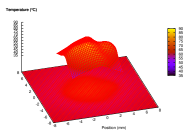
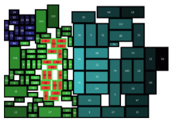
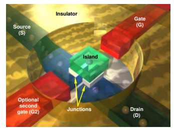
Temperature-Aware and Low-Power Design and Synthesis of Integrated Circuits and Systems

Robert P. Dick

http://robertdick.org/talp
L477 Tech
847–467–2298
Department of Electrical Engineering and Computer Science
Northwestern University







Definitions

- Temperature: Average kinetic energy of particle
- Heat: Transfer of this energy
- Heat always flows from regions of higher temperature to regions of lower temperature
- Particles move
- What happens to a moving particle in a lattice?

Acoustic phonons

- Lattice structure
- Transverse and longitudinal waves
- Electron–phonon interactions
  - Effect of carrier energy increasing beyond optic phonon energy?

Optic phonons

- Minimum frequency, regardless of wavelength
- Only occur in lattices with more than one atom per unit cell
- Optic phonons out of phase from primitive cell to primitive cell
- Positive and negative ions swing against each other
- Low group velocity
- Interact with electrons
- Importance in nanoscale structure modeling?

Nanostructure heat transfer

- Boundary scattering and superlattices
- Quantum effects when phonon spectra of materials do not match
  - Splitting

Why do wires get hot?

- Scattering of electrons due to destructive interference with waves in the lattice
- What are these waves?
- What happens to the energy of these electrons?
- What happens when wires start very, very cool?
- What is electrical resistance?
- What is thermal resistance?

Why do transistors get hot?

- Scattering of electrons due to destructive interference with waves in the lattice
- Where do these waves come from?
- Where do the electrons come from?
  - Intrinsic carriers
  - Dopants
- What happens as the semiconductor heats up?
  - Carrier concentration increases
  - Carrier mobility decreases
  - Threshold voltage decreases

Power consumption trends

- Initial optimization at transistor level
- Further research-driven gains at this level difficult
- Research moved to higher levels, e.g., RTL
- Trade area for performance and performance for power
- Clock frequency gains linear
- Voltage scaling ${V_{DD}}^2$ – very important

Power consumption in synchronous CMOS

$$P = P_{SWITCH} + P_{SHORT} + P_{LEAK}$$
$$P_{SWITCH} = C \cdot {V_{DD}}^2 \cdot f \cdot A$$
$$\dagger\, P_{SHORT} = \frac{b}{12}(V_{DD} - 2 \cdot V_T)^3 \cdot f \cdot A \cdot t$$
$$P_{LEAK} = V_{DD} \cdot (I_{SUB} + I_{GATE} + I_{JUNCTION} + I_{GIDL})$$

C : total switched capacitance — V_{DD} : high voltage
f : switching frequency — A : switching activity
b : MOS transistor gain — V_T : threshold voltage
t : rise/fall time of inputs

† P_{SHORT} usually $\leq 10\%$ of P_{SWITCH}

Smaller as $V_{DD} \rightarrow V_T$

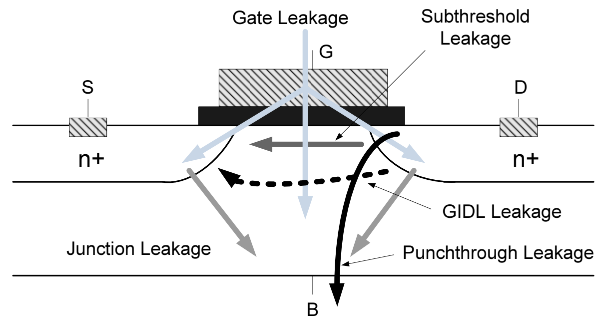
Adiabatic charging

- Voltage step function implies $E = C{V_{CAP}}^2/2$
- Instead, vary voltage to hold current constant: $E = C{V_{CAP}}^2 \cdot RC/t$
- Lower energy if $T > 2RC$
- Impractical when leakage significant

Wiring power consumption

- In the past, transistor power $\gg$ wiring power
- Process scaling $\Rightarrow$ ratio changing
- Conventional CAD tools neglect wiring power
- Indicate promising areas of future research

Leakage



Subthreshold leakage current

$$I_{subthreshold} = A_s \frac{W}{L} {v_T}^2 \left(1 - e^{\frac{-V_{DS}}{v_T}}\right) e^{\frac{(V_{GS}-V_{th})}{n v_T}}$$

- where A_s is a technology-dependent constant,
- V_{th} is the threshold voltage,
- L and W are the device effective channel length and width,
- V_{GS} is the gate-to-source voltage,
- n is the subthreshold swing coefficient for the transistor,
- V_{DS} is the drain-to-source voltage, and
- v_T is the thermal voltage.

