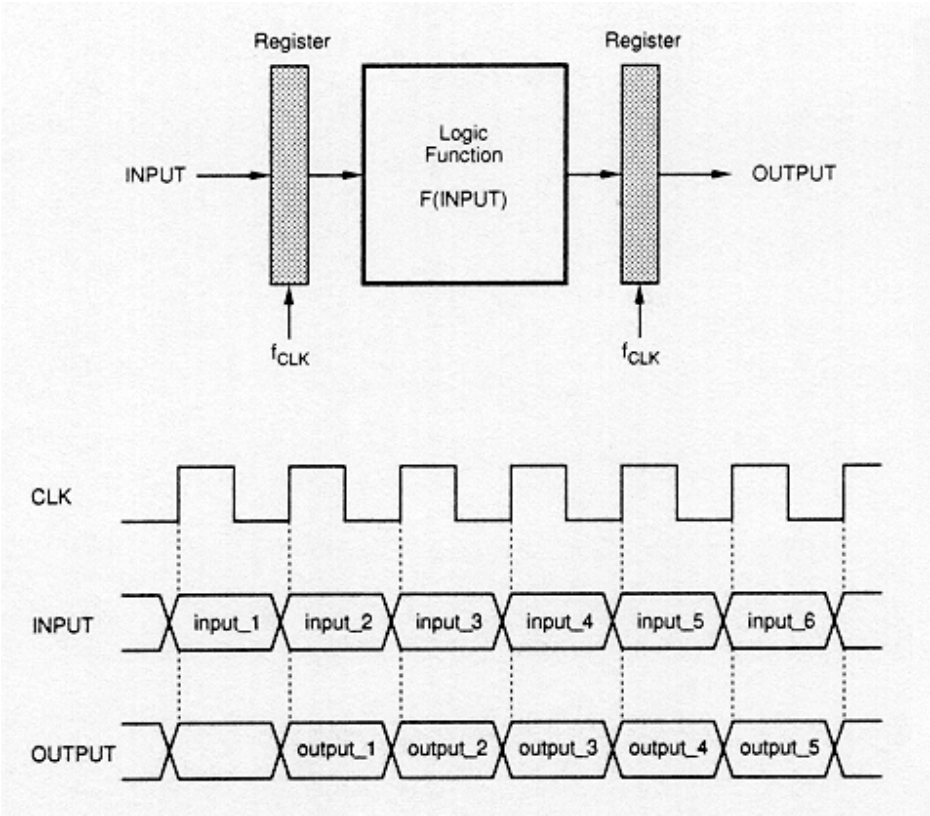
Power Delay Product Optimum

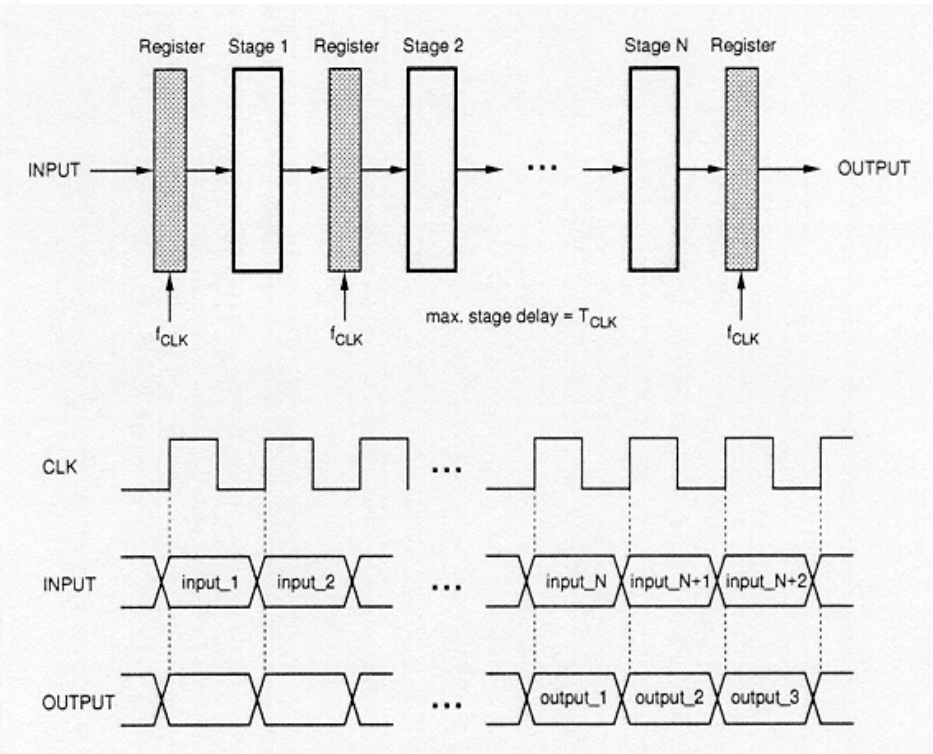
- **Just like Vt scaling vs. power supply there is diminishing returns for sizing**
  - Preceding curve shows delay vs. power
    - Obtained by modifying the size of the gate to analyze delay and power
    - By decreasing W/L, delay goes up but power goes down
      - After a while, decreasing W/L increases delay tremendously without lowering power
    - By increasing W/L, delay goes down but power goes up
      - After a while, increasing W/L costs you tremendously in power without lowering delay
    - Optimal point where slope of curve is -1

Pipeline Approach to Voltage Scaling

- **Start with a single design with two registers**
  - Consider the logic in between allows freq = fmax
- **Now break the logic into N separate parts with equal delay**
  - Separate each part by a register
  - Logic will be several times faster (New fmax = N x Old fmax)
    - Vdd can be lowered in order slow down logic to fit original fmax freq
  - However, additional capacitance of each register has been added.
- **Power savings could be as much as 80% once all things are considered**

