Synthesis for Low Power: A New VLSI Design Paradigm for DSM

Ajit Pal

Professor
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
INDIA -721302
Outline

• Why Low Power?
• Sources of power dissipation
  – Dynamic Power
  – Static Power
• Degrees of Freedom
• Low Power Techniques
  – Supply Voltage Scaling Techniques
  – Minimizing switching capacitances
  – Leakage Power reduction Techniques
Why Low-power?

• Until recently performance has been synonymous with circuit speed or processing power, e.g. MIPS or MFLOPS.
• Implementation involved Area-Time tradeoff. Power Consumption = k.A.f, where k = 0.063 W/cm².MHz, A is the area in cm² and f is the frequency in MHz.
• Power consumption were of secondary concern.
Why Low-power?

- Contemporary high performance processors consume heavy power
- Cost associated with packaging and cooling such devices is prohibitive
- Low-power methodology to be used to reduce cost of packaging and cooling

<table>
<thead>
<tr>
<th>Processor</th>
<th>Clock (MHz)</th>
<th>Technology (mm)</th>
<th>$V_{dd}$ (Volt)</th>
<th>Peak Power (Watt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ultra Sparc</td>
<td>167</td>
<td>0.45</td>
<td>3.3</td>
<td>30</td>
</tr>
<tr>
<td>Intel Pentium</td>
<td>200</td>
<td>0.50</td>
<td>3.3</td>
<td>26</td>
</tr>
<tr>
<td>Alpha 21064</td>
<td>200</td>
<td>0.50</td>
<td>3.3</td>
<td>30</td>
</tr>
<tr>
<td>Alpha 21164</td>
<td>300</td>
<td>0.45</td>
<td>3.3</td>
<td>50</td>
</tr>
<tr>
<td>Alpha 21264</td>
<td>667</td>
<td>0.35</td>
<td>2.0</td>
<td>72</td>
</tr>
<tr>
<td>Alpha 21364</td>
<td>1000</td>
<td>0.25</td>
<td>1.5</td>
<td>100</td>
</tr>
</tbody>
</table>
• Lead processor power increases every generation
• Compactions provide higher performance at lower power
Why Low-power?
Why Low-power?

- Emergence of portable computing and communication equipment, such as laptops, palmtops, cell-phones, etc. Growth rate of these portable equipment are very high.
- As these devices are battery operated, battery life is of primary concern. Unfortunately, the battery technology has not kept up with the energy requirement of the portable equipment.
- Commercial success of these products depend on weight, cost and battery life.
- Low power design methodology is very important to make them commercially viable.
Why Low-power?

- Reliability is closely related to power dissipation – Every 10°C rise in temperature roughly doubles the failure rate

Onset temperatures of various failure mechanism

- Thermal runway
- Gate dielectric
- Junction diffusion
- Electromigration diffusion
- Electrical parameter shift
- Package related failure
- Silicon interconnect fatigue

°C above normal operating temperature
Why Low-power?

- According to an estimate of the U.S. Environmental Protection Agency (EPA), 80% of the power consumption by office equipment are due to computing equipment and a large part from unused equipment.
- Power is dissipated mostly in the form of heat. The cooling techniques, such as AC transfer the heat to the environment.
- To reduce adverse effect on environment efforts such as EPA’s *Energy Star* program leading to power management standard for desktops and laptops has emerged.
- Drive towards Green PC
Sources of Power Dissipation

- CMOS has emerged as the technology of choice for low power applications and is likely to remain so in the near future.
- The sources of power dissipation in CMOS circuits
  - Dynamic Power
  - Static
Dynamic Power Dissipation

- Dynamic Power
  - Switching power
  - Short-circuit power
  - Glitching power

- Static Power
  - Due to reverse-biased junction diode currents when the transistors are off
  - Due to sub-threshold leakage current
Dynamic Power Dissipations

