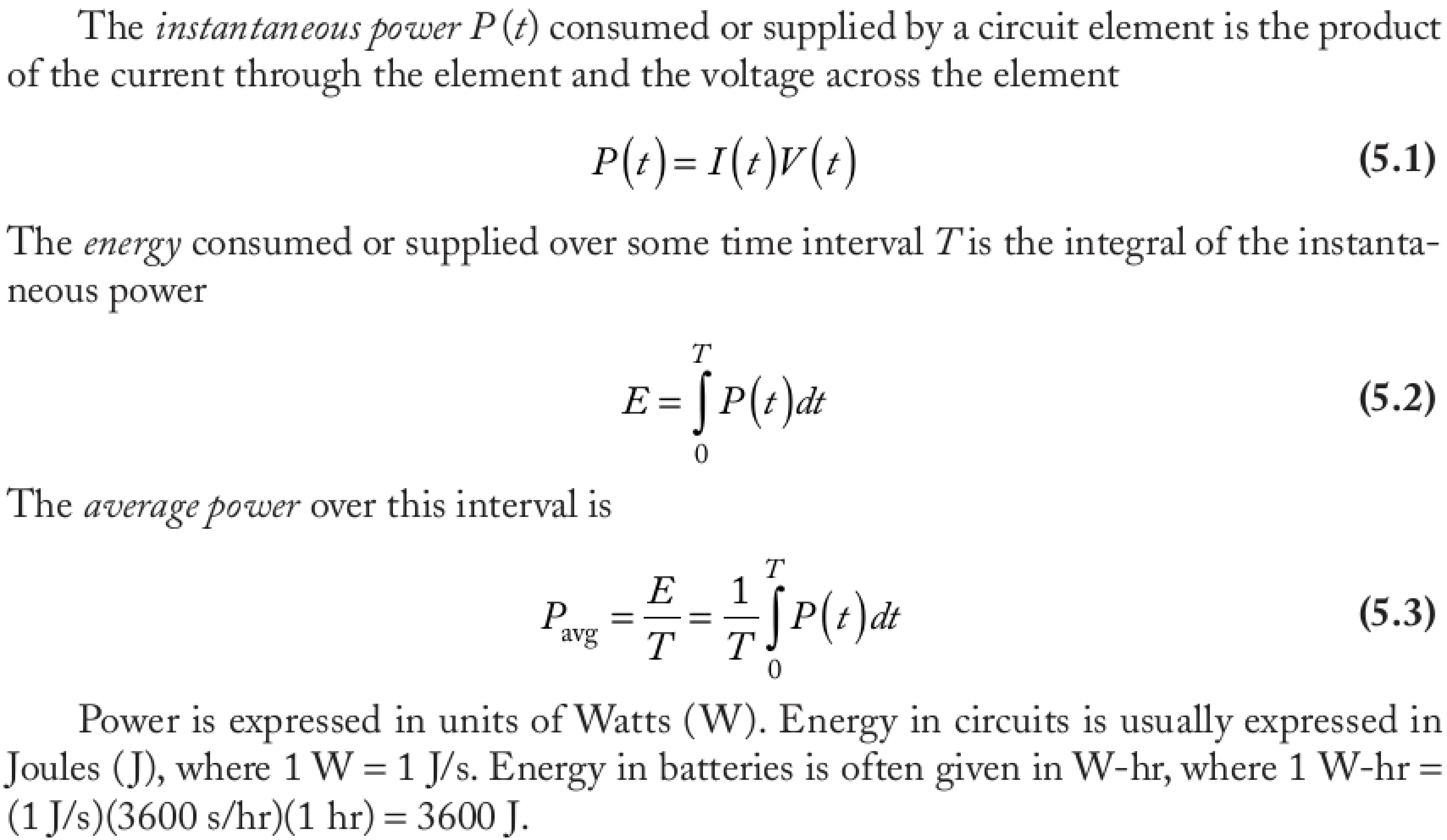
Design and Implementation of VLSI Systems
Lecture06

PERFORMANCE ESTIMATION

Power in circuit element



The *instantaneous power* $P(t)$ consumed or supplied by a circuit element is the product of the current through the element and the voltage across the element

$$P(t) = I(t)V(t) \tag{5.1}$$

The *energy* consumed or supplied over some time interval T is the integral of the instantaneous power

$$E = \int_0^T P(t)\,dt \tag{5.2}$$

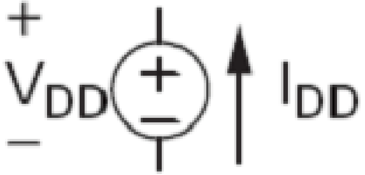
The *average power* over this interval is

$$P_{\text{avg}} = \frac{E}{T} = \frac{1}{T}\int_0^T P(t)\,dt \tag{5.3}$$

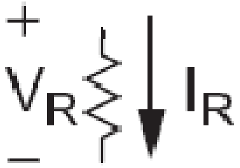
Power is expressed in units of Watts (W). Energy in circuits is usually expressed in Joules (J), where 1 W = 1 J/s. Energy in batteries is often given in W-hr, where 1 W-hr = (1 J/s)(3600 s/hr)(1 hr) = 3600 J.

Power in circuit element

$$P_{VDD}(t) = I_{DD}(t)V_{DD}$$

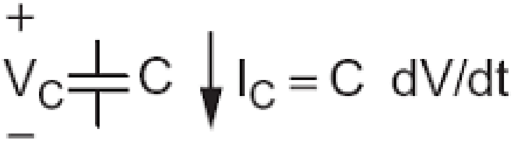


$$P_R(t) = \frac{V_R^2(t)}{R} = I_R^2(t)R$$



$$E_C = \int_0^\infty I(t)V(t)dt = \int_0^\infty C\frac{dV}{dt}V(t)dt$$

$$= C\int_0^{V_C} V(t)dV = \tfrac{1}{2}CV_C^2$$



Source of power dissipation

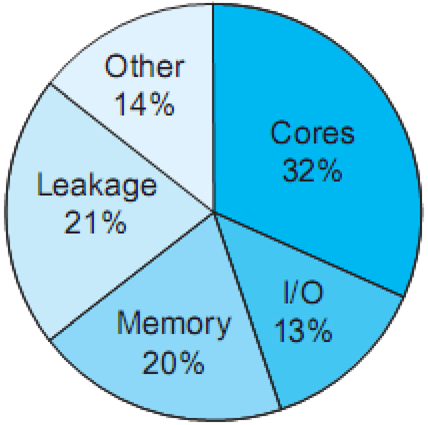
- Power dissipation in CMOS circuits comes from two components:
- Dynamic dissipation due to
  - charging and discharging load capacitances as gates switch
  - “short-circuit” current while both pMOS and nMOS stacks are partially ON
- Static dissipation due to
  - subthreshold leakage through OFF transistors
  - gate leakage through gate dielectric
  - junction leakage from source/drain diffusions
  - contention current in ratioed circuits

