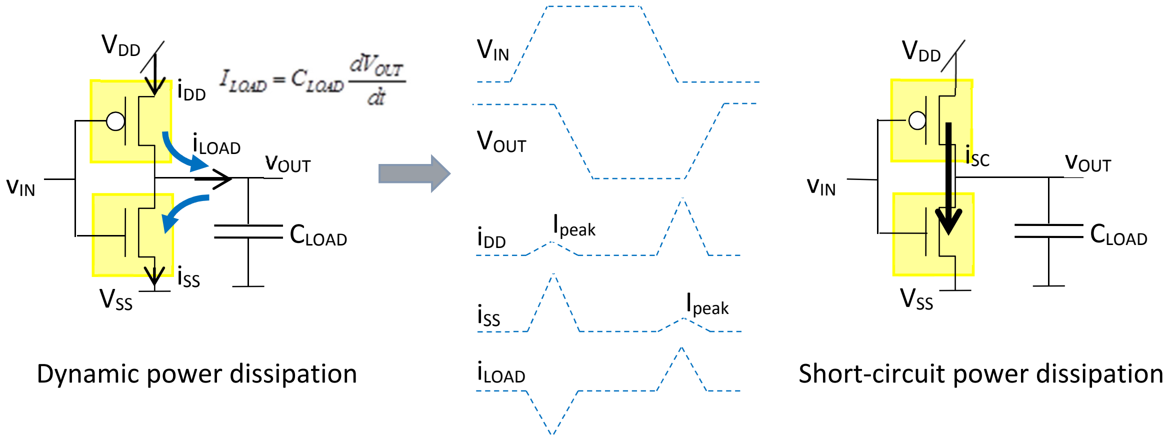
Chapter 7: Power dissipation and low-power design

In this chapter we will discuss the two main modes of power dissipation in CMOS circuits, i.e. dynamic and static power. Dynamic power is dissipated when circuit nodes are switching, while static power is dissipated in idle circuits due to leakage currents. After a short discussion of these basic modes of power dissipation, we will review some of the most common techniques for low-power design.

A. Dynamic power dissipation

A complex CMOS circuit contains billions of capacitive circuit nodes, nodes that are charged and discharged during switching thereby dissipating power. All these node capacitors are charged through p-channel devices in the pull-up network, and discharged through n-channel devices in the pull-down network. This is illustrated in Fig. 7.1 for the case of an inverter. In the middle of Fig. 7.1, the delayed output voltage response, V_{OUT}, with respect to the input voltage, V_{IN}, is shown together with the current through the p-channel device, I_P, the current through the n-channel device, I_N, and the current through the load capacitor, I_{LOAD}. The current flow illustrated in these graphs provides the most basic insight into how power is dissipated when a capacitor is charged and discharged between the voltage supply rails.



Dynamic power dissipation

Short-circuit power dissipation

Fig. 7.1. Dynamic power dissipation.

The energy per low-to-high transition, when a load capacitor C_L is charged to the supply voltage V_{DD} through a p-channel pull-up MOSFET, is given by

$$Energy/transition = \int_{0}^{T} i_{DD}(t)V_{DD}dt = CV_{DD}\int_{0}^{V_{DD}} dV = CV_{DD}^{2} \qquad (7.1)$$

where T is the cycle time. We can then describe the dynamic power as the energy per transition multiplied by the number of low to high transitions, i.e. the switching frequency, f_{clock}. By adding all the capacitances that are being charged, we obtain an effective capacitance, C_{eff}, and the following model for the dynamic power dissipation:

$$P_{dynamic} = f_{clock}C_{eff}V_{DD}^{2}. \qquad (7.2)$$

As an example, a 10 W integrated circuit operating at 1 GHZ with a supply voltage of one volt has an effective capacitance of 10 nF. Half of the power needed to charge the capacitor is wasted in the pull-up device, while the other half is stored in the capacitor as can be seen from this derivation,

$$\text{Energy stored in capacitor C} = \int_0^T i_{DD}(t) V dt = C \int_0^{V_{DD}} V(t) dV = \frac{1}{2} C V_{DD}^2 . \tag{7.3}$$

