Low Power System Design

ATTENTION: This cheat sheet is only a simple transcript so far, it is partially unstructured and may contain bad mistakes.

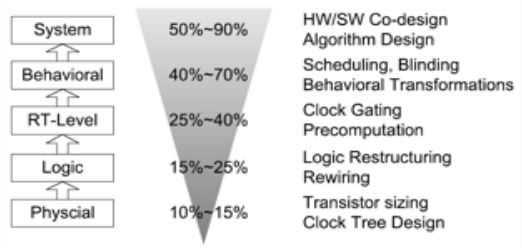
1. Introduction

1.1. Low Power Motivaton

- Longer Battery Life, or smaller battery size
- Decrease of working temperature of the device
  active cooling requires additional power
  thermal desing power limits performance
  device reliability, longer chip life
- Lower operation cost
- reduce size, thiner wires
  higher current → higher cross-talk → more decoupling capacitors
- Power devices from energy harvesters

1.2. System-level energy optimization

RTL or higher level. Requires accurate system model



1.3. CMOS Circuits

High-k metal for gates.
Apply bulk voltage to lower threshold voltage V_{th}

1.3.1 Short Channel Effect

Horizontal and vertical electric fields interact
threshold voltage decreases → higher leakage current
velocity of electrons saturate

1.3.2 MOS Capacitance

Thiner oxide layer → higher oxide capacitance → more charge required to control the gate
higher current or lower switcing frequency (slower charging of the capacitor)
C_{g} has been constant for 25 because t_{ox} scaled with L

1.3.3 Propagation delay

because of parasitic capacitance. time from 50% input voltage to 50% output voltage.

1.4. Power and Energy

Power $P(t) = I_{\text{dd}} \cdot V_{\text{dd}}$
Energy $E(t) = \int_0^T P(t)\,\mathrm{d}t$

1.4.1 Dynamic Power

During logic transition: switching power, short-circuit power, glitch power

1.4.2 Static Power

Regardless of the logic transition
Low V_{dd}: less dynamic power but more leakage power.

1.5. Dynamic Power Consumption

Switching power: Charge and discharge capacitance
$E_{0\to1} = C_L V_{\text{DD}}^2$
one half of the energy is dissipated by heat, the other is stored in the capacitor
$P_{\text{switch}} = f_{\text{s}} C_{\text{L}} V_{\text{DD}}^2$
Short-Circuit Power Consumption:
$P_{\text{short}} = \frac{\mu C_{\text{ox}}}{12} \frac{W}{L} (V_{\text{dd}} - 2V_{\text{th}})^3 \tau f$

1.6. Static Power Consumption

due to scaling, modern transistors are leaking everywhere. subthreshold current: $I_{\text{sub}} = I_s \exp \frac{q(V_{GS} - V_{th} - V_{\text{offset}})}{n k_{\text{B}} T} + (1 - \exp)$ additional: gate tunneling, PN-junction leakage

1.6.1 Drain induced barrier lowering (DIBL)

short channel effect, voltage across drain-to-body PN junction, more electrons leak across the PN junction
Gate leakage: tunneling through the gate to body, drain, and source
junction leakage: band-to-band tunneling from P-side valence band to N-side conduction band
Hot carrier injunction: electrons gain enough thermal energy to jump into the oxide, there they are trapped (aging of transistor)

1.7. Leakage Power Reduction

Lower V_{DD}, voltage scaling, coolinglicon on insulater technology, dual gate design, input vector control (adjust gates, when transistors are off), MTCMOS (transistors to switch parts off)

1.8. Alpha Power Model

Because of the short channel effect: $I_{\text{DS}} = K_{\text{meas}} \frac{W}{L} (V_{\text{GS}} - V_{\text{th}})^{\alpha}$
Originally $\alpha = 2$ but right now it is ≈ 1.25 and in the future will approach to 1. Transistor become more linear.

2. Power Estimation

2.1. Power Estimation Methods

Using lab equipments
Using onboard shunt resistors with ADCs
Simulation (with SPICE)

2.2. Probability Model

Signal probability $\mathrm{P}(g = 1)$: Probability that a signal g will be at high level.
Signal activity $A(g) = \lim_{T\to\infty} \frac{n_g(T)}{T}$: value how often the signal changes.