Source of power dissipation

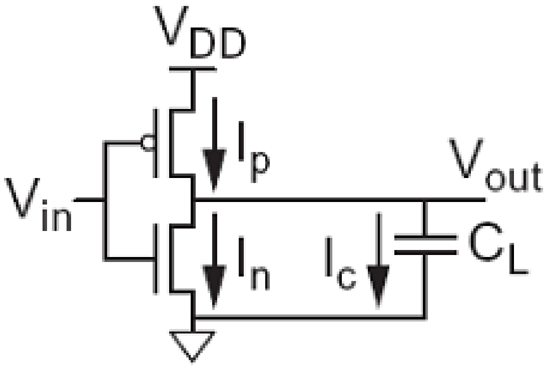
$$P_{\text{dynamic}} = P_{\text{switching}} + P_{\text{short circuit}}$$

$$P_{\text{static}} = \left(I_{\text{sub}} + I_{\text{gate}} + I_{\text{junct}} + I_{\text{contention}}\right)V_{DD}$$

$$P_{\text{total}} = P_{\text{dynamic}} + P_{\text{static}}$$



Charging a capacitor



- When the gate output rises
  - Energy stored in capacitor is

$$E_C = \tfrac{1}{2} C_L V_{DD}^2$$

  - But energy drawn from the supply is

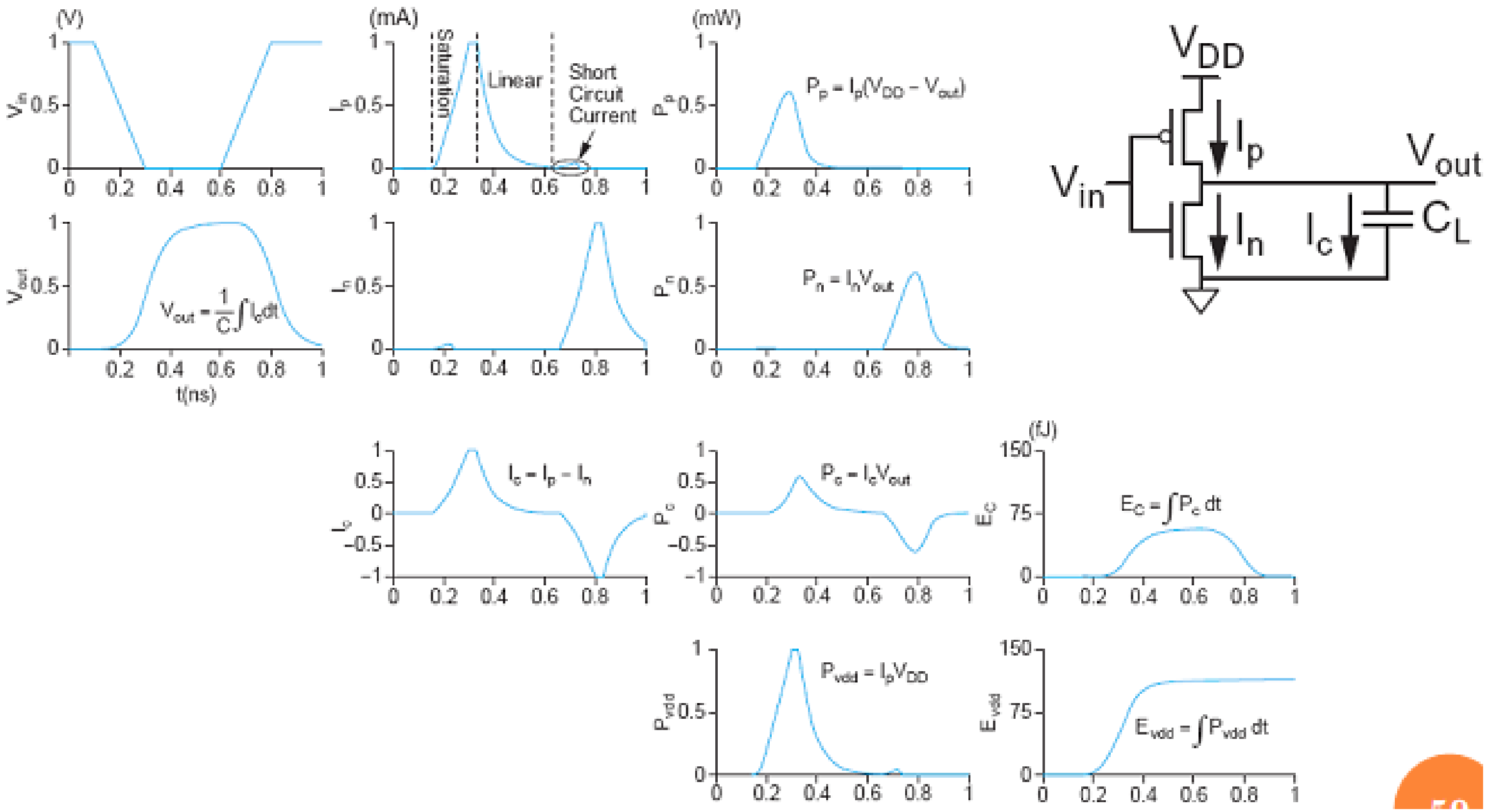
$$E_{VDD} = \int_0^\infty I(t) V_{DD} dt = \int_0^\infty C_L \frac{dV}{dt} V_{DD} dt$$

$$= C_L V_{DD} \int_0^{V_{DD}} dV = C_L V_{DD}^2$$

  - Half the energy from V_{DD} is dissipated in the pMOS transistor as heat, other half stored in capacitor
- When the gate output falls
  - Energy in capacitor is dumped to GND
  - Dissipated as heat in the nMOS transistor

Switching waveforms

- Example: $V_{DD} = 1.0$ V, $C_L = 150$ fF, $f = 1$ GHz



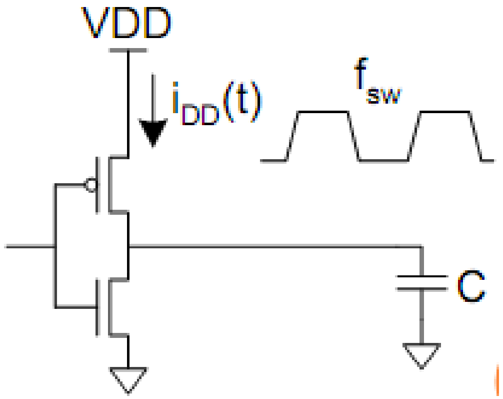
Switching power

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

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

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

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



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 power:

$$P_{\text{switching}} = \alpha C V_{DD}^{2} f$$

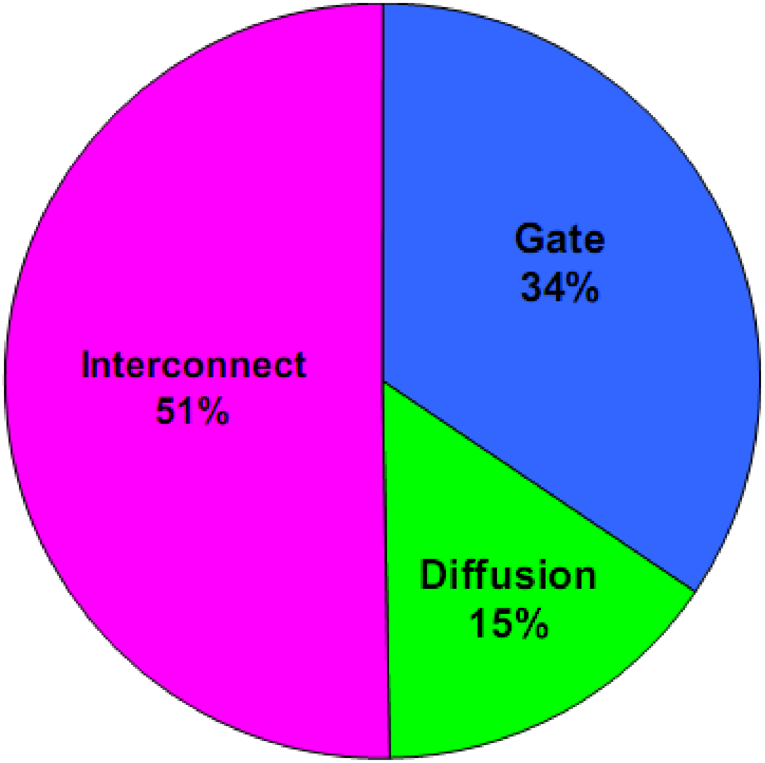
Short circuit currents

- When transistors switch, both nMOS and pMOS networks may be momentarily ON at once
- Leads to a blip of “short circuit” current.
- < 10% of dynamic power if rise/fall times are comparable for input and output
- We will generally ignore this component

Power dissipation sources

- $P_{total} = P_{dynamic} + P_{static}$
- Dynamic power: $P_{dynamic} = P_{switching} + P_{shortcircuit}$
  - Switching load capacitances
  - Short-circuit current
- Static power: $P_{static} = (I_{sub} + I_{gate} + I_{junct} + I_{contention})V_{DD}$
  - Subthreshold leakage
  - Gate leakage
  - Junction leakage
  - Contention current

Dynamic power breakup



Total dynamic Power

[source: Intel'03]

Dynamic power example

Example:
A digital system-on-chip in a 1 V 65 nm process (with 50 nm drawn channel lengths and Q= 25 nm) has 1 billion transistors, of which 50 million are in logic gates and the remainder in memory arrays. The average logic transistor width is 12 Q and the average memory transistor width is 4 Q. The memory arrays are divided into banks and only the necessary bank is activated so the memory activity factor is 0.02. The static CMOS logic gates have an average activity factor of 0.1. Assume each transistor contributes 1 fF/µm of gate capacitance and 0.8 fF/µm of diffusion capacitance. Neglect wire capacitance for now (though it could account for a large fraction of total power). Estimate the switching power when operating at 1 GHz.

Dynamic power reduction

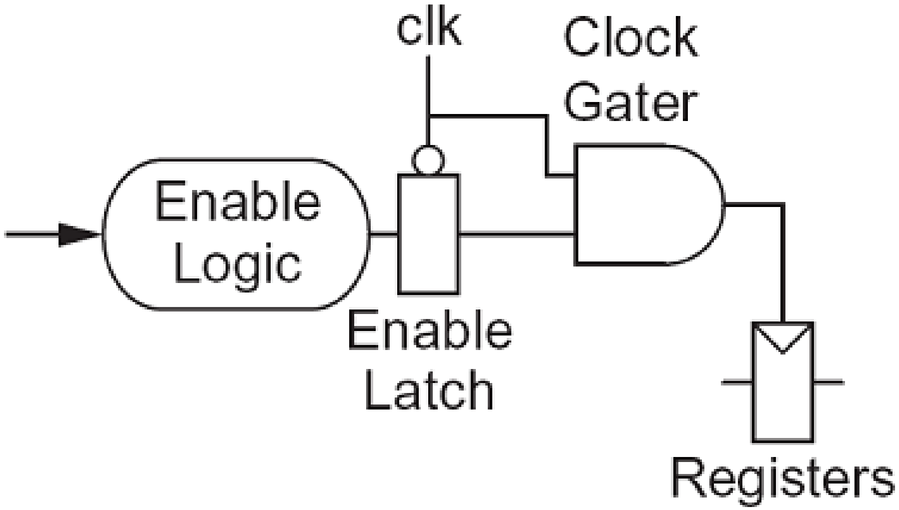
- $P_{\text{switching}} = \alpha C V_{DD}{}^{2} f$
- Try to minimize:
  - Activity factor
  - Capacitance
  - Supply voltage
  - Frequency

Activity factor

- The activity factor is a powerful and easy-to-use
- Turn off a circuit => activity factor and dynamic power go to zero.
- Blocks turned off by stopping the clock => clock gating.
- When a block is on, the activity factor is 1 for clocks
- The activity factor of a logic gate can be estimated by calculating the switching probability.

Clock gating

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



Activity factor estimation

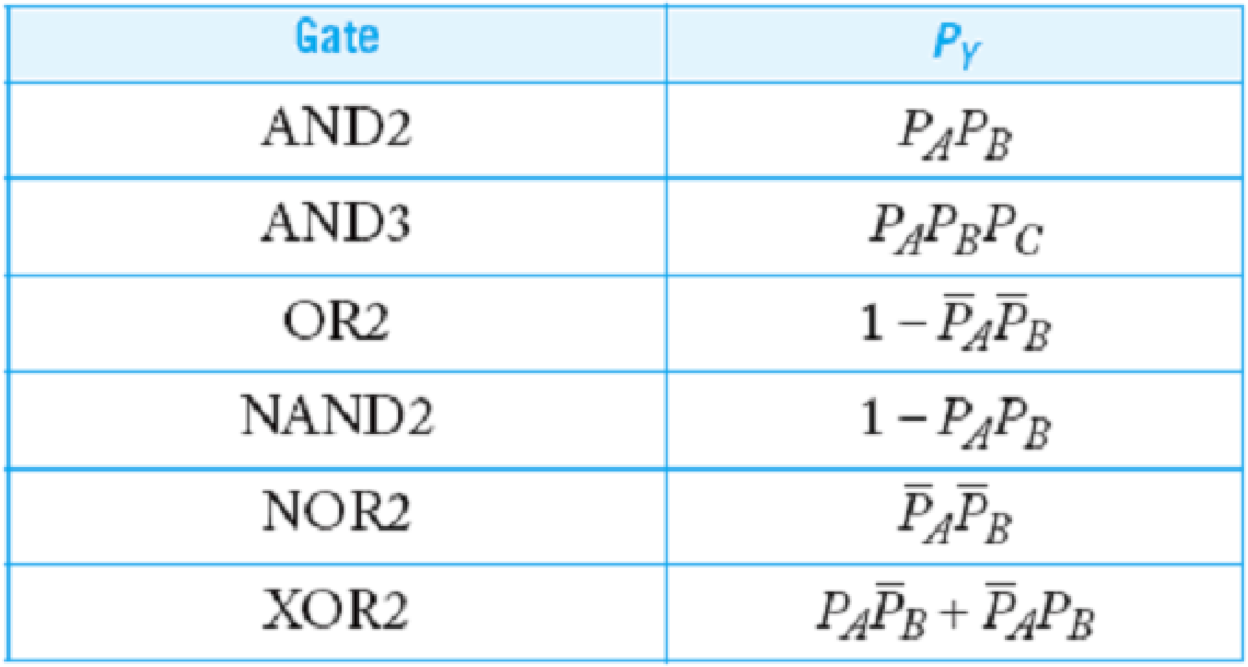
Switching Probability:

- Activity factor of a node is the probability that it switches from 0 to 1.
- This probability depends on the logic function.
- Analyze probability that each node is 1 => can estimate the activity factors

Activity factor estimation

- Let P_i = Prob(node i = 1)
  - $\bar{P}_i = 1 - P_i$
- $\alpha_i = \bar{P}_i * P_i$
- 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 $\alpha \approx 0.1$

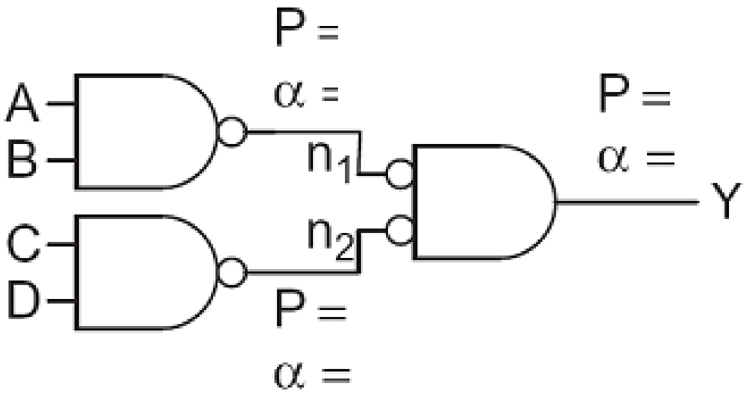
Switching probability



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

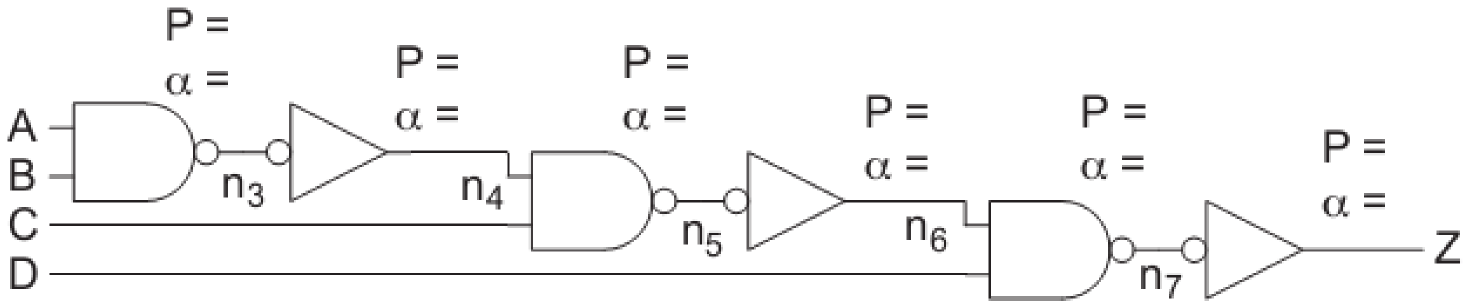
Example

- 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



Example

Figure shows a 4-input AND gate built using a chain of gates. Determine the activity factors at each node in the circuit assuming the input probabilities $P_A=P_B=P_C=P_D=0.5$.



Capacitance

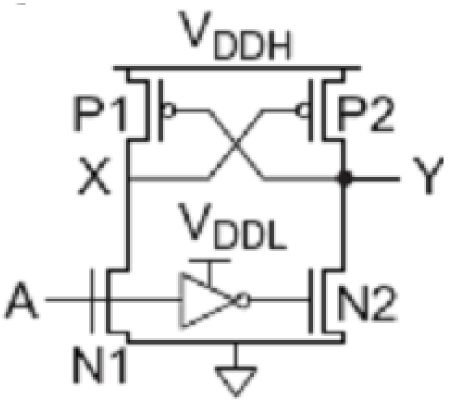
- Gate capacitance
  - Fewer stages of logic
  - Small gate sizes
- Wire capacitance
  - Good floorplanning to keep communicating blocks close to each other
  - Drive long wires with inverters or buffers rather than complex gates

Voltage/Frequency

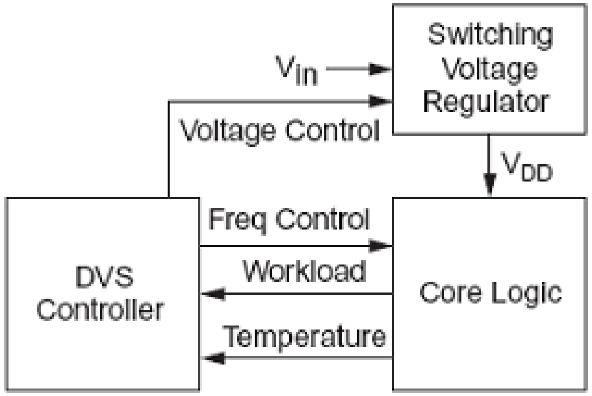
- Voltage has a quadratic effect on dynamic power.
- Many transistors operate in velocity-saturated => lower power supply may not reduce performance as much as long-channel models predict.
- The chip may be divided into multiple voltage domains, where each domain is optimized for the needs of certain circuits.
- For example, a system-on-chip might use a high supply voltage for memories to ensure cell stability, a medium voltage for a processor, and a low voltage for I/O peripherals running at lower speeds.
- Voltage can be adjusted based on operating mode; for example, a laptop processor may operate at high voltage and high speed when plugged into an AC adapter, but at lower voltage and speed when on battery power.
- If the frequency and voltage scale down in proportion, a cubic reduction in power is achieved. For example, the laptop processor may scale back to 2/3 frequency and voltage to save 70% in power when unplugged.

Voltage/Frequency

- Run each block at the lowest possible voltage and frequency that meets performance requirements
- Voltage Domains
  - Provide separate supplies to different bl
  - Level converters required when crossing from low to high V_{DD} domains



- Dynamic Voltage Scaling
  - Adjust V_{DD} and f according to workload



V_{in}

Switching Voltage Regulator

Voltage Control

V_{DD}

DVS Controller

Freq Control

Workload

Temperature

Core Logic

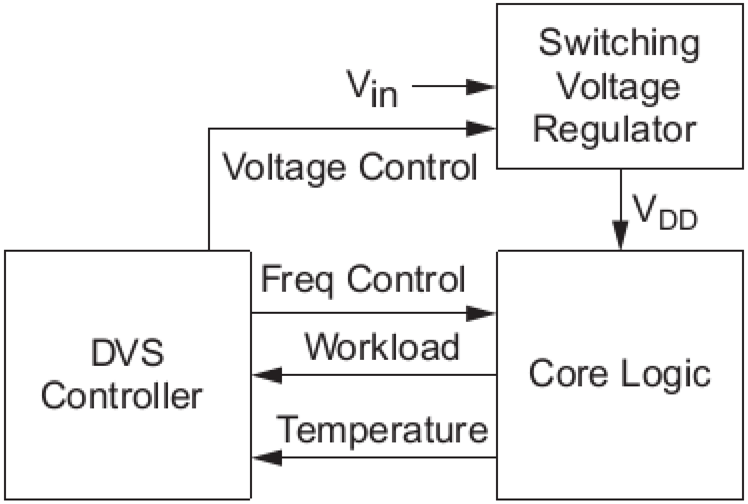
Dynamic voltage scaling (DVS)

- Many systems have time-varying performance requirements.
- For example: a video decoder requires more computation for rapidly moving scenes than for static scenes.
- Can save large amounts of energy by reducing the clock frequency to the minimum sufficient to complete the task on schedule, then reducing the supply voltage to the minimum necessary to operate at that frequency.