- **Switching Power**
  - Due to the charging and discharging of load and parasitic capacitors

\[
P_{\text{dynamic}} = \alpha_L \cdot C_L \cdot V_{DD}^2 \cdot f + \sum_i \alpha_i \cdot C_i \cdot V_{DD} \cdot (V_{DD} - V_T)
\]
Switching Power

During transition of the output from 0 to $V_{dd}$, the energy drawn from the power supply is given by

$$E_{0 \rightarrow 1} = \int_{0}^{V_{dd}} p(t) dt = \int_{0}^{V_{dd}} V_{dd} i(t) dt$$

$$i(t) = C_L \frac{dV}{dt}$$

Substituting this we get

$$E_{0 \rightarrow 1} = V_{dd} \int_{0}^{V_{dd}} C_L dV = C_L V_{dd}^2$$

If a square wave of repetition frequency $f \ (I/T)$ is applied at the input then the power dissipated per unit time is given by

$$P_d = \frac{1}{T} C_L V_{dd}^2 = C_L V_{dd}^2 f$$

Ajit Pal  IIT Kharagpur
Dynamic Power Dissipation

- Short Circuit Power Dissipation
  - As input changes slowly, power dissipation takes place even when there is no load or parasitic capacitor. This is known as the short circuit current.
  - Note that the short circuit power dissipation is greatly affected by the power supply scaling and is also proportional to the frequency and rise/fall time of the input signal.

\[ I_{sc} = \frac{1}{12} \cdot \frac{k \tau f}{V_{DD}} \cdot (V_{DD} - V_T)^3 \]
Short Circuit Power Dissipation
Short Circuit Power Dissipation

\[ I_{\text{mean}} = 2 \times \frac{1}{T} \left[ \int_{t_1}^{t_2} i(t)dt + \int_{t_2}^{t_3} i(t)dt \right] \]

Because of symmetry we may write

\[ I_{\text{mean}} = \frac{4}{T} \left[ \int_{t_1}^{t_2} i(t)dt \right] \]

For the nMOS transistor is operating in the saturation region

\[ I_{\text{mean}} = \frac{4}{T} \left[ \int_{t_1}^{t_2} \frac{\beta}{2} V_{\text{in}}(t) - V_t)^2 \right] dt \]

Contd...
Short Circuit Power Dissipation

\[ I_{mean} = \frac{2 \beta}{2} \int_{\frac{\tau}{V_{dd}}}^{\frac{\tau}{V_{t}}} \left( \frac{V_{dd}}{\tau} t - V_t \right)^2 dt \]

This results in

\[ I_{mean} = \frac{\beta}{12V_{dd}} (V_{dd} - 2V_t)^3 \frac{\tau}{T} \]

Short circuit power is given by

\[ P_{sc} = V_{dd} \cdot I_{mean} = \frac{\beta}{12} (V_{dd} - 2V_t)^3 \tau \cdot f. \]

Ajit Pal    IIT Kharagpur
Glitching Power

Output waveform showing glitch at output $O_2$
Leakage Current Mechanisms of Deep-submicrometer Transistors

Ajit Pal    IIT Kharagpur
Static Power Dissipation

- $I_1 =$ Reverse-bias p-n junction diode leakage current
- $I_2 =$ Band-to-band tunneling current
- $I_3 =$ Subthreshold leakage current
- $I_4 =$ Gate Oxide tunneling current
- $I_5 =$ Gate current due to hot-carrier injection
- $I_6 =$ Channel punch-through
- $I_7 =$ Gate induced drain-leakage current
Reverse Biased Leakage

nMOS is ON

pMOS is OFF

V_{out} = "0"

Drain leakage

Reverse leakage current

p-type substrate

Ajit Pal    IIT Kharagpur
p-n Junction Reverse-Biased Current

nMOS inverter and its physical structure

Ajit Pal    IIT Kharagpur
The current for one diode is given by:

\[
I_{rdlc} = A J_s \left( e^{\frac{q V_d}{n K T}} - 1 \right)
\]

