On Leakage Currents

Sources and Reduction for Transistors, Gates, Memories and Digital Systems

Wolfgang Nebel, Domenik Helms
Agenda

- Introduction
  - CLEAN project
  - Motivation
- Leakage Physics
- Variation
- Leakage Control
- Memory Leakage
- Estimation
- Conclusion
CLEAN Project: Consortium

- ST Microelectronics (I/F)
- Infineon Technologies (D)

- BullDAST (I)
- ChipVision Design Systems (D)

- University of Catalunya (E)
- Politecnico di Torino (I)
- OFFIS (D)
- WUT (P)
- LETI (F)

- EDAC (D), DTU (DK), COREP (I), BME (H)

IDM = integrated device manufacturer
SME = small and medium sized enterprise
CLEAN Project : Challenges

- **Power** consumption of nanoelectronics is a **limiting factor**

- **Leakage** power will soon become the **dominating** part

- Nanometric processes have large **process variations**
  - leakage & performance variation
  - severe impact on yield

- **EDA** support for low leakage today is extremely **poor**
CLEAN Project: Goals and Keys

- Attack leakage for **65nm** and **45nm** CMOS with innovative tools and circuit solutions
  - SoC-level
  - micro-architectural-level
  - RT-level
  - cell’s library level
  - back-end.

- Develop **prototype EDA tools** to implement **missing parts** of today’s flows

- Integrate and **field-test** those tools into **industrial flows**
CLEAN Project: WP1 Modelling and Design

- **Device level leakage models**
  - analysis of process variability and temperature on 65nm & 45nm

- **Circuit & gate level leakage models**

- **RT block & IP component macro modelling**

- Develop **circuit structures** for
  - sleep-transistor cells
  - voltage anchor cells
  - ABB support

- Develop low-leakage **memory/cache architectures**
CLEAN Project: WP2 Synthesis and Optimization

- Physical synthesis for distributed sleep transistor insertion
- RTL leakage management based on voltage anchoring
- Tools for architectural level ABB-islands
- Architectural level synthesis for power gating
Methodologies, Flows and Tools Specification and Integration

- Definition of EDA requirements for the leakage optimization

- Definition of the flow requirements

- Integration of the WP2 methodologies into
  - PowerOpt (alias ChipVision ORINOCO 2.0)
  - PowerChecker (Bulldast EDA Toolsuite)
CLEAN Project: Prototype Tools Integration

Front-End – ChipVision PowerOpt

Design Description (Java, C++, SystemC)

Leakage Modeling and Optimization

HW/SW Partitioning

Architectural Exploration and (IP) Block Selection

Block Design/Synthesis

Block Integration and Communication Design

RTL Planning/Optimization (DataPath, Memory, Buses, Clock, Power Distribution, Test Structures)

Physical Synthesis

System Level Power Estimator

Behavioral Power Optimization

RTL Power Optimization

Gate/Transistor-Level Power Optimization

BullDAST PowerChecker – Back-End

Proven Seamless FE-BE Integration

CLEAN Project: Prototype Tools Integration

<08/08/13> ISLPED Tutorial
Agenda

- Introduction
  - CLEAN project
  - Motivation
- Leakage Physics
- Variation
- Leakage Control
- Memory Leakage
- Estimation
- Conclusion
What Moore’s law really said: For each technology generation, there is a best number of components with the best per components price.

Reason: The price of a system does not depend on the number of components, thus more components means less cost per component. But above a certain limit, the yield is going down, exponentially increasing the cost again.

Moore’s observation: For each new technology, the best number of components and the best price per component both develop exponentially.
Moore’s law for the Intel processor family:

Also frequency developed exponentially (at least till 2002)
But also the (unwanted) power density (power per unit area) developed exponentially.

Transistor count and frequency have no limit (the more the better), but for power density, there is a bound, we should certainly not pass

(hot plate $\approx 10\text{W/cm}^2$, nuclear fuel rod surface $\approx 100\text{W/cm}^2$, rocket nozzle $\approx 1000\text{W/cm}^2$, sun’s surface $\approx 60000\text{W/cm}^2$)
Motivation: A Timeline of Problems

Scaling has always been done for the sake of transistor count and frequency. All!! other major technology changes have always been done due to power:

- Going from NMOS to CMOS ➔ because of short circuit currents
- Going from constant voltage scaling CVS to constant field scaling CFS (field = voltage per size) ➔ because of dynamic power
- The end of the ‘gigahertz race’ ➔ because of subthreshold leakage (CFS would need threshold voltage scaling, but subthreshold leakage does not allow any further scaling)
- By now, we left the era of mean physics (with various leakage effects) and entered the era or ‘large atoms’:
  - Today’s technology is characterized by variations and tomorrow’s technology will face unpredictable variations and aging (=degradation)
As motivation:
The number of elements, that we need for CMOS production.
In the 1980’s, only 6 elements were needed to build CMOS systems (Si = semiconductor, O= isolation, B,P=doping, Al=interconnect, H=chemical passivation)
In the 1990’s, only 8 new materials were added, all were improving the systems behavior (Cu=better interconnect, Ge=higher mobility, Ta,W= better conductor→semiconductor interface, N=higher dielectricity in the oxide,…)
Nowadays, we use nearly each element (short of the radioactive ones).
It's a well known fact, that there are 3 sources of power consumption:

The dominating one – the dynamic power – is resulting from the energy needed to charge and discharge the load capacitance when doing a transition

There are two reasons why short circuit currents – occurring in transition when both transistors are open – are usually not regarded.

at first: For steep input flanks, the load capacitance buffers the short circuit – limiting the power to some 5-10% of the dynamical one

and then: in terms of power estimation, knowing the input slope, the short circuit currents can be modeled by an equivalent additional load capacitance

Leakage currents in contrast behave completely different:

- they occur even if there is no transition
- they basically flow anywhere
- they do not behave like an equivalent capacitance but an equivalent resistance
- Leakage is not just one effect, but a collection of different effects:

-> next page: collection of these effects
There are 6 different leakage effects (explained for an NMOS transistor):

If the channel is closed: Drain is typically at high, all other terminals at low potential.
We observe:
- subthreshold currents through the channel
- GIDL currents from the gate/drain overlap region to the substrate
- tunneling through the gate oxide from the drain to the gate
- punchthrough from drain to source when both pn junctions touch each other
- pn-junction leakage as known from diodes

If the channel is open, source and drain have the same potential. Typically Source, drain and gate are at high and bulk at low potential. Now we see:
- tunneling from gate to bulk
- junction leakage from drain to bulk

when switching we see hot carrier injection carried by electrons ballistically traversing the gate oxide to the gate and thus carrying a current from gate to channel
Prognosis of the development of the 3 major power effects (for a single transistor disregarding interconnect!).