Activity factor $\alpha \in [0, 1]$
Clock: $\alpha = 1$, Dynamic gates: $\alpha = \frac{1}{2}$, Static gates: $\alpha \approx 0.1$

Modeled with strict static sochastic process $g(t)$

2.2.1 Logic

NOT: $p_{\text{out}} = 1 - p$
AND: $p_{\text{out}} = p_1 p_2$
OR: $p_{\text{out}} = 1 - (1 - p_1)(1 - p_2)$

2.2.2 Binary Decision Diagram

Shannons Expansion: $\mathrm{P}(g) = \mathrm{P}(A)\,\mathrm{P}(g_{A=1}) + \mathrm{P}(\overline{A})\,\mathrm{P}(g_{A=0})$

2.3. Switching Activity

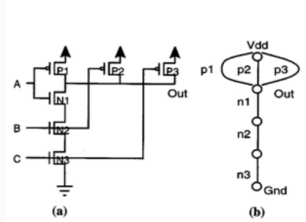
Delay paths cause 8% to 20% of dynamic power.
Static transition probability: $\mathrm{P}_{0\to1} = \mathrm{P}(out = 0) \cdot \mathrm{P}(out = 1) = p_0(1 - p_0)$
Boolean difference: $\frac{\partial f}{\partial x} = f_{x=1} \oplus f_{x=0}$

2.3.1 Reduce Switching Activity

Different coding: fewer bit transition between states (gray coding)
Gate minimization
Avoid glitches: reduce unecessary transitions
Power down: turn off parts

2.4. Leakage Current

CMOS 3 input NAND Gate:
Crucial is the series or paralles connected resistance of the off transistors



State (ABC)	Leakage Current (nA)	Leaking Transistors
000	0.095	N1, N2, N3
001	0.195	N1, N2
010	0.195	N1, N3
011	**1.874**	**N1**
100	0.185	N2, N3
101	**1.220**	**N2**
110	**1.140**	**N3**
111	**9.410**	**P1, P2, P3**

For efficient leakage estimation only look at the dominant states.

2.5. Energy characterization of CPUs

Suitable for high-level energy reduction
Relative base cost in nJ for different ops(and, sub, eor)
Variable cost in pJ for stage and registers

2.6. Energy State Maschine

Static power consumption:

- Leakage current
- Static current flow from VDD to ground when output is 0
- Proportional to duration of stay
- Dependent on the clock frequency

Dynamic power consumption:

- Charging and discharging load capacitors
- Short-circuit current
- Proportional to the number of clock cycles
- Independent to the clock frequency

Other:
Power Estimation by using Performance Counters
Architectural Simulators

3. Circuit-Level Low-Power Techniques

3.1. Dont Care Optimization

exploit "dont care" values in logic to reduce transition probability
$F = AB + \overline{AB}$ $F' = A + \overline{B}$ $\mathrm{P}(A) = \frac{1}{3}$ $\mathrm{P}(B) = \frac{1}{2}$
$\mathrm{P}(F) = 1 - (1 - \mathrm{P}(AB)(1 - \mathrm{P}(\overline{AB})) = \mathrm{P}(AB) + \mathrm{P}(\overline{AB}) + \mathrm{P}(A)\,\mathrm{P}(B)\,\mathrm{P}(\overline{A})\,\mathrm{P}(\overline{B}) = \frac{5}{9}$
$\mathrm{P}(F') = 1 - (1 - \mathrm{P}(A)(1 - \mathrm{P}(\overline{B})) = \mathrm{P}(A) + \mathrm{P}(\overline{B}) - \mathrm{P}(A)\,\mathrm{P}(\overline{B}) = \frac{2}{3}$

3.1.1 Logic Factorization

Reduce literal count to minimize the number of transistors being used to represent the target logic.

3.1.2 Technology Mapping

Hide nodes with high switching activity inside the gates. Select the library with same function but different capacitances while meeting the delay constraints

3.1.3 State Encoding

Graycode, but also one can add aditional bits to reduce Hamming distance between transition. Example: additional inverter bit on bus lines. If more than half of the bits will change all bits can be inverted to reduce Hamming distance.

