ECE5461: Low Power SoC Design

Tae Hee Han: than@skku.edu
Semiconductor Systems Engineering
Sungkyunkwan University
Course Information

- Objectives
  - This course covers all major aspects of low-power design of SoCs, and addresses emerging topics related to future design. It explores the many different domains and disciplines that impact power consumption from system-level to device level.

- Lecture Schedule
  - Mon. /Wed. 9:00 ~ 10:15 AM

- References for this course
  - Findlay Shearer, Power Management in Mobile Devices, Newnes, 2007
  - Liming Xiu, VLSI Circuit Design Methodology Demystified, Wiley Inter-Science, 2008
## Course Schedule

<table>
<thead>
<tr>
<th>Schedule</th>
<th>Contents</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>Basic Concept, Introduction</td>
<td></td>
</tr>
<tr>
<td>Week 2</td>
<td>Battery Aware Power Management / System-level Power Estimation</td>
<td></td>
</tr>
<tr>
<td>Week 3~4</td>
<td>System level power optimization</td>
<td></td>
</tr>
<tr>
<td>Week 5</td>
<td>Algorithm / Architecture level power optimization</td>
<td></td>
</tr>
<tr>
<td>Week 6~7</td>
<td>Logic/Circuit/Device level Power Reduction</td>
<td></td>
</tr>
<tr>
<td>Week 8</td>
<td>Term paper assign</td>
<td></td>
</tr>
<tr>
<td>Week 9</td>
<td>Logic/Circuit/Device level Power Reduction</td>
<td></td>
</tr>
<tr>
<td>Week 10~12</td>
<td>Case studies</td>
<td></td>
</tr>
<tr>
<td>Week 13~15</td>
<td>Term paper discussion</td>
<td></td>
</tr>
<tr>
<td>Week 16</td>
<td>Final Exam.</td>
<td></td>
</tr>
</tbody>
</table>
Grading System

- Homework/Term Paper: 50%
- Attendance: 10%
- Final Exam: 30%
- Etc. : 10%
Basic Concept & Introduction
Sad fact:

Computers turn electrical energy into heat. Computation is a byproduct.

Air or water carries heat away, or chip melts.
55 W-hour battery stores the energy of 1/2 a stick of dynamite.

If battery short-circuits, catastrophe is possible ...
What Consumers Care About

- Users want more features in their mobile devices:
  - MP3, Camera, Video, GPS...
- Convenient form factor, affordable price
- But also need long battery life
  - Battery technology is not evolving fast enough!
  ➔ Need to manage power consumption
Smart Devices and ICT Convergence
Needs vs. Reality

- **Algorithmic/Application Complexity**
- **Processor Performance (Moore’s Law)**
- **Battery Capacity**

The diagram shows the progression of technology advances from 1G to 4G, with annotations for 2G, 3G, and 4G, indicating the evolution of technology over time.
Problem Statement: Explosive Growth of SoC Complexity

- Massive feature integration: Driving SoC complexity to the extreme
  - Multiple processors
    - CPU processor
    - DSP processor
    - Graphic processor
  - Many high-performance engines
    - Video cores
    - DMA engines

Distributed Heterogeneous Architectures
What Drives the Widening Power Gap

- **Performance and style dictates design**
  - To be smarter, more performance and functionalities are needed
  - Demand for portability equates to space limitations on batteries

- **Slow Battery Research and Development**
  - Cutting edge battery R&D is focused elsewhere (hybrid vehicle technology)
    - Development focuses on delivering higher discharge vs. power conserving batteries

- **Convergence is primary contributor**
  - Users continue to demand more applications
  - Operators derive increasing revenue streams from application based services
Energy and Power

- **Energy**: ability to do work
  - Most important in battery-powered systems

- **Power**: energy per unit time
  - Important even in wall-plug systems --- *power becomes heat*

- **Power draw increases with...**
  - Vcc
  - Clock speed
  - Temperature
Why are Power & Energy Important?

