ECE260B – CSE241A
Winter 2010

Low power implementation
A system perspective

Website: http://cseweb.ucsd.edu/classes/wi10/cse241a/

Low power implementation : Metrics

- User experience prospective:
  - For mobile devices:
    - Active time of the device: Time interval of performing a well defined set of tasks (defined use mode: audio play , voice call, web browsing, video playback, etc) between two battery charges
    - Standby time of the device: Time interval between two battery charges when the device is fully functional ready to be activated but does not perform any functional user driven tasks.
- For electrical powered devices:
  - Efficiency: Power consumption for performing a defined set of tasks relative to performance metrics:
    - mW/Mhz
  - Average power consumption
  - Peak power consumption

Low power implementation : Metrics

- Power consumption in digital systems:
  - $P_{total} = P_{active} + P_{leakage}$

    $P_{active} = P_{internal} + P_{switching} = P_{internal} + \alpha C V^2 f$

  V – voltage

  f – frequency

  C – capacitive load

  α – activity factor

Low power implementation: Design synergy

- Low power implementation in the modern system on chips today requires a holistic and concurrent approach which includes collaboration between:
  - System level design
  - Architectural design
  - Software Hardware co-design
  - IP design:
    - Circuit design
    - Physical implementation of the IP
  - Physical design (chip/block level)
  - Power verification and modeling
  - Silicon correlation and validation

System optimization

- Power delivery network optimization:
  - On die vs on board (PCB) voltage regulators
  - Voltage regulators efficiency
  - Voltage rails definition
  - System level power management:
    - Adaptive voltage scaling (AVS)
    - Dynamic clock frequency and voltage scaling (DCVS)
    - Static voltage scaling (SVS)

- Analog vs digital processing system level optimization
  - Optimization at the system with the goal of moving most of the signal processing (data transformation) in the digital domain. The power consumption in the digital domain is scalable with the process technology scaling and with the system use mode requirements.
  - Digitally assisted analog processing

Architectural optimization

- Memory hierarchy
  - On die vs. off die memory
  - Cache size (miss penalty)
  - Cache hierarchy (architecture)
  - Address space definition
- Processor architecture
  - Von Neumann , Harvard
  - VLIW (high IPC)
  - 16bit, 32bit, 64 bit instruction architecture (IA) (Code compression)
  - In order vs out of order execution
  - Superscalar implementation
  - Multi thread implementation
  - Scalability : Single core vs. Multi core
  - Application specific IA optimization
    - FFT cores
    - Multipliers, adders ,shifters

Architectural optimization

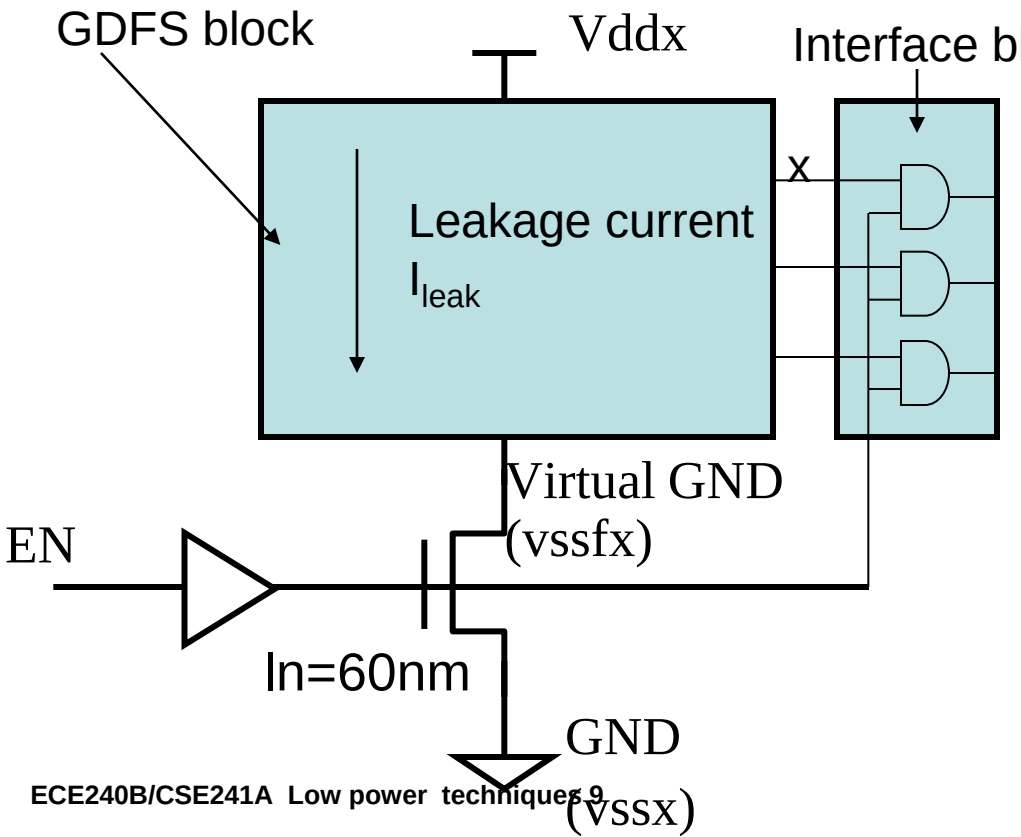
- Hardware accelerators:
  - Graphic 2D, 3D
  - Video encoder/decoder (720p, 1080p)
  - Multimedia display
  - Audio + DSP (digital signal processing unit)
  - Modem baseband
- Bus architecture
  - AHB implementation (Advanced high performance bus)
  - AXI
  - Fabric (high speed, high bandwidth interconnect):
    - Bandwidth
    - Latency
    - Power
- Clocking architecture:
  - PLL's
  - Frequency planning
  - Clock architecture
  - Synchronous vs asynchronous clocks

Architectural optimization

- IO interfaces
  - DDR (LPDDR), SDIO
  - PCI-X, USB, MIPI, HDMI, GPIO
- Engineering system level design and optimization (ESL):
  - Algorithmic driven hardware implementation and optimization
  - System level power modeling
- Hardware software co-design and optimization

