

Low Power Techniques for SoC Design: basic concepts and techniques

Estagiário de Docência
M.Sc. Vinícius dos Santos Livramento

Prof. Dr. Luiz Cláudio Villar dos Santos

Embedded Systems - INE 5439
Federal University of Santa Catarina

September, 2014

Outline

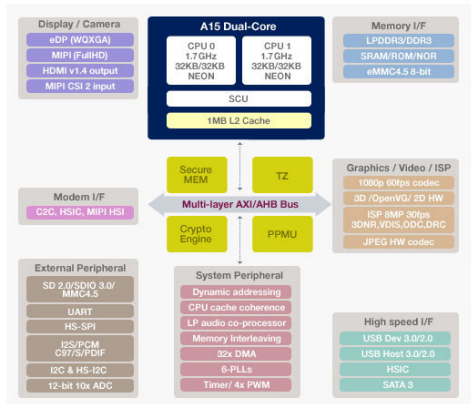
1. Motivation
2. Basic Concepts
 - Power vs. Energy
 - Dynamic and Static Power
 - Trends on Total Power Consumption
3. Standard Low Power Design Techniques
 - Clock Gating
 - Gate Level Optimization
 - Multi V_{th}
 - Multi V_{dd}
4. Advanced Low Power Design Techniques
 - Power Gating
 - Voltage and Frequency Scaling

Motivation



- ▶ Portable mobile devices (PMDs) comprise one of the fastest growing segments of the electronics market
- ▶ PMDs integrate a number of computationally-intensive functionalities
- ▶ Since PMDs are powered by batteries, energy is a major problem
- ▶ To tackle the energy issue a number of techniques are used throughout software and hardware design flow

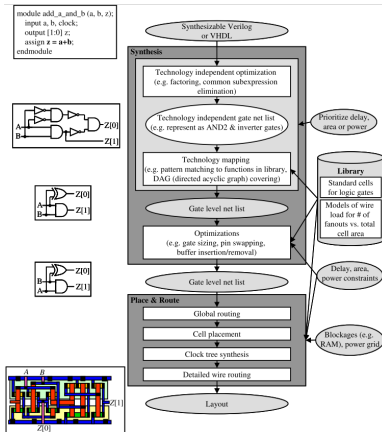
Motivation



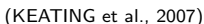
(SAMSUNG, 2014)

- ▶ PMDs are complex systems of hardware and software known as System-on-Chip (SoC)
- ▶ An example of contemporary SoC is the Samsung Exynos 5 Dual used by Google Nexus 10 and Samsung Galaxy Tab II
- ▶ The Exynos 5 Soc is implemented in CMOS 32 nm and comprises 2x ARM Cortex-A15 processor and others complexes blocks
- ▶ This course focuses on low power design techniques for embedded hardware

Embedded hardware design flow

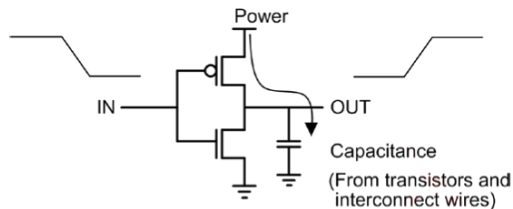


- ▶ The embedded hardware design flow is based on libraries of pre-characterized gates known as standard cell libraries
- ▶ It starts from a RTL description and ends up with a layout ready for manufacturing
- ▶ Several steps are performed (some iteratively) so as to achieve the design functional and non-functional objectives as area, delay and power



- ▶ Delay (s)
 - ▶ Performance metric
- ▶ Energy ($Joule$)
 - ▶ Efficiency metric: effort to perform a task
- ▶ Power (J/s or $Watt$)
 - ▶ Energy consumed per unit time
- ▶ Power Density (W/cm^2)
 - ▶ Power dissipated per unit of area

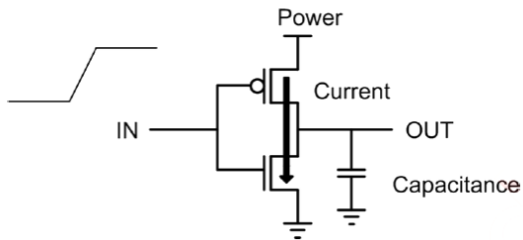
Dynamic (switching) Power



(KEATING et al., 2007)

- ▶ Energy / transition (J)
 - ▶ $C_L \times V_{dd}^2$ consumed from source
 - ▶ $\frac{1}{2} \times C_L \times V_{dd}^2$ dissipated during output transition $0 \rightarrow 1$
 - ▶ $\frac{1}{2} \times C_L \times V_{dd}^2$ dissipated during output transition $1 \rightarrow 0$
- ▶ P_{dyn} (W)
 - ▶ $\frac{1}{2} \times C_L \times V_{dd}^2 \times f_{clock} \times \alpha$
- ▶ Switching activity (α)
 - ▶ $0 \leq \alpha \leq 1$

Dynamic (short circuit) Power



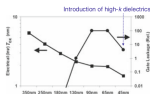
(KEATING et al., 2007)

- ▶ Energy / transition (J)
 - ▶ Short circuit power occurs when both the NMOS and PMOS transistors are on
 - ▶ $P_{sc} = t_{sc} \times V_{dd} \times I_{peak} \times f_{clock} \times \alpha$
 - ▶ t_{sc} is the time duration of the short circuit current
 - ▶ I_{peak} is the total switching current
 - ▶ As long as the ramp time (slew) of the input signal is kept short, the short circuit current occurs for only a short time during each transition

Static (leakage) Power

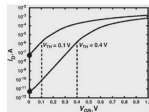


Scaling leads to gate-oxide thickness of a couple of molecules



Causes gates to leak!

four orders of magnitude



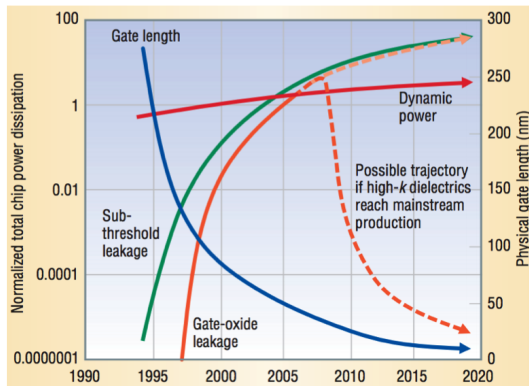
Leakage: sub-threshold current for $V_{GS} = 0$



- ▶ Transistors are imperfect switches
- ▶ Main sources of static power are gate and sub-threshold leakage
- ▶ Gate leakage
 - ▶ Tunneling currents through thin gate oxide (SiO_2)
- ▶ Sub-threshold leakage
 - ▶ Current that flows from drain to source when transistor is off
 - ▶ $I_{sub} = \mu C_{ox} V_t^2 \frac{W}{L} \cdot e^{\frac{V_{gs} - V_{th}}{nV_t}}$
 - ▶ Threshold voltage v_{th} depends exponentially on $V_{gs} - V_{th}$

(RABAEY, 2009)

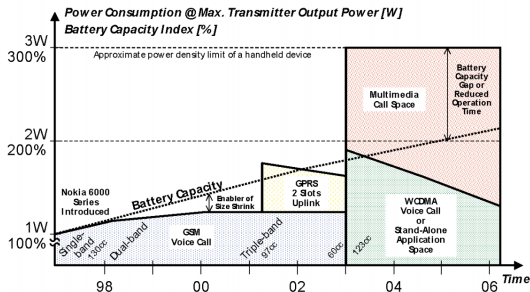
Trends on Total Power Consumption



(KIM et al., 2003)

- ▶ Dynamic power slightly increases
 - ▶ Power per transistor has reduced
 - ▶ Number of transistor in a chip has increased
- ▶ Gate leakage increases exponentially
 - ▶ Controlled through the use of high-k transistors from 45nm on
- ▶ Sub-threshold leakage
 - ▶ Threshold voltage v_{th} depends exponentially on $V_{gs} - V_{th}$
 - ▶ $V_{gs} - V_{th}$ has reduced in recent technologies
 - ▶ Multi V_{th}

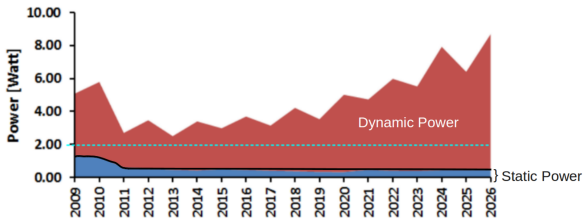
Trends on Power Requirements for Mobile on 2004



(NEUVO, 2004)

- ▶ There is a gap between battery capacity and power consumption
- ▶ Power consumption limit fixed: 3W

Trends on Power Requirements for Mobile on 2011



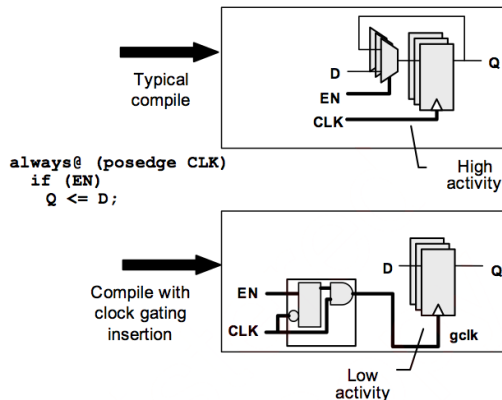
(CARBALLO; B., 2011)

- ▶ A SoC with 48.8M logic gates using low-power techniques dissipates 3.5W in 2011
- ▶ In 2026 the number of gates grows to 1995.5M and the power increases to 8.22W
- ▶ Power consumption limit reviewed: fixed at 2W until 2026
- ▶ **LOW POWER DESIGN TECHNIQUES ARE OF UTMOST IMPORTANCE!**

Low Power Design Techniques

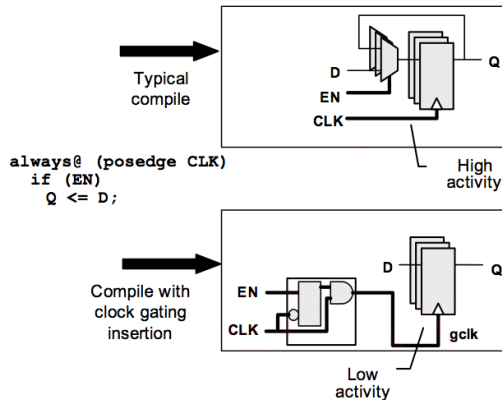
- ▶ Dynamic Power Reduction
 - ▶ Clock gating
 - ▶ Gate Level Optimization
- ▶ Static Power Reduction
 - ▶ Multi V_{th}
- ▶ Total Power Reduction
 - ▶ Multi V_{dd}
 - ▶ Power gating
 - ▶ Dynamic Voltage and Frequency Scaling

Impact of Clock Gating



- ▶ $P_{dyn} : \frac{1}{2} \times C_L \times V_{dd}^2 \times f_{clock} \times \alpha$
- ▶ 50% or more dynamic power can be spent in the clock tree buffers since they have high switching activity
- ▶ A significant ammount of dynamic power is dissipated by flip-flops
- ▶ Clock gating turns clocks to idle modules resulting in ZERO activity

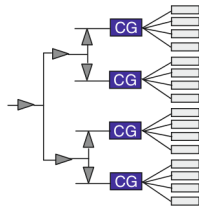
Clock Gating Within the Synthesis Flow



- ▶ Most standard cell libraries include clock gating cells
- ▶ Modern design tools support automatic clock gating e.g., Synopsys Design Compiler
- ▶ Small area overhead
- ▶ No change to RTL is required to implement clock gating
- ▶ Clock gating is inserted without changing the logic function

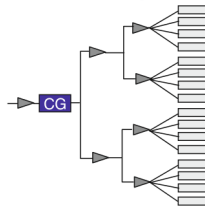
Clock Gating Within the Synthesis Flow

Power savings



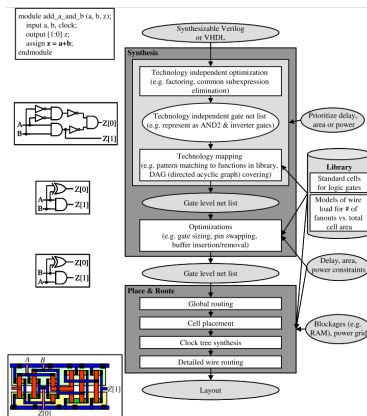
(KEATING et al., 2007)

Simpler skew management, less area



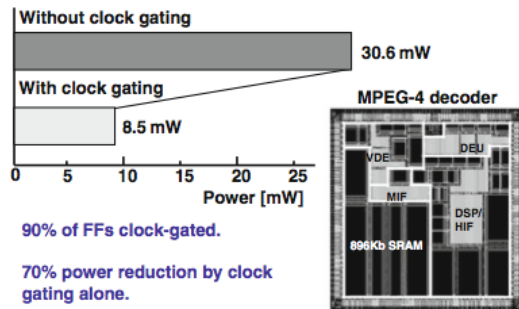
- ▶ Clock tree consumes a lot of dynamic power
- ▶ Trade off between fine and coarse-grain clock gating
 - ▶ Fine-grain allows for turning off specific blocks. It comes at the expense of more area and skew
 - ▶ Coarse-grain allows for higher power savings due to clock buffers. On the other hand, modules cannot be turned off as often

Clock Gating Within the Synthesis Flow



(CHINNERY; KEUTZER, 2008)

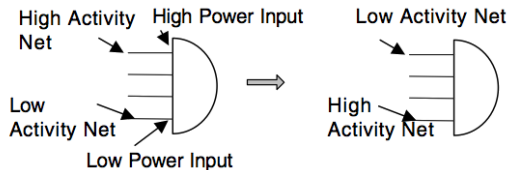
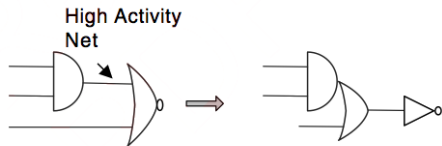
Exampe of Clock Gating



- ▶ Use of clock gating on an MPEG4 decoder
- ▶ Gating 90% of flip-flops
- ▶ From 30.6mW to 8.5mW: 70% of dynamic power reduction

(RABAEY, 2009)

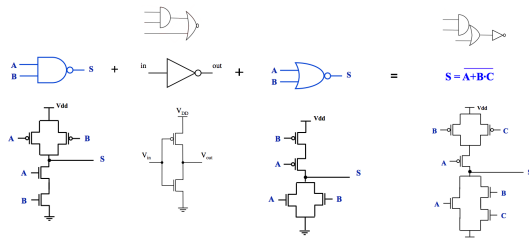
Impact of Gate Level Optimization



- ▶ A number of logic optimizations are performed during the design flow
- ▶ Modern design tools (e.g., Synopsys Design Compiler) perform a number of logic optimization so as to optimize area, power or delay
- ▶ Example of techniques are: Technology mapping, logic restructuring, gate sizing and buffer

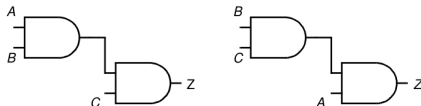
(KEATING et al., 2007)

Technology Mapping and logic restructuring



- ▶ An AND gate with high activity followed by a nor gate can be replaced by a complex AND-OR gate plus an inverter
- ▶ Total number of transistors reduced from 10 to 6
- ▶ Complex gates present intrinsic capacitances substantially smaller than inter-gate routing capacitances of a network of simple gates.
- ▶ A smaller output capacitance reduces the gate dynamic power

Reducing Switching Activity



$$P_{(A=1)} = 0.5$$

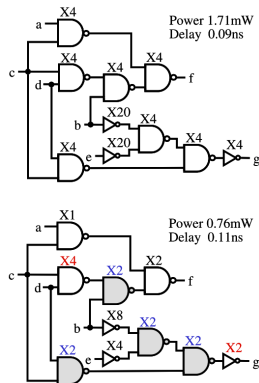
$$P_{(B=1)} = 0.2$$

$$P_{(C=1)} = 0.1$$

(RABAEY; CHANDRAKASAN; NIKOLIC, 2002)

- ▶ Input reordering can effectively reduce switching activity (for pins with high transition rate) and thereby dynamic power
- ▶ Even though the activity at pin Z is the same for both cases, a simply reordering of inputs can reduce switching activity by $\approx 78\%$
- ▶ In the first circuit, activity is equal $(1 - 0.5 \times 0.2)(0.5 \times 0.2) = 0.09$
- ▶ In the second circuit, activity is equal $(1 - 0.2 \times 0.1)(0.2 \times 0.1) = 0.0016$

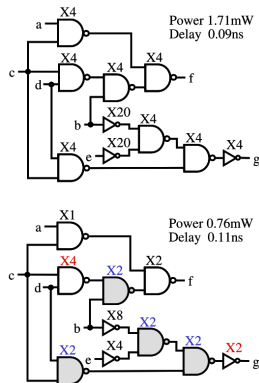
Gate Sizing and Buffer Insertion



- ▶ Gate sizing is an important optimization technique used for different objectives. The idea is to select the size (increase or decrease drive strength) of the gates so as to reduce the delay on critical paths or reduce power on non critical paths
- ▶ In buffer insertion, the tool can insert buffers rather than increasing the drive strength of the gate itself

(CHINNERY; KEUTZER, 2008)

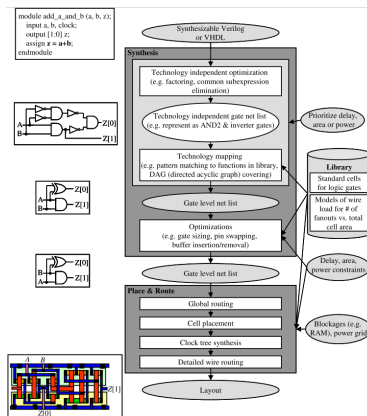
Gate Sizing and Buffer Insertion



- ▶ For example, suppose a target delay of 0.11 ns. Since the initial synthesis achieved a delay of 0.11 ns, some gates can be sized to reduce power
- ▶ In the example, simply sizing 4 gates reduced power by 55%

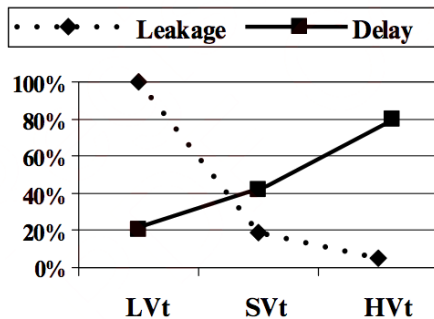
(CHINNERY; KEUTZER, 2008)

Gate Level Optimization Withing the Synthesis Flow



(CHINNERY; KEUTZER, 2008)

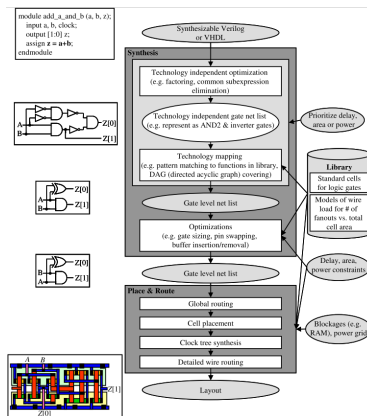
Impact of Multi V_{th}



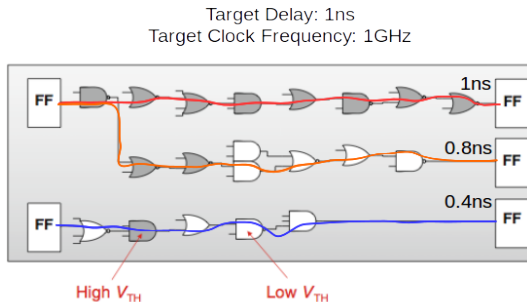
(KEATING et al., 2007)

- ▶ $I_{sub} = \mu C_{ox} V_t^2 \frac{W}{L} \cdot e^{\frac{V_{gs} - V_{th}}{nV_t}}$
- ▶ $I_{ds} = \mu C_{ox} \frac{W}{L} \cdot \frac{(V_{gs} - V_{th})^2}{2}$
- ▶ Sub-threshold leakage depends exponentially on V_{th}
- ▶ Delay has a much weaker dependence on V_{th}
- ▶ Standard cell libraries offer two or three versions of cells with different V_{th}
- ▶ Modern design tools support automatic V_{th} assignment e.g., Synopsys Design Compiler

Multi V_{th} Withing the Synthesis Flow



Example of Multi V_{th}



(RABAEY, 2009)

- ▶ In a circuit, different paths have different delays.
- ▶ A positive slack means that the signal is ready at input of FF before the target delay
- ▶ One approach would be synthesize for high performance using low V_{th} gates and then swapping non-critical gates for high V_{th}
- ▶ Another approach would be synthesize for low power using high V_{th} gates and then swapping critical gates for low V_{th}

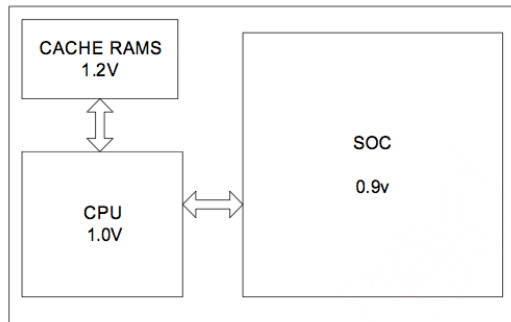
Example of Multi V_{th}

	High- V_{TH} Only	Low- V_{TH} Only	Dual- V_{TH}
Total Slack	-53 ps	0 ps	0 ps
Dynamic Power	3.2 mW	3.3 mW	3.2 mW
Static Power	914 nW	3873 nW	1519 nW

(RABAEY, 2009)

- ▶ Experiment performed jointly by Toshiba and Synopsys to evaluate the impact of two different V_{th}
- ▶ Using only high- V_{th} degrades the performance
- ▶ Using only low- V_{th} increases static power 4.2x w.r.t. high- V_{th}
- ▶ The dual- V_{th} strategy leaves timing and dynamic power unchanged, while reducing the static power by half w.r.t low- V_{th}

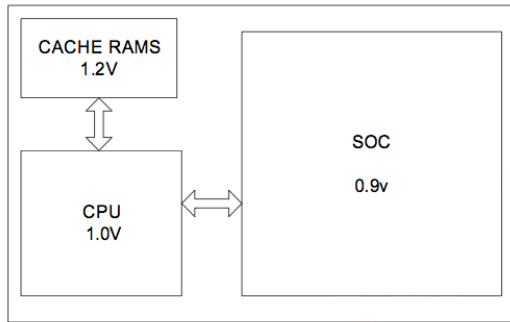
Impact of Multi V_{dd}



(KEATING et al., 2007)

- ▶ $P_{dyn} = \frac{1}{2} \times C_L \times V_{dd}^2 \times f_{clock} \times \alpha$
- ▶ $I_{ds} = \mu C_{ox} \frac{W}{L} \cdot \frac{(V_{gs} - V_{th})^2}{2}$
- ▶ Reducing V_{dd} provides a quadratic reduction on P_{dyn} but increases the delay of gates
- ▶ Reduce V_{dd} on non-critical paths
- ▶ Power benefit without compromising the system performance

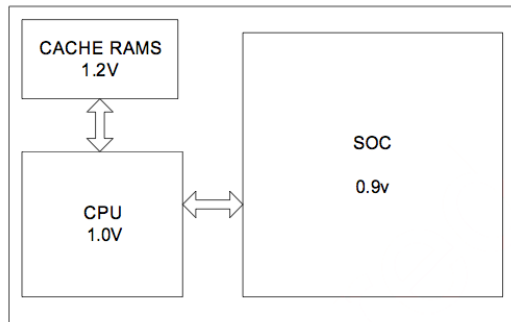
Impact of Multi V_{dd}



(KEATING et al., 2007)

- ▶ One challenge of power gating is interfacing signals between blocks
 - ▶ A signal from a low- V_{dd} to a high- V_{dd} block may cause a very slow transition thereby resulting in large short-circuit currents
 - ▶ A logic '1' in low- V_{dd} may not be enough to achieve '1' in high- V_{dd}
- ▶ To overcome such problems, level shifters are placed between the voltage islands
- ▶ Contemporary standard cell libraries offer level shifters cells

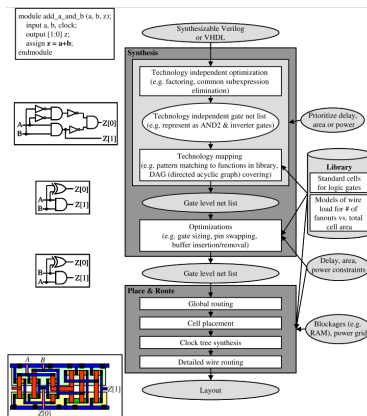
Impact of Multi V_{dd}



(KEATING et al., 2007)

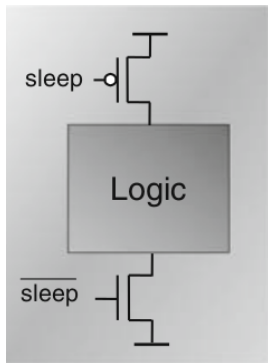
- ▶ Multi V_{dd} is not performed automatically by modern design tools
- ▶ The choice of voltage islands, as well as insertion of must be decided by the designers
- ▶ Other challenges of Multi V_{dd} include power planning, timing analysis, voltage regulators, etc

Multi V_{dd} Withing the Synthesis Flow



(CHINNERY; KEUTZER, 2008)

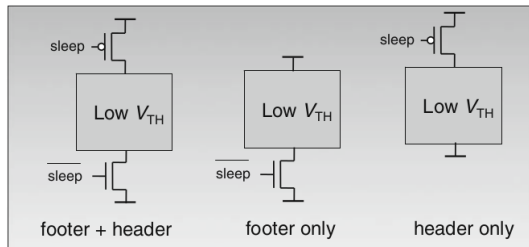
Impact of Power Gating



(RABAEY, 2009)

- ▶ $P_{stat} = \mu C_{ox} V_t^2 \frac{W}{L} \cdot e^{\frac{V_{gs} - V_{th}}{nV_t}}$
- ▶ $P_{dyn} = \frac{1}{2} \times C_L \times V_{dd}^2 \times f_{clock} \times \alpha$
- ▶ Static power
 - ▶ Dissipated even on standby or sleep mode
 - ▶ Even more important on battery powered portable devices
- ▶ Turning off the power of an idle block reduces static power to \approx ZERO

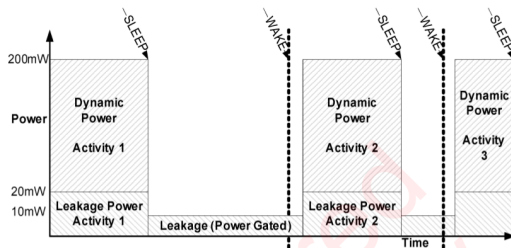
Impact of Power Gating



(RABAEY, 2009)

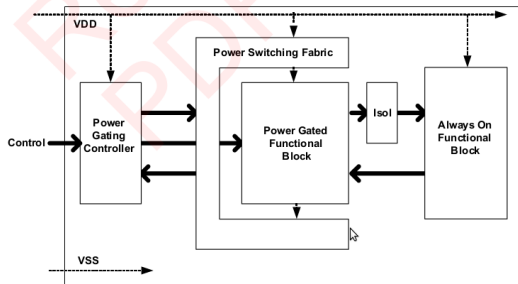
- ▶ The idea is to use on-off switches to disconnect the module from the supply rails
 - ▶ Header (PMOS) transistor is connected to V_{dd}
 - ▶ Footer (NMOS) transistor is connected to GND and is more area-efficient than PMOS
 - ▶ Using both is more effective in reducing leakage since it exploits stacking effect independently of the input patterns

Impact of Power Gating



(KEATING et al., 2007)

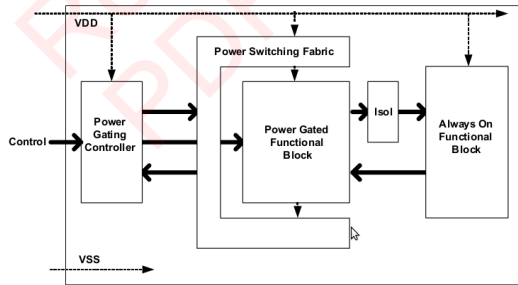
- ▶ Activity profile for a sub-system using power gating
- ▶ Some information must be retained during sleep
 - ▶ Some flip-flops must retain data during sleep mode
 - ▶ After wakeup the data must be restored
 - ▶ There is a leakage overhead to retain data during sleep



(KEATING et al., 2007)

- ▶ A simplified view of an SoC that uses internal power gating
- ▶ In this example only V_{dd} is switched, whereas GND is directly provided to the entire chip
- ▶ The power gating controller controls switches that provide power to the power gated block

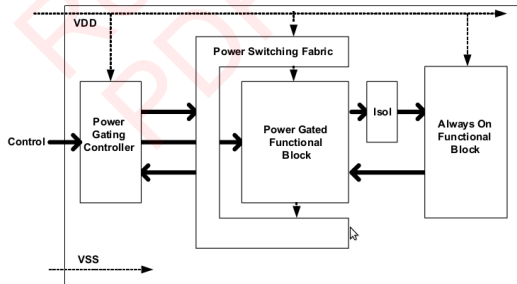
Example of Power Gating



(KEATING et al., 2007)

- ▶ One challenge of power gating is interfacing signals between blocks
- ▶ The signal from/to a power up/down block must be isolated
- ▶ To overcome such problems, isolation cells are placed between the blocks
- ▶ Contemporary standard cell libraries provide isolation cells

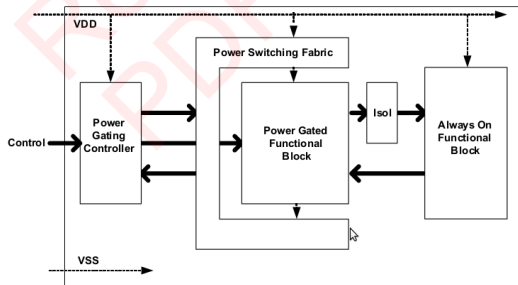
Example of Power Gating



(KEATING et al., 2007)

- ▶ Another challenge of power gating is how to retain the internal state of the block during power down/up
 - ▶ It is common to use retention registers to store the internal state
 - ▶ The choice of a retention strategy is crucial to determine the amount of time to power down/up, as well as the leakage consumption during sleep mode
- ▶ Contemporary standard cell libraries provide retention registers

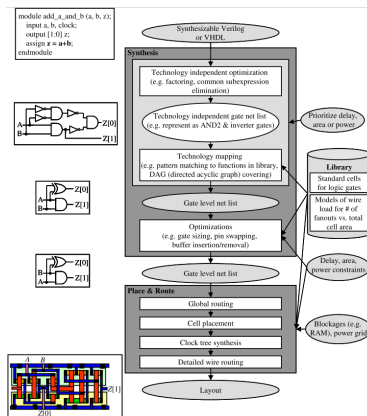
Example of Power Gating



(KEATING et al., 2007)

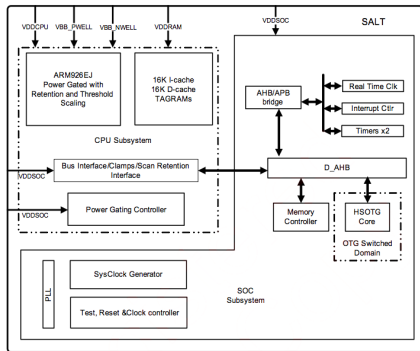
- ▶ In practice, power gating is a challenging task during the design flow
- ▶ Designers must clearly define which blocks can be powered down as well as define the power up and power down sequence
- ▶ The state retention plan must be carefully studied
- ▶ Modern design tools do not place automatic isolation cells nor insert retention registers

Power Gating Withing the Synthesis Flow



(CHINNERY; KEUTZER, 2008)

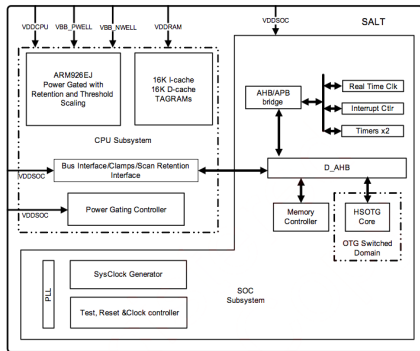
A Real Use Case Example of Power Gating: Salt 90nm



- ▶ Implemented in 90nm and contains an ARM processor
- ▶ The idea is to evaluate the impact on power and performance of using:
 - ▶ Clock gating
 - ▶ Power gating
 - ▶ Different V_{dds}
 - ▶ Different clock frequencies

(KEATING et al., 2007)

The Salt SoC

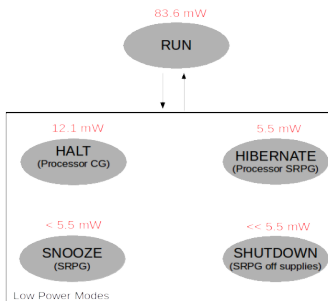


(KEATING et al., 2007)

► The project uses four low-power modes

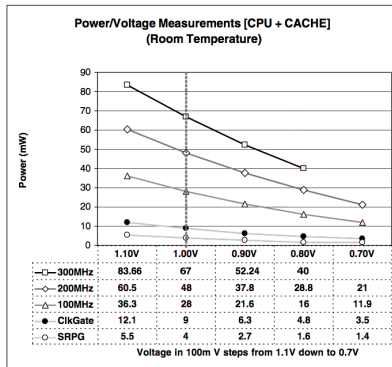
- Halt turns off the clocks to the processor
- Snooze turns off the internal power supply to the processor with state retention, but cache memories remain powered up. This mode allows fast power up
- Hibernate turns off the external power supply to the processor but cache memories remain powered up
- Shutdown turns off the external power supply to the processor and caches

Power State Machine for Salt



- ▶ The project uses four low-power modes
 - ▶ Halt turns off the clocks to the processor
 - ▶ Snooze turns off the internal power supply to the processor with state retention, but cache memories remain powered up. This mode allows fast power up
 - ▶ Hibernate turns off the external power supply to the processor but cache memories remain powered up
 - ▶ Shutdown turns off the external power supply to the processor and caches

Measurement and Analysis Post-Silicon of Salt Project

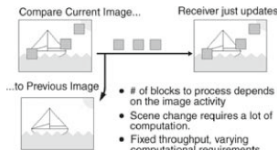


(KEATING et al., 2007)

- ▶ Evaluate power at different operation modes
- ▶ Nominal V_{dd} is 1.0v.
- ▶ V_{dd} steps in 10%: from 110% to 70%
- ▶ Three different clock frequencies being the nominal 300MHz
- ▶ The first three measurements show the dynamic power for different frequencies/ V_{dds}
- ▶ Clock gate and Save Restore Power Gating measurements show the leakage power

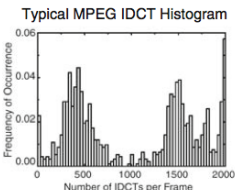
Dynamic Voltage and Frequency Scaling

Example: Video Compression



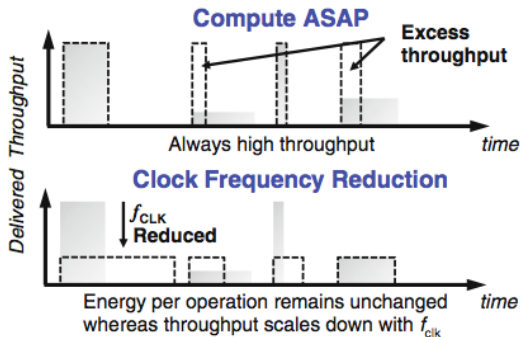
True also for voice processing, graphics, multimedia, and communications

(RABAEY, 2009)



- ▶ Workload can vary a lot over time
- ▶ For instance, the motion compensation block of a video compression that computes how much a video frame differs from the previous one
 - ▶ A fast moving car chase scene has a lot of computation
 - ▶ A nature landscape varies little over time
- ▶ The IDCT histogram shows that the computational effort can vary 2-3 orders of magnitude

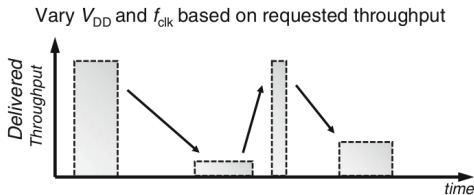
Dynamic Voltage and Frequency Scaling



(RABAEY, 2009)

- ▶ Adjusting only the frequency reduces power but leaves the energy per operation unchanged
- ▶ Therefore, the amount of work that can be performed by the battery remains the same
- ▶ A more effective way of exploiting the workload variation is to adjust simultaneously frequency and supply voltage

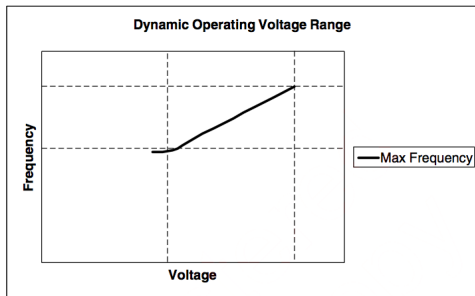
Dynamic Voltage and Frequency Scaling (DVFS)



(RABAEY, 2009)

- ▶ Impact on dynamic and static power
 - ▶ $P_{dyn} : \frac{1}{2} \times C_L \times V_{dd}^2 \times f_{clock} \times \alpha$
 - ▶ $I_{sub} = \mu C_{ox} V_t^2 \frac{W}{L} \cdot e^{\frac{V_{gs}-V_{th}}{nV_t}}$
 - ▶ $I_{ds} = \mu C_{ox} \frac{W}{L} \cdot \frac{(V_{gs}-V_{th})^2}{2}$
- ▶ The idea is to dynamically adjust the voltage and frequency of block according to the workload
- ▶ Dynamic Voltage and Frequency Scaling has a set of voltage and frequency values that are dynamically switched
- ▶ DVFS not only reduces power but reduces the energy per operation as well

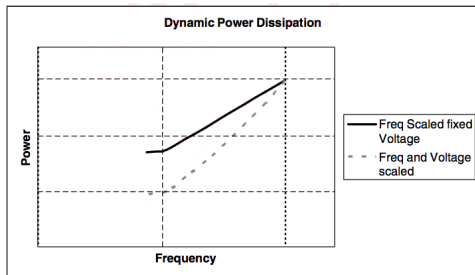
Voltage and Frequency Scaling Opportunity using DVFS)



(KEATING et al., 2007)

- ▶ $P_{dyn} : \frac{1}{2} \times C_L \times V_{dd}^2 \times f_{clock} \times \alpha$
- ▶ There is a region of operation where frequency increases monotonically over voltage within some limits
 - ▶ Maximum voltage specified for the technology (process)
 - ▶ Minimum voltage in which the circuitry runs safe
- ▶ Therefore, the designer can explore different pairs (V_{dd}, f_{clock}) during design time

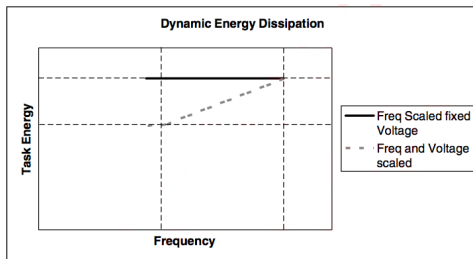
Power and Energy Reduction Opportunities using DVFS)



(KEATING et al., 2007)

- ▶ $P_{dyn} : \frac{1}{2} \times C_L \times V_{dd}^2 \times f_{clock} \times \alpha$
- ▶ There is a different power dissipation relationship between reducing frequency with and without reducing supply voltage
- ▶ The gap between the two curves equals the power saving achievable between the minimum and maximum operating voltages
- ▶ DVFS allows more than a linear power reduction

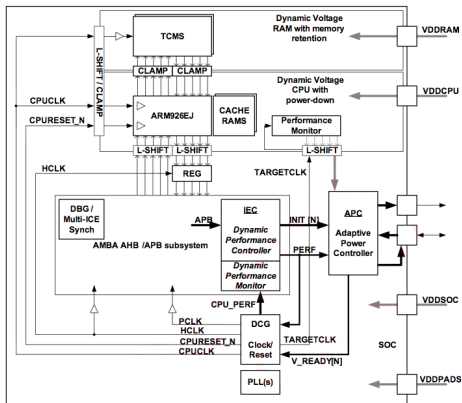
Power and Energy Reduction Opportunities using DVFS)



(KEATING et al., 2007)

- ▶ $P_{dyn} : \frac{1}{2} \times C_L \times V_{dd}^2 \times f_{clock} \times \alpha$
- ▶ Energy is the integration of power over the time taken to complete a task
- ▶ Ignoring leakage, reducing frequency at half, halves the dynamic power but takes twice as long to complete the task
- ▶ Scaling voltage reduces quadratically the dynamic power, allowing to reduce energy as well

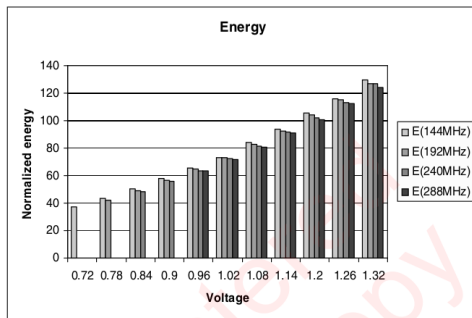
A Real Use Case Example of DVFS: ULTRA926 130nm



- ▶ A chip using an ARM processor developed in conjunction with Synopsys
- ▶ It uses a nominal supply voltage of 1.2V
- ▶ The idea is to evaluate the energy savings of applying DVFS
- ▶ Other low power techniques are also used such as power gating and clock gating

(KEATING et al., 2007)


A Real Use Case Example of DVFS: ULTRA926 130nm





(KEATING et al., 2007)


- ▶ Histogram plotting the energy consumption data for different pairs (V_{dd} , f_{clock})
- ▶ From 60% to 110% of max 1.2V.
From 50% to 100% of max 288MHz
- ▶ Some pairs fail to attend the required performance e.g., (0.72V, 192MHz), (0.78V, 240MHz)
- ▶ Note only scaling frequency results in almost the same energy value
- ▶ It is possible to observe a close-to-linear energy relationship for different voltage ranges


References I


 CARBALLO, J.-A.; B., K. A. Itrs chapters: Design and system drivers. In: *Future Fab International (36)*. [S.l.: s.n.], 2011. p. 45–48.

 CHINNERY, D.; KEUTZER, K. *Closing the power gap between ASIC & custom: tools and techniques for low power design*. [S.l.]: Springer, 2008.


 KEATING, M. et al. *Low power methodology manual: for system-on-chip design*. [S.l.]: Springer Publishing Company, Incorporated, 2007.


 KIM, N. S. et al. Leakage current: Moore's law meets static power. *Computer*, v. 36, p. 68–75, December 2003.

 NEUVO, Y. Cellular phones as embedded systems. In: IEEE. *Solid-State Circuits Conference, 2004. Digest of Technical Papers. ISSCC. 2004 IEEE International*. [S.l.], 2004. p. 32–37.

 RABAEY, J. *Low power design essentials*. [S.l.]: Springer, 2009.

References II

 RABAEY, J. M.; CHANDRAKASAN, A. P.; NIKOLIC, B. *Digital integrated circuits*. [S.l.]: Prentice hall Englewood Cliffs, 2002.

 SAMSUNG. *Exynos 5 Dual*. 2014. Disponível em: <http://www.samsung.com/global/business/semiconductor/product/application/detail?productId=7668>.

Low Power Techniques for SoC Design: basic concepts and techniques

Estagiário de Docência
M.Sc. Vinícius dos Santos Livramento

Prof. Dr. Luiz Cláudio Villar dos Santos

Embedded Systems - INE 5439
Federal University of Santa Catarina

September, 2014