When the output goes low, the energy that was stored in the capacitor is dissipated in the pull-down device. Besides the current, i_{LOAD}, needed to charge the capacitor there is also a small “short-circuit” current, i_{SC}, flowing between the V_{DD} and V_{SS} rails during a short time, t_{SC}, of the switching period when both pull-up and pull-down networks are conducting at the same time. Assuming that the average “short-circuit” current is half of its peak value, I_{peak}, and since the logic gate is “shorted” twice per cycle, the “short-circuit” energy loss per cycle is given by

$$\text{"Short circuit" energy loss} = \int_0^T i_{SC}(t) V_{DD} dt \approx I_{peak} V_{DD} t_{SC} . \tag{7.4}$$

There are models available to estimate both the switching time, i.e. the time while the input signal changes from V_{TN} to V_{DD}+V_{TP}, or vice versa, and to relate the peak current to the maximum current, I_{ON}. Typically, I_{peak} is in the range of 10-20% of I_{ON}. As stated above, the short-circuit power is typically less than 10% of the total dynamic power.

As we will discuss later, methods for low power design include reducing the clock frequency, the effective capacitance, and the supply voltage. As we shall see, there is a variety of architectural and logic design techniques for minimizing switching activity.

B. Static power dissipation due to leakage

Power is dissipated also when a chip is not switching. Power dissipated due to MOSFET leakage currents is called static power. As illustrated in Fig. 7.3, there are three main sources of static leakage currents in a CMOS logic gate:

- Sub-threshold leakage from drain to source in a MOSFET operating in weak inversion.
- Gate leakage currents flowing through the insulator to the substrate due to tunneling through the insulator barrier. In modern technologies, high-k dielectric materials have replaced traditional silicon dioxide insulators to keep gate leakage in check. This appears to be the only effective way of reducing gate leakage.
- Drain junction leakage currents flowing from drain to substrate through the reverse-biased drain junctions.

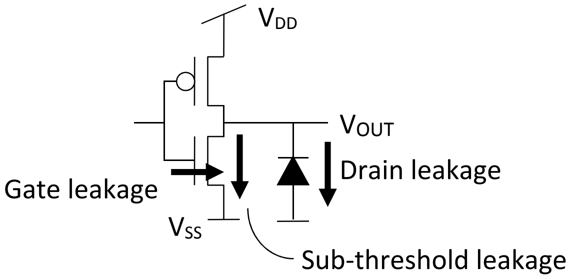


Fig. 7.2. Static power dissipation.

The most important leakage current is the subthreshold current occurring when a CMOS gate is not fully turned off. The current-voltage characteristic shown in Fig. 7.3 indicates that if we scale the threshold voltage along with the supply voltage to maintain the gate voltage overdrive, the leakage current I_{OFF} increases. A variety of techniques exist to minimize static leakage, including turning power off to idle blocks, so called power gating, using high-V_T cells wherever performance goals allow in non-critical timing paths, and to exploit the stacking effect to be explained in a minute. The relationship between subthreshold leakage and delay for different threshold voltages is illustrated in Fig. 7.4. This figure shows that leakage can be reduced by increasing the threshold voltage, but the price to pay for lower leakage is longer gate delays due to reduced gate voltage overdrives.

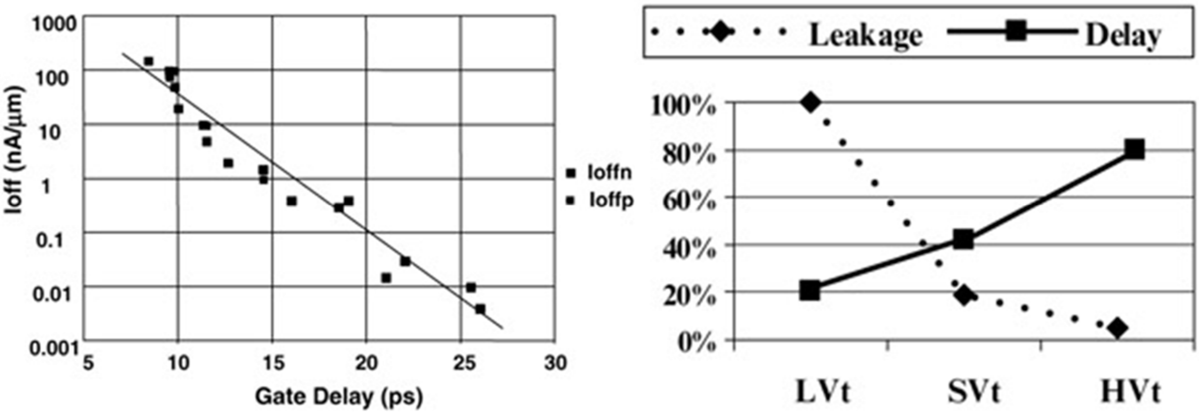


Fig. 7.4. Relationship between subthreshold leakage and delay for different threshold voltages.