For bulk CMOS, gate leakage will be skyrocketing. Thus the introduction of high-k devices (which entered the market in 2008) was mandatory. High-k devices will be discussed later. They can remove gate-leakage for good.

From 2010 on, ultra thin body devices will be produced keeping the subthreshold leakage under control.

The introduction of dual gate devices will drastically reduce the leakage for the cost of higher dynamic power (dual gate means dual capacitance to charge).
Motivation: Memory Driver

Prognosis of the leakage and dynamic power development for SRAM and DRAM. Dynamic power depends on the number of accesses, thus it is given in power per MHz. Static power depends on the number of transistors, thus it is given in power per MByte.
Agenda

- Introduction
- Leakage Physics
- Variation
- Leakage Control
- Memory Leakage
- Estimation
- Conclusion
Subthreshold Leakage

- Classical $I_{sd}(V_{gate})$ behaviour for 130nm
- For 45nm the off current rises by $10^6$.
  → subthreshold current in off-state $I_{off}$
- Threshold voltage is $V_{ds}$ dependent.
  → increasing $V_{ds}$ decreases $V_{th}$
  → drain induced barrier lowering (DIBL)

1) Characteristic plot for source-drain current vs. gate voltage for 130nm
   Below a certain threshold voltage:
   subthreshold slope of about 80mV/decade
   Resulting: $I_{off}=10$ fA/µm Weak Inversion and Junction Leakage

2) Trend for smaller nodes: $V_{th}$ is sinking, thus $I_{off}$ is rising exponentially
   Resulting: $I_{off}=30$ nA/µm
   Factor 100 per generation  OR  10 per year

3) DIBL: drain voltage biases threshold voltage -> higher off-current
   Resulting: $I_{off}=300$ nA/µm

4) GIDL: a drain gate voltage lowers the drain’s junction barrier.
   thus higher $I_{off}$ due to junction leakage
Subthreshold Leakage

\[ I_{\text{subth}} \approx \alpha T^2 \frac{W}{L} e^{-\beta \frac{V_{\text{th}}}{T}} \quad \text{if gate is locking and } V_{\text{ds}} \text{ is high} \]

\[ V_{\text{th}} = V_{\text{FB}} + \Phi_{S}(T) + \gamma \left( \sqrt{\Phi_S + V_{bs}} - \sqrt{\Phi_S} \right) - \frac{\left( V_{\text{bi}}(T) - \Phi_S \right) + V_{\text{ds}}/2}{\cosh(L/l_c) - 1} \]

\[ + \alpha (V_{bs}) \frac{\Phi_S T_{ox}}{W + \Delta W} \quad - k_{\text{retro}} V_{bs} + k_{\text{halo}} (L) \sqrt{\Phi_S} + \Delta_{\text{DITS}}(V_{ds}, T) \]

Off-current is exponentially dependent on threshold voltage and the threshold voltage again depends on several parameters. Thus every effect on threshold voltage is also an effect on leakage.

**Simplified equation from BSIM4 manual:**
- first two: zero bias threshold
- next the body effect coefficient. Explained later. Note now: if bulk source voltage is 0, body effect is 0. A positive source voltage results in a positive term thus threshold is higher and leakage lower
- next factors: describe the effect of non uniform doping. We discuss this later on (example: body effect increases \( V_{th} \) with \( V_{bs} \) and retrograd well reduces is again – thus \( V_{th} \) is stabilized. The same holds for \( V_{ds} \) & \( L \))
- the negative DIBL-factor basically depends linearly on \( V_{ds} \) and exponentially on the channel length.

Diagram: The shorter the channel, the more the drain influences the effective channel length
- The last parameter describes the increase of threshold due to a parasitic channel at the sides of the transistor (not very important)
Gate Leakage

Tunneling: Electrons can pass potential barriers, higher than their energy (classically impossible).

Tunneling current exponentially depends on barrier height $V$ and width $T$ and on carrier’s mass $m_{eff}$.

- Energy diagram of a Poly-Si transistor:
  - Leakage current can be carried by tunneling electrons or holes.
    - direct tunneling: from gate to channel
    - Fowler-Nordheim: from gate to oxide

- Carriers leak to source, drain and channel
- Channel leakage is sub-divided into source, drain and bulk leakage

This diagram shows textbook example of tunneling effect at rectangular potential barrier:
Classically impossible, an electron on the left can pass a barrier higher than its energy with a certain probability.

The resulting current density exponentially depends on the thickness of the barrier, the energy difference between electron and barrier and the electrons effective mass.

If we look at a band diagram of gate and channel we see a similar picture:
Electrons as holes can tunnel through the insulating oxide – called direct tunneling or can tunnel into the insulators conduction band – called FN-tunneling

In the overlap region, the tunneling can carry current from the gate to source and drain directly.
The current tunneling to the channel will go to source drain or bulk
There are 3 mechanisms transporting electrons from the p to the n region:

- As known from diodes, we have randomly drifting carriers and electron-hole generation at imperfections.
- As soon as build in potential plus applied reverse bias are higher than the band gap, electrons can directly tunnel from the p-side’s valence band to the n-side’s conduction band.
- This will typically result in very low additional junction-currents.
Smaller barrier means exponentially higher BTBT:
- tech scaling: steeper doping profiles
- Gate Induced Drain Leakage (GIDL)
- GIDL: field from drain to gate reduces junction width
- Body biasing also influences the BTBT current

But a special effect can drastically increase this BTBT:
The potential difference between drain and gate makes the pn-junction steeper and thus the tunneling distance smaller, and thus the current exponentially higher.