Where,

- \( Js \) = reverse saturation current density, \( V_d \) = diode voltage,
- \( n \) = emission co-efficient of the diode
- \( q \) = charge of an electron \((1.602 \times 10^{-19})\),
- \( K \) = Boltzmann constant \((1.38 \times 10^{-23} \text{ J} / \text{k})\),
- \( T \) = temperature in °K

Then total static power dissipation due to diode leakage current for 1 million transistors is given by:

\[
P = V_{dd} \sum_{i=1}^{10} I_{di} \approx 0.01 \mu W
\]
High electric field across reverse-biased p-n junction causes significant current known as BTBT current, which dominates the p-n junction leakage current
Band-to-Band Tunneling Current

The tunneling current density is given by

$$J_{b-b} = A \frac{E V_{app}}{E_g^{1/2}} \exp \left( -B \frac{E_g^{3/2}}{E} \right)$$

Where,

$$A = \frac{\sqrt{2m^* q^3}}{4\pi^3 \eta^2}$$

and

$$B = \frac{4\sqrt{2m^*}}{3q\eta}$$

Ajit Pal   IIT Kharagpur
Sub-threshold Leakage Current

Static power due sub-threshold leakage current:

\[ I_{\text{sub}} = A e^{\frac{q}{n'} kT} \left( V_G - V_S - V_{\text{tho}} - \delta' V_S + \eta V_{DS} \right) \left( 1 - e^{\frac{-q V_{DS}}{kT}} \right) \]

- This current increases drastically with temperature
- It also increases as threshold voltage is scaled down along with the power supply voltage for better performance.
Various mechanism which affect the subthreshold leakage current are:

- Drain induced Barrier Lowering
- Body effect
- Narrow-width effect
- Effect of channel length and Vth Roll off
- Effect of temperature
Contributions of Various Power Dissipations

- Switching power: 80%-90%
- Leakage power: 10%-30%
- Short-circuit power: 0%-5%
Why Leakage Power is an Issue?

- In stand-by application leakage component becomes significant % of total power
- Leakage current approaches 10% of total power in sub deep micron technology
Why Leakage Power is an Issue?

Leakage power is becoming a large component of total power dissipation.

Ajit Pal  IIT Kharagpur
Three degrees of freedom inherent in the low-power design space:

- Supply voltage
- Physical capacitance
- Switching activity

Optimizing power consumption invariably involves reducing one or more of three parameters

The supply voltage has the most dominating effect on power dissipation because of quadratic relationship
Low-Power Design Methodology

- Low-power design methodologies are to be applied throughout the design process from system-level to layout-level, gradually refining or detailing the abstract specification or model of the design.

- Starting with the system specification the following steps are performed to get the layout:
  - System Specification => System-level Design
  - Behavioral Description => High-level Synthesis
  - Structural RTL Description => Logic Synthesis
  - Logic-level netlist => Layout Synthesis => Layout
Power Reduction at Different Levels

Power optimization approaches at the high-level are significant since research results indicate that higher levels of abstraction have greater potential for power reductions.
Supply Voltage Scaling

- Device feature size scaling
- Architectural level approaches
  - Parallelism
  - Pipelining
- Voltage scaling using high-level transformations
- Dynamic voltage scaling
Supply Voltage Scaling for Low Power

- A factor of two reduction in supply voltage yields a factor of four decrease in energy.
- Theoretical lower limit of supply voltage for CMOS circuit is 0.2V.
- Unfortunately, as supply voltage is lowered delays increases leading to dramatic reduction in performance.
Device Feature Size Scaling

1. \( t'_{ox} = \frac{t_{ox}}{S} \)
2. \( N_D' = N_D \times S \)
3. \( L' = \frac{L}{S} \)
4. \( X_j' = \frac{X_j}{S} \)
5. \( W' = \frac{W}{S} \)
6. \( N_A' = N_A \times S \)