- Battery life for mobile devices
- Reliability at high temperatures
- Power density (cooling)
  - Limits compaction & integration
- Cost
  - Energy cost
  - Cost of power delivery, cooling system, packaging
- Environmental issues
  - IT responsible for 0.53 billion tons of CO₂ in 2002
Metrics

- **Energy (Joules) = Power (Watts) × Time (sec)**
  - Power is limited by infrastructure (e.g., power supply)
  - Energy: what the utilities charge for or battery can store

- **Power density = power/area**
  - The major metrics for the cooling system

- **Combined metrics**
  - How to tradeoff performance for power savings
  - TPS/W, energy × delay (EDP), energy × delay² (EDP²), ...
Recall: Charge-based Digital Logic

Key principles in the charge based digital logic

- **Representation of digital states**
  - Logic “0”: No Charge in the capacitor
  - Logic “1”: Charge stored in the capacitor

- **Change of digital state**
  - Charge/dis-charge capacitor through a resistor

\[ V_{\text{out}} = \frac{Q}{C} \]
Power Consumption in ICs

- Dynamic or active power consumption
  - Charging and discharging capacitors
  - Depends on switching activity

- Short circuit currents (Can be ignored)
  - Short circuit path between supply rails during switching
  - Depends on the size of the transistors

- Leakage current or static power consumption
  - Leaking diodes and transistors
  - Gets worse with smaller devices and lower Vdd
  - Gets worse with higher temperatures

\[ E = \int_{0}^{t} (C V_{DD}^2 f_c + V_{DD} I_{lk}) \, dt \]

Total Power Dissipation

\[ \int_{0}^{t} V_{DD} I_{\text{leak}} \, dt \]

Static Power Dissipation

\[ \int_{0}^{t} CV_{DD}^2 f_c \, dt \]

Dynamic Power Dissipation

- Minimize \( I_{\text{leak}} \) by:
  - Reduce the voltage
  - Use fewer transistors
  - Use lower leakage transistors

- Minimize \( I_{\text{switch}} \) by:
  - Reduce the voltage
  - Decrease switching cap
  - Lower switching activity
Power Wall

- Moore's Law
  - Transistor density increases every 18~24 months

- CMOS Power
  - Total Power = \( V^2 \cdot f \cdot C \cdot \alpha + V \cdot I_{\text{leakage}} \)
  - Drastic increase in leakage current and decrease in noise margin prevent the voltage scaling around 1V

**Limitations in Processor Performance**
- Memory Wall
- ILP Wall
- Power Wall

**Not only Battery, but also Heat!**
- Intel 80386 consumed \( \sim 2 \text{ W} \)
- 3.3 GHz Intel Core i7 consumes 130 W
- Heat must be dissipated from 1.5 x 1.5 cm chip
- This is the limit of what can be cooled by air
Pollack’s Rule: "performance increase due to μ-architecture advances is roughly proportional to [the] square root of [the] increase in complexity“

Implications (in the same technology)
- New μ-Arch consumes about 2-3x die area of the last μ-Arch, but provides 1.5-1.7x performance
Reducing Power/Energy

- An interdisciplinary issue
  - Circuits, architecture, software, systems

- Key high-level ideas
  - Reduce redundant work/components
  - Turn off unused components
  - Pick implementation that best matches constraints
    - E.g., don’t use a 3GHz processor if 1GHz would do
Reducing Power/Energy: Another Option

STMicroelectronics’ Platform 2012

GOPS/mm² – GOPS/W

1

General-purpose Computing
CPU

3

Throughput Computing
GPGPU

6

Mixed
HW

> 100

SW

1GOPS/mW

Platform 2012

HW IP

Closing The Accelerator Efficiency Gap with Agile Customization
Holistic Approach for entire energy delivery chain
Energy/Power Flow in Mobile Device

- **Battery**
  - Li+/Li-Poly
    - 3V ~ 4.2V
  - NiMH/Akamine
    - 0.9V ~ 1.5V
  - Solar/Fuel cell
  - Role
    - Main / Backup
  - Type
    - Primary
    - Secondary
  - Management
    - Protector
    - Gas gauge
    - Security

- **Power Supply**
  - Adaptor
    - 5V, 12V, ...
    - ± 10% Tolerance
    - Xmer/Switching
    - Car outlet
  - USB Port
    - 4.5V ~ 5.25V
    - $I_{\text{max}}$: 500mA (USB 2.0) ~ 900mA (USB 3.0)

- **Power Conversion**

- **Load**
  - Standard IC
    - 5V, 3.3V, 2.5V, 1.8V, 1.1V, ...
    - $V_{\text{core}}$
  - RF IC/Device
    - Low noise required
  - Display
    - LED Lighting/Flash
    - LCD/OLED/EL Bias
    - CCFL supply
  - Motor/Inductive
    - Vibrator
    - HDD
  - USB Host
    - Ports
  - Others
    - Tuners, ...