=> dynamic voltage scaling (DVS) or dynamic voltage/frequency scaling(DVFS)

Dynamic voltage scaling (DVS)

- Figure shows a block diagram for a basic DVS system.
- DVS controller takes information from system about workload and die temperature.
- Determines supply voltage and clock frequency sufficient to complete the workload on schedule or to maximize performance



- A switching voltage regulator efficiently steps down Vin from a high value to the necessary V_{DD}.
- The core logic contains a phase-locked loop or other clock synthesizer to generate the specified clock frequency.
- The DVS controller determines the operating frequency, then chooses the lowest supply voltage suitable for that frequency.
- One method of choosing voltage is with a precharacterized table of voltage vs. frequency.

Dynamic voltage scaling (DVS)

- Define: rate - fraction of maximum performance required to complete the workload in a specified amount of time.
- Figure plots energy against rate.
- If the rate is less than 1, the clock frequency can be adjusted down, or the system can run at full frequency until the work is done, then stop the clock and go to sleep

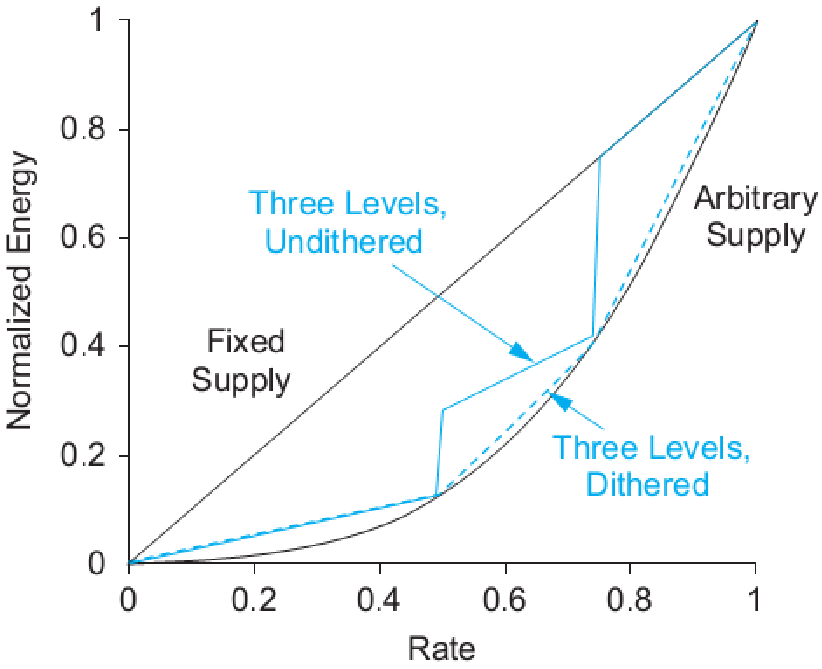


FIGURE 5.18 Energy reduction from DVS

- Without DVS, the energy varies linearly with the rate.
- With ideal DVS, the voltage could also be reduced at lower rates.
- Assuming a linear relationship between voltage and frequency, the energy is proportional to the rate cubed, giving much greater savings at lower rates. Operating at half the maximum rate costs only one-eighth of the energy

Frequency

- Dynamic power is directly proportional to frequency, so a chip obviously should not run faster than necessary.
- Reducing the frequency allows downsizing transistors or using a lower supply voltage, which has an even greater impact on power.
- The performance can be recouped through parallelism, especially if area is not as important as power.
- Even if multiple voltage supplies are not available, a chip may still use multiple frequency domains so that certain portions can run more slowly than others.
- For example, a microprocessor bus interface usually runs much slower than the core. Low frequency domains can also save energy by using smaller transistors.

Short circuit current

- Short-circuit power dissipation occurs as both pullup and pulldown networks are partially ON while the input switches
- It increases as the input edge rates become slower
- It decreases as load capacitance increases
- Short-circuit current is a small fraction (<10%) of current to the load
- Gates with balanced input and output edge rates have low short-circuit power.
- Short-circuit power is strongly sensitive to the ratio $v=V_t/V_{DD}$.
- If v>0.5, short-circuit current is eliminated entirely because the pullup and pulldown net-works are never simultaneously ON.
- For v=0.3 or 0.2, short-circuit power is typically about 2% or 10% of switching power
- In nanometer processes, V_t can scarcely fall below 0.3 V without excessive leakage, and V_{DD} is on the order of 1 V, so short-circuit current has become almost negligible.

Static power

- Static power is consumed even when a chip is not switching.
- Prior to the 90 nm node, leakage power was of concern primarily during sleep mode because it was negligible compared to dynamic power.
- In nanometer processes with low threshold voltages and thin gate oxides, leakage can account for as much as a third of total active power.
- This section briefly reviews each source of static power.
- It then discusses power gating, which is a key technique to reduce power in sleep mode.
- Subthreshold leakage is usually the dominant source of static power
- Techniques for leakage reduction: multiple threshold voltages, variable threshold voltages, and stack forcing.

Static power

Subthreshold Leakage:

- Subthreshold leakage current flows when a transistor is supposed to be OFF.
- For V_{ds} exceeding a few multiples of the thermal voltage (V_{ds}>50 mV), it can be simplified to

$$I_{\text{sub}} = I_{\text{off}} 10^{\frac{V_{gs}+\eta\left(V_{ds}-V_{DD}\right)-k_{\gamma}V_{sb}}{S}}$$

- where I_{off} is the subthreshold current at V_{gs}=0 and V_{ds}=V_{DD}, and S is the subthreshold slope.
- I_{Off}: about 100 nA/µm for typical low-Vt devices to below 1 nA/µm for high-Vt devices.
- η is the DIBL coefficient, around 100 mV/V for a 65 nm transistor
- k_{γ} is the body effect coefficient, which describes how the body effect modulates the threshold voltage.
- Raising the source voltage or applying a negative body voltage can further decrease leakage.

Static power

Subthreshold Leakage:

- The leakage through two or more series transistors is dramatically reduced on account of the stack effect
- Figure: two series OFF transistors with gates at 0 volts
- The drain of N2 is at V_{DD}, so the stack will leak.
- Middle node voltage Vx settles to a point that each transistor has the same current.
- If Vx is small, N1 will see a much smaller DIBL effect and will leak less.
- As Vx rises, Vgs for N2 is negative => reduce leakage.

=> the series transistors leak less.

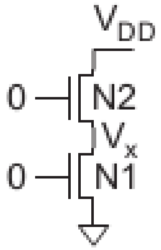


FIGURE 5.20 Series OFF transistors demonstrating the stack effect

Static power

Subthreshold Leakage:

- This can be demonstrated mathematically by solving for Vx and Isub, assuming that Vx>50 mV.

$$I_{sub} = \underbrace{I_{off} 10^{\frac{\eta(V_x - V_{DD})}{S}}}_{N2} = \underbrace{I_{off} 10^{\frac{-V_x + \eta((V_{DD} - V_x) - V_{DD}) - k_\gamma V_x}{S}}}_{N1}$$

$$V_x = \frac{\eta V_{DD}}{1 + 2\eta + k_\gamma}$$

$$I_{sub} = I_{off} 10^{\frac{-\eta V_{DD}\left(\frac{1+\eta+k_\gamma}{1+2\eta+k_\gamma}\right)}{S}} \approx I_{off} 10^{\frac{-\eta V_{DD}}{S}}$$

- Use typical values and V_{DD}=1.0 V, stack effect reduces subthreshold leakage by a factor of about 10.
- Stacks with three or more OFF transistors have even lower leakage

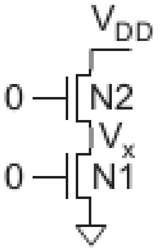
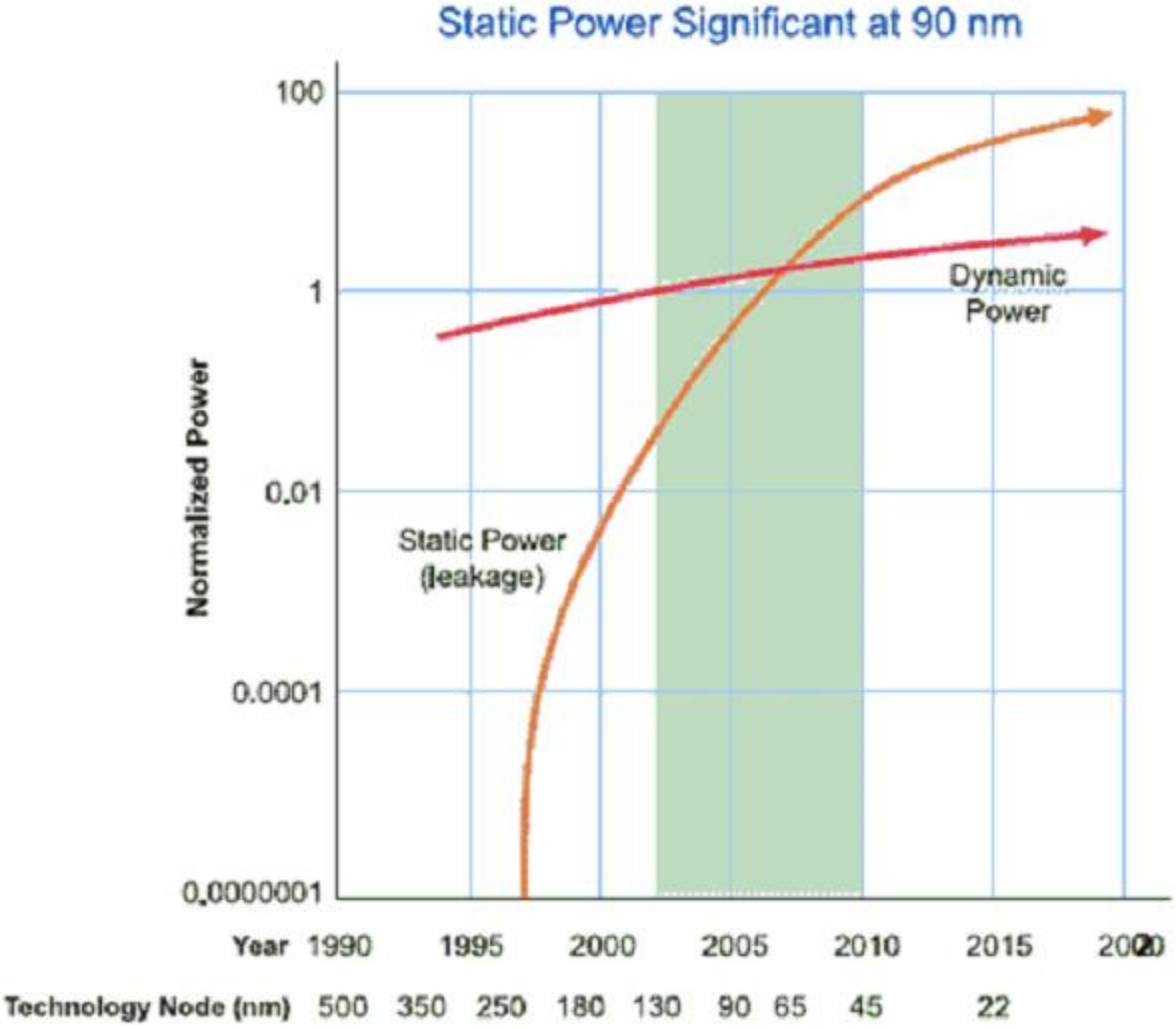


FIGURE 5.20
Series OFF transistors demonstrating the stack effect

Increasing Static Power at Shrinking Process Nodes.



Static power example

- Revisit power estimation for 1 billion transistor chip
- Estimate static power consumption
  - Subthreshold leakage
    - Normal V_t: 100 nA/μm
    - High V_t: 10 nA/μm
    - High V_t used in all memories and in 95% of logic gates
  - Gate leakage 5 nA/μm
  - Junction leakage negligible

Gate leakage

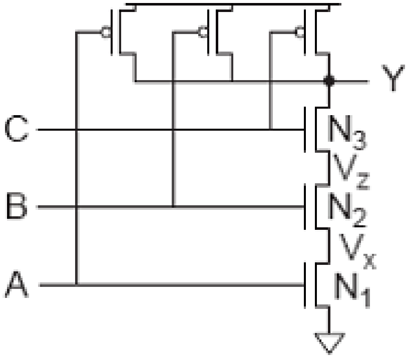
- Extremely strong function of t_{ox} and V_{gs}
  - Negligible for older processes
  - Approaches subthreshold leakage at 65 nm and below in some processes
- An order of magnitude less for pMOS than nMOS
- Control leakage in the process using $t_{ox} > 10.5$ Å
  - High-k gate dielectrics help
  - Some processes provide multiple t_{ox}
    - e.g. thicker oxide for 3.3 V I/O transistors
- Control leakage in circuits by limiting V_{DD}