Before proceeding to discuss low power design, let us just mention the stacking effect. As discussed before, when two series MOSFETs are ON, the channel lengths effectively add thereby reducing the ON current to half. The internal node voltage is a few tenth of a volt, with most of the voltage drop across the top saturated MOSFET. The reason for the ON current decrease is the decrease of the gate-source voltage of the top MOSFET due to V_X. The effect occurs, but with more dramatic current reductions, when the two series MOSFETs are OFF. Due to the drain-induced barrier-lowering (DIBL) effect[6], the subthreshold leakage current can be written,

$$I_{sub} = I_{OFF} e^{\frac{\sigma(V_x - V_{DD})}{nV_t}} = I_{OFF} 10^{\frac{\sigma(V_x - V_{DD})}{S}}, \tag{7.5}$$

where I_{sub}=I_{OFF} for V_X=V_{DD} according to definition, and where the subthreshold swing S=nV_tln10. Since the two subthreshold currents through MOSFETs N1 and N2 must be equal, we obtain

$$I_{sub} = \underbrace{I_{off} 10^{\frac{\sigma(V_x - V_{DD})}{S}}}_{N1} = \underbrace{I_{off} 10^{\frac{-V_x + \sigma((V_{DD} - V_x) - V_{DD})}{S}}}_{N2}, \tag{7.6}$$

an equation that yields an internal node voltage

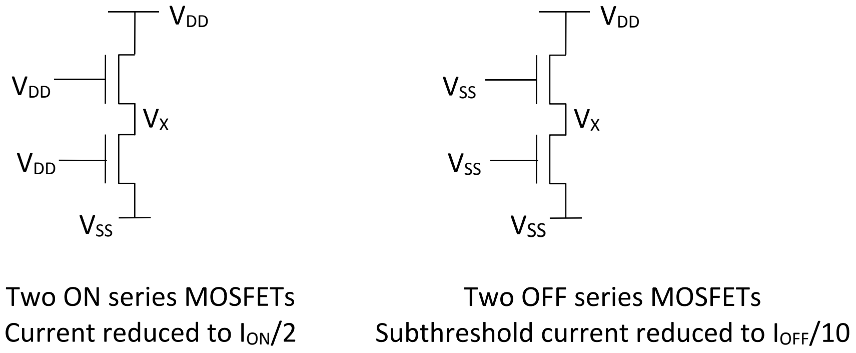
$$V_x = \frac{\sigma V_{DD}}{1 + 2\sigma} \approx \sigma V_{DD} \approx 0.1 \text{ V}, \tag{7.7}$$

[6] Due to drain-induced barrier-lowering, the threshold voltage decreases with increasing drain to source voltages, $V_T = V_{TO} - \sigma V_{DS}$.

for a typical V_T reduction of 0.1 V when V_{DS} increases from V_{SS} to V_{DD}. This internal node voltage and the resulting reverse bias of the gate of the N2 MOSFET has a strong influence on the subthreshold leakage as found by inserting (7.7) into (7.5),

$$I_{sub} = I_{off} 10^{\frac{-\sigma V_{DD}\left(\frac{1+\sigma}{1+2\sigma}\right)}{S}} \approx I_{off} 10^{\frac{-\sigma V_{DD}}{S}} \approx I_{off} 10^{-V_{DD}} \approx I_{off} /10 , \qquad (7.8)$$

for a typical subthreshold swing of 100 mV/decade. The stacking effect is illustrated in Fig. 7.4.



Two ON series MOSFETs
Current reduced to $I_{ON}/2$

Two OFF series MOSFETs
Subthreshold current reduced to $I_{OFF}/10$

Fig. 7.4. Illustration of the stacking effect

C. Low power design methods

In this section we will describe a number of low power design methods, including clock gating, gate sizing, and multi-VDD for reducing dynamic power, and the power gating and VTCMOS methods for reducing static power.

