Introduction to Embedded Systems Research: Power, Energy, and Temperature

Robert Dick

dickrp@umich.edu
Department of Electrical Engineering and Computer Science
University of Michigan
Outline

1. Deadlines and announcements
2. Power and temperature definitions and fundamentals
3. Thermal analysis
4. Power models for embedded systems
## Deadlines and Announcements I

From now on, deadlines and announcements will come at the start of lecture slides.


I was quite ill recently. I’m catching up on feedback/evaluations.


4 March: Project checkpoint 1.

11 March: Midterm exam.
30 March: Project checkpoint 2.

21 Apr: Project deadline.

10:30am–12:30pm 29 Apr: Final exam.
Presentations feedback

Private feedback on presentations in office hours?
Context


Brief lecture on RTOSs and CPS.


Lecture on power, energy, and temperature.

Outline

1. Deadlines and announcements
2. Power and temperature definitions and fundamentals
3. Thermal analysis
4. Power models for embedded systems
Temperature: Average kinetic energy of particle.

Heat flow: 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

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.
Nanostructure heat transfer

Boundary scattering.

Quantum effects when phonon spectra of materials do not match.
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 $V_{DD}^2$ – 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

\[ \dagger P_{SHORT} \text{ usually } \leq 10\% \text{ of } P_{SWITCH} \]

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.
In the past, transistor power $\gg$ wiring power.

Process scaling $\Rightarrow$ ratio changing.

Conventional CAD tools neglect wiring power.
Leakage

- Gate Leakage
- Subthreshold Leakage
- Junction Leakage
- GIDL Leakage
- Punchthrough Leakage

Deadlines and announcements
Power and temperature definitions and fundamentals
Thermal analysis
Power models for embedded systems
Subthreshold leakage current

\[ I_{\text{subthreshold}} = A_s \frac{W}{L} v_T^2 \left( 1 - e^{-\frac{V_{DS}}{v_T}} \right) e^{\frac{(V_{GS} - V_{th})}{nv_T}}, \]

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.

Simplified subthreshold leakage current

\[ V_{DS} \gg v_T \text{ and } v_T = \frac{kT}{q}. \] 

\[ q \text{ is the charge of an electron. Therefore, equation can be simplified to} \]

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

(1)
Exponential?
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.

Temperature-aware leakage estimation

- Power estimation at reference temperature (using PrimerPower, HSPICE, etc.)
- Leakage power
- Dynamic power
- Detailed IC power profile
- Chip-package thermal analysis
- Detailed IC thermal profile
- Leakage power analysis

Until leakage power & temperature profiles converge
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 $P_{LEAK} > P_{SWITCH}$?
Outline

1. Deadlines and announcements
2. Power and temperature definitions and fundamentals
3. Thermal analysis
4. Power models for embedded systems
R(C) model

Partition into 3-D elements (diagram 2-D for simplicity)

Thermal resistance $\leftrightarrow$ Resistance

Heat flow $\leftrightarrow$ Current

For dynamic: Heat capacity $\leftrightarrow$ Capacitance
Problem definition

\[ C \frac{dT(t)}{dt} = AT(t) + PU(t) \]

- A is the thermal conductivity matrix
- Steady-state: Initial temperature and C unnecessary
- Dynamic: Transient temperature analysis, must also consider heat capacity
Thermal analysis infrastructure overview

3-D chip/package/ambient heat capacity and thermal conductivity profiles

Power profile

Initial 3-D temperature profile and hybrid oct-tree (optional)

Multigrid incremental solver

Spatial hybrid oct-tree refinement

Initial 3-D temperature profile and hybrid oct-tree

Spatial hybrid oct-tree refinement

Initialize/update discrete event simulator queue

Process one pending event

Adapt neighboring element step sizes

Sample period reached?

Short time scale dynamic thermal analysis

Long time scale dynamic thermal analysis

System response coefficient calculation

Dynamic thermal profile calculation

Sample period reached?

Spatial hybrid oct-tree refinement

Moment matching for chip/package

Dynamic thermal profile calculation

N

Y

N

Y

Y

N

Y

Converged?

Y

N

Adapt profile based on k(T)

3-D thermal profile (and hybrid oct-tree)

Thermal gradient conditions satisfied?