This effect is called Gate induced drain leakage.

Diagram: Drain to bulk current of 45nm NMOS:
Behavior separates to left and right part.
Left: Thermal dependence
Right: gate voltage dependence
below the GIDL voltage (of 0.8V), no GIDL influence
- thermal behavior of drifting, electron hole generation
Above GIDL voltage: BTBT is dominating.
- Drain voltage and oxide thickness determine to bulk current
Leakage Physics: Conclusion

- Dynamic short circuit
- Sub-threshold leak
- Gate leak
- Pn-junction leak

Positive correlation:
- Temp
- $L$
- $T_{ox}$
- $V_{BB}$

Negative correlation:
- Delay
- $V_{DD}$
- $V_{DD}$
- Short circuit

Conclusion:
- Negative correlation
- Positive correlation
Agenda

- Introduction
- Leakage Physics
- Variation
- Leakage Control
- Memory Leakage
- Estimation
- Conclusion
**Variation:** Variation Model

- Each parameter $p$ of each transistor can be described as

$$p = p_{\text{nom}} + \delta p_{\text{glb}} + \delta p_{\text{det}}(x, y) + \delta p_{\text{rnd}} + \delta p_{\text{degrad}}(t, OC)$$

- $p_{\text{nom}}$: nominal parameter value

- $\delta p_{\text{glb}}$: global variation (per die)
  - average $p$ on this instance (wafer or die) minus factory average

- $\delta p_{\text{det}}$: deterministic variation
  - depends on the position on the instance (wafer or die)

- $\delta p_{\text{rnd}}$: random variation
  - unpredictable (only distribution function may be given)

- $\delta p_{\text{degrad}}$: degradation (aging)
  - depends on time and operation conditions (OC)

The actual value of one parameter (e.g. the oxide thickness) can vary due to several reasons:

- $p_{\text{nom}}$ is the nominal parameter value, the designer wanted to have
- $p_{\text{glb}}$ is the difference between the $p_{\text{nom}}$ and the de facto average parameter value (combination of fab, lot, wafer and die-to-die variation)
- $p_{\text{det}}$ results from die-to-die variation and leads to a non-uniform, but deterministic variation (e.g.: 1nm less in lower left corner and 2nm more in upper right corner)
- $p_{\text{rnd}}$ variation due to intra-die
Example of impact of process variability on 65nm CMOS

On current is a measure of the device speed (double on-current means double performance).

Here you can see high correlation between on-current (speed) and off-current (energy consumption).

Later, we will exploit this correlation (the fact, that faster devices also have higher leakage power).
Variation: Local Variation

- Random parameter variations → smaller scale than inter-die variations
- Example: $10^8$ Transistors – 100 CP’s – 20 gates on each CP

\[
P = \sum_{i=1}^{10^8} P_{\text{transistor}} = 10^8 \cdot \bar{P}_{\text{transistor}}
\]

\[
D = \max\left(\sum_{j=1}^{100} D_{\text{transistor}}\right)
\]

20 is not a large number

Law of large numbers helps

As presented in this slide, the variation of the input parameters changes the average value of the result.

If you just compute the average leakage by using average length,…. you will underestimate the average.

Reason: As the I(L) dependency is non-linear, the leakage of the average length is smaller than the average leakage of all lengths.

The lower formula gives the correction due to the variation
Variation: Sources of Variability

Random discrete dopants

Line edge roughness

Poly-Si granularity

Layer thickness

OPC variations

Interface roughness

CMP effects
**Variation: Reliability Issues**

### Electromigration

**Caused by:** high temperature, high current density in interconnect
**Results in:** high temperature, high current density in interconnect → runaway problem, sudden death

![Electromigration Image]

### Hot electron oxide degradation

**Caused by:** high $V_{DD}$, very low $V_{BB}$ (strong reverse ABB)
**Results in:** oxide charge $\rightarrow V_{th}$ increase $\rightarrow$ oxide breakdown

![Hot electron oxide degradation Image]

### Oxide breakdown

**Caused by:** oxide charge, dirt in oxide, radiation, mesh errors
**Results in:** sudden runaway $\rightarrow$ oxide failure

![Oxide breakdown Image]

### Negat. Bias Temper. Instability

**Caused by:** hydrogen passivation of PMOS channel lost to the oxide
**Results in:** $V_{th}$ increase

![Negat. Bias Temper. Instability Image]
## Variation: Conclusion

<table>
<thead>
<tr>
<th>Process</th>
<th>Environment</th>
<th>Temporal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Global</strong></td>
<td>L and W, Layer thickness, R’s, Doping, $V_{\text{body}}$</td>
<td>Temp range, $V_{\text{DD}}$ range</td>
</tr>
<tr>
<td><strong>Deterministic</strong></td>
<td>OPC, Phase shift, Layout mediated strain, Well proximity</td>
<td>Self-heating, IR drops</td>
</tr>
<tr>
<td><strong>Stochastic</strong></td>
<td>Random dopants, LER, Poly Si granularity, interface roughness, high-k morphology</td>
<td></td>
</tr>
<tr>
<td><strong>Across-chip</strong></td>
<td>Line width due to pattern density effects</td>
<td>Thermal hot spots due to non-uniform power dissipation</td>
</tr>
</tbody>
</table>

**Well proximity effect:** Some of the ions scattered out of the edge of the photoresist are implanted in the silicon surface near the mask edge, altering the threshold voltage of those devices.

**Pattern density effects:** Dishing & erosion of CMP
Agenda

- Introduction
- Leakage Physics
- Variation
- Leakage Control
  - Transistor engineering
  - Power gating
  - ABB / DVS
- Memory Leakage
- Estimation
- Conclusion
Well engineering by non uniform doping [NUD]

Retrograde well (vertical NUD)
- below the surface:
  - reduce doping concentration
  - increase carrier mobility
- deep in the well:
  - increase doping concentration
  - limit channel depth

Halo doping & pocket implant (lateral NUD)
- increase channel doping at the pn/junction
- steeper pn-junction profile
- less short channel effects

At the technology level, leakage can be reduced by doing sophisticated doping – not uniformly distributed – called well engineering

In the vertical direction:
- doping concentration below the surface can be reduced increasing carrier mobility
- deep below the channel the doping is increased limiting the channel depth and shielding punchthrough currents
- is called retrograde well

In the lateral direction:
- introduce a halo around source and drain having high doping
- the junction profile gets steeper, the channel length gets stabilized reducing SCE
- most of the channel has high carrier mobility

In addition: highly p-doped further pockets can be implanted shielding S/D directly below the surface -> pocket implant

Compare to the definition of the threshold voltage to see the impact of NUD
Oxide dielectricity

- Gate capacitance is performance constrained
- Capacitance is kept high by reducing $T_{ox}$
- But gate oxide may run ‘out of atoms’
- High gate tunneling

\[
C_{ox} \propto k_{SiO_2} \cdot \frac{LW}{T_{ox}}
\]

- Idea: Introduce high-k insulators
- Effective electrical $T_{ox}$ is small
  \[T_{ox}^{eff} = T_{ox} k_{SiO_2} / k_{material}\]
- Physical $T_{ox}$ stays large enough to prevent tunneling

- $k(SiO_2) = 3.9$  $k(SiNO) = 4.1-4.2$  $k(\text{Hf}_2\text{O}_2) = 50$
- But: Problems with lattice sizes
- Gate insulator is added atomic layer by layer

Today in some processes: SiNO nitrided gate oxide = oxinitride
- Increases $k$ from 3.9 to 4.1-4.2
- $T_{ox}$ + 5 to 10 %
Zr=Zirconium $\rightarrow$ $k=40$
Hf=Hafnium $\rightarrow$ $k=72$
**Technology: FD UTB SOI**

- Fully-depleted Ultra Thin Body SOI
- Increased **electrostatic control** due to increased gate-channel coupling
- Advantages of SOI technologies
  - Reduced junction capacitances
  - Immunity to single event upset and to latch-up
  - Increase layout density (no wells)

---

Idea of FD UTB SOI:

1) Ultra thin channel → depletion layer capacitance is lower → higher slope (see subthreshold section)

2) No doping in channel needed → no doping variation
FinFETs

- Create channel as a fin
- Two gates control the channel

- Better channel control
  - higher switching speed
  - lower subthreshold leakage

AMD 15nm FinFET gate

Gate delay: 300fs (3300GHz)

Available: 2009?

TSMC: FinFET already for 65nm for critical devices [EETimes]

FinFET: The channel is constructed as a fin. Thus both sides can be controlled by the source.

As the channel is controlled from both sides, it can be switched on or off more effectively
  → the slope can not be reduced below the fundamental limit of 63mV/decade, but can be brought very close
Technology: Strained Silicon

Strained silicon:

Electron speed: 70% faster
Circuit performance: 35% higher


Idea here: increase the mobility of the carriers

➤ faster electrons means higher performance
Overview over technology development

PDSOI: Partially depleted SOI
Agenda

- Introduction
- Leakage Physics
- Variation
- Leakage Control
  - Transistor engineering
  - Power gating
  - ABB / DVS
- Memory Leakage
- Estimation
- Conclusion


**Power Gating: Overview**

- Sleep transistor insertion
  - low leakage sleep transistors
  - high performance logic
  - implementation as PMOS or NMOS or both
  - decreases leakage in standby mode

**BUT:**
- increases leakage in active mode
- increases dynamic power
- increases area demand
- reduces switching speed
- reduces the maximum stack depth

If not compensated by higher \( V_{DD} \)

To reduce the leakage if the circuit is dysfunctional, we can put a high Vth sleep transistor in series to the circuit.

This will result in very high leakage savings (subthreshold and gate leakage) if the power switches are locking, but if the component is not idle, the application of power-gating has some disadvantages:
**Power Gating: Application**

- **Per gate**
  - only modified *cell library* needed
  - fine grained controllable
  - high *area* overhead

- **Per row**
  - only additional *gating cell* needed
  - delay penalty *hard* to predict

 Basically 3 different implementation styles:

Power-gating each standard cell creates a huge area overhead (needs a lot of gating transistors and a lot of control interconnect).

But to apply this, we do not have to modify our design tools (only the library).

Doing power-gating per stdcell-row is much more efficient in terms of control and area overhead, but standard tools can no longer handle the design. Even full custom new tools will have to significantly modify their implementation, as the delay of a cell is no longer a local problem (mapping is typically done before layout, but now timing depends on the layout, and mapping of course depends on the delay). This will result in a huge overdesign or in much more complex design tools.

The power ring solution is only mend to add power gating to old IP’s which can not be changed.
**Power Gating: Interfacing Problem**

- If \( \text{Sleep}_A = 0 \) => G’s output may be floating.
- Several solutions:
  - let output of A be 1
  - let input of B be not dominant
  - fix output with voltage anchors

---

This slide shows a problem when doing power gating:

Usually, not the entire system sleeps, but some parts are always awake.

Interconnect from a sleeping region to a wake region may have huge problems because the voltage on this interconnect will very slowly go from 0 to 1 (NMOS gating) or 1 to 0 (PMOS gating) over 1000’s of cycles leading to extreme short circuits in the wake region.
**Power Gating: State Retention**

- **Problem**
  - How to store states of registers within a gated region?

- **Solution**
  - Balloon State Retention Flip Flop

There is an additional interfacing problem: It is (as described) important not to produce short circuit by floating inputs. But additionally, when we power gate large regions (including registers), we will lose the state of these registers.

Balloon Flip Flops fulfill both, they work as voltage anchors, and they can save the register state on sleep.

Balloon Flip Flops consist of 3 latches: Two are fast and work as Flip Flop when awake. The third one is slow and leakage saving and store the state when sleep. The black latch is not power gated.
**Power Gating: Ground bounce**

What’s a ground bounce and why does it happen?