※ EL: Electro Luminescence
※ CCFL: Cold Cathode Fluorescent Lamp
Anatomy of a Handset

CPU

Application Tasks

MMI (Man-Machine Interface)

PROTOCOL STACK

L1 SW

DSP

DATA I/O

LCD, Camera, Etc

L1 DBB HARDWARE

ABB

RF

Digital Baseband

PM IC

ADC

DAC

NOR or NAND Flash

SRAM or SDRAM

Application Processor

Duplexor/Switch

LNA

Filter

VCO

Filter

PA

PM IC
Anatomy of a Handset: Another View

Applications Framework

- Database (UI, Phonebook, Security, Java, Browser, Messaging, Multimedia playing)

Modem Applications Abstraction Layer

- Multimode Protocol Service IF

Multimedia Device Framework

- Transport Service
- Platform Devices
- Multimedia Engine
- Media Devices
- File System
- Data Format
- Media Codec

Application Protocol Framework

- IP System

Cellular Protocol Stack: Multimode Protocol

- 2G
- 3G
- 4G

Traffic Manager

- Tools Agent Framework
- System Framework
- Physical Layer Framework

Hardware

- HW Accelerator & Device Interface
- AV codec
- Display
- PM
- Positioning
- Broadcast
- PAN
- LAN
- MAN
- RF
- Cellular Radio Interface
- 2G/3G/4G RF
Gross Power Consumption Exceeds Thermal Dissipation Capability of Mobile Device

- Large plastic communicator, open
- 100cc plastic clamshell, open
- 100cc plastic monoblock
- 100cc metal monoblock
- Small metal communicator, open

- Power conversion
- Mass Memory
- Audio
- Camera
- Display+backlight
- Local Connectivity
- Apps Engine
- Cellular BB
- Cellular RF

2002 - 2006

Cooling required
Inside Smartphone: Apple iPhone 4

- Infineon 3G Baseband
- Dialog Power Management
- A4 Application Processor
- Accelerometer
- Gyroscope
- Numonyx NOR and mobile DDR
- Samsung flash memory
- TI touchscreen controller
- Cirrus Logic audio codec
- GSM/GRPS Front End Module

iFixit teardown of iPhone 4: logic board
Nvidia Tegra2 Multi-Core SoC

- 8 Dedicated Processors
- Highest CPU Performance - (ARM Cortex-A9@1GHz × 2)
- HD 1080p Video
- GeForce® Graphics
- Ultra Low Power
Nvidia Tegra 4i

- New Quad core ARM Cortex-A9 R4 @2.3GHz
- Integrated i500 HSPA/LTE Modem
- “4 plus 1” Companion Core
- ULP GeForce 60 GPU cores
- Computational photography architecture
- Image signal processor
- Video engine
ARM big.LITTLE Architecture for Low Power

**Android**
- Android Power Management

**Linux**
- Linux Power Management

**Task Migration**
- Hypervisor Mode

**big.LITTLE Processing Platform**
- Cortex-A15
- Cortex-A7

**big.LITTLE Effect**
- Peak Performance
- Energy

More Performance
Less Energy

<table>
<thead>
<tr>
<th>Cortex-A8</th>
<th>Cortex-A8</th>
<th>2x Cortex-A9</th>
<th>4x Cortex-A9</th>
<th>2x Cortex-A15</th>
<th>2x Cortex-A57</th>
</tr>
</thead>
<tbody>
<tr>
<td>65 nm</td>
<td>45 nm</td>
<td>40 nm</td>
<td>32 nm</td>
<td>28 nm</td>
<td>20 nm</td>
</tr>
</tbody>
</table>

32
Samsung Exynos 5410 Octa - for Galaxy S4

Heterogeneous CPU operation
- Two Heterogeneous Quad-core CPUs for
  - Can be switched based on task and work loads.
  - Efficient power consumption with Maximized performance.