<table>
<thead>
<tr>
<th>Year</th>
<th>1985</th>
<th>1987</th>
<th>1989</th>
<th>1991</th>
<th>1993</th>
<th>1995</th>
<th>1997</th>
<th>1999</th>
<th>2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature</td>
<td>2.5</td>
<td>1.7</td>
<td>1.2</td>
<td>1.0</td>
<td>0.8</td>
<td>0.5</td>
<td>0.35</td>
<td>0.25</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Ajit Pal    IIT Kharagpur
## Constant Field Scaling

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Before Scaling</th>
<th>After Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel length</td>
<td>$L$</td>
<td>$L' = L/S$</td>
</tr>
<tr>
<td>Channel width</td>
<td>$W$</td>
<td>$W' = W/S$</td>
</tr>
<tr>
<td>Gate oxide thickness</td>
<td>$t_{ox}$</td>
<td>$t'<em>{ox} = t</em>{ox}/S$</td>
</tr>
<tr>
<td>Junction Depth</td>
<td>$x_j$</td>
<td>$x'_j = x_j/S$</td>
</tr>
<tr>
<td>Power supply voltage</td>
<td>$V_{DD}$</td>
<td>$V'<em>DD = V</em>{DD}/S$</td>
</tr>
<tr>
<td>Threshold Voltage</td>
<td>$V_{TO}$</td>
<td>$V'<em>TO = V</em>{TO}/S$</td>
</tr>
<tr>
<td>Doping Densities</td>
<td>$N_A$</td>
<td>$N'_A = N_A S$</td>
</tr>
<tr>
<td></td>
<td>$N_D$</td>
<td>$N'_D = N_D S$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quality</th>
<th>Before Scaling</th>
<th>After Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gate Capacitance</td>
<td>$C_g$</td>
<td>$C'_g = C_g/S$</td>
</tr>
<tr>
<td>Drain Current</td>
<td>$I_D$</td>
<td>$I'_D = I_D/S$</td>
</tr>
<tr>
<td>Power Dissipation</td>
<td>$P$</td>
<td>$P' = P/S^2$</td>
</tr>
<tr>
<td>Power Density</td>
<td>$P/Area$</td>
<td>$P'/Area' = (P/Area)$</td>
</tr>
<tr>
<td>Delay</td>
<td>$t_d$</td>
<td>$t'_d = t_d/S$</td>
</tr>
<tr>
<td>Energy</td>
<td>$E = P.t_d$</td>
<td>$E' = \frac{P}{S^2} \times \frac{t_d}{S} = \frac{P.t_d}{S^3} = \frac{1}{S^3} E$</td>
</tr>
</tbody>
</table>

Ajit Pal  IIT Kharagpur
### Constant Voltage Scaling

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Before Scaling</th>
<th>After Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel length</td>
<td>$L$</td>
<td>$L' = L / S$</td>
</tr>
<tr>
<td>Channel width</td>
<td>$W$</td>
<td>$W' = W / S$</td>
</tr>
<tr>
<td>Gate oxide thickness</td>
<td>$t_{ox}$</td>
<td>$t'<em>{ox} = t</em>{ox} / S$</td>
</tr>
<tr>
<td>Junction Depth</td>
<td>$X_j$</td>
<td>$x'_j = x_j / S$</td>
</tr>
<tr>
<td>Power supply voltage</td>
<td>$V_{DD}$</td>
<td>$V'<em>{DD} = V</em>{DD}$</td>
</tr>
<tr>
<td>Threshold Voltage</td>
<td>$V_{to}$</td>
<td>$V'<em>{to} = V</em>{to}$</td>
</tr>
<tr>
<td>Doping Densities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_A$</td>
<td>$N'_A = N_A \cdot S^2$</td>
<td></td>
</tr>
<tr>
<td>$N_D$</td>
<td>$N'_D = N_D \cdot S^2$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quality</th>
<th>Before Scaling</th>
<th>After Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gate Capacitance</td>
<td>$C_g$</td>
<td>$C'_g = C_g / S$</td>
</tr>
<tr>
<td>Drain Current</td>
<td>$I_D$</td>
<td>$I'_D = I_D \cdot S$</td>
</tr>
<tr>
<td>Power Dissipation</td>
<td>$P$</td>
<td>$P' = P \cdot S$</td>
</tr>
<tr>
<td>Power Density</td>
<td>$P / \text{Area}$</td>
<td>$P' / \text{Area}' = S^3 P / \text{Area}$</td>
</tr>
<tr>
<td>Delay</td>
<td>$t_d$</td>
<td>$t'_d = t_d / S^2$</td>
</tr>
</tbody>
</table>