Steady-state thermal analysis
Thermal analysis infrastructure overview

- Multigrid incremental solver
- Spatial hybrid oct-tree refinement
- Steady-state thermal analysis
- Thermal gradient conditions satisfied?
  - Y: Adapt profile based on k(T)
  - N: Initial 3-D temperature profile and hybrid oct-tree

N

Y
Thermal analysis infrastructure overview

Initial 3-D temperature profile and hybrid oct-tree

Initialize/update discrete event simulator queue

Process one pending event

Adapt neighboring element step sizes

Sample period reached?

Y

Short time scale dynamic thermal analysis

Dynamic thermal profile (and)

Y

Initialize/update discrete event simulator queue

Spatial hybrid oct-tree refinement

Moment matching for chip/pack

System response coefficient calculation

Dynamic thermal profile calculation

Converged?

N
Thermal analysis infrastructure overview

Adapt neighboring element step sizes

Spatial hybrid oct-tree refinement

Moment matching for chip/package

System response coefficient calculation

Dynamic thermal profile calculation

Sample period reached?

Y

N

Short time scale dynamic thermal analysis

3-D thermal profile (and hybrid oct-tree)
Thermal analysis infrastructure overview

1. **3-D chip/package/ambient heat capacity and thermal conductivity profiles**
2. **Power profile**
3. **Initial 3-D temperature profile and hybrid oct-tree (optional)**

**Flowchart:****
- **Multigrid incremental solver**
- **Spatial hybrid oct-tree refinement**
- **Thermal gradient conditions satisfied?**
  - *Y*: Next stage
  - *N*: Prior stage
- **Adapt profile based on k(T)**
- **Converged?**
  - *Y*: Next stage
  - *N*: Prior stage
- **Initialize/update discrete event simulator queue**
- **Process one pending event**
- **Adapt neighboring element step sizes**
  - *Y*: Next stage
  - *N*: Prior stage
- **Sample period reached?**
  - *Y*: Next stage
  - *N*: Prior stage
- **Short time scale dynamic thermal analysis**
  - *Y*: Next stage
  - *N*: Prior stage
- **Spatial hybrid oct-tree refinement**
- **Dynamic thermal profile calculation**
  - *Y*: Next stage
  - *N*: Prior stage
- **System response coefficient calculation**
- **Moment matching for chip/package**
- **Long time scale dynamic thermal analysis**
  - *Y*: Next stage
  - *N*: Prior stage
Steady-state thermal analysis

Basis: Multigrid analysis

Fast, multi-resolution relaxation method for matrix solving.

1. Iterative solver (relaxation) on fine grid.
2. Coarsen and propagate residual upward.
3. Iterative solver for error at coarser level.
4. Correct fine-grained solution based on coarse-grained error.
5. Iterative solver for error at fine level.

- Main challenge: Too slow for repeated use on large structures, especially 3-D chip-package modeling.
- Observation: Steepness of thermal gradients vary across IC.
Spatial adaptation can improve performance w.o. loss of accuracy.
Hybrid oct-tree

- Reduce element count by merging when $\Delta T < \epsilon$
- Conventional oct-tree inefficient for chip-package model
- Anisotropic thermal gradients
- We generalize to hybrid oct-tree
- Arbitrary partitioning on each axis
Hybrid oct-tree

Level 3
Level 2
Level 1
Outline

1. Deadlines and announcements
2. Power and temperature definitions and fundamentals
3. Thermal analysis
4. Power models for embedded systems
General case

Many components.

Each may have many power management/activity states.

System-wide power consumption depends on the specific combination of component states.

How many samples?

- 10 components.
- 5 states, each.
- \(5^{10} \approx 10\)-million system-wide states.

How to get enough samples to characterize?
Independence assumption for embedded system power modeling

What if a component’s power consumption were mostly independent of the power management/activity states of other components?

How many samples?

- 10 components.
- 5 states, each.
- 50 samples of interest.

The assumption is often correct.

When it is not, can treat the two interdependent components as a single component.
Practical embedded system power estimation

1. For each component.
   1. Put all other components in lowest power state.
   2. Measure component power consumption in each state.

2. Can manually use measurements to build expression for system-wide power consumption.

3. Also works for incomplete sampling by using linear regression to find the relationship between each state variable and the system-wide power consumption.
Applying the power model

Estimate/measure the proportion of time each component spends in each state.

Sum the products of time proportions and component–state power consumptions to get system-wide average power consumption.

This is often inaccurate for instantaneous power consumption.
Not good for power supply provisioning or thermal design.

Often very accurate over timescales of minutes.
Good for battery lifespan estimation.