Roberto Zafalon  
Director, EU R&D Projects

**Leakage Aware Design for Next Generation's SOCs**

Date’09 workshop, April 24th 2009  
Designing for Embedded Parallel Computing  
Platforms: Architectures Session

---

**Outline**

- Market Application rush  
- Basics of CMOS Leakage Power consumption

- Why bothering for low power systems?  
  - Technology Scaling, Trends & Roadmap  
  - Leakage Aware design strategies  
  - Cost of heat removal: packaging and reliability  
  - Memory architectures  
  - Increased market share of mobile electronics  
  - Limitations of battery technology

- Conclusion
30 Years of Electronics Industry CAGR

- **Semic. Capex:** 17%
- **Semic. Market:** 15%
- **Electronic Systems:** 8%
- **WW GDP:** 3.4%

Market Application rush

- **1 TOPS/W**
- **100 GOPS/W**
- **5 GOPS/W**

Year of Introduction

- 2005
- 2007
- 2009
- 2011
- 2013
- 2015

- 3D TV
- 3D gaming
- 3D project projection display
- Autonomous driving
- 3D ambient interaction
- Sign recognition
- A/V streaming
- Sign language recognition
- Gesture recognition
- Expression recognition
- Hand gesture recognition
- Sign language decoding
- Face recognition
- Gesture recognition
- Face recognition
- Adaptive route
- Image dictation
- Gesture detection
- Gbit radio
- UWB
- Structured encoding
- H.264 encoding
- H.264 decoding
- Structured decoding
- Collision avoidance
- Autonomous driving
- Mobile base-band
- Auto personalization
CMOS Roadmap: 3 main showstoppers

Pat Gelsinger, CTO Intel Corp.
Quote from DAC’04 Keynote:

*Power is the only limiter !!*

CMOS Roadmap: 3 main showstoppers:

1. Subthreshold Leakage Current (I_{off})
2. Huge Process Variation Spread
3. Interconnect Performance and Signal Integrity

A further quote, to start with...

Roberto Zafalon, Low Power System Design mngr, STMicroelectronics
CLEAN-IP General Project Manager

Quote from CLEAN Press Release published by EETIMES on Jan 2006:

“*Semiconductor industry urges to overcome the technology shortcomings for 65nm and below, and in particular, process variability and unreliability, as well as leakage currents,*”

“*Industry needs to decrease power consumption of nanoelectronic devices, increase design productivity and thus make the raised SoC’s complexity manageable.*”
Why bothering for low power systems?

- **Practical market issue:**
  - Increasing market share of mobile, asking for longer cruising life
  - Limitations of battery technology

- **Economic issue:**
  - Reducing packaging costs and achieving energy savings

- **Technology issue:**
  - Enabling the realization of high-density chips
    (heat poses severe constraints to reliability)

Electronic Technology Today: CMOS Convergence

- CMOS technology dominates in modern ICs.
• The theoretical limit for transistor gate length on silicon is around 1.5nm.
  – Today’s 65nm CMOS process has a gate length of 42nm: i.e. **28X larger** than the theoretical limit!
  – In 32nm, the gate length is 21nm i.e. **14X above limit**
• The gate delay determines the fundamental speed of the logic. The theoretical limit is 0.04ps
  – Today’s 65nm logic NAND2 is ~1ps, i.e. **24X slower**!
• Transistor density, i.e. the number of device which can be squeezed into a chip, reaches the limit around 1.8 billion Tx per cm².
  – Today’s 65nm CMOS device is **7.5X larger**! (i.e. 750Kgate/mm² = 2.4M Tx/mm² = 240M Tx/cm²)
• Performance as measured by clock speed, fell off Moore’s Law during the last decade, thanks to Multi Processors computing architectures.

**Basics of CMOS Power Consumption**

• Power consumption of a CMOS gate: 
  \[ P = P_{SW} + P_{SC} + P_{Lk} \]
  where:
  • \( P_{SW} \) = Switching (or dynamic) power.
  • \( P_{SC} \) = Short-circuit power.
  • \( P_{Lk} \) = Leakage (or stand-by) power.
• In older technologies (0.25um and above), \( P_{Lk} \) was marginal w.r.t. switching power:
  – Switching power minimization was the primary objective.
• In deep sub-micron processes, \( P_{Lk} \) becomes critical:
  – Leakage accounts for around 5-10% of power budget at 180nm;
  this grows to 20-25% at 130nm and to 35-60% at 65nm.
Leakage Currents in Bulk CMOS

- $I_{\text{sub}}$: Subthreshold current.
- $I_{gs}, I_{gb}, I_{gd}$: Gate oxide tunneling.
- $I_{jbs}, I_{jbd}$: Junction reverse current.
- $I_{\text{GIDL}}, I_{\text{GIDL}}$: Gate induced D,S leakage.
- $I_{\text{i}}$: Impact ionization current.