Ajit Pal  IIT Kharagpur
Architecture-Level Approaches: Parallelism

- Parallel processing can be an important technique for reducing power consumption in CMOS circuits.
- Key approach is to trade area for power while maintaining the same throughput.
- In simple terms, if the supply voltage is reduced by half, the power is reduced by one-fourth and performance is lowered by half.
- The loss in performance can be compensated by parallel processing.

Ajit Pal   IIT Kharagpur
Architecture-Level Approaches: Parallelism

- Example: Two 16 bit registers supplies two operands to a Adder. Delay of the critical path of the adder is 10 nsec. Operating frequency = 100 MHz
- The estimated dynamic power of the circuit is

$$P_{\text{ref}} = C_{\text{ref}} \cdot V_{\text{ref}}^2 \cdot f_{\text{ref}}.$$
Architecture-Level Approaches: Parallelism

- Here the multiplier has been duplicated twice, but the input registers have been clocked at half the frequency of $f_{\text{ref}}$. This helps to reduce the supply voltage such that the critical path delay is not more than 20 nsec.

- The estimated dynamic power is

$$P_{\text{par}} = 2.2 C_{\text{ref}} \left( \frac{V_{\text{ref}}}{2} \right)^2 \times \frac{f_{\text{ref}}}{2}$$
In this realization, instead of 16-bit addition 8-bit addition is performed in each stage. The critical path delay through the 8-bit adder stage is about half that of 16-bit adder stage. Therefore, the 8-bit adder will operate at a clock frequency of 100 mHz with a reduced power supply voltage of $V_{\text{ref}}/2$.

Estimated power is:

$$P_{\text{pipe}} = C_{\text{pipe}} \cdot V_{\text{pipe}}^2 \cdot f_{\text{pipe}} = \left(1.15C_{\text{ref}}\right)\left(\frac{V_{\text{ref}}}{2}\right)^2 \cdot f = 0.28P_{\text{ref}}.$$
Architecture-Level Approaches: Pipelining and Parallelism

- Here, more than one parallel structure is used and each structure is pipelined. Both power supply and frequency of operation are reduced to achieve substantial overall reduction in power dissipation.

Estimated power

\[ P_{pipe} = \left(2.5C_{ref}\right) \left(0.4V_{ref}\right)^2 \left(\frac{f_{ref}}{2}\right) \]

Ajit Pal    IIT Kharagpur
Dynamic Voltage Scaling (DVS)

<table>
<thead>
<tr>
<th>Frequency (f) MHz</th>
<th>Voltage $V_{dd}$</th>
<th>Relative power</th>
</tr>
</thead>
<tbody>
<tr>
<td>700</td>
<td>1.65</td>
<td>100</td>
</tr>
<tr>
<td>600</td>
<td>1.60</td>
<td>80.59</td>
</tr>
<tr>
<td>500</td>
<td>1.50</td>
<td>59.03</td>
</tr>
<tr>
<td>400</td>
<td>1.40</td>
<td>41.14</td>
</tr>
<tr>
<td>300</td>
<td>1.25</td>
<td>24.60</td>
</tr>
<tr>
<td>200</td>
<td>1.10</td>
<td>12.70</td>
</tr>
</tbody>
</table>