Low power techniques: Power gating

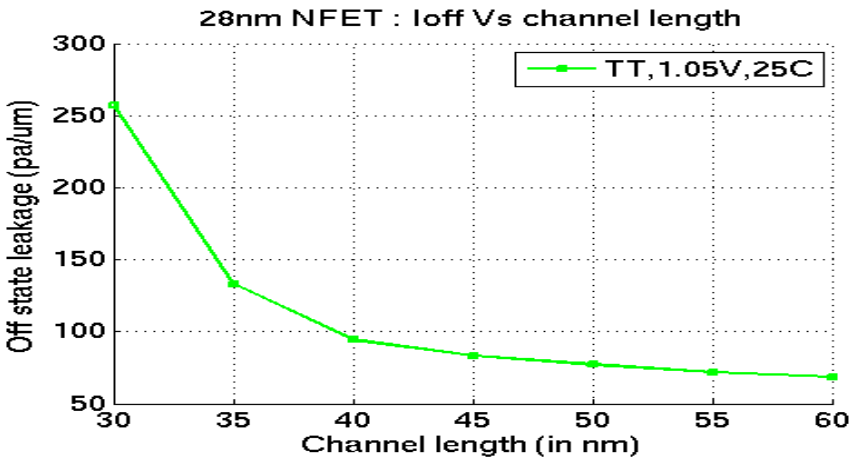
- Widely use in all the portable devices today:
  - Global distributed foot-switch GDFS
  - Global distributed head-switch GDHS
- Main goal to eliminate the current leakage (reduce leakage power) in standby mode.

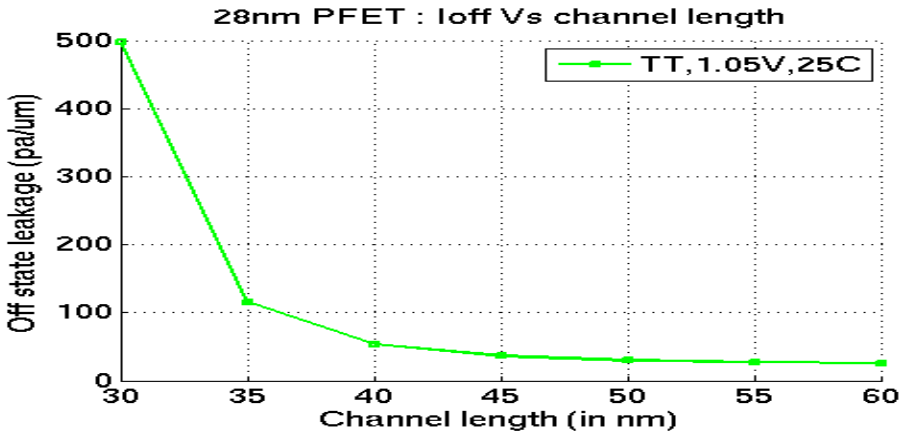


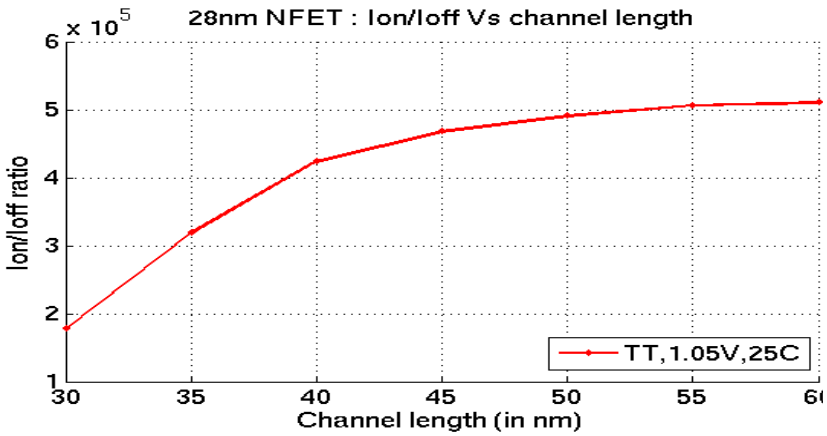
- Leakage savings. (+)
- Voltage droop (-)
- Area overhead. (-)
- Routing resources (-)
- Leakage savings mode : High Roff (~ GΩ)
- Functional mode : Low Ron
- Isolation cells required for the output signals of the block in sleep mode to avoid undefined logic state propagation
- 10X to 1000X leakage saving

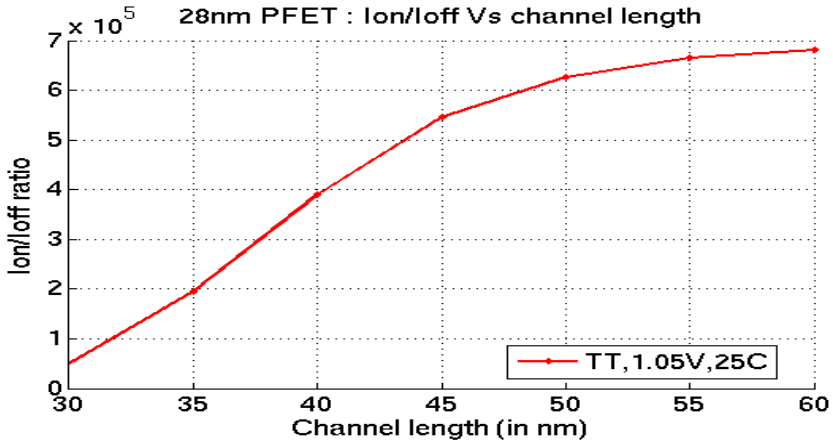
Low power techniques: Power gating

- Chanel length modulation:

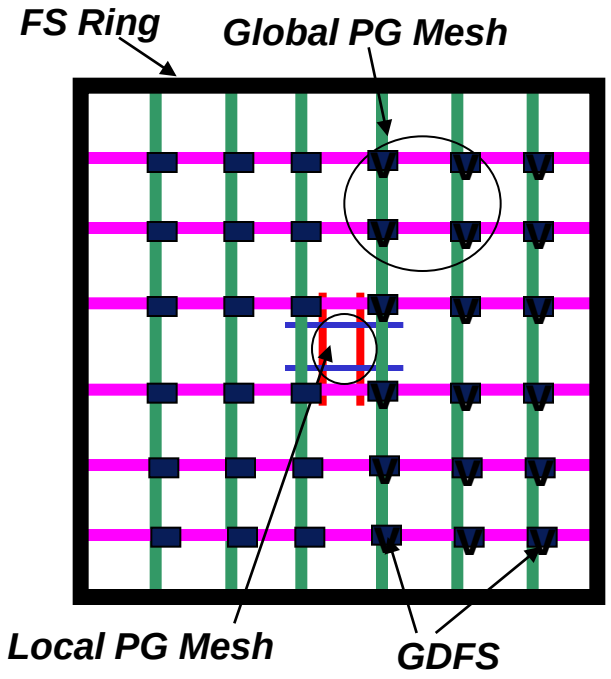


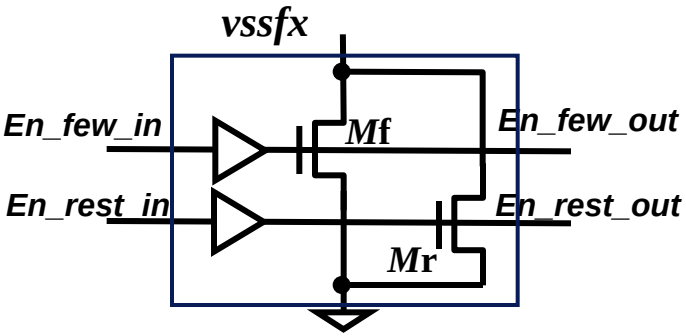






Low power techniques: Power gating

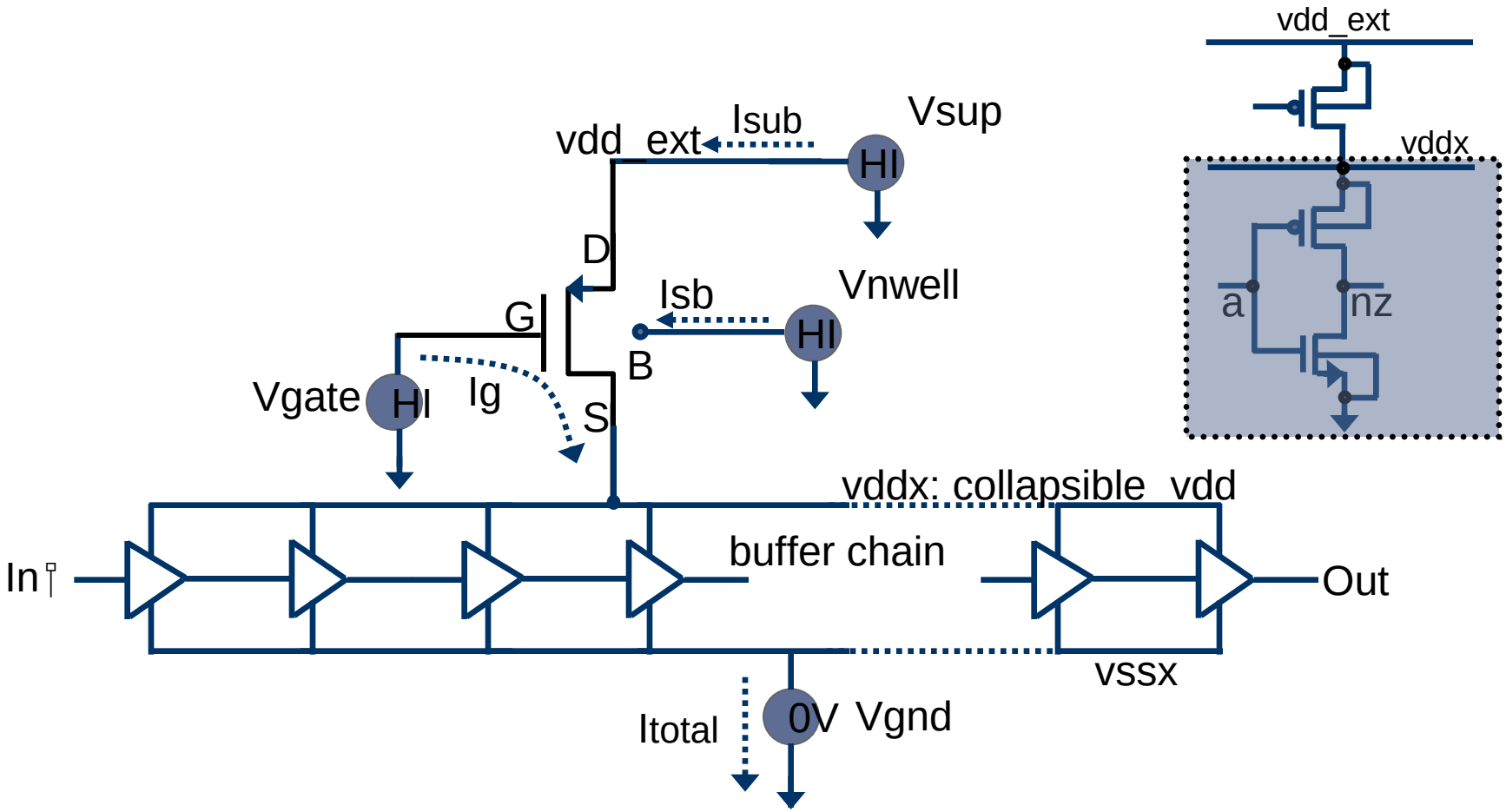




- Global distributed FS/HS
  - Can be modeled as an additional resistance between global and local power mesh
  - Does not break global mesh
  - Needs sleep control signal distribution
  - Suitable for large size macros
- FS/HS ring
  - Smaller cost on sleep control distribution
  - Larger IR drop compared with GDFS, especially for flip-chip case
  - IR drop increases quicker when the size of the block increases (cubic w.r.t. the length)
  - Suitable for small size macros without memories or small hard macros which have the memories on a different power rail than the power gated logic

