Low Power System-On-Chip Design
Chapter 13:
Retention Register Design

Ismo Hänninen
Chapter 13: Retention Register Design

- Overview

- Retention Register Types
  - Single Control Live Slave
  - Dual Control Balloon
  - Single Control Balloon

- Retention Register Relative Layouts

- Memory Retention Methods
  - Reduced VDD Retention
  - Source-Diode Biasing
  - Source Biasing

- Memory Retention Latency Reduction
Overview

- Storing the data during block power-down (Power Gating)
  - Retention capable flip-flops
  - Or larger memories
Retention Register Types

Single Control Live Slave
Dual Control Balloon
Single Control Balloon
Single Control Live Slave

- Standard master-slave flip-flop with minimal modifications
  - Master-latch power-gated, fast Low-\(V_T\) transistors
  - Slave-latch always powered-on, slow High-\(V_T\) transistors
  - Three-state buffer (T1) between latches, isolating slave input
  - AND-gate for isolating clock input, during power-down
  - Control signal NRETAIN
...Single Control Live Slave

- **Usage:**
  - Control signal changed only, when clock is inactive (low)
  - Control signal always driven
  - Possible set/reset signals isolated
  - Clock has to be low, before restoring state

- **Advantages:**
  - Minimal area
  - Single control signal

- **Disadvantages:**
  - Slow slave-latch (low leakage High-$V_T$ transistors)
    ⇒ CLK to Q output delay increased
  - Longer input data hold-time needed (due to slow clock gating AND), but also good balance of rise and fall times needed (fast gates)
  - Making sure clock is low, when restoring state
Dual Control Balloon

- Standard master-slave flip-flop with added retention latch
  - Master and slave latches power-gated, *fast* Low-$V_T$ transistors
  - Retention latch (shadow latch) always powered-on, *slow* High-$V_T$ transistors
  - Three-state buffer (T1) isolating retention latch input
  - Control signals SAVE, NRESTORE
Dual Control Balloon

Usage:
- SAVE control signal always driven, **pulsed** to latch data into shadow register
- NRESTORE control signal may float during power-off, **pulsed** to restore data
- Clock can be low or high, before restoring state
- Possible set/reset signals do not affect shadow register, when holding data

Advantages:
- Minimal leakage power (minimal shadow latch and control transistors)
- Almost no performance loss, compared to non-retaining flip-flop (only very little increase on transistor loading)
- Clock phase independent restore

Disadvantages:
- Area increased, compared to live-slave (due to third latch)
- Two control signals, requiring two buffer networks, also adds complexity
Single Control Balloon

- Standard master-slave flip-flop with added retention latch
  - Master and slave latches power-gated, fast Low-$V_T$ transistors
  - Retention latch (shadow latch) always powered-on, slow High-$V_T$ transistors
  - Three-state buffer (T1) isolating retention latch input
  - Control signal NRET (save during runtime, restore on edge)

Institute of Digital and Computer Systems / TKT-9636
Department of Computer Systems / TKT-9626

Ch13: Retention Register Design
...Single Control Balloon

Usage:
- Control signal always driven
- Clock can be low or high, before restoring state
- Possible set/reset signals do not affect shadow register, when holding data

Advantages:
- Almost minimal leakage power (minimal shadow latch and control transistors)
- Almost no performance loss, compared to non-retaining flip-flop
  (only very little increase on transistor loading)
- Clock phase independent restore
- System-level dynamic power reduced, since only one control signal network

Disadvantages:
- Area increased, compared to live-slave (due to third latch)
- Some (small) additional dynamic power, since shadow register follows slave-latch state (during run-time)
Retention Register Relative Layouts

- Standard D-type register
  (scan-testable, minimal output buffer)

- Balloon D-type register
  (scan-testable, minimal output buffer)
Memory Retention Methods

*Reduced VDD Retention*
*Source-Diode Biasing*
*Source Biasing*
Memory Retention Methods

- Often, FIFOs’, caches, and other memories are allowed to reset after power-down (losing contents).

- For minimum wake-up latency, also memory contents have to be retained:
  - Large memories are too expensive to implement by using retention register approach (replacing standard, optimized memory cells).
  - Needs a solution to reduce the basic cell leakage (without corrupting memory contents).

- Three basic approaches for SRAM.
Reduced VDD Retention

- Memory is given own power supply, with two modes (VDD):
  - Normal supply voltage
  - Reduced (0.5-0.6 V) sleeping voltage

- Reduced voltage saves leakage power
  - Memory contents stay uncorrupted, but cannot be accessed

- Simple implementation
  - No memory circuit change needed
  - A switchable, dedicated power supply added
Source-Diode Biasing

- Biasing the SRAM cell source voltage
  - Reduced operating voltage (in addition to lowered VDD)
  - Reverse body bias reduces sub-threshold leakage (NMOS substrate voltage is ground)
- Simplest biasing:
  - Diode in source supply of cell
  - Bypass control switch
- Problem: Diode threshold sets fixed bias for the process, difficult to optimize
Source Biasing

- Biasing the SRAM cell source voltage
  - Reduced operating voltage (in addition to lowered VDD)
  - Reverse body bias reduces sub-threshold leakage (NMOS substrate voltage is ground)

- **Dedicated source bias supply** (replacing the diode)
- Optimal operating voltage and bias in sub-1V designs
Memory Retention Latency Reduction

- For memories, which are not often accessed (but always retained), if **short access latency** is required

- Block-based retention and wakeup
  - Memory divided into **banks**, individual control of sleep mode
  - Normally, bank mode is **sleep**
  - Read/write access **actives** only one bank (address decoder)
  - Reduced wakeup latency, since smaller power blocks
    (smaller virtual VSS network, per bank)
  - Compromise bank size, speed vs. column overhead
    (small banks cause area and power overhead in column sense-amps)

- Row-based retention and wakeup
  - Biased voltage net for each **row**, individual control of sleep mode
  - Normally, row mode is **sleep**
  - Read/write access **actives** only one row
  - Very fast, since a row is relatively small
  - Reduced control overhead, by using row-grouping
Conclusion

- Retention flip-flops easily replace standard flip-flops
  - Performance loss (live slave)
  - Or area penalty (balloon)

- For retained memories, the only practical way to save energy during sleep is reduced operating voltage
  - VDD and VSS moved near each other
  - Sleep/wakeup latency reduction with block/row grouping