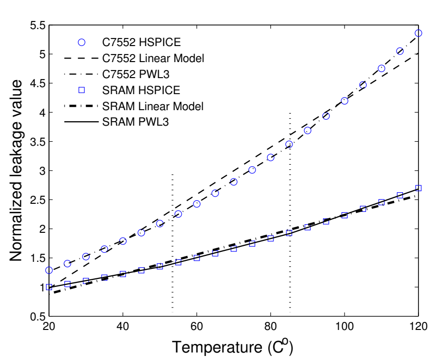
A. Chandrakasan, W.J. Bowhill, and F. Fox. *Design of High-Performance Microprocessor Circuits*. IEEE Press, 2001

Simplified subthreshold leakage current

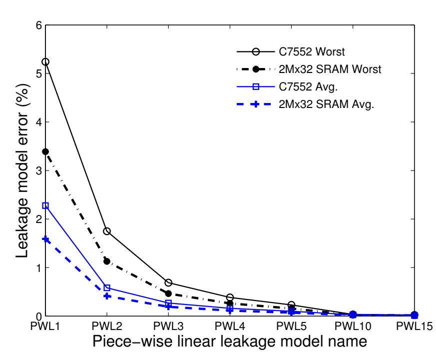
$V_{DS} \gg v_T$ and $v_T = \frac{kT}{q}$. q is the charge of an electron. Therefore, equation can be simplified to

$$I_{subthreshold} = A_s \frac{W}{L} \left(\frac{kT}{q}\right)^2 e^{\frac{q(V_{GS}-V_{th})}{nkT}} \quad (1)$$

Exponential?



Piece-wise linear error



Gate leakage

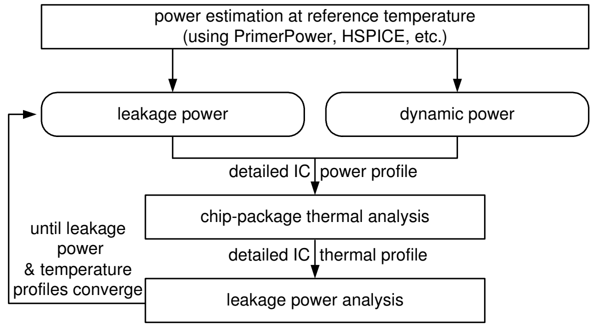
Caused by tunneling between gate and other terminals.

$$I_{gate} = WLA_J \left(\frac{T_{oxr}}{T_{ox}}\right)^{nt} \frac{V_g V_{aux}}{T_{ox}^2} e^{-BT_{ox}(a-b|V_{ox}|)(1+c|V_{ox}|)}$$

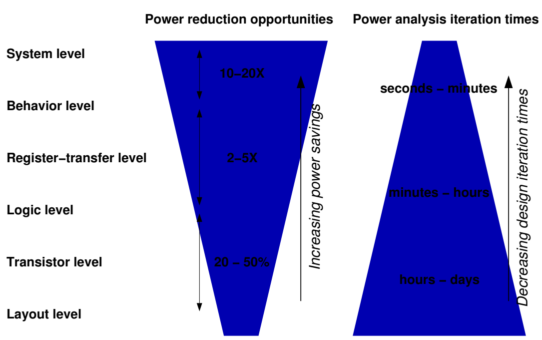
- where A_J, B, a, b, and c are technology-dependent constants,
- nt is a fitting parameter with a default value of one,
- V_{ox} is the voltage across gate dielectric,
- T_{ox} is gate dielectric thickness,
- T_{oxr} is the reference oxide thickness,
- V_{aux} is an auxiliary function that approximates the density of tunneling carriers and available states, and
- V_g is the gate voltage.

K. M. Cao, W. C. Lee, W. Liu, X. Jin, P. Su, S. K. H. Fung, J. X. An, B. Yu, and C. Hu. BSIM4 gate leakage model including source-drain partition. In *IEDM Technology Dig.*, pages 815–818, December 2000

Temperature-aware leakage estimation



Design level power savings



From Anand Raghunathan

Power consumption conclusions

- Voltage scaling is currently the most promising low-level power-reduction method: V^2 dependence.
- As V_{DD} reduced, V_T must also be reduced.
- Sub-threshold leakage becomes significant.
- What happens if $P_{LEAK} > P_{SWITCH}$?
- Options to reduce leakage (both expensive):
  - Liquid nitrogen – diode leakage
  - Silicon-on-insulator (SOI) – I_{SUB}

Reading assignment

- G. Chen, R. Yang, and X. Chen. Nanoscale heat transfer and thermal-electric energy conversion. *J. Phys. IV France*, 125:499–504, 2005
- An article relevant to your chosen mini-project topic