Low power techniques: Power gating

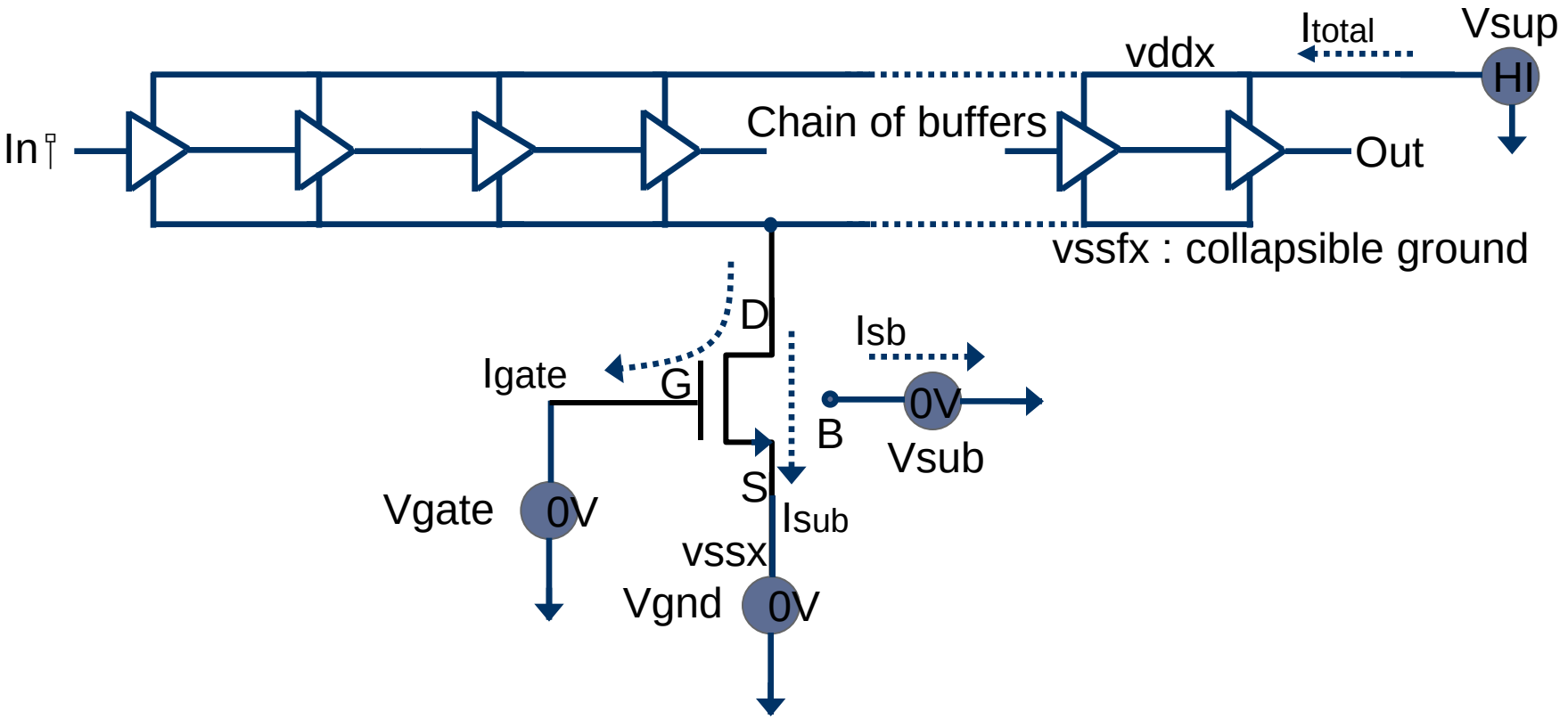
- During design and optimization of the header cell we need to take into consideration all the leakage current sources in OFF state.



- Standard cell PMOS well terminal connected to local power (vddx).

Low power techniques: Power gating

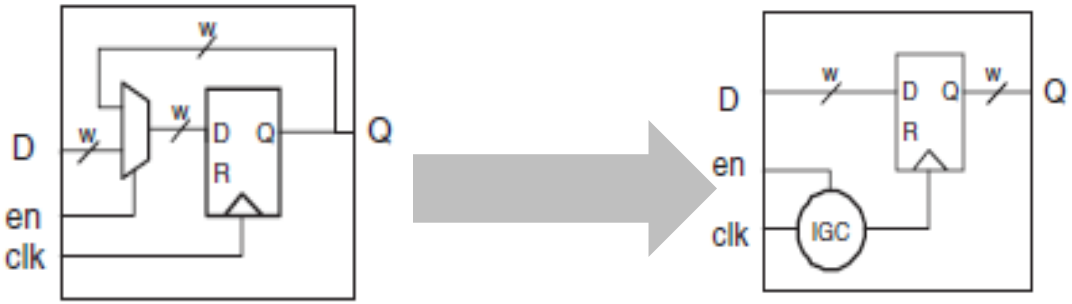
- During design and optimization of the header cell we need to take into consideration all the leakage current sources in OFF state.

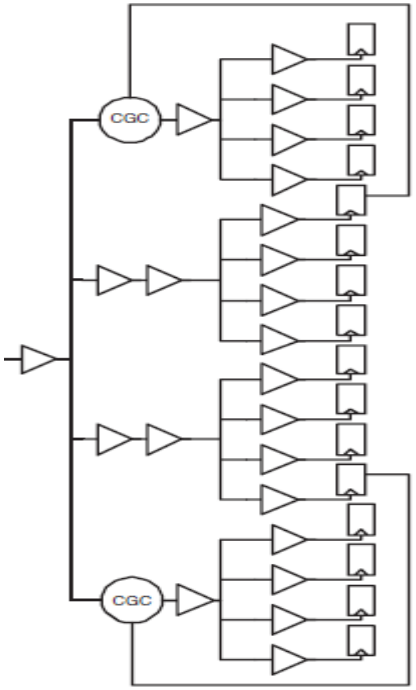


- Junction leakage contributes substantially more to static dissipation than sub-threshold leakage in deep sub-micron process nodes.

Low power techniques: Clock gating

- Clock gating technique is used extensively to reduce the active power on the clock tree

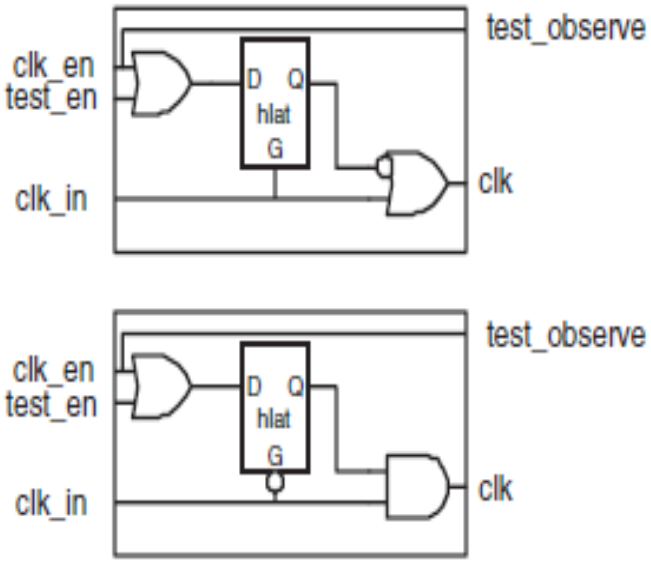




- The clock gating cells are inserted in the design:
  - Architectural definition stage
  - During logical implementation
  - During logical synthesis (RTL to gate)
  - During physical placement of the design and clock tree synthesis

Low power techniques: Clock gating

- There are multiple types of clock gating cells (circuit implementation):
  - Clock halt high
  - Clock halt low