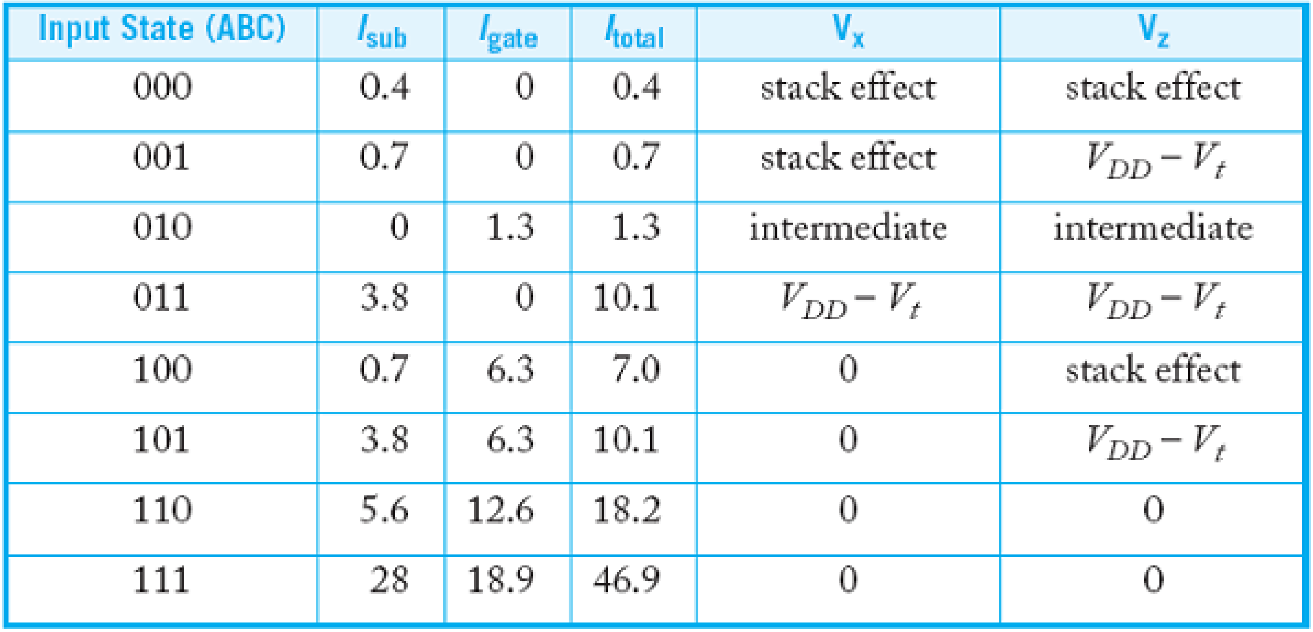
NAND3 leakage example

- 100 nm process

$I_{gn} = 6.3$ nA $\qquad I_{gp} = 0$

$I_{offn} = 5.63$ nA $\qquad I_{offp} = 9.3$ nA





Input State (ABC)	I_{sub}	I_{gate}	I_{total}	V_x	V_z
000	0.4	0	0.4	stack effect	stack effect
001	0.7	0	0.7	stack effect	$V_{DD} - V_t$
010	0	1.3	1.3	intermediate	intermediate
011	3.8	0	10.1	$V_{DD} - V_t$	$V_{DD} - V_t$
100	0.7	6.3	7.0	0	stack effect
101	3.8	6.3	10.1	0	$V_{DD} - V_t$
110	5.6	12.6	18.2	0	0
111	28	18.9	46.9	0	0

Junction leakage

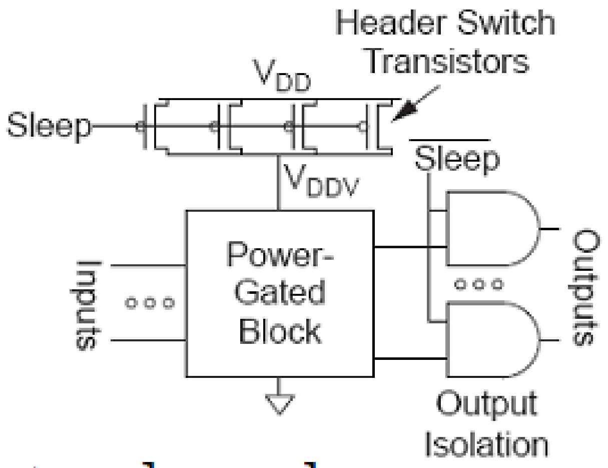
- From reverse-biased p-n junctions
  - Between diffusion and substrate or well
- Ordinary diode leakage is negligible
- Band-to-band tunneling (BTBT) can be significant
  - Especially in high-V_t transistors where other leakage is small
  - Worst at $V_{db} = V_{DD}$
- Gate-induced drain leakage (GIDL) exacerbates
  - Worst for $V_{gd} = -V_{DD}$ (or more negative)

Leakage control

- Leakage and delay trade off
  - Aim for low leakage in sleep and low delay in active mode
- To reduce leakage:
  - Increase V_t: *multiple V_t*
    - Use low V_t only in critical circuits
  - Increase V_s: *stack effect*
    - *Input vector control* in sleep
  - Decrease V_b
    - *Reverse body bias* in sleep
    - Or forward body bias in active mode

Power gating

- Turn OFF power to blocks when they are idle to save leakage
  - Use virtual V_{DD} (V_{DDV})
  - Gate outputs to prevent invalid logic levels to next block



- Voltage drop across sleep transistor degrades performance during normal operation
  - Size the transistor wide enough to minimize impact
- Switching wide sleep transistor costs dynamic power
  - Only justified when circuit sleeps long enough

Power gating

Example:

- A cache in a 65 nm process consumes an average power of 2 W. Estimate how wide should the pMOS header switch be if delay should not increase by more than 5%?

Low power architectures

- VLSI design used to be constrained by the number of transistors that could fit on a chip.
- Extracting maximum speed from each transistor maximized overall performance.
- Many designs have become power constrained and the most energy-efficient design is the highest performer.

=> shift to multicore processors.

Low power architectures

Microarchitecture

- Energy-efficient architectures take advantage of the structured design principles of modularity and locality
- Processor performance grows with the square root of the number of transistors.
- Complex, sprawling processors to extract the last bit of instruction-level parallelism from a problem => inefficient energy.
- Microarchitectures are moving toward larger numbers of simpler cores seeking to handle task and data-level parallelism.
- Smaller cores have shorter wires and faster memory access.
- Memories have a much lower power density than logic because activity factors are small and simplified leakage control.
- If a task can be accelerated using either a faster processor or a larger memory, the memory is often preferable.
- Memories now comprise more than half the area of many chips

Low power architectures

Microarchitecture

- Special-purpose functional units can offer 10 times better energy efficiency than general-purpose processors.
- Accelerators for compute-intensive applications such as graphics, networking offload these tasks from the processor.
- Heterogeneous architectures, combining regular cores, specialized accelerators, and large amounts of memory, are of growing importance.
- Commercial software has historically lagged at least a decade behind hardware advances such as virtual memory, memory protection, 32- and 64-bit datapaths, and robust power-management.

Parallelism and Pipelining

- Effective ways to reduce power consumption
- Replacing a single functional unit with N parallel units allows each to operate at 1/N the frequency.
- A multiplexer selects between the results.
- The voltage can be scaled down, offering quadratic savings in energy at the expense of doubling the area.
- Replacing a single functional unit with an N-stage pipelined unit reduces amount of logic in a clock cycle at expense of more registers.
- Voltage also can be scaled down.
- The two techniques can be combined for even better energy efficiency.