The DAG after unrolling and using distributivity and constant propagation

No voltage scaling
Ideal DVS

Work load ($r$)

Normalized Energy

Processor voltage

CPU clock frequency
Dynamic Voltage Scaling (DVS)

Basic scheme of DVS

DC/DC Converter to be used in DVS

Ajit Pal  IIT Kharagpur
Minimizing Switched Capacitance

- Hardware Software Tradeoff
- By choice of suitable data representation (Bus Encoding)
- Two’s complement Vs Sign Magnitude
- Architectural optimization
- Logic styles
Hardware Software Tradeoff

- Same functionality can be either realized by hardware or by software or by a combination of both.

  - **Hardware-based approach:**
    - Faster
    - Costlier
    - Consumes more power

  - **Software-based approach:**
    - Cheaper
    - Slower
    - Consumes lesser power
Conventional superscalar out-of-order CPUs use hardware to create and dispatch micro-ops that can be executed in parallel.

Ajit Pal    IIT Kharagpur
Ajit Pal    IIT Kharagpur

Superscaler Versus VLIW approach

A compiler generates long instructions having multiple operations meant for different functional units

Ajit Pal    IIT Kharagpur
Transmeta’s Crusoe Processor

- Long instruction word, called molecule, can be 64 or 128 bits long
- A molecule can contain up to 4 RISC like instructions, called atoms
- All atoms get executed in parallel
- Molecules are executed in order
The Code Morphing software mediates between x86 software and the Crusoe Processor.

Ajit Pal    IIT Kharagpur
It is fundamentally a dynamic translation system

- A program that compiles instructions for instruction set architecture into instructions for another ISA
- Here, x86 code is compiled into VLIW code

Code Morphing s/w insulates x86 programs from the h/w engine’s native instruction set

- The native instruction set can be changed arbitrarily without affecting any x86 software at all
- Only Code Morphing s/w needs to be ported
Comparison of Heat Dissipation

A Pentium III processor plays a DVD at 105°C

A Crusoe processor model TM5400 plays a DVD at 48°C

Ajit Pal  IIT Kharagpur
Bus Encoding

- Communicating data bits in an appropriately coded form can reduce the switching activity

- Goals of coding
  - Remove undesired correlation among information bits, or
  - Introduce controlled correlation

- Coding for reduced switching activity falls under the second category
  - Introducing sample to sample correlation such that total number of bit transitions is reduced
Two chips are connected using $m = 2^n$ wires, and a $n$-bit data word is encoded by placing \('1'\) on the $i^{th}$ wire, where $0 \leq i \leq 2^n - 1$ is the binary value corresponding to the bit pattern and \('0'\) on the remaining $m-1$ wires.

Guarantees precisely one $0 \rightarrow 1$ and one $1 \rightarrow 0$ bit transition when a different data word is sent.

Number of wires required increases exponentially with the word size of the data.
Gray Coding Vs Binary Coding

- A gray code sequence is a set of numbers in which adjacent numbers only have one bit difference.

<table>
<thead>
<tr>
<th>Decimal Value</th>
<th>Binary Code</th>
<th>Gray Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0000</td>
<td>0000</td>
</tr>
<tr>
<td>1</td>
<td>0001</td>
<td>0001</td>
</tr>
<tr>
<td>2</td>
<td>0010</td>
<td>0011</td>
</tr>
<tr>
<td>3</td>
<td>0011</td>
<td>0010</td>
</tr>
<tr>
<td>4</td>
<td>0100</td>
<td>0110</td>
</tr>
<tr>
<td>5</td>
<td>0101</td>
<td>0111</td>
</tr>
<tr>
<td>6</td>
<td>0110</td>
<td>0101</td>
</tr>
<tr>
<td>7</td>
<td>0111</td>
<td>0100</td>
</tr>
<tr>
<td>8</td>
<td>1000</td>
<td>1100</td>
</tr>
<tr>
<td>9</td>
<td>1001</td>
<td>1101</td>
</tr>
<tr>
<td>10</td>
<td>1010</td>
<td>1111</td>
</tr>
<tr>
<td>11</td>
<td>1011</td>
<td>1110</td>
</tr>
<tr>
<td>12</td>
<td>1100</td>
<td>1010</td>
</tr>
<tr>
<td>13</td>
<td>1101</td>
<td>1011</td>
</tr>
<tr>
<td>14</td>
<td>1110</td>
<td>1001</td>
</tr>
<tr>
<td>15</td>
<td>1111</td>
<td>1000</td>
</tr>
</tbody>
</table>
Gray Coding Vs Binary Coding