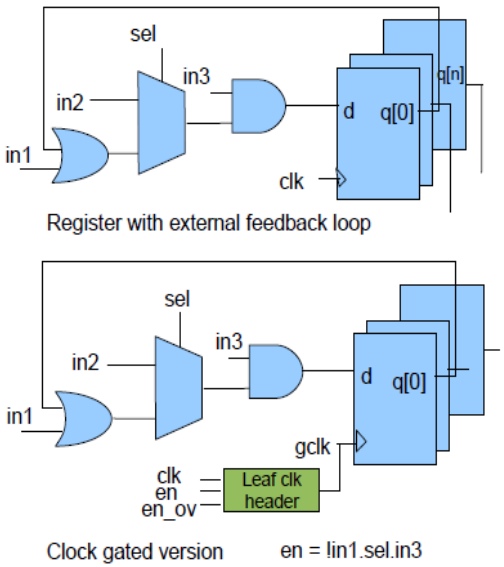


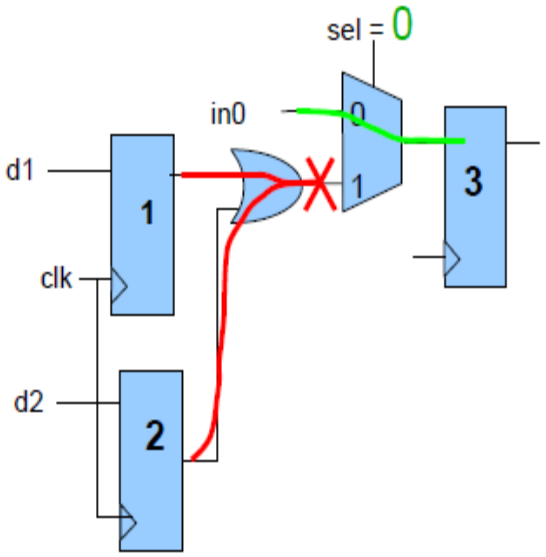
•Clock gating insertion in the clock tree will impact:
- •Insertion delay
- •Skew of the clock tree
- •Active Power:
  - •If a CGC cell is seldom enabled (no gating) it can have a negative impact on the overall active power of the clock tree

•There are multiple types of clock gating strategies which can be implemented in a design :
- •Combinational clock gating
- •Sequential clock gating

Low power techniques: Clock gating

- Combinational clock gating





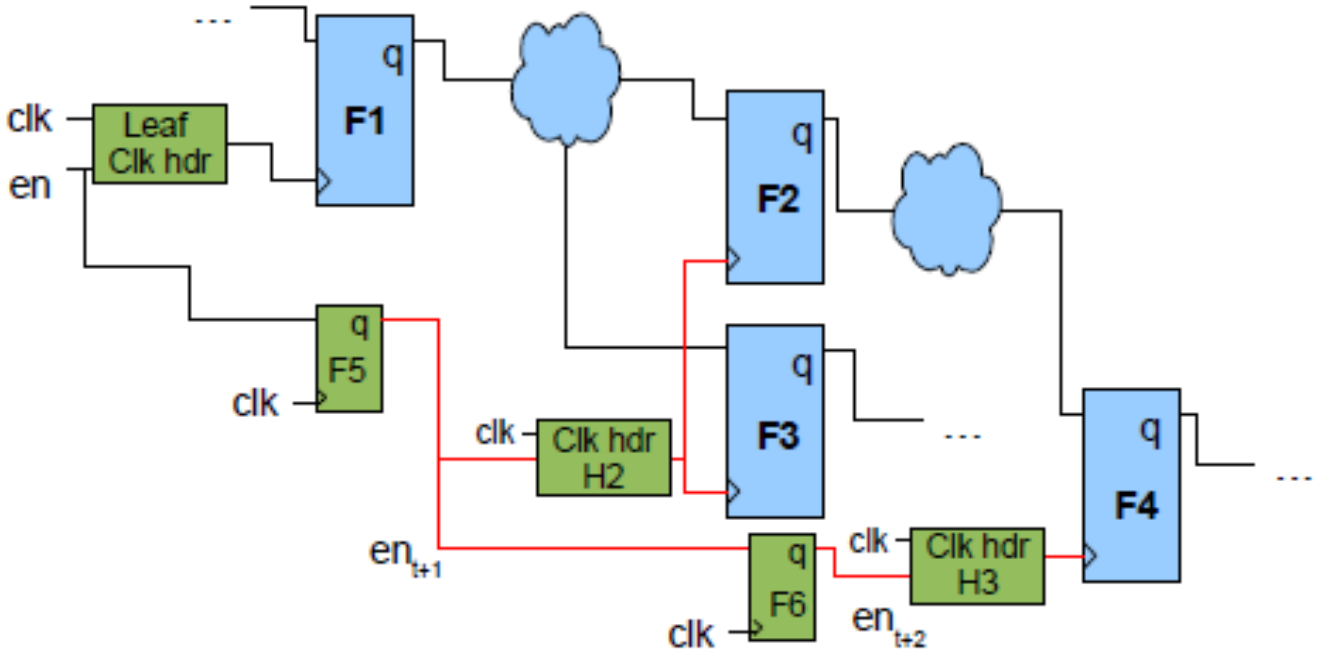
When en=(!in1)&(sel)&(in3) data from the flop output q0
Is feedback to input d of the same flop. This represent an opportunity for combinational clock gating

Observability-based clock gating.
When sel=0 flops 1 and 2 are not observable for flop 3 .
When sel=0 we can apply clock gating for flops 1 and 2

Courtesy of Krishnan Sundaresan, Aravind Oommen, Doug Meserve, Hemango Das, Jaewon Oh, Mohd Jamil
“A tool for exploring advanced RTL Clock Gating Opportunities in Microprocessor Design”

Low power techniques: Clock gating

- Sequential clock gating



Clock gating propagation. When F1 is gated no new data will propagate downstream to F2 and F3. Next clock cycle we can gate F2 and F3. We introduce in the design the staging flops F5 and F6.

Courtesy of Krishnan Sundaresan, Aravind Oommen, Doug Meserve, Hemango Das, Jaewon Oh, Mohd Jamil
“A tool for exploring advanced RTL Clock Gating Opportunities in Microprocessor Design”

Low power techniques: AVS

- Adaptive voltage scaling:
  - Low power technique used to take advantage of the silicon process variation
  - IC's silicon can be FF (fast nmos, fast pmos), TT or SS.
  - We design the IC at the SS corner. The TT and FF silicon can meet our performance requirements at lower voltage.
  - AVS is a circuit implementation on the die which replicates the critical paths in the IC design and compares the delay on these paths with the required delay value imposed by the clock frequency of the system f.
  - If the measured delay is smaller than the required delay AVS will instruct the voltage regulator to reduce the voltage supply gradually with a given step until the delta between the required delay and measured delay are in the AVS margin of error.
  - AVS requires silicon calibration and works in tandem with DCVS
  - AVS has a start-up condition defined by the software:
    - Frequency target
    - Starting voltage

Low power techniques: DCVS

- Dynamic clock frequency and voltage scaling:
  - Low power technique used to adjust the frequency of operation and the voltage supply of the system to the software application needs.
  - Using silicon measurements a (f, V) look-up table is created. In the look-up table we define pairs of frequency of operation and the corresponding voltage supply.
  - DCVS can be hardware or software controlled. Most of the implementations are software (SW) due to simplicity of implementation.
  - DCVS can be easily implemented for CPU's, DSP's (IP's) which have a dedicated power domain (power rail). If in the same power domain we have multiple functional blocks (CPU, Graphic core, Bus etc) in order to perform DCVS we will need to find a common denominator in terms of pairs of frequencies (performance requirements) and voltage supply for all the IP's, which does not happen very often. (diminishing return)
  - DCVS provides the biggest power saving for IP's which have dedicated power rails.
  - For DCVS and AVS a very important quality metric is system latency. How long it takes to adjust the frequency and voltage of the IP ? The lower is the latency the bigger are of power savings.

Low power techniques: SVS

- Static voltage scaling:
  - During design phase we define a set of voltage supplies values which will correspond with a set of predefined use modes of the device:
    - SVS mode: V1
    - Normal mode: V2
    - Turbo mode: V3
  - The design will be timed closed during design phase to all the predefined corners (PV1T), (PV2T) , (PV3T) and modes (SDC1), (SDC2), (SDC3). (SDC = Synopsys Design Constraints)
  - SVS does not require silicon calibration
  - SVS is software controlled. The transition between different modes of operation is controlled by software (SW) and/or the Power Management Unit implemented on the die.
  - The opportunity of frequency and voltage scaling is bounded to the predefined use modes of the device.(small number of modes)

Building blocks optimization (IP's)

- A major portion of power optimization is implemented today in the IP blocks:
  - Memories
  - Standard cells
  - IO's (Pads)
  - PLL's
  - Mix signal blocks : ADC's, DAC's
  - Standard interfaces: DDR, USB, MIPI, SDIO, HDMI
- Active power and leakage power optimization at the IP level is critical for the overall power efficiency of the system
- The power optimization at the IP level targets both active power and leakage power:
  - Power optimization driven by performance requirements
  - Smart shut-off hardware implementation
  - Low voltage optimization

Building blocks optimization: Standard cells

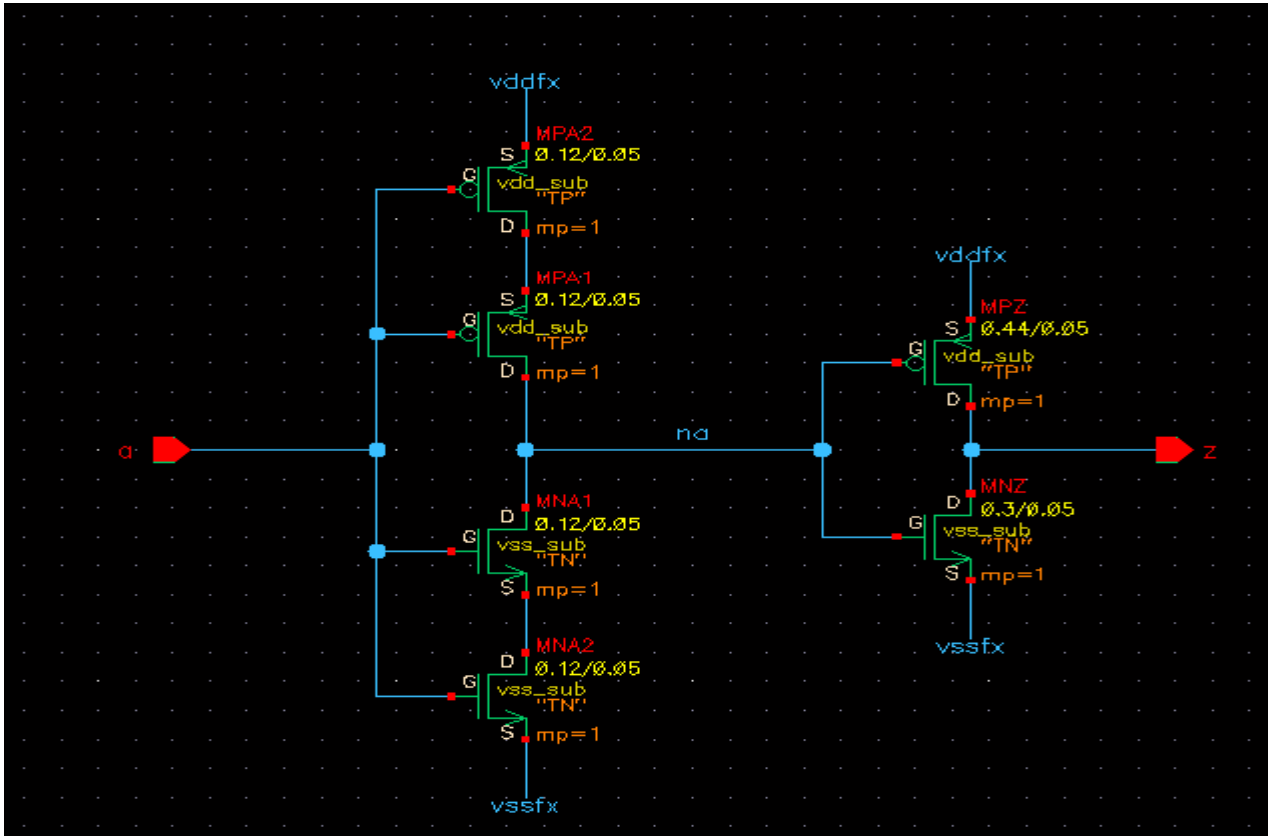
- Multi voltage threshold libraries:
  - High VT (low leakage, lowest performance)
  - Nominal VT (higher leakage, higher performance)
  - Low VT (highest leakage, highest performance)
  - The difference in leakage between High VT lib vs. Low VT lib can be 10X to 30X
  - The difference in performance is 20% to 40%
- Multi channel length libraries:
  - Min channel length library
  - Longer channel length footprint compatible library
  - 40% to 2X reduction in leakage between the library variants
  - 10% to 20% performance penalty

Building blocks optimization: Standard cells

- Multi track architecture libraries:
  - 8 tracks library: high density low performance blocks, low leakage, low active power
  - 10 tracks library: lower density higher performance blocks, higher leakage library
  - 12 tracks library: high performance library, highest leakage library
- High drive strength granularity libraries:
  - Providing high granularity libraries enables the synthesis tool to select the optimum drive strength cells for a given load (smaller area, leakage power and internal power)
- Library customization:
  - Multiple β=Wp/Wn ratio cells optimized for power*delay metric
  - Custom combinational and sequential cells optimized for power and performance

Building blocks optimization: Standard cells

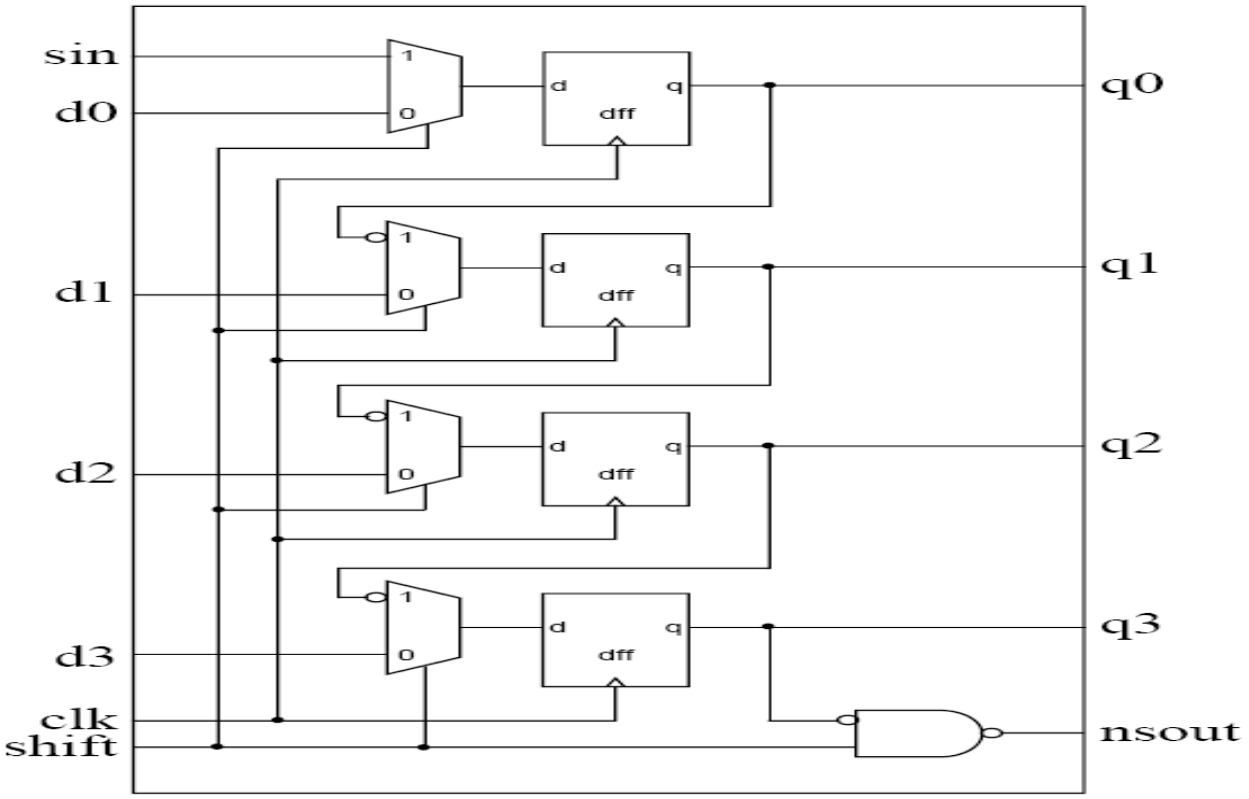
- Leakage optimization using stacked devices:



- 8x lower leakage compared to nominal channel length, 3x higher delay
- Topology used in delay cells (hold fixing)

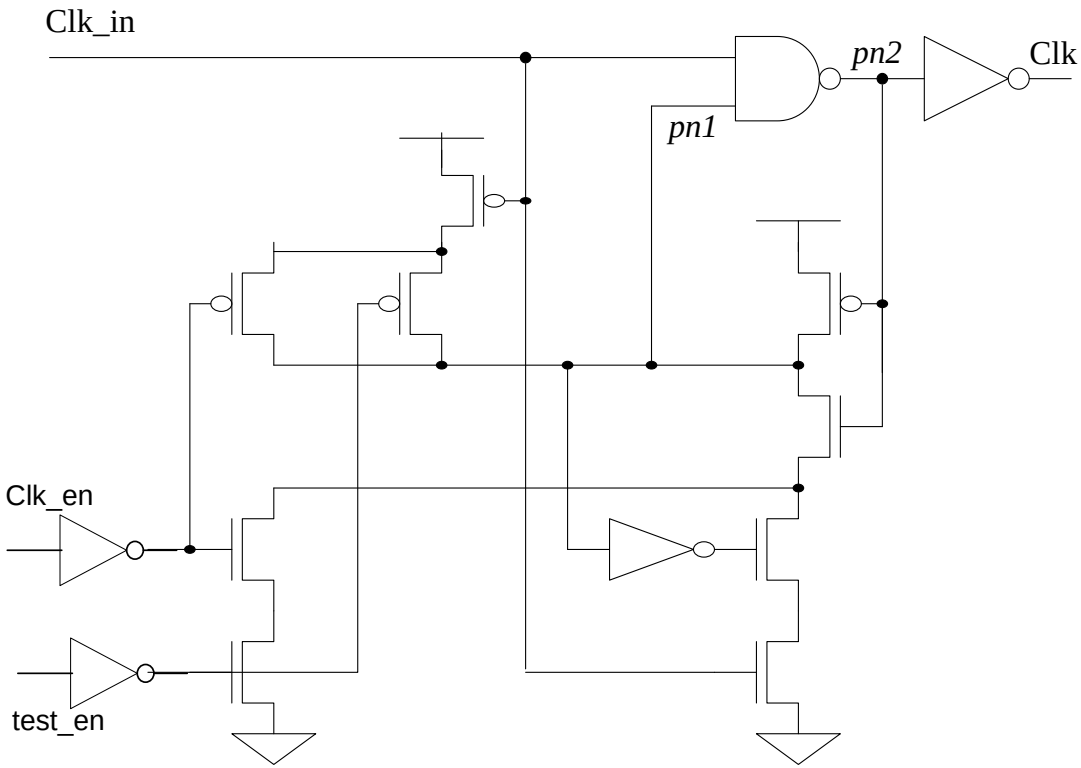
Building blocks optimization: Standard cells

- Array of latches and flops:
  - Major impact on active power for the clock tree implemented at the block/top level. All the flops/latches are sharing the internal clock. We will have less buffers in the clock tree implemented at the block/top level.



Building blocks optimization: Standard cells

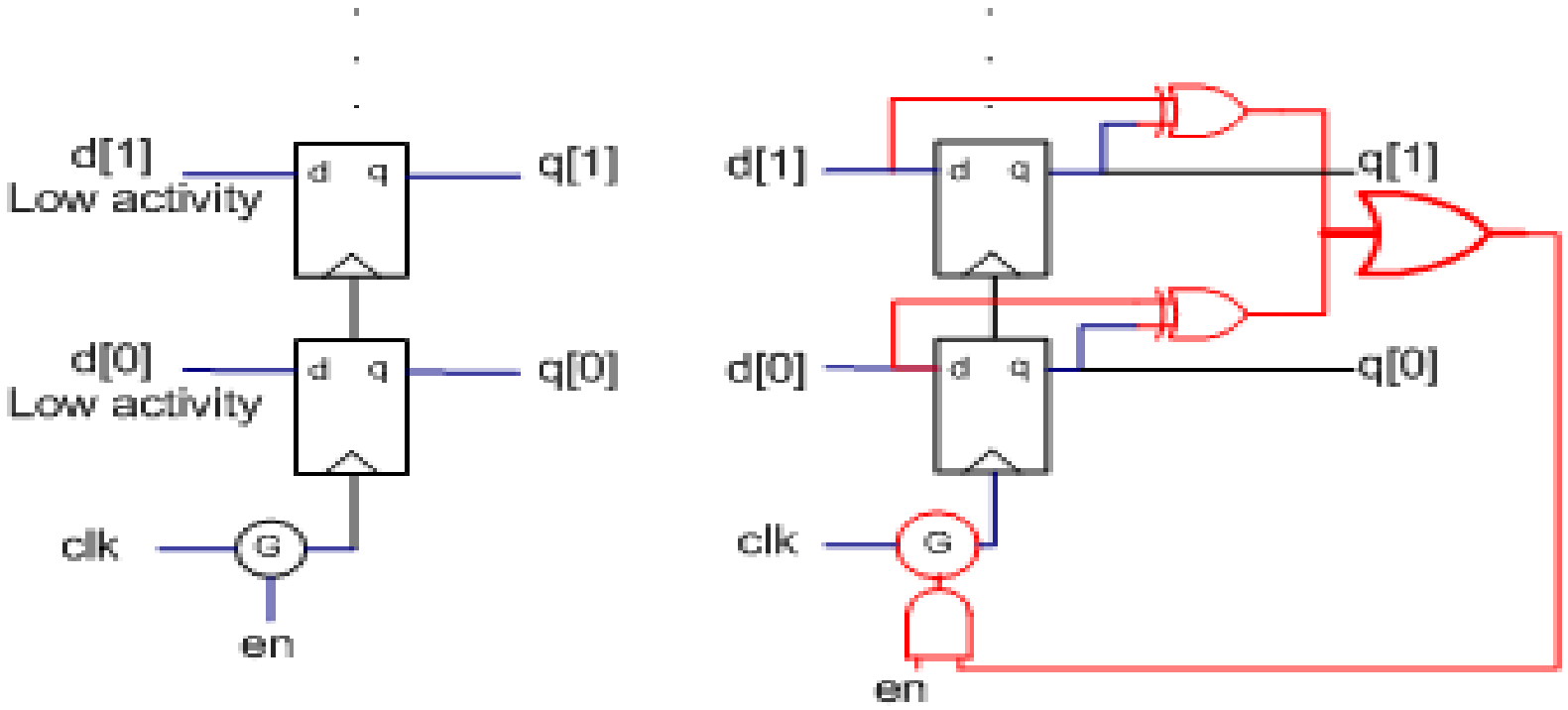
- Low power clock gating cells:



- 40% lower power (enabled), 60% lower power (gated) than a standard CGC implementation

Building blocks optimization: Standard cells

- Conditional capture flops:
  - Used for low activity flops
  - Area penalty and added DFT complexity



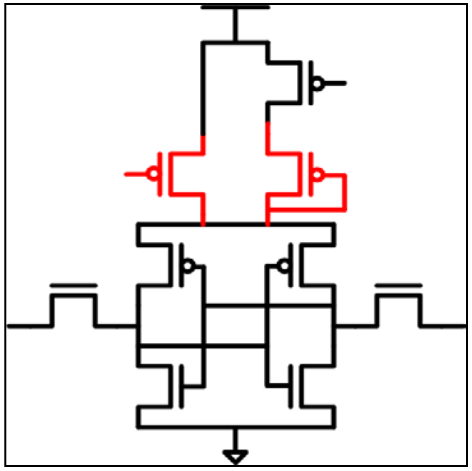
Building blocks optimization: Memories

- Memory design:
  - Performance driven implementation:
    - Low/medium performance (low leakage), smaller bit cell
    - High performance (higher leakage), bigger area bit cells
  - Low power techniques used in memory design:
    - Power gating:
      - Power gate the periphery
      - Power gate the bit cell array
    - Source biasing during inactive mode
    - Selective power gating for the inactive memory banks
    - Split power rail:
      - The periphery and the memory array have different power rail supplies

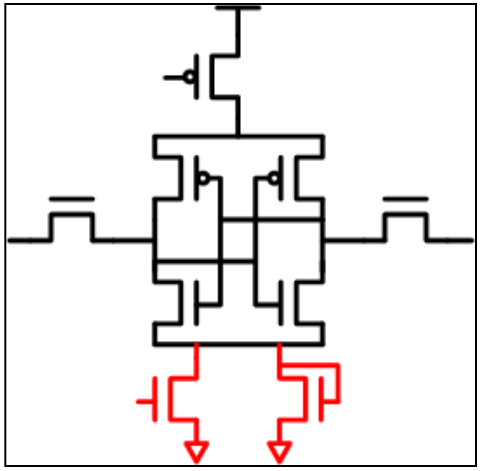
Building blocks optimization: Memories

- Source biasing of the bit cells memory array:
  - Head-switch or foot-switch implementation
  - Dynamic biasing when we do not perform any read or write operations

VDD diode



Source biasing



Physical design optimization

- Physical design optimization:
  - Power mesh creation
  - Clock tree synthesis
  - Power gating insertion (GDFS, GDHS)
  - Voltage noise analysis and optimization
  - Routing pitch selection
  - PVT corner selection
  - Timing closure methodology
- Logical and physical synthesis:
  - Vt selection
  - Clock gating insertion
  - Flop/latch tray clustering

Power verification

- Power verification methodology:
  - RTL verification:
    - Power verification and profiling
  - Power linting
  - Gate netlist power verification (post place and route)
  - PDN analysis:
    - Voltage noise analysis
    - Frequency domain analysis
  - Silicon power correlation:
    - Leakage power
    - Clock power
    - Use modes and concurency
- System level power modeling