- Power grid **not an ideal conductor**, it consists of:
  - Resistive parts
  - Capacitive parts
  - Inductive parts

- During wakeup:
  - Large but slowly changing current during wakeup
  - Resistive part is dominating
  - Result: Reduced voltage

---

<08/08/13> ISLPED Tutorial 46
Agenda

- Introduction
- Leakage Physics
- Variation
- Leakage Control
  - Transistor engineering
  - Power gating
  - ABB / DVS
- Memory Leakage
- Estimation
- Conclusion
Adaptive body biasing (ABB)
- also called VTCMOS (variable threshold CMOS)

In standard CMOS:
- the body voltage for NMOS devices is at 0V
- the body voltage for PMOS devices is at $V_{DD}$

Idea of ABB:
- contact the well of NMOS and PMOS
- vary the well potential
- modify the effective threshold voltage

The idea of ABB is simple. Force the occurrence of the body effect (see subthreshold slides) by applying a body potential which is not 0 for NMOS and not $V_{DD}$ for PMOS.
### ABB / DVS: Background ABB 2/2

- **Analysis:** Vary $V_{\text{ABB}}$ and measure $V_{\text{TH}}$
  - Result for NMOS (and PMOS):
    - an ABB voltage higher than 0V (lower than $V_{\text{DD}}$) will result in a lower effective $|V_{\text{TH}}|$,
    - increased component speed
    - increased subthreshold leakage
  - Effect **weakens** with smaller technology
  - But with high-k ABB will get import again
    - Higher $T_{\text{OX}}$ decreases BTBT current (GIDL effect)

### Spice simulation:

Subthreshold current depends on the body voltage.

In the diagram, you can see:
1. The subthreshold current is rising with shrinking device sizes
2. The effect of forward VBB is stronger than the effect of backward VBB
3. The effect of VBB is lowered with shrinking device sizes
ABB / DVS: Limits of Reverse ABB

- RBB reduces the threshold current
- Increases GIDL to substrate

- The total current has a technology dependent minimum

The BTBT (Band to band tunnelling from drain to substrate) limits the applicability of the RBB.

The minimum of the total current $I_{tot} = I_{sub} + I_{BTBT}$ is (Analysis) at $dI_{tot}/dV_{BB} = 0 \Rightarrow dI_{sub}/dV_{BB} = - dI_{BTBT}/dV_{BB}$

The absolute value of the optimal VBB is technology dependent (here for different doping depths 17nm red curve to 20nm blue curve)

For the 70nm technology the max VBB was 0.1V, and for the 50nm it can be below 0.05V because the effect of BTBT rises with technology.
ABB / DVS: Yield Optimization

- **DVS vs. ABB**
  - DVS more potential because of wide application range

- **Power / Performance**
  optimum with ABB and DVS
  - Values depend on the technology and the system

- **ABB / DVS for yield optimization**
  - varying power and performance in systems due to process variation
  - power and delay limits determine yield
  - ABB / DVS used to move systems closer to the limits

\[ \text{Finding best } V_{BB}/V_{DD} \text{ working point} \]

\[ \text{Design time tuning} \]

\[ \text{Working point depends on measurement} \]

\[ \text{Test time tuning} \]
ABB / DVS: Runtime Adaption

- **ABB example of usage [Tschanz02]:**
  - for a given circuit find the critical path
  - duplicate the critical path (with non blocking inputs)
  - and check functional correctness

ABB voltage can be chosen in several ways:

Fixed (in design time, the ABB is chosen) – is only done for FBB – will be explained later

Once (while testing the chip, the VBB is chosen optimally) – will reduce variance

Adaptively (while runtime) – this can also reduce variability due to temperature and Vdd noise. A replica of the critical path is measured, the VBB is always high enough to ensure timing of the critical path.
# Power Gating vs. ABB comparison

<table>
<thead>
<tr>
<th>ABB</th>
<th>Power gating</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Low benefit</td>
<td>+ High benefit</td>
</tr>
<tr>
<td>+ State preserving</td>
<td>- Loosing state</td>
</tr>
<tr>
<td>+ Compensates $V_{th}$ fluctuation</td>
<td>- Does not compensate</td>
</tr>
<tr>
<td>+ Additional hardware only once</td>
<td>- Additional transistor in series:</td>
</tr>
<tr>
<td></td>
<td>slower, larger, lower yield</td>
</tr>
<tr>
<td>additional interconnect only</td>
<td>(-)For per row gating:</td>
</tr>
<tr>
<td>(+)Only needs modified library:</td>
<td>Timing is hard to validate</td>
</tr>
<tr>
<td></td>
<td>+ Conventional well structure</td>
</tr>
<tr>
<td>+ IP can be reused</td>
<td>- Flip-flops have to be replaced</td>
</tr>
<tr>
<td>- Needs triple well</td>
<td></td>
</tr>
</tbody>
</table>

**Additional notes:**
- **Conventional design tools**
- **IP can be reused**
- **Needs triple well**

---

<08/08/13> ISLPED Tutorial 53
Agenda

- Introduction
- Leakage Physics
- Variation
- Leakage Control
- Memory Leakage
- Estimation
- Conclusion
Cell Design: Leakage in SRAM cells

- Subthreshold leakage occurs when:
  - Transistor is off
  - \( V_{ds} \) is non-zero
- Cell state (stored value) determines exactly which transistors “leak”

![SRAM Cell Diagram](image)

- \( N1, P1 = \text{Internal Leakage} \)
- \( N4 = \text{Bitline Leakage} \)
- \( \rightarrow = \text{Gate Leakage} \)
- \( \leftarrow = \text{BL2WL Gate Leakage} \)
Replace leaky transistors with high-$V_t$ (HV) transistors.
Cell Design: DVT memory cells

Navid Azizi, Farid N. Najm, and Andreas Moshovos, Low-Leakage Asymmetric-Cell SRAM, IEEE Transactions on VLSI, pp. 701-715, Aug. 03

<08/08/13> ISLPED Tutorial 57
Memory Architectures: Cache Decay

- Example: Power down each cache line after 1024 idle cycles
  - 10 bit counter per line needed ➞ huge area & power penalty
- Solution: Power down after 4 ticks (1 tick = 256 cycles)
  - 2 bit counter per line needed
  - shutdown after 769-1024 cycles (phase dependent) (769=1024-255)

[Kaxiras et al.]
**DRI: Dynamically Resizable I-Cache**

Dynamically estimates and adapts to the required instruction-cache size and turns off a selected set of cache lines.