Technology Scaling

- Smaller geometries
  - Higher device density:
    - Smaller gate capacitance, yet many more gates/chip
    - Higher switched capacitance $\Rightarrow$ Higher switching power.
  - Higher clock frequencies:
    - Higher switching power
  - Lower supply voltages:
    - Lower switching power, but also lower speed $\Rightarrow$ Lower threshold voltages $\Rightarrow$ Exponential leakage
- Consequence:
  - Power density increases as technology scales!
Squeezing costs of computing cores

**ARM 9**
180 nm
11.8 mm²

- 130 nm, 5.2 mm²
- 90 nm, 2.6 mm²
- 65 nm, 1.4 mm²
VDD (no more) scaling is increasing the «power crisis»

Evolution of VDD (LSTP)

Regular Decrease in 10 years
From 5V to 1.2V (x 0.7 per node)

1.2V plateau
1V plateau?

Power Trend for microprocessors

- Power density in Intel’s microprocessors:

Nuclear Reactor
Rocket Nozzle
Sun’s surface
### CMOS Logic Tech Overview

**Density vs technology**

![Graph showing density vs technology with labels for different technologies and power consumption.](image)

- **Source:** STMicroelectronics

---

### Gate total power

**Graph showing gate total power vs duty cycle.**

- **Source:** STMicroelectronics

---

- **Gate total power**
  - 65LP/G
  - 45GS
  - 32LP
  - 32GP

- **Duty Cycle [%]**
  - 1%
  - 5%
  - 10%
  - 20%
  - 50%
  - 90%

- **Total power per gate [nW]**
  - 0
  - 500
  - 1000
  - 1500
  - 2000
  - 2500
  - 3000
  - 3500
  - 4000

- **F=500MHz**
90/65/45nm Speed vs Leakage

Source: STMicroelectronics

Ioff/Ion for 32LP, 65LP and 65GP

Source: STMicroelectronics
Technology Scaling

- Increasing contribution of leakage power:
  - Example: ASICs [source: STMicroelectronics]
  - Example: Microprocessors [source: Intel].

SoC Requirements for MP platforms (1)

- Processing performance is expected to grow more than 200x in the next 15 years.
SoC Requirements for MP platforms (2)

- # PE per chip; Processing Performance; ND2’s max switching frequency (normalized to 2007)

![Graph showing power consumption over years]

MOS Capacity by Dimensions

![Bar chart showing MOS capacity by dimensions]

Dynamic vs. Leakage Power

Source: ITRS Roadmap

Possible trajectory for high-k dielectrics

Semiconductor’s Challenge

Moore’s Law at Work!
**Leakage crisis: Is it a technology issue only?**

- **Trends:**
  - nominal Vdd getting stable around 1V
  - MOS’s Vth linearly scales to keep constant speed
  - But… leakage grows exponentially with Vth reduction !!
  - sub-threshold current from 100 to 1000 pA/um
  - gate leakage to become larger than sub-threshold
  - total static power from 21E-12 to 60E-12 W/Transistor

- SOI has major disadvantages w.r.t. sub-threshold reduction!

---

**“Leakage Aware” design strategy includes**

**A. Gate/Circuit-level techniques**
- Use of multiple Vth
  - Dual-Vth design.
  - Mixed-Vth (MVT) CMOS design.
  - MTCMOS.
  - Sleep transistor insertion/Voltage islands
  - State retention FFs

**B. Techniques for memory circuits**
- Cell state (stored value) determines exactly which transistors “leak”
  - **State-preserving** techniques:
    - Only suitable choice for non-cache memories (e.g., scratchpad).
  - **State-destroying** techniques:
    - Suitable for caches (can invalidate values).

**C. Architectural techniques**
- Adaptive Body Biasing (ABB).
- Adaptive Voltage Scaling (AVS).
- Vth hopping.
- Multiple VBB
Low Leakage Memory Approaches

• Leakage reduction techniques can be broadly classified in terms of how memory state is managed:
  – **State-preserving** techniques:
    • Memory cell value is preserved when in low-leakage state.
    • Suitable choice for non-cache memories (e.g., scratch-pad).
  – **State-destroying** techniques:
    • Memory cell value is **NOT** preserved when in low-leakage state.
    • Suitable only for caches (can invalidate values).

• Tradeoff between:
  – Residual leakage paid to preserve the state.
  – Restoring the lost state from higher levels of the memory hierarchy.
Low Leakage Memory Approaches (cont.)

- **Circuit-level techniques:**
  - Modify internal structure of SRAM cells.
    - Transistor size, P/N ratio, $V_{th}$, body bias.
    - Additional transistors.
    - Precharge policy tuning
  - May possibly require specialized process (e.g., different $Tox$, Halo doping, multiple $V_{th}$).

- **Architectural techniques.**
  - Use system level information to determine conditions to drive portions of memory into low-leakage state.
  - Portions of Memory: bit lines, blocks, regions, etc.

Spatio-Temporal-Value Cache

- **Partitioned Architecture (Outcome of CLEAN):**
SoC Design Grand Challenges
(source: ITRS 2007)

• MANAGEMENT OF OVERALL POWER
  – Due to the Moore’s law, power management is the primary issue across most application segments.
  – Needs to be addressed across multiple levels, especially system, design, and process technology.

• MANAGEMENT OF LEAKAGE POWER
  – Leakage currents increase by 10x per tech node.
  – From system design requirements & improvements in CAD design tools, down to leakage and performance requirements for new architectures.

Subthreshold Leakage vs. Temperature

![Graph showing subthreshold leakage vs. temperature for different process nodes (0.10 μm, 0.13 μm, 0.18 μm, 0.25 μm). The graph plots leakage current (Ioff) in nA/μm against temperature in Celsius (Temp). The graph illustrates the increasing leakage current with higher temperatures and finer process nodes.](image)
Thermal map of a Multi Processor SoC

Chip floorplan

Steady state temperature

Some hot spots in steady state:
§ Silicon is a good thermal conductor (only 4x worse than Cu) and temperature gradients are likely to occur on large dies
§ Lower power density than on a high performance CPU (lower frequency and less complex HW)

Thermal Management Challenge

BGA Normalized cost vs. thermal enhancement

- **BGA package rough** (Cost-performance ÷ High-performance)
  - max power density = 50–60 W/cm²
  - Cost per pin = 0.25–1.1 ¢/pin (~ 90 pins/cm²)
  - Max pincount = 500–2500+
Increased share of Mobile Phone Subscribers

- Cellular Phones: GSM+CDMA
  - The fastest growing communication technology of all time.
- The billionth subscriber user was connected in Q1 2002

![Bar chart showing millions of subscribers over time from 1998 to Q2-08.]

Mobile Phones Regional Split at Q2-2008

- 3665 M subscribers as of Q2-2008
- Mobile Broadband Network (HSPA) subscribers has reached 50 M from 11 M on 2007 (i.e. 4 M/Month growth rate).

GSM Regional Statistics Q2-2008

![Pie chart showing regional distribution.]

- Asia Pac: 1547
- Europe West: 412
- Europe East: 280
- Americas: 323
- Africa: 413
- Middle East: 413
- USA/CND: 492
Cellular Phone’s standby current

Nomadik™: ST’s example of Mobile Multi-Media driver

STn8820
Nomadik™: a flagship design for ultra low power!

... and ... not Only Mobile!

- 20% of electrical energy consumed in Amsterdam is used for Telecom
- In the US, Internet is responsible for 9% of the electrical energy consumed nation-wide
  - This grows to 13% with all computer applications
- Transfering 2 MBytes of data through the internet consumes the energy of 1 pound of coal (1 pound=0.453 Kg)

Source: 2000 CO2 conference, Amsterdam, NL
Conclusion

• Semiconductor market is still CMOS dominated:
  – Switching and leakage power.
• Leakage will become dominant for technology nodes below 65nm.
  – Leakage power optimization must be addressed from both technology and design points of view.
• Many circuit-level techniques have been investigated recently:
  – Not yet fully supported by commercial EDA tools.
• Higher-level approaches are still in their infancy:
  – Results are promising.

• The electronics industry calls for a REVOLUTION!
Industry’s Needs

- Ultra low power systems
- Ultra low power cognitive radio
- Energy scavenging
- Micro-Nano systems
- System in Package
- System On Wafer
- ...