Reducing dynamic power

Clock gating. A significant fraction of the dynamic power in a chip is dissipated in the network distributing the clock. Up to 50% or even more of the dynamic power can be spent in the clock buffers. Clock buffers not only have the highest toggle rate in the system, but there are lots of them, and they often have a high drive strength to minimize clock delay. In addition, the flops receiving the clock dissipate some dynamic power even if the input and output remain the same.

The most common way to reduce power dissipated in the clock distribution network is to turn clocks off when they are not required. This approach is known as *clock gating*. Modern design tools support automatic clock gating, which means that they can identify where clock gating can be inserted without changing the function of the logic. Today most libraries include specific integrated clock gating (ICG) cells like the one shown in Fig. 7.5; cells that are recognized by the synthesis tools. The combination of explicit clock gating cells and automatic insertion makes clock gating a simple and reliable way of reducing power. No change in the RTL code is required to implement this style of clock gating [8].

Gate sizing under a delay constraint. In many cases we are willing to increase delay to save energy. We can do this by sizing gates or sub-blocks in our design since the RC product is not really constant independent of gate sizing. In order to fully understand how sizing works we must realize that there is a non-scalable part of the capacitance. The RC product being constant is only a zeroth order approximation. This leads to an energy/delay relationship shown in Fig. 7.6. Sometimes this relationship is simplified across a limited delay range assuming a constant energy-delay product.

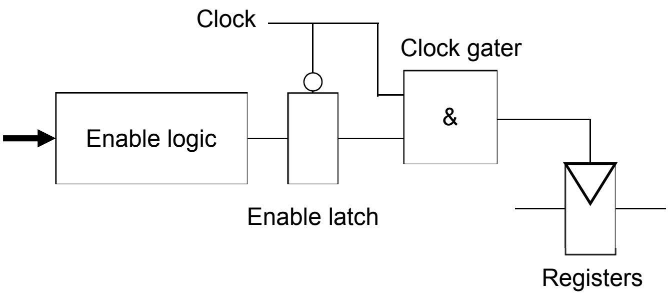


Fig. 7.5. Integrated clock gating cell (ICG) [8].

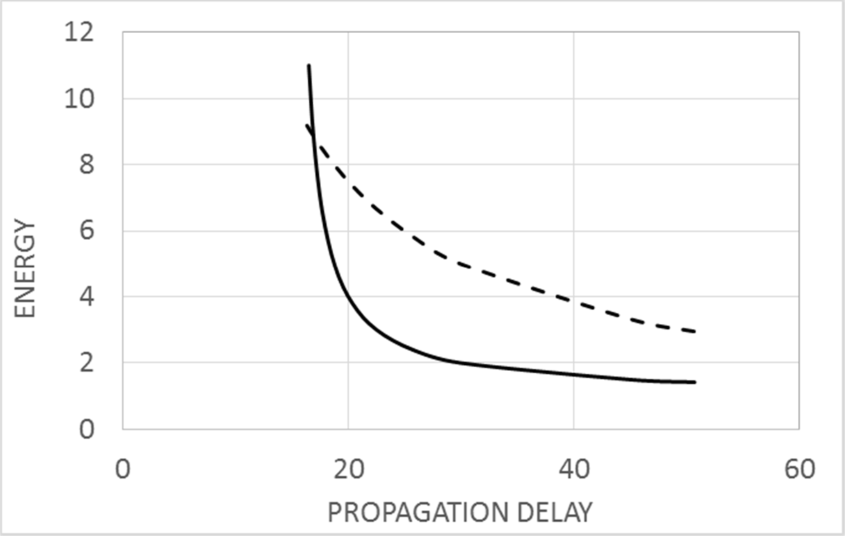


Fig. 7.6. Energy vs. delay. Simplified model assuming constant energy-delay product (dashed line).

Multi-VDD. Because of the quadratic dependence of power on voltage, decreasing the supply voltage is a highly leveraged way to reduce dynamic power. But because the speed of a gate decreases with decreases in supply voltage, this approach needs to be done carefully. System-on-chip designers can take advantage of this approach to use a a lower voltage supply for blocks that do not need to run particularly fast, such as peripherals, than for more speed-critical blocks. This approach is known as *multi-voltage*, or multi-VDD.

Multi-VDD is illustrated in Fig. 7.7 showing a multi-voltage architecture with three different voltage domains. Here, the cache memory runs at the highest voltage because they are on the critical timing path. The CPU is run at a high voltage because its performance determines system performance. But it can be run at a slightly lower voltage than the cache and still have the overall CPU subsystem performance determined by the cache speed. The rest of the chip can run at a lower voltage still without impacting overall system performance. Often the rest of the chip is running at a much lower frequency than the CPU as well. Thus, each block in the system is running at the lowest voltage consistent with meeting system timing constraints.

However, mixing blocks of different VDD supplies adds complexity to the design – not only do we need to add I/O pins to supply the different power rails, but we also need a more complex power grid and we need level shifters on signals running between blocks. For high to low VDD conversion simple inverters will do as level shifters, while for low to high signal conversion more complicated level shifters are needed as shown in Fig. 7.8.

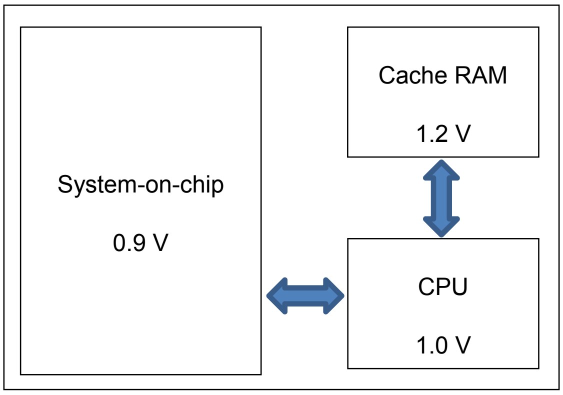


Fig. 7.7. Multi-VDD chip architecture [Weste & Harris].

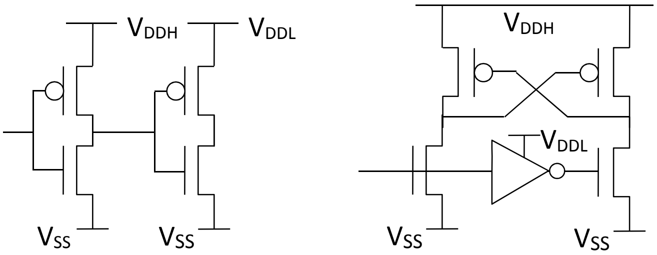


Fig. 7.8. High-to-low and low-to-high level converters [Weste & Harris].

Dynamic Voltage scaling (DVS). For processors, we can provide a variable supply voltage; during tasks that require peak performance, we can provide a high supply voltage and correspondingly high clock frequency. For tasks that require lower performance, we can provide a lower voltage and slower clock. This approach is known as *dynamic voltage scaling* (DVS) and is illustrated in Fig. 7.9.

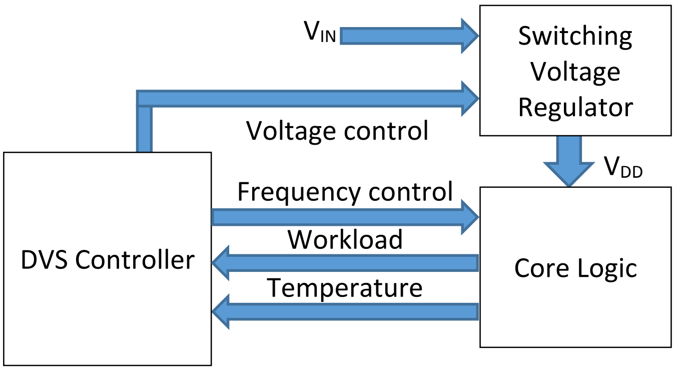


Fig. 7.9. Dynamic voltage scaling [Weste & Harris].

Reducing static power

The easiest way to reduce static power dissipation during sleep mode is to turn off the power supply to the sleeping blocks. This method is called power gating. Another method to reduce static leakage is to use multiple threshold voltages so that low-VT MOSFETs can be used in timing critical paths, while high-VT MOSFETs will reduce leakage in non-critical timing paths. There is also a similar technique called Variable Threshold CMOS (VTCMOS) where the threshold voltage can be changed by biasing the substrate.