Coarse-grain: entire portions of cache are turned on/off

The number of lines in sleep mode is controlled by periodically examining the miss rate

**Miss rate < pre-determined value**
- Put to sleep another part of the cache

**Miss rate > acceptable value**
- Activate more lines are
Memory Architectures: Drowsy memories

- When accessing a drowsy line, its state is lost
  - Gating circuitry prevents access to drowsy lines

Simple policy: Send drowsy signal each N cycles
No-Access policy: Send drowsy signal N cycles after last access

Comparable performance, Leakage reduction > 80%
Memory Architectures: Conclusion

Replace SRAM cell
-> slight performance reduction
-> high leakage savings

Power down lines after long idle

Make memory drowsy and wake up busy lines
Agenda

- Introduction
- Leakage Physics
- Variation
- Leakage Control
- Memory Leakage
- Estimation
- Conclusion
Estimation: What leakage is modelled?

- Subthreshold leakage
  - \( V_{ds}, V_{bs} \)
- Gate tunneling
  - \( V_{gs} \)
- Gate induced drain leakage
  - \( V_{gd}, T_{ox} \)

Variables:
- \( V_{gs}, V_{ds}, V_{bs} \)
- \( L_{ch}, \ldots \)
- \( T_{ox} \)
**Estimation: Leakage and delay under PTV variation**

- **Global (inter die)**
  - Example: all wafer $T_{\text{ox}}$

- **Local random (intra die)**
  - Example: dopant distribution

- **Local deterministic**
  - Example: well proximity

---

*ISLPED Tutorial*
Estimation: Why RTL modelling?

- SystemC
- behavioural synthesis
- solution
- simulation
- evaluation
- modification
- f_{target}
- \Sigma
- RT synthesis
- inter-die
- parametric yield
- Verilog + UPF
- adaptive solution

- data
  0010 0101 0110 1110
  1001 1100 0000 0100
  1010 0001 1001 1101
  1010 0010 1011 0010

- floorplan
- temperature
- voltage

- E_{dyn}
- area
- T_{crit}
- P_{leak}

SystemC

<08/08/13> ISLPED Tutorial 65
Estimation: The leakage model

\[
\ln\left(\frac{I_{\text{sub}}}{W}\right) = \ln(\frac{\chi}{L}) - \beta V_{\text{TH}}(L, T_{ox}, N_{ch})/T_{ox} \\
= \ldots \approx \alpha_0 + \alpha_1 \sqrt{\frac{N}{L}} + \alpha_2 \frac{1}{L} \sqrt{\frac{T}{N}} + \alpha_3 \frac{1}{L^2} \sqrt{\frac{T}{N}}
\]

\[I_{\text{component}}(\text{Temp}, V_{DD}, V_{BB}, L_{ch}, T_{ox}, N_{dep})\]

- Determined by characterization
- Depend on Temp, V_{DD}, V_{BB}
- Analytical process parameter dependence

\[
I_{\text{component}} = \beta_0 \cdot I_{\text{NMOS sub}}^{\text{NMOS}} + \gamma_0 \cdot I_{\text{PMOS sub}}^{\text{PMOS}} + \beta_1 \cdot I_{\text{NMOS gate}}^{\text{NMOS}} + \gamma_1 \cdot I_{\text{PMOS gate}}^{\text{PMOS}} + \beta_2 \cdot I_{\text{NMOS stack}}^{\text{NMOS}} + \gamma_2 \cdot I_{\text{PMOS stack}}^{\text{PMOS}}
\]
Estimation: The delay model

- Inverter model
- Gate model
- RTL model

Characterization depends on slope, C_{load}, V_{DD}, V_{BB}, temp, L_{ch}, T_{ox}, N_{dep}

\[ d_{gate, input} = f(d_{inv}, C_{load}, temp) \]

\[ d_{component} = f(d_{inv}, temp) \]
Estimation: Process Variation Engine

Inter-die variation data → Monte Carlo → select typical sets → binning → intra-die variation data

model

pre-integrate

<table>
<thead>
<tr>
<th>µL</th>
<th>µT</th>
<th>µN</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

leak t_{crit}

Models have increased expectation value

Estimation: Process Variation Engine
How the model is integrated into the tool:

The tool reads SystemC description and produces RTL netlist + dynamic power prediction + RT level floorplan
Use this information to compute a temperature and $V_{DD}$ map. Decide about an optimization strategy. Now all information except for the variation is available. Variation data comes from the variation engine (see later).

All data for the models is available and leakage is computed.
Leakage is re-iterated in the thermal & electrical mapping.
Estimation: Conclusion

- **Model strengths:**
  - accurate leakage (std dev<10%) and delay (std dev<5%) model
  - inter-die variation is accurately handled even for correlated parameters
  - resulting models are fast, compact and accurate

- **Limitations:**
  - only valid for a small die area, where PTV gradients are negligible (thus each RT component needs its own model)

- **Under development:**
  - expectation value is good for intra-die effect on leakage but SSTA is needed for intra-die delay modeling
Agenda

- Introduction
- Leakage Physics
- Variation
- Leakage Control
- Memory Leakage
- Estimation
- Conclusion
The subthreshold slope is one of the major design metrics. The slope defines the ratio between on current (=speed) and off current (=leakage). If we could reduce the slope, we could easily reduce the threshold voltage to gain performance. Because of the slope, the oxide thickness has to always be aggressively scaled down. This is boosting the gate leakage. Thus subthreshold slope combines the system performance with the two most important leakage sources.

\[ I_{\text{subth}} \approx \alpha \frac{W}{L} e^{-\beta \frac{V_{\text{th}}}{S-T}} \]

\[ S = 1 + \frac{C_{\text{dm}}}{C_{\text{ox}}} \]

\[ I_{\text{gate}} \propto LWV_{\text{DD}}^2 T_{\text{ox}}^{-2} \exp(-\beta T_{\text{ox}} V_{\text{ox}}) \]
Global variation: each parameter that can vary in production and that has an influence on the threshold voltage is contributing to the variation of subthreshold leakage and performance.
Recent Problem: **Variation**