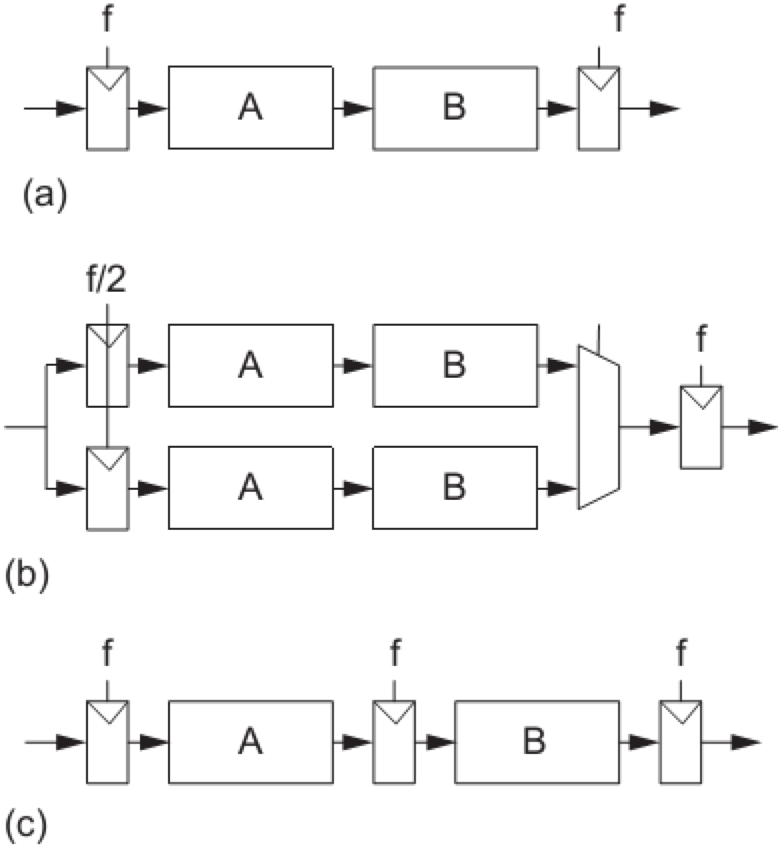


FIGURE 5.29 Functional units: (a) normal, (b) parallel, (c) pipelined

Parallelism and Pipelining

- When leakage is unimportant, parallelism offers a slight edge because the multiplexer has less overhead than the pipeline registers.
- Perfectly balancing logic across pipeline stages can be difficult.
- If leakage is a substantial fraction of total power, pipelining preferable because parallel hardware has N times leakage
- Now V_{DD} is closer to the best energy-delay point, the potential supply reduction and energy savings are diminishing.
- Parallelism and pipelining remain primary tools to extract performance from the vast transistor budgets now available.

Parallelism and Pipelining

- Chip designers must turn off portions of the chip when they are not active by applying clock and power gating.
- Many chips employ a variety of power management modes giving a trade-off between power savings and wake-up time.
- For example, the Intel Atom processor operates at a peak frequency of 2 GHz at 1 V, consuming 2 W.
- The power management modes are shown in Figure 5.30.
- Low frequency mode: clock drops as slow as 600 MHz while the power supply reduces to 0.75 V.
- Sleep mode C1: core clock is turned off and the level 1 cache is flushed and power-gated to reduce leakage, but the processor can return to active state in 1 microsecond.
- Sleep mode C4: PLL is also turned OFF.
- Sleep mode C6, core and caches are power-gated to reduce power to < 80 mW, but wake-up time rises to 100 microseconds.
- For a typical workload, processor can spend 80–90% of its time in C6 sleep mode, reducing average power to 220 mW.

Parallelism and Pipelining

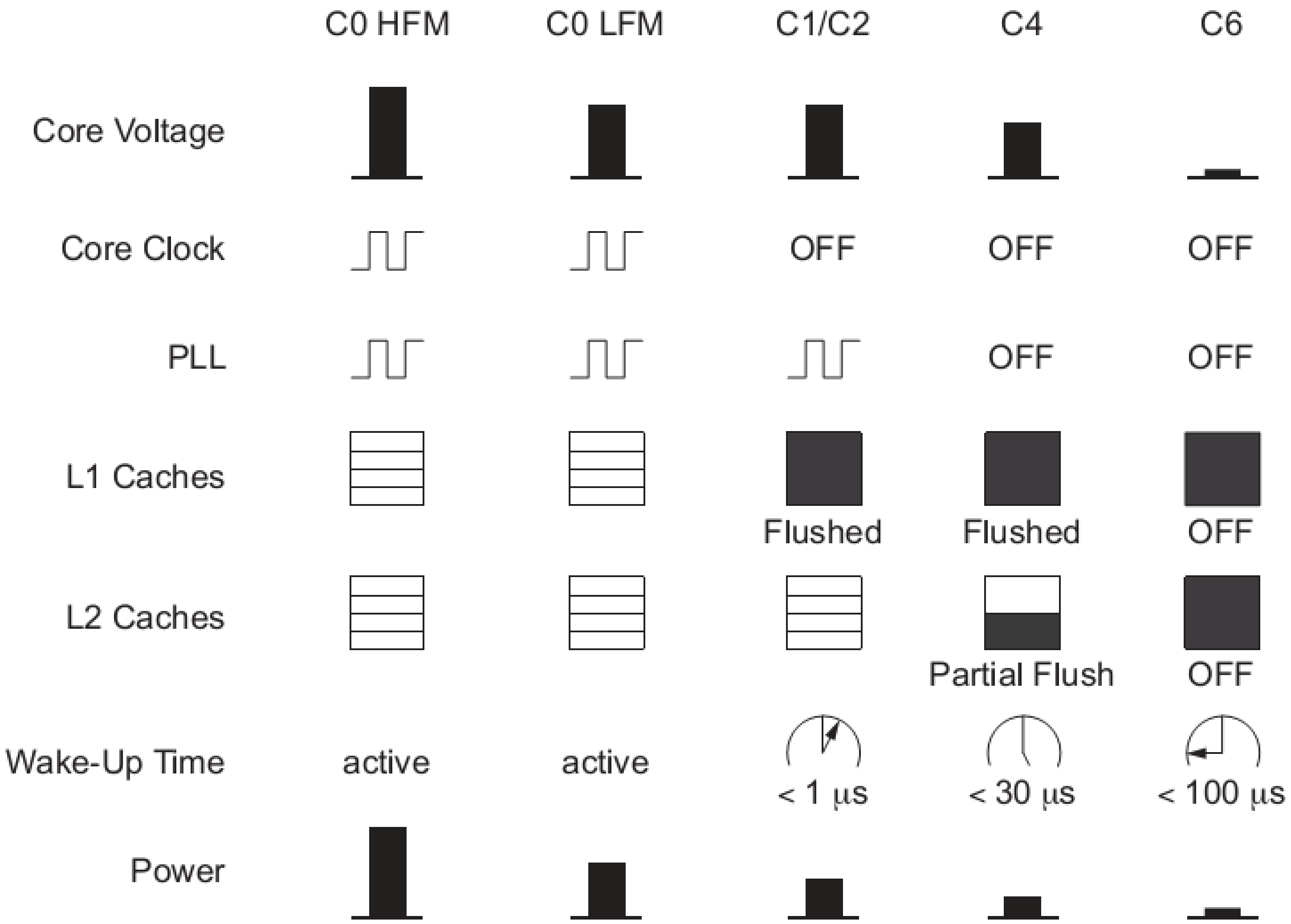


FIGURE 5.30 Atom power management modes (© 2009 IEEE.)

Homework

- Chapter 5: 1, 2, 3, 4, 5, 6, 7, 8, 10