Power gating is a technique where the power supply is shut down to blocks of logic. In this case, the logic block receives its power from a virtual V_{DD} rail, V_{DDV}. When the block is active, the header switches are ON thereby connecting V_{DDV} to V_{DD}. When the block goes to sleep, the header switch turns OFF, allowing V_{DDV} to float and gradually sink toward 0.

Isolation cells are used to prevent power gated block outputs from floating which could cause short circuit currents in the input circuitry of any powered block, see fig. 7.10. As the name suggests, these cells isolate the power gated block from the normally-ON block. Isolation of the signals of a switchable module is essential to preserve design integrity. Usually, a simple OR or AND logic can function as an output isolation device, defining outputs as either high or low. Isolation control signals are provided by the power gating controller, or power management unit (PMU). In a power-gated multi-VDD circuit implementation, Enable Level Shifter (ESL) cells that are a combination of a level shifter and an isolation cell can be used, see Fig. 7.10b.

Power gating affects design architecture more than clock gating, and introduces a number of design issues. Power gating uses low-leakage (high-VT) PMOS transistors as header switches to shut off power supplies to parts of a design in standby or sleep mode. Inserting sleep transistors splits the power network into a permanent power network connected to the power supply, and a virtual power network, V_{DDV}, that drives the cells and can be turned off. The header switch requires careful sizing. It should add minimal delay to the circuit during active operation, and should have low leakage during sleep. The transition between active and sleep modes takes some time and energy, so power gating is only effective when a block is turned off long enough. When a block is gated, the state must either be saved or reset upon power-up. The quality of the complex power network is critical to the success of a power gated design. Power gating can be implemented using cell-based (fine grain) approaches, or distributed coarse-grain approach.

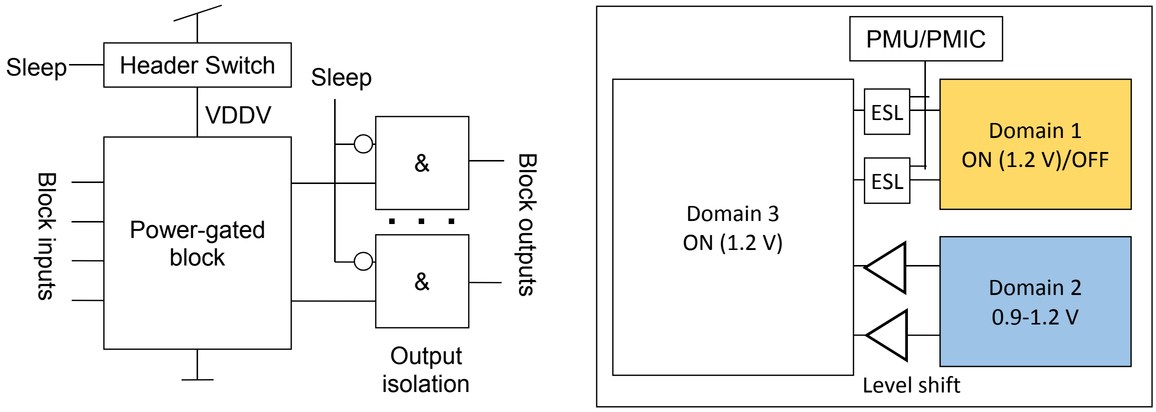


Fig. 7.10. Power gated Multi-VDD chip architecture with level shifters and ESL cells. [8].

The coarse-grained approach implements the grid style sleep transistors which drives cells locally through shared virtual power networks. This approach is less sensitive to PVT[7] variation, introduces less IR-drop variation, and imposes a smaller area overhead than the cell-based implementations. In coarse-grain power gating, the power-gating transistor is a part of the power distribution network rather than part of the standard cell.

[7] Process-Voltage-Temperature

There are two ways of implementing a coarse-grain structure, ring-based and column-based. An example of a ring-based sleep transistor implementation is shown in Fig. 7.11. In a ring-based implementation, power gates are placed around the perimeter of the module that is being switched off as a ring. Special corner cells are used to turn the power signals around the corners. In a column-based implementation, the power gates are inserted within the module with the cells abutted to each other in the form of columns. The global power is in the higher layers of metal, while the switched power is in the lower layers.