Pipeline Approach





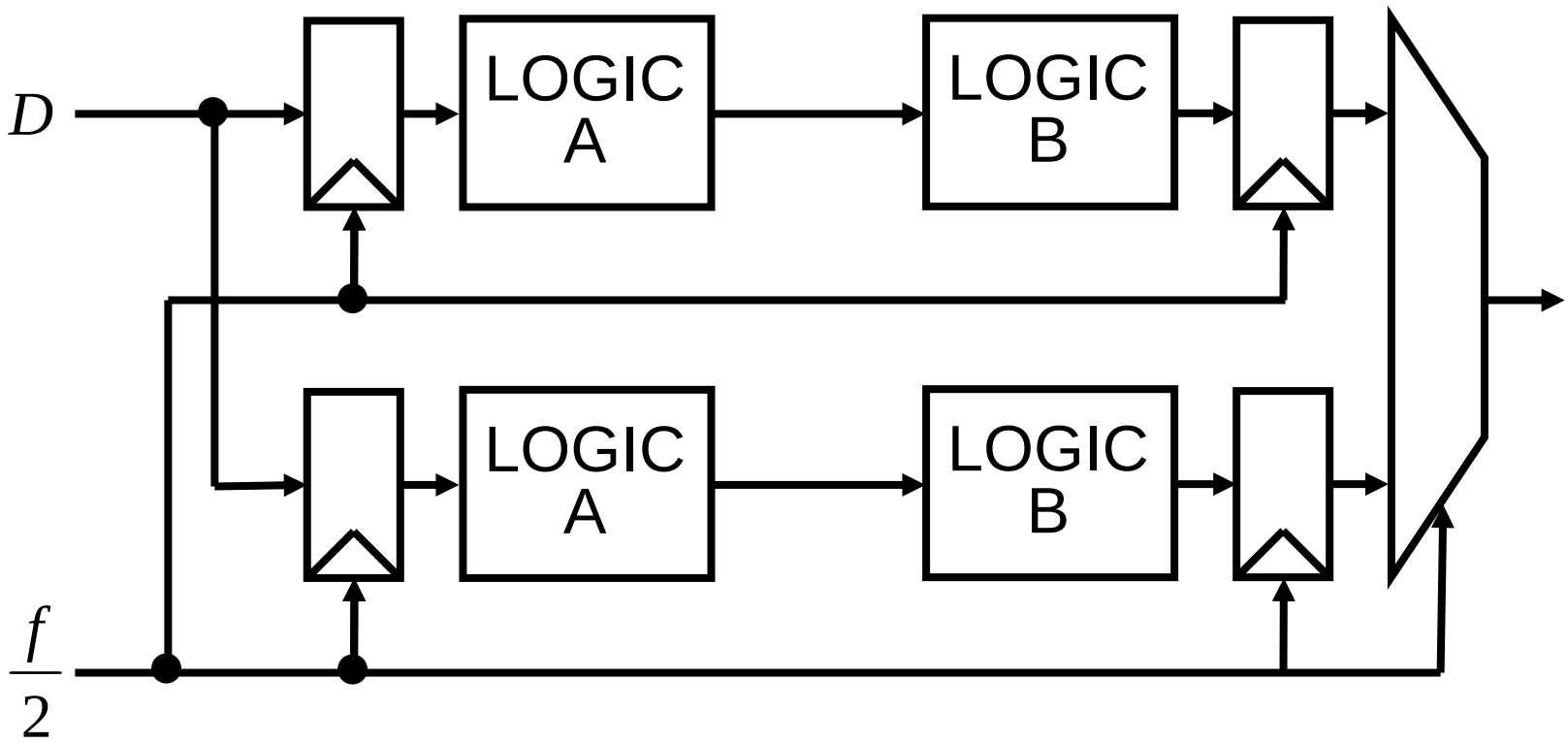
Single Register **Multiple Registers**

- **Tradeoff power for a little more area and more latency by reducing voltage to meet fixed throughput**

Hardware Replication (Parallelism)

- **Create N redundant paths for data/logic**
- **Input data sent to all path inputs**
  - Outputs from the multiple paths arrive at same time
- **Have clock to each input register at F_{clk}/N**
- **Use mux to select from all outputs**
- **To reduce power**
  - Reduce power supply voltage for each path
    - You can afford the slower speed since replication speeds up total circuit performance
  - Gate clocks (turn them off) for unused paths

Parallelization Driven Voltage Scaling



- **Parallelize computation up to N times**
- **Reduce clock frequency by factor N**
- **Reduce voltage to meet relaxed frequency constraint**

Tradeoffs of Parallelization

- **Amount of parallelism in application may be limited**
- **Extra capacitance overhead of multiple datapaths**
  - N times higher input loading
  - N-to-1 selector on output
  - Lower clock frequency somewhat offset by higher clock load
- **Consumes more area, devices, more leakage power especially in deep submicron**
- **Voltage reduction typically results in dramatic power gains**
  - ~3X power reduction

Summary

- **Various causes of power dissipation**
  - Switching, short circuit, leakage current
- **Reducing power dissipation**
  - Voltage scaling – Decreases dynamic power quadratically, other power linearly
  - Technology Scaling – Reduces capacitances
  - Transistor Sizing – Make sure you are on the correct part of the power delay tradeoff curve
  - Pipeline approach
  - Hardware replication (parallelism) approach

Next Topic: Wires

- **On-chip resistance and capacitance**
  - Delay estimation
  - Buffering and repeater insertion