3.1.4 Retiming

The process of repositioning registers (FFs) in a pipelined circuit (while maintaining I/O functionality). Block the glitch propagation to the large load cap. Move registers to nodes with higher switching activity.

3.1.5 Clock Gating

Provide a way to selectively stop the clock. Force the circuit to make no switching whenever the computation at the next cycle is unnecessary.
Design gated-clock distribution circuit with minimum routing overhead
Improvement: Power reduction of about 30% on standard benchmarks with random test vectors.
Major limitation is representing explicitly FSM tables with many states

3.2. Body Bias Techniques

Lowest acceptable threshold voltage is determined by

- Sub-threshold leakage current
- Die-to-die and within-die threshold voltage variations
- Doping concentration in the channel area

Reverse Body Biasing (RBB):
Apply a negative voltage across the source-to-substrate p-n junction. Threshold voltage changes due to the body effect.

Forward Body Biasing (FBB):
Apply a positive voltage across the source-to-substrate p-n junction

3.2.1 Adaptive RBB

Dynamically varies the body bias voltage depending upon local speed and power requirement.
Effective in reducing variations (supply voltage, temperature and die-to-die process parameters).
Technology scaling may result in losing control of the charge distribution in the channel area. Effectiveness of the RBB technique is reduced due to a weaker body effect with technology scaling

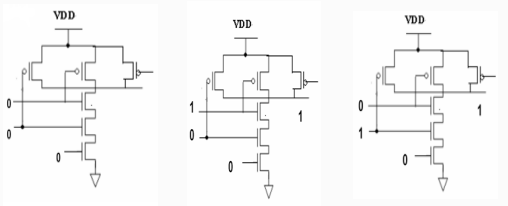
3.2.2 Bidirectional Body Bias (BBB)

Beyond 50 nm technology, bidirectional body bias circuit technique is desirable.
Increase the circuit speed – FBB technique
Reduce the circuit speed and leakage power – RBB technique

3.3. Generalized Multiple V_{th} Problem

Single value for V_{DD}, several values for V_{th}
Dual V_{th}: high and low, most poplar case
Gate-based assignment approach.
Speed up critical path: low V_{th}, leads increased leakage, dont speed up too much
Rest transistors: high V_{th}

3.4. Input Vector Control (IVC)



Least subthreshold leakage Least gate leakage Largest gate leakage

IVC during sleep mode. Advantage over power gating: less capacitor charging, technology scaling has no negative effects
From 10% to 55% reduction in the leakage is expected

3.5. MTCMOS

Multi-threshold CMOS: sleep transistor insertion to use both high-V_{th} and low-V_{th} cells in a logic block. A low-V_{th} block is gated with high-V_{th} power switches that are controlled by SLEEP signal

Delay $T_{\text{dly}} \approx \frac{C_L V_{\text{DD}}}{2 I_j}$

Virtual GND $V_x = \frac{1}{2}\beta(V_{\text{DD}} - V_{\text{th}} - V_x)^2$

Limitations: Area overhead (sleep transistors), slower operation, wake-up delay, process modification for dual V_{th}, ground bounce due to high current spikes

4. System Level LP Techniques

Systems designed to deliver peak performance but it is not needed most of the time.
Decision Methods: Time-out, prediction, stochastic

4.1. Dynamic Power Management (DPM)

Components may self manage state transitions.
Power manager (PM) implemented mostly in software.
Break even time T_{BE} most important factor: Minimum idle time for amortizing the cost of component shutdown

$$T_{\text{BE}} = \frac{t_{10}P_{10} + t_{01}P_{01} - (t_{10}+t_{01})P_{\text{sleep}}}{P_{\text{active}} - P_{\text{sleep}}}$$

transitions: 10 from active to sleep, 01 from sleep to active

Idle time long enough, shut down time short enough, transition power low enough, sleep current low enough

4.1.1 Estimation

Timeout, Predictive, Stochastic
Controlled Markov Processes (CMP)
Component: service provider, Workload: service requester

4.2. Dynamic Voltage Scaling (DVS)

Instead of active and idle switching, finish a task as close to its deadline as possible. Reduce voltage and frequency (DVFS). Fast switching within tenth of microseconds
When f is going up: increase voltage first
When f is going down: decrease f first

4.2.1 Inter-Task vs. Intra-task DVS(II)

Inter-Task: Scaling occurs at the start of a task

Intra-Task: Different frequency is set for each sub-task
Many decision points → prediction of ex. path
Average Prediction better than Worst Case Prediction

4.3. Wireless Sensor Networks

Receiver has to wake up before the sender according to oscilator accuracy.
BLE

4.3.1 Low Power Techniques

Transmission Power adjustment, duty cycle adjustment, general CPU power techniques, payload compression
Energy efficient routing: weighted graph for transmission costs through the net (take also battery levels into account)

5. Architectural-Level Low-Power Design

5.1. Interconnect

Interconnect heavily affects power consumption. Interconnects have to run through all the chips, all capacity lines must be driven Capacity high More current

5.1.1 encoding

simplest example: bus invert coding
one more invert bit line. Instead of flipping all bits, just toggle invert bit. Calculate Hamming distance, halfes maximum switching.

More: Redundancy in space or time: remember previous bits.
Reduce supply voltage

crosstalk: capacitance between wires
add shield wire

Low Swing Bus (Lower Voltage)
Bus segmentation: drive only parts of a bus system
Adiabatic Busses: Reuse charge on buses

5.2. Multimedia Management

Battery Life is important for Apple (20h vs real 8h)
Multi level decoding for audio
Leave out higher frequency

5.2.1 Video Application

Watermark Video with workload information

5.2.2 3D Game Application

Technique 1: Predict next frame on previous frames
Technique 2: Understand structure information in frames
Technique 3: PID controller
If prediction is bad, add technique 2!

5.3. Low Power Memories

Splitting Memories into smaller sub-systems and activating only the needed memory circuits in each access.
Example: banked cache
Scratch Pad Memory: Let software decide memory hierarchy for optimization. Ideal for specialized embedded systems
Trace Cache: store instructions in execution order instead of compiled order.
Dynamic Direction Prediction-Based Trace Cache: Using branch prediction to decide where to fetch instructions from
Selective Trace Cache: Identifying frequently executed "hottraces and store them. **Dynamically Resizable Instruction (DRI) Cache:** It can deactivate its individual sets on demand by gating their supply voltages
Cache Decay: deactivates if it has not been accessed for a pre-determined amount of time
Drowsy Cache: Data is retained, Gated Precharging in 100cycle window.

6. Student Reports

6.1. Energy Characterization Models

Offline: CPU and Memory access energy by hamming distance of address and data of instructions
Runtime: Read periodically the HPCs, Leakage in sleep states, temperature
Advantages: no knwoledge of internal CPU structure required

6.2. Flip Flop Retiming

Invented to balance pipeline stages.
Inserting a Flip-Flop can prevent propagation of an unwanted switching by glitches (delayed switching)
Phase shifted clock signal for Flip-Flops
Disable circuit parts

6.3. Architectural-Level Low-Power Design

Low Level Techniques
Low Power Flip-Flop: reduce switching, reduce charging, use rising and falling clock edge

6.4.

Short circuit analysis delay power consumption find optimal transistor size, optimal gate size path balancing: balancing path delays will reduce glitches

6.5.

Multimedia: clock gating

6.6. Low Power Wireless Sensor Networks

Energy Harvesting: controllable or not, predictable or not

6.7.

Activity in one node depends of switching of previous node. Estimate switching propability

6.8.

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$

6.9. Parker and McClusky's algorithm

6.10. Shannon Expansion

$f = x_i \cdot f_{x_i} + \overline{x}_i \cdot f_{\overline{x}_i} = (x_i + f_{\overline{x}_i}) \cdot (\overline{x}_i + f_{x_i})$
$\overline{f} = x_i \cdot \overline{f}_{x_i} + \overline{x}_i \cdot \overline{f}_{\overline{x}_i}$