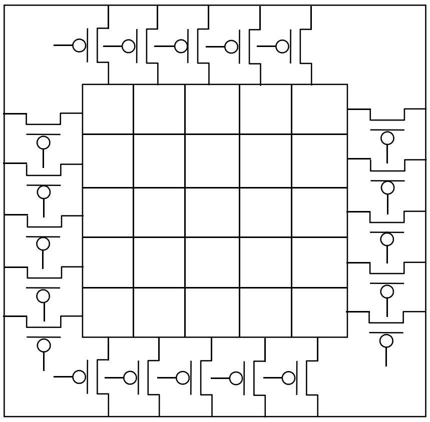


Fig. 7.11. Ring-based sleep transistor implementation [8]

Gate sizing depends on the overall switching current of the module at any given time. Since only a fraction of circuits switch at any point of time, power gate sizes are smaller as compared to the fine-grain switches. Dynamic power simulation using worst-case vectors can determine the worst-case switching for the module and hence the size. The IR drop can also be factored into the analysis. The simultaneous switching capacitance, that is the circuit node capacitance that can be switched simultaneously without affecting the integrity of the power network, is a major consideration in coarse-grain power gating implementation.

State retention power gating (SRPG) is a technique that allows the voltage supply to be reduced to zero for the majority of a block's logic gates while maintaining the supply for the state elements of that block. Using the SRPG technique, when in the inactive mode, power to the combinational logic is turned off and the sequential stays on. SRPG can thereby greatly reduce power consumption when the application is in stop mode, yet it still accommodates fast wake-up times.

Since the state of the digital logic is stored in the flip-flops, if the flip-flops are kept on a constantly powered voltage grid, the intermediate logic can be put onto a voltage grid that can be power gated. When the voltage is reapplied to the intermediate logic, the state of the flip-flops will be re-propagated through the logic and the system can start where it has left off as illustrated in Figure 7.12.

Retention registers are special low-leakage flip-flops used to hold the main register data of a power gated block. Thus the internal states of the power-gated block can be retained and loaded back to it when the block is reactivated. As just mentioned, retention flops are special flops with multiple power supplies, where part of the registers are always powered. The retention mechanism is controlled by the power gating controller. See Fig. 7.13. [Wikipedia on power gating]

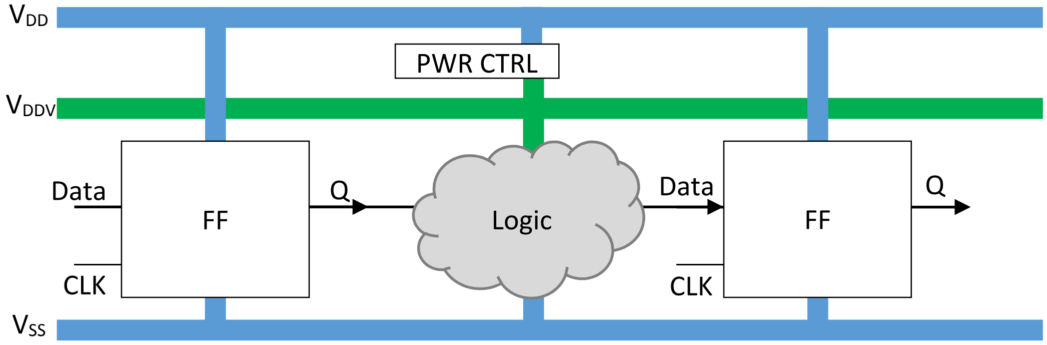


Fig. 7.12. State Retention Power Gated Cell (SRPG)

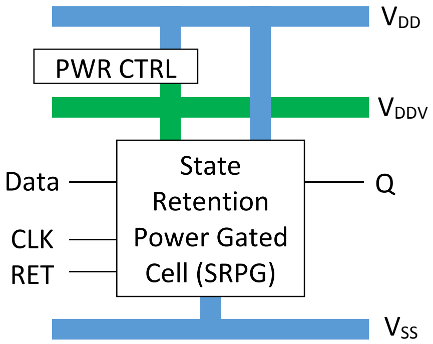


Fig. 7.13. State Retention Power Gated Cell (SRPG)

The retention flop has the same structure as a standard master-slave flop. However, the retention flop has a balloon latch that is connected to the true-VDD. With the proper series of control signals before sleep, the data in the flop can be written into the balloon latch. Similarly, when the block comes out of sleep, the data can be written back into the flip-flop.

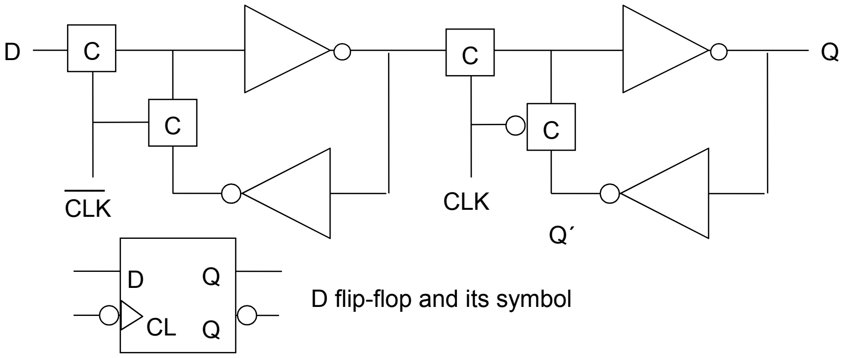


Fig. 7.13. D-type flip-flop and its symbol.

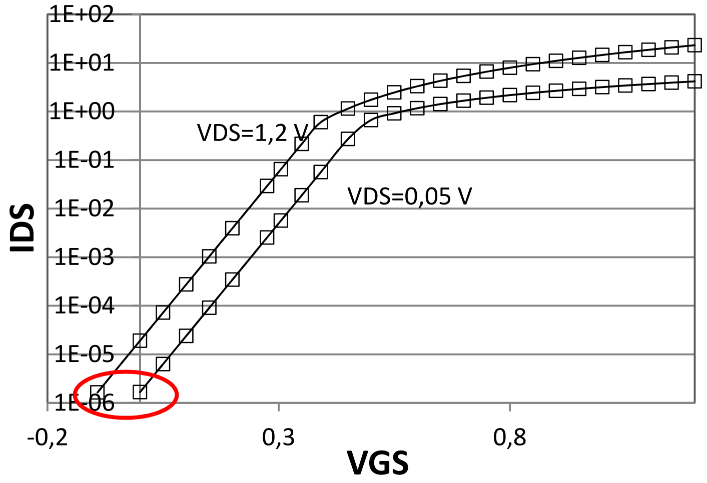
The advantages of SRPG include shutdown leakage savings, which can be independent of process variations. It allows for faster system power-on because the state is preserved in the slave latch. Disadvantages include increased area and die size; timing penalties such as increased signal and clocking delays; increased routing resources (power routing for V_{SS} and a power-gating signal tree with on buffers); specialized library models for SRPG cells; additional power overhead in the active mode; and impacts to functional verification, physical integration, and design for test (DFT).

	Domain 1	Domain 2	Domain 3
Mode 1	1.2 V	1.2 V	1.2 V
Mode 2	1.2 V	0.9 V	1.2 V
Mode 3	OFF	1.2 V	1.2 V
Mode 4	OFF	0.9 V	1.2 V

	Domain 1	Domain 2	Domain 3
Mode 1	1.2 V	1.2 V	1.2 V
Mode 2	1.2 V	0.9 V	1.2 V
Mode 3	OFF	1.2 V	1.2 V
Mode 4	OFF	0.9 V	1.2 V

Table 7.6. Multi-Mode operation

Stacking effect In this section shall also briefly discuss the stacking effect. When two MOSFETs in series are turned OFF, the bottom one, MN1, will be in its linear region and the top one, MN2, be saturated. The internal node voltage will give the top MOSFET a negative gate voltage acting to reduce the leakage of the saturated one to the lower leakage level of the bottom one.



References

1. Special cells required for Multi-Voltage Design, Posted by Godwin Maben on April 15th, 2007. https://blogs.synopsys.com/magicbluesmoke/2007/04/15/special-cells-required-for-multi-voltage-design/
2. A Practical Guide to Low-Power Design - User Experience with CPF, Available through the Power Forward Initiative, https://www.si2.org/?page=1061 More about the Cadence Power Forward Initiative, http://www.cadence.com/Alliances/power_forward/pages/default.aspx,