- It is useful when the data is sequential and highly correlated, like Instruction addresses.
- No of bit transitions is limited to 2 for sequential data.
- For random data, the no of transitions for binary and gray code were approximately equal.
Comparison of the temporal activity

<table>
<thead>
<tr>
<th>Benchmark Program</th>
<th>Instruction Address</th>
<th>Data Address</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Binary Coded</td>
<td>Gray Coded</td>
</tr>
<tr>
<td>Fastqueens</td>
<td>2.46</td>
<td>1.03</td>
</tr>
<tr>
<td>Qsort</td>
<td>2.64</td>
<td>1.33</td>
</tr>
<tr>
<td>Reducer</td>
<td>2.57</td>
<td>1.71</td>
</tr>
<tr>
<td>Circuit</td>
<td>2.33</td>
<td>1.47</td>
</tr>
<tr>
<td>Semigroup</td>
<td>2.68</td>
<td>1.99</td>
</tr>
<tr>
<td>Nand</td>
<td>2.42</td>
<td>1.57</td>
</tr>
<tr>
<td>Boyer</td>
<td>2.76</td>
<td>2.09</td>
</tr>
<tr>
<td>Browse</td>
<td>2.51</td>
<td>1.64</td>
</tr>
<tr>
<td>Chat</td>
<td>2.43</td>
<td>1.54</td>
</tr>
</tbody>
</table>

Bit transitions per instruction executed

Ajit Pal  IIT Kharagpur
Temporal Transition Activity Comparison for Instruction Addresses

<table>
<thead>
<tr>
<th>Application</th>
<th>Bit Transitions Per Instruction Executed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chat</td>
<td>1.54, 2.43</td>
</tr>
<tr>
<td>Browse</td>
<td>1.64, 2.51</td>
</tr>
<tr>
<td>Boyer</td>
<td>1.99, 2.76</td>
</tr>
<tr>
<td>Nand</td>
<td>1.57, 2.42</td>
</tr>
<tr>
<td>Semigroup</td>
<td>1.99, 2.68</td>
</tr>
<tr>
<td>Circuit</td>
<td>1.47, 2.33</td>
</tr>
<tr>
<td>Reducer</td>
<td>1.71, 2.57</td>
</tr>
<tr>
<td>Qsort</td>
<td>1.33, 2.64</td>
</tr>
<tr>
<td>Fastqueens</td>
<td>1.03, 2.46</td>
</tr>
</tbody>
</table>

Ajit Pal    IIT Kharagpur
Temporal Transition Activity Comparison for Data Addresses

- Chat: 1.2, 1.32
- Browse: 1.32, 1.4
- Boyer: 1.25, 1.76, 1.72
- Nand: 1.16, 1.16, 1.25
- Semigroup: 1.2, 1.38, 1.34
- Circuit: 1.18, 1.33
- Reducer: 1.18, 1.47
- Qsort: 1.25, 1.32
- Fastqueens: 0.85, 0.91

Bit Transitions per Instruction Executed

Ajit Pal    IIT Kharagpur
Bus Inversion Coding

- It is a redundant coding scheme where $m = n + 1$
- If the $i^{th}$ data word is $S_i$, then either $S_i$ or $\sim S_i$ is transmitted depending on which would result in fewer no of bit transitions
- An extra bit $P$ encodes the polarity of the data word
- The coding technique works better for smaller values of $n$
  - For $n=2$, switching activity reduction is 25%
  - For $n=32$, switching activity reduction is 11%
Bus Inversion Coding

• For larger value of $n$, the $n$-bit data bus can be divided into smaller groups and each group is coded independently by associating polarity bit with each of the group.

• When a 32-bit bus is divided into 4 groups of 8-bits, it gives 18.3% reduction in switching activity as opposed to 11%.
Bus Inversion Coding

Predicted Reduction in Switching Activity

Ajit Pal  IIT Kharagpur
The Gray coding provides an asymptotic best performance of a single transitions for each address generated when infinite streams of consecutive addresses are considered.

However, the code is optimum only in the class of irredundant codes, i.e. codes that employ exactly n-bit patterns to encode a maximum of $2^n$ words.

By adding some redundancy to the code, better performance can be achieved by adapting the T0 encoding scheme, which requires a redundant line INC.

The T0 code provides, zero transition property for infinite streams of consecutive addresses.

**Encoding**

\[
(B(t), INC(t)) = \begin{cases} 
B(t-1), & \text{if } t > 0, \\
b(t), & \text{otherwise}
\end{cases}
\]

**Decoding**

\[
b(t) = \begin{cases} 
 b(t-1) + S & \text{if } INC = 1 \text{ and } t > 0 \\
 B(t) & \text{if } INC = 0
\end{cases}
\]
Reducing Glitching Activity

- Static designs can exhibit spurious transitions due to finite propagation delays from one logic block to the next.
- “Extra” transitions can be minimized by
  - Balancing all signal paths
  - Reducing logic depth

Figure: Reducing the glitching activity

Ajit Pal  IIT Kharagpur
Logic Styles for High Performance and Low Power

- Potential Logic Styles
  - Static CMOS Logic
  - Dynamic CMOS Logic
  - Pass-Transistor Logic (PTL)
- Experimental Results
- Conclusions
Static CMOS Logic

- Advantages
  - Ease of fabrication
  - Good noise margin
  - Robust
  - Lower switching activity
  - Good input/output decoupling
  - No charge sharing problem
  - Availability of matured logic synthesis tools and techniques

- Disadvantages
  - Larger number of transistors (larger chip area and delay)
  - Spurious transitions (glitch) due to finite propagation delays resulting in extra power dissipation and incorrect operation
  - Short circuit power dissipation
  - Weak output driving capability
  - Large number of standard cells requiring substantial engineering effort for technology mapping
Dynamic CMOS Logic

- **Advantages**
  - Combines the advantages of low power of static CMOS and low chip area of pseudo-nMOS
  - Reduced number of transistors compared to static CMOS (n+2 versus. 2n)
  - Faster than static CMOS logic
  - No short circuit power dissipation
  - No spurious transition and glitching power dissipation

- **Disadvantages**
  - Higher switching activity
  - Not as robust as static CMOS logic
  - Clock skew problem in cascaded realization
  - Suffers from charge sharing problem
  - Matured synthesis tools are not available
Pass-Transistor Logic

Advantages
- Lower area due to smaller number of transistors and smaller input loads
- Ratio-less PTL allows minimum dimension transistors and hence makes area efficient circuit realization
- No short circuit current leading to lower power dissipation

Disadvantages
- Increased delay due to long chain of pass-transistors
- Multi-threshold voltage drop
- Dual-rail logic to provide all signals in complementary form
- There is possibility of sneak path
Problems in PTL Synthesis

- Multi-threshold voltage drop
- Sneak path
- Long chain of pass transistors

\[ T = 0.69 \frac{n(n+1)}{2} R.C_L \]
Experimental Results

- Static CMOS circuits have been realized using Berkeley SIS tool (script.rugged to optimize the netlist and technology mapping with 44-2.genlib and option of minimum area)
- A large number of benchmark circuits are realized using the three logic styles with C/C++ programming in Sