Low Power CMOS Process Technology

Scott Crowder

IBM, SRDC, East Fishkill, NY
Outline

• What is Low Power?
• Process For Low Standby Power
• Process For Low Active Power
• Other Low Power Requirements
• Conclusions
Power Components in Digital CMOS

- **Standby Power**
  - Power when no function is occurring
  - Critical for battery driven
  - Can be reduced through circuit optimization
  - Temperature dependent leakage current dominates power

- **Active Power**
  - Switching power plus passive power
  - Critical for higher performance applications

- **Other Sources of Power Consumption**
  - Analog & I/O Power
    - Some circuit configurations have current flow even w/o switching
  - Dynamic Memory Refresh Power
    - eDRAM, 1T SRAM require periodic refresh of data
Power Components (simple version)

Power = Switching Power + Passive Power

\[ f \cdot C \cdot V \cdot V \]

\[ I(V,T) \cdot V \]

Frequency

Capacitance

Voltage

Current

Voltage
Power Trends

Handheld Technology
High Vt Devices, 25°C

Passive Power (picoWatts/Micron)

<table>
<thead>
<tr>
<th>180nm</th>
<th>130nm</th>
<th>90nm</th>
<th>65nm</th>
<th>45nm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Desktop Processor Technology
Base Devices, 10% Activity, 105°C

Power for 10 x 10 mm chip (Watts)

<table>
<thead>
<tr>
<th>180nm</th>
<th>130nm</th>
<th>90nm</th>
<th>65nm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What is Low Power in CMOS?

• Depends on product application
• Examples
  #1: Handheld phone chip
  #2: Server microprocessor chip
Example #1: Handheld Application

• Key Attributes:
  – Battery Powered
  – In off-state significant portion of time
  – Standby condition at room temperature
  – Lawsuit if chip gets too hot
  – Standby dominates power requirements
    • Leakage current at room temperature dominates power

• What is Low Power Optimized Design?
  – Want highest performance @ given leakage current
Example #2: Server Application

• Key Attributes:
  – Powered by electrical grid
  – Almost always on
  – Power condition at operating temperature
  – Lower power reduces packaging & cooling expense
  – Switching power dominated power requirements in past
    • Want lower capacitance, voltage at maximum frequency
  – Passive power at operating temperature now a major component
    • Want lower high temperature leakage current & lower voltage

• What is Low Power Optimized Design?
  – Want switching power to dominate
  – Want to maximize performance & modulate voltage to meet power requirements
One More Major Component: Cost

• Critically important to technical solution

• Example #2: Server microprocessor
  – Higher performance = higher end product price
  – Will add reasonable cost to enable performance
    • Higher cost process: strain engineering, added metal levels
    • Flip-chip packaging, heat sinks & fans

• Example #1: Handheld application
  – Market will not allow higher cost solutions
  – Lower power specifications
  – Focus on lowest cost process that meets market demand
Low Standby Power
Components of Passive Power

- **Major components**
  - **Source leakage current**
    - Sub threshold off-state current
  - **Gate current**
    - Tunneling current thru gate dielectric
  - **Well current**
    - Band-to-band tunneling current to well
IS: Source Sub threshold Leakage

- **Strong Vth dependence**
- **Strong Temperature dependence**
Under scaling Supply Voltage

- Higher performance at given leakage power by raising Vdd
- Vdd scaling has slowed dramatically for low standby power
Methods to Modulate IS Leakages

- Reverse Body Bias or Active Well

Added well implant to separate well and substrate
Reverse Body Bias

Body bias dependence limited by IG & IW leakage

Leakage dominated by IS but low body effect
IG: Gate Leakage

- Function of gate dielectric material and thickness
- Low temperature dependence -> dominates at low T

Kraus, et al, IEEE TED Vol 52, No. 6 2005, page 1141
Methods to Reduce Gate Leakage: N

• Increased Nitrogen concentration reduces gate leakage

Kraus, et al, IEEE TED Vol 52, No. 6 2005, page 1141
Methods to Incorporate Nitrogen

- **Plasma Nitridation**
  - SiO₂ base oxide
  - N plasma
  - More N at top surface

- **Thermal Nitridation**
  - SiO₂ base oxide
  - 700-1000°C NH₃, NO
  - N throughout

- **N-Implant**
  - Oxidation
  - More N at bottom

---

**iedm**

Process Technology/Scott Crowder
## Methods to Reduce Gate Leakage: High-K

![High-k/Metal Gate Stack: \( T_{\text{inv}} = 14.5\text{A}, T_{\text{oxGL}} = 10\text{A} \)]

### Benefits
- >1000x lower IG
- Lower \( T_{\text{inv}} \) at same IG to enable device scaling

### Issues
- \( V_{\text{th}} \) control (WF engineering)
- Charge trapping
- Mobility degradation
- Increased complexity
Mobility

- Acceptable long channel mobility with Hf-based oxides
- Requires optimization of HfSiO(N)/SiO(N) composition

[Source: M Frank et al., VLSi-TSA05]
Gate Workfunction Tuning

- Optimized high-k dielectric stack reduces pFET Vt shift by 200~300 mV
- Selective use of ultra-thin interface/surface modification layers

M. M. Frank et al VLSI-TSA 2005
IW: Tunneling & Trap-Assisted Current

• Band to band tunneling increases with doping levels & steeper gradients

• Gate field enhanced tunneling increases with thinner gate dielectric

• Reduction of trap assisted current varies with cause
Methods to Reduce Junction Leakages

- Leakage modulated primarily by transistor doping profiles

NFET Junction Leakage vs. Halo Angle

PFET Junction Leakage vs. S/D Energy

Ref: Hook, ACEED 2005
Low Active Power
Components of Active Power

Power \sim f \times C \times V \times V + I(V,T) \times V

Higher Transistor Performance at Same Leakage

\downarrow

Same Fmax at Lower Voltage

\downarrow

Lower Active Power
Performance Scaling

Traditional performance scaling: Shorter Gate Length
Higher Gate Capacitance (IG)
Shallower Junctions (IW)
Lower Threshold Voltage (IS)
Lower Supply Voltage

All did not scale in 65nm technology

Higher performance with: Higher Supply Voltage
Higher Inversion Capacitance
Higher Mobility
Mobility Innovation: Stress Memory

Amorphous layer
By implant

Si is crystallized and stress is “memorized”

Control

Tensile Nit stressor

[Chan, CICC ’05]

nFET Ion (µA/µm)
nFET Ioff (nA/µm)
Mobility Innovations: Strained Liner

[NMOS] Un-stress tensile

[b) PMOS] Un-stress Compressive

[Chan, CICC ’05]
Mobility Innovations: Embedded SiGe

- Introduces compressive stress
- Lowers contact resistance
Mobility Innovations: <100> PFET
Mobility Innovations: <100> PFET

- Significant boost in PFET current without added process steps
- Loss of ability to increase current with compressive stress

[Chan, CICC ’05]
Mobility Innovations: Hybrid Orientation

• 110 surface: 2x higher PFET $\mu$
  >2x lower NFET $\mu$

• Hybrid Orientation
  • 110 for PFET
  • 100 for NFET

Process Challenges:
- Low defect epitaxy
- Planarization
- Trench isolation
Inversion Capacitance: Metal Gates

Metals 3-5 nm thick on 500 nm SiO₂, anneal 3 °C/s up to 1000 °C in FG
Binary/ternary alloys: W₂N*, Ta₂N*, TaN*, TaSiN*, RuO₂**, CoSi₂, NiSi

- Metal gate incorporation to remove gate depletion
- Holy Grail: Band edge workfunction & thermally stable
Metal Gate Integration Options: FuSi

“One Step”
- Metal (Ni,Co)
- Polysilicon
- Silicide

Anneal, Strip, Anneal

“Two Steps”
- Metal Deposition
- Metal (Ni,Co)
- Polysilicon
- Silicide

Anneal, Strip, Anneal

NiSi

[Kedzierski, IEDM '02]
Challenges for FuSi

• Dual Workfunction creation
  – Alloys
  – Doping
  – Phase change

• Workfunction control & stability
• Variation with gate density & size
Metal Gate Integration Options: Replacement Gate

Challenges:
- Integration of different workfunction metals: high complexity
- High-k deposition on small gate area
- ALD probably required for gate metal
Metal Gate Workfunction Stability

- Workfunction of integrated metal gates has strong dependence on deposition & annealing conditions

![Deposition and Annealing Diagram](image)

- Potential root causes of workfunction shifts
  - Metal induced gap states (MIGS)
  - Oxide charge (O vacancies) in HfO₂
  - Oxidation of interfacial metal
  - Hf-diffusion into metal (WF change)
  - Dipole layers at SiO₂/HfO₂ interface

[E. Cartier, VLSI '05]
Memory Limitation to Voltage Scaling

- SRAM cell stability limits scaling of array Vdd
  - Vth mismatch due to dopant fluctuation in best case is constant with scaling
  - Lack of gate dielectric scaling has caused it to increase
  - Reduction of all other components of mismatch critical

- Design Options
  - Larger cells
  - Separate array Vdd
  - Alternative cells/memory
    - 8T SRAM
    - eDRAM

Butterfly curve comparison for scaled cells [VLSI 2005]
Components of Switching Power

Power \sim f \ast C \ast V \ast V

Capacitance

Transistor

Metallization

\( C_{cg}, C_{mg}, C_{g}, C_{ov}, C_{d, \text{per}}, C_{d, \text{area}}, C_{d, \text{per}}, C_{\text{vert}}, C_{\text{hor}} \)
Capacitance Reduction: SOI

- Reduction of area diffusion capacitance to negligible level
- Reduction of perimeter capacitance with film scaling
- Additional AC drive current benefit from gate coupling
Options to Gate Capacitance Scaling

• Need better electrostatic control to scale $L_G$
  – Higher gate inversion capacitance
    • High-K gate dielectric material to reduce EOT
    • Metal gate material to eliminate gate depletion
  – Fully depleted double gate structures
    • Better channel control through structural change

• $L_G$ must decrease more than $C_{inv}$ increases
Capacitance Scaling with Metal Gate

- $L_G$ must decrease to prevent capacitance increase

- Short channel effect degradation when metal gate workfunction deviates too far from band edge
  - Lose portion of performance benefit and most of power benefit
One Double Gate Structure: FinFET

Process Challenges:
- Fin creation and control
- Source/drain resistance engineering
- Contact formation and control
BEOL Capacitance Scaling

Performance \sim R \times C

Capacitance \sim k_{\text{eff}} \times h / \text{space}

Resistance \sim \rho / (h \times w)

Reduce inter-level dielectric constant => lower capacitance

Reduce resistance => lower height => lower capacitance
Capacitance Scaling: Reducing BEOL $k$

- Multiple porous dielectric options with $k < 2.4$
  - Spin-on and CVD deposition options
- All materials have lower material strength as $k$ decreases
  - Optimization of BEOL stack & packaging required


Capacitance Scaling: Scaling $h$

- Copper will continue to be material for low $\rho$ metallization

**Assumptions**
- Grain size = 2x line width
- Smooth surfaces
- Mean free paths: $\text{Al}=15 \text{ nm}$; $\text{Cu}=39 \text{ nm}$; $\text{Au}=40 \text{ nm}$; $\text{Ag}=45 \text{ nm}$

Capacitance Scaling: Scaling h

- Scaling requires thinner liner & seed films
  - 45nm dimensions demonstrated
  - Porous dielectrics must be compatible with thin liner to have benefit

- ALD liner & Cu seed reduction provide opportunity for additional scaling

ITRS 45nm PVD liners demonstrated

Cu BEOL using Ru liner

[Edelstein, 2005]
Passives Scaling: BEOL Capacitors

- Benefit of scaling: better “free” BEOL capacitors
- Good tolerance due to averaging over many levels
Passives Scaling: Decoupling Capacitors

- Gate capacitor scaling not possible with low passive power
- >10x reduction in area & >100x reduction in leakage with trench capacitors
Conclusions: Handheld Low Power

- Standby power dominated

- Choices:
  - Add strain engineering without gate scaling
  - Add high-K dielectrics to enable gate scaling
  - Add design complexity to modulate leakage

- Decision will be driven by lowest cost solution
Conclusions: Plugged In Low Power

- Active Power Dominated
- Performance elements added as developed

Choices:
  - Lower supply voltage
    - May require new memory choice
  - Add design complexity to reduce both switching and passive power
    - Voltage islands, clock gating, etc…