**global variation**
threshold voltage of all transistors is deviating

**influence**

\[
V_{th}\downarrow = \text{performance} \uparrow \quad & \quad \& \quad \text{leakage} \uparrow \\
V_{th}\uparrow = \text{performance} \downarrow \quad & \quad \& \quad \text{leakage} \downarrow \\
\]

**\(V_{th}\)-correlation for one die**

within fab: identical \(V_{th}\)  
wafer-to-wafer: identical \(V_{th}\)  
die-to-die: \(V_{th}\) gradients  
within die: random \(V_{th}\)  

Different sources of variation occur at different production steps and have thus a different influence range (entire lot, one wafer, single die, or each transistor individually)

In fact there are also intermediate effect (for instance resulting in variation gradients over the die)
Recent Problem: **Running out of Atoms**

Our technology is running out of atoms soon.

The random dopant effect will in 45nm result in high statistical fluctuation of the threshold voltage (only 100 dopants are not always distributed smoothly).

In 32nm, each single doping atom may have significant influence on the transistor.
Each material we use in production shows a certain granularity (larger than single atom/molecule) size. The metal clusters for instance will lead to a line edge roughness of approximately 5nm for each metal line. For a 90nm interconnect, this effect may be only additional noise, but for a 22nm line, a +-5nm variation is significant.
Recent Problem: **Running out of Atoms**

The oxide thickness is already the smallest structure we have to control. In 65nm, typical thickness is at 1.5nm (which is 5 molecule layers). With high-k, this size is becoming larger again (approx 10-15nm).
Degradation (=aging of the components) is the next problem that we will have to face.

Electro-migration: At high temperatures and current densities, the metal ions follow the momentum of the electrons.

Problem: At the thinnest position the current density and temperature both are highest. The thinner the connection, the faster its material vanishes.

→ Drastical end with a runaway situation (see sudden northwood death syndrome)
## Recent problems: Degradation

<table>
<thead>
<tr>
<th>Electromigration</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Caused by:</strong> high temp, high current density</td>
</tr>
<tr>
<td><strong>Results in:</strong> high temp, high current density → runaway problem, sudden death</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hot Electron Degradation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Caused by:</strong> high $V_{DD}$, very low $V_{BB}$</td>
</tr>
<tr>
<td><strong>Results in:</strong> oxide charge $\rightarrow V_{th}$ increase $\rightarrow$ oxide breakdown</td>
</tr>
</tbody>
</table>

Hot electron degradation (there are 4 different effects causing hot electron degradation).

Electrons can enter the oxide and are trapped there pre-charging the channel, and thus influencing the threshold voltage.

Systems can recover from hot electron degradation. The change is not permanent.
### Recent problems: Degradation

<table>
<thead>
<tr>
<th><strong>emission</strong></th>
<th><strong>Caused by:</strong> high temp, high current density</th>
<th><strong>Results in:</strong> high temp, high current density → runaway problem, sudden death</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>hot electron degradation</strong></td>
<td><strong>Caused by:</strong> high $V_{DD}$, very low $V_{BB}$</td>
<td><strong>Results in:</strong> oxide charge → $V_{Th}$ increase → oxide breakdown</td>
</tr>
<tr>
<td><strong>oxide breakdown</strong></td>
<td><strong>Caused by:</strong> oxide charge, dirt in oxide, radiation, mesh errors</td>
<td><strong>Results in:</strong> sudden runaway → oxide failure</td>
</tr>
</tbody>
</table>

Due to charges, dirt, mesh errors or radiation impact in the oxide, the current can punch through the oxide, resulting in a permanent device failure.
**Recent problems: Degradation**

<table>
<thead>
<tr>
<th>Degradation Type</th>
<th>Caused by</th>
<th>Results in</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electromigration</td>
<td>high temp, high current density</td>
<td>high temp, high current density (\rightarrow) runaway problem, sudden death</td>
</tr>
<tr>
<td>Hot Electron Degradation</td>
<td>high (V_{DD}), very low (V_{BB})</td>
<td>oxide charge (\rightarrow V_{th}) increase (\rightarrow) oxide breakdown</td>
</tr>
<tr>
<td>Oxide Breakdown</td>
<td>oxide charge, dirt in oxide, radiation, mesh errors</td>
<td>sudden runaway (\rightarrow) oxide failure</td>
</tr>
<tr>
<td>Negative Bias Temperature Instability</td>
<td>hydrogen passivation of PMOS channel lost to the oxide</td>
<td>(V_{th}) increase</td>
</tr>
</tbody>
</table>

The passivation between channel and oxide (hydrogen atoms) can vanish (attracted by the gate).

The effect just occurs in PMOS devices when the transistor is conduction (bulk at VDD and gate at ground) and the bulk\(\rightarrow\)gate voltage is negative.
Which anti-leakage techniques are already available?

We already have well engineering since 2002 in order to control the threshold voltage dependence to other parameters (well engineering is still being further developed)
Recent Transistor Design

MOSFET structure
1. poly-silicon gate
2. buffer oxide layer
3. gate sidewall isolation
4. gate oxide
5. metal drain contact
6. elevated source
7. source
8. shallow source extension
9. channel
10. halo doping
11. pocket implant
12. retrograde well
13. field oxide (FOX)

well engineering 2002
stabilize $V_{th}$
6, 8: reduce $R_{sd}$
10: reduce $V_{th}(L_{ch})$ dep.
11: reduce $V_{th}(V_{DD})$ dep.
12: reduce $V_{th}(V_{be})$ dep.

strained silicon 2004
70% faster electrons
35% higher performance

Since 2004, we can control (increase) the mobility of the majority carriers (electrons or holes) to increase the performance without threshold voltage reduction.
## Recent Transistor Design

### MOSFET structure
1. poly-silicon gate
2. buffer oxide layer
3. gate sidewall isolation
4. gate oxide
5. metal drain contact
6. elevated source
7. source
8. shallow source extension
9. channel
10. halo doping
11. pocket implant
12. retrograde well
13. field oxide (FOX)

### well engineering 2002

#### stabilize $V_{th}$
- 6,8: reduce $R_{sd}$
- 10: reduce $V_{th}(L_{ch})$ dep.
- 11: reduce $V_{th}(V_{DD})$ dep.
- 12: reduce $V_{th}(V_{be})$ dep.

- Integrated 45 nm CMOS process
- High performance
- Low leakage
- Meets reliability requirements
- Manufacturable in high volume

### high-k 2008

- high slope is needed
- high slope needs high $C_{ox}$
- $C_{ox}$ sinks with L & W scaling
- $T_{ox}$ can no longer be reduced → introduce HfO$_2$

### strained silicon 2004

- 70% faster electrons
- 35% higher performance

Since 2008, also high-k technology is available (e.g. Intel Penryn) to control gate leakage.
Which anti-leakage techniques are about to come?

Multi-gate devices such as FinFET will combine SOI, high-k and an increased electrical control over the channel.

With FinFETs the subthreshold slope will be brought very close to the fundamental limit enabling further threshold voltage and thus also frequency scaling.
**Future Transistor Design**

**multi-gated transistors**

increasing electrostatic control approach the 63mV limit

**UTB FD SOI**

fully depleted ultra thin body SOI

- ultra-thin: SOI layer thinner than depletion width
- channel is undoped → no wells
- higher integration density
- no doping variation

Fully depleted ultra thin body devices will enable bulk CMOS at least down to the 32nm node. The undoped channels avoid random dopant variation and increase the integration density, as no wells are needed.
Power gating in principal is very easy: The logic is implemented in fast, but high leakage devices. If the component is idle, a slow, but low leakage gating transistor in series to the gate cuts the leakage currents paths.
Gate Level Techniques: **Power Gating**

The problems of power gating are mostly technologically:

Depending on the gating transistor size, area overhead, remaining leakage, performance degradation, wakeup time and supply noise can be traded off.

At wakeup, the gates are in an intermediate, high short circuit state for a very long time.

Power gating can be done gatewise (high area overhead, but with a modified library, all design tools can be reused)

or rowwise (conceptually better, but especially the timing of each cell now depends on the activity of all neighbor cells)
Gate Level Techniques: **Power Gating**

**easy concept**
- use fast, low $V_{th}$ transistors for logic
- use low leakage, high $V_{th}$ transistors to cut leakage path on idle

**technical problems**
- gating transistor sizing:
  - supply noise and wakeup delay
  - floating outputs
  - row- or cell based gating

**state retention**
- sequentials are losing state at power down voltage anchors, balloon latches, dual $V_{DD}$

---

When using power gating, the interface between gated and not gated parts have to be stabilized.

When going to sleep, the state of the sequentials has to be stored and then restored, when waking up.

For both problems, design for power gating at the behavioural level can avoid or at least reduce the overhead for interfacing and state retention.
Above the gate level, the most promising technique is adaptive leakage management. The idea is always to either speed up slow systems for the cost of leakage power or the slow them down, reducing the leakage.

The plot here show a system’s power and performance distribution before… [see next page]
on-chip binning

exploit correlation of leakage and performance adaptively slow down the system to save leakage

...and after application of ABB with AVS, all systems would exactly meet its performance constraint for the least possible leakage.
Above the gate level, the most promising technique is adaptive leakage management. The idea is always to either speed up slow systems for the cost of leakage power or the slow them down, reducing the leakage.

The available techniques are adaptive voltage scaling, adaptive body biasing, or a combination of both.

Depending on the application time, different effects can be exploited:
At design time, these techniques are not adaptive, nevertheless, by identifying the best body bias/supply working point, leakage can be reduced without a performance degradation (thus, this sweetspot tuning should always be regarded as son as body bias control is available)
At test time, ABB/AVS techniques can exploit the strong leakage-delay correlation of global process variations. Testtime tuning has a low overhead, as measurements of leakage and delay can be made off-chip by the testing equipment.
ABB/AVS at boot-time can also exploit aging margins in young systems to reduce leakage and thus slow down the (thermal and current driven) aging effects. With bot-time tuning, onchip monitors for delay (and temperature) have to be introduced
At runtime, also margins for thermal and voltage variations can be used to reduce leakage as long as the system is not in its worst case condition. But of course, runtime tuning also has the largest implementation overhead. The system has to be made ready to cope with a runtime adoption of the supply voltage.
Depending on the application time, different effects can be exploited:

At design time, these techniques are not adaptive, nevertheless, by identifying the best body bias/supply working point, leakage can be reduced without a performance degradation (thus, this sweetspot tuning should always be regarded as soon as body bias control is available).

At test time, ABB/AVS techniques can exploit the strong leakage-delay correlation of global process variations. Testtime tuning has a low overhead, as measurements of leakage and delay can be made off-chip by the testing equipment.

ABB/AVS at boot-time can also exploit aging margins in young systems to reduce leakage and thus slow down the (thermal and current driven) aging effects. With bot-time tuning, onchip monitors for delay (and temperature) have to be introduced.

At runtime, also margins for thermal and voltage variations can be used to reduce leakage as long as the system is not in its worst case condition. But of course, runtime tuning also has the largest implementation overhead. The system has to be made ready to cope with a runtime adoption of the supply voltage.
Final conclusions:
The leakage currents will be kept under control, but we have to sacrifice all frequency scaling.

From the technology side, transistors will be made less sensible to parameter variation, and the remaining global variation can be reduced by adaptive techniques.

Some statistic variations will be bypassed (e.g. random dopant effect). Some effects are not really statistic (as well proximity effect or OPC related disturbances). With better estimation software, they may be made deterministic and can then be avoided by design (If well proximity effect is always increasing a certain transistor’s threshold voltage, we can reduce the threshold by design. The final device will have exactly the desired properties.).

But especially the line edge roughness will remain a critical source of statistical variation.

Degradation will be the next problem, we will have to face.