<table>
<thead>
<tr>
<th></th>
<th>1st Quad-core CPU</th>
<th>2nd Quad-core CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture</td>
<td>ARM v7a</td>
<td>ARM v7a</td>
</tr>
<tr>
<td>Process</td>
<td>Samsung 28nm HKMG</td>
<td>Samsung 28nm HKMG</td>
</tr>
<tr>
<td>Frequency</td>
<td>200MHz~1.8GHz+</td>
<td>200MHz~1.2GHz+</td>
</tr>
<tr>
<td>Area</td>
<td>19mm^2</td>
<td>3.8mm^2</td>
</tr>
<tr>
<td>Power-ratio</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>L1 Cache size</td>
<td>32KB I/D cache</td>
<td>32KB I/D cache</td>
</tr>
<tr>
<td>L2 Cache size</td>
<td>2MB Data cache</td>
<td>512KB Data cache</td>
</tr>
</tbody>
</table>

Die Photo

Big CPU
High performance for compute intensive applications

Little CPU
Low power execution of majority workloads
Galaxy S4 Teardown

- Qualcomm WCD9310 audio codec
- Qualcomm MDM9215M 4G GSM/UMTS/LTE modem
- ARM Holdings MBG965H
- Qualcomm PM8917 power management
- Broadcom BCM4335 Single-Chip 5G Wi-Fi MAC/Baseband/Radio
- Samsung K3QF2F200E 2 GB LPDDR3 RAM + Snapdragon 600 APQ8064T 1.9 GHz Quad-Core CPU lurks below)
- Toshiba THGBM5G7A4JBA4W 16 GB eMMC (eMMC integrates a NAND flash memory and a controller chip in a single package)
Galaxy S4 Teardown

- Qualcomm WTR1605L seven-band 4G LTE RF transceiver
- Broadcom 20794S1A standalone NFC chip
- Silicon Image 8240BO MHL 2.0 transmitter
- Maxim MAX77803 microcontroller
- SWA GNF09
- Qualcomm PM8821 power management IC
- Skyworks 77619 power amplifier module for quad-band GSM/EDGE
A 20nm Scenario (High-end Processor)

Assume $V_{DD} = 1.2V$
- FO4 delay < 5 ps
- Assuming no architectural changes, digital circuits could be run at 30 GHz
- Leading to power density of 20 kW/cm$^2$ (??)

Reduce $V_{DD}$ to 0.6 V
- FO4 delay $\approx$ 10 ps
- The frequency is lowered to 10 GHz
- Power density reduces to 5 kW/cm$^2$ (still way too high)

This means:
- A 2cm$^2$ processor consumes 10 kW
- A bound of 100W requires only 1% to be active ⇒ dark silicon

Ref: S. Borkar (Intel)
The “Dark Silicon” Problem

<table>
<thead>
<tr>
<th>Node</th>
<th>45nm</th>
<th>20nm</th>
<th>10nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>2009</td>
<td>2013</td>
<td>2017</td>
</tr>
<tr>
<td>Area</td>
<td>1</td>
<td>~1/4</td>
<td>~1/16</td>
</tr>
<tr>
<td>Peak freq</td>
<td>1</td>
<td>1.6</td>
<td>2.4</td>
</tr>
<tr>
<td>Power</td>
<td>1</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>@ 45nm freq</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>@ peak freq</td>
<td>1</td>
<td>1</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Exploitable Si (equivalent Power)

Note: Precise scaling details don’t matter as much as the general observation

Source: Rob Aitken (ARM)
## How Much Energy in the Air?

<table>
<thead>
<tr>
<th>Source</th>
<th>Energy (μW/cm³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solar (outdoor)</td>
<td>15,000</td>
</tr>
<tr>
<td>Air flow</td>
<td>380</td>
</tr>
<tr>
<td>Human power</td>
<td>330</td>
</tr>
<tr>
<td>Vibration</td>
<td>200</td>
</tr>
<tr>
<td>Temperature</td>
<td>40</td>
</tr>
<tr>
<td>Pressure Var.</td>
<td>17</td>
</tr>
<tr>
<td>Solar (indoor)</td>
<td>10</td>
</tr>
</tbody>
</table>

**Vibrations**

**Thermal**

**Air Flow**
Cutting Edge SoC Design looks like the Rocket Science

**DFV**
: Design For Verification

**DFM**
: Design For Manufacturability

**DFP**
: Design For (Low) Power

**DFT**
: Design For Testability

**Payload**
: Pure Functional Implementation efforts

- **Overhead**
Observation #1: Design Challenges

- Technology shrink leads to critical design challenges
Observation #2: Design Complexity

- Complexity outpaces Design Productivity

[Source: SEMATECH]
Observation #3: Power Densities

- Power densities affects packaging, cooling, reliability, speed, ...

![Graph showing power densities over time with various labels like Hot Plate, Nuclear Reactor, Rocket Nozzle, and Sun’s Surface.](image-url)
Observation #4: Battery Technology Limitation

- Less than 10% technology improvement is expected for battery for next 10 years

Source: Dataquest
Observation #5: Leakage Power

- Leakage may ruin Moore’s Law (it is worse than expected), threatening the success of CMOS by ITRS

[ ISLPED 04, Ray Bryan ]
Observation #6: Chip I/O Bottleneck

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Logic trans/chip (M)</td>
<td>60</td>
<td>235</td>
<td>925</td>
<td>3,650</td>
<td>14,400</td>
</tr>
<tr>
<td>Signal pins/chip</td>
<td>1024</td>
<td>1024</td>
<td>1280</td>
<td>1408</td>
<td>1472</td>
</tr>
</tbody>
</table>

Simultaneously Switching Noise

[Source: SEQUENCE]
## Observation #7: Interconnect Challenge

### Impact of interconnect has to be considered in early design stage

- Helps faster convergence – tight correlation with the backend
- Produces more efficient designs – lower area, power
- Design flow becomes more predictable
- Improves performance – higher frequency

**Source: Synopsys (2012)**

<table>
<thead>
<tr>
<th></th>
<th>2005</th>
<th>2010</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
<td>130 to 90 nm</td>
<td>65, 45, 32 nm</td>
<td>28, 20, 14 nm</td>
</tr>
<tr>
<td>Wire length (m/cm²)</td>
<td>1,019</td>
<td>2,222</td>
<td>3,143</td>
</tr>
<tr>
<td>Important new effects</td>
<td>Route topology</td>
<td>Layer awareness, Coupling capacitance</td>
<td>Resistive shielding, Much less resistance on higher metal layers</td>
</tr>
</tbody>
</table>

- **Wire length**
  - **2005**: 1,019 m/cm²
  - **2010**: 2,222 m/cm²
  - **2012**: 3,143 m/cm²

- **Process**
  - **2005**: 130 to 90 nm
  - **2010**: 65, 45, 32 nm
  - **2012**: 28, 20, 14 nm

- **Important new effects**
  - **2005**: Route topology
  - **2010**: Layer awareness, Coupling capacitance
  - **2012**: Resistive shielding, Much less resistance on higher metal layers
Observation #9: Process Variation

Source: Borkar, De Intel®
Observation #10: Cost of Chip Development


Source: Xilinx, IBS(2011)
Observation #11: CMOS Limitation

More than Moore: Diversification
- Analog/RF
- Passives
- HV Power
- Sensors Actuators
- Biochips

Interacting with people and environment
Non-digital content System-in-Package (SiP)

Combining SoC and SiP: Higher Value Systems

Beyond CMOS

Source: ITRS 2011
Observation #12: Design Methodology Change

Traditional Chip Design

Mobile Chip Design

- OMAP2420:
- Five Power Domains (MCU Core, DSP Core, Graphic Accelerator, Peripheral, Alive logic)
- 40x Leakage Power Savings

[Source: Stork (TI), DAC 2006]
Observation # 13: Noise Isolation

- Digital switching noise propagates through the substrate

[Source: TI, Stork, DAC 2006]
Power Profile Optimization

Dynamic Power Profile

Static Power Profile

Optimized Static Power Profile

System Workload
Classification of Low Power Techniques

Methods
- Reducing Activity
- Reducing Capacitance
- Scaling Supply Voltage
- Scaling Threshold Voltage

Techniques
- Voltage Scaling
- Instruction-Level Optimization
- Control-Data-Flow Transformation
- Dynamic Power Management
- Approximate Signal Processing
- Memory Optimization
- Hardware-Software Partitioning
- Parallelism/Pipelining
- Don’t-care Optimization
- Path Balancing
- Factorization
- Technology Decomposition/Mapping
- Encoding
- Retiming
- Gated Clocks
- Pre-computation
- Transistor/Interconnect Sizing
- Transistor Reordering
- Threshold Voltage Scaling

Overheads
- Negligible
- Area
- Speed
- Noise

Note:
- System Level
- Architecture Level
- Logic Level
- Circuit/Device Level
Summary: Reducing Power @ All Design Levels

- Algorithmic level
- Compiler level
- Architecture level
- Organization level
- Circuit level
- Silicon level

Important concepts:
- Lower Vdd and freq. (even if errors occur) / dynamically adapt Vdd and freq.
- Reduce circuit
- Exploit locality
- Reduce switching activity, glitches, etc.

\[ P = \alpha \cdot f \cdot C \cdot V_{dd}^2 \]

\[ E = \int P \, dt \Rightarrow E / \text{cycle} = \alpha \cdot C \cdot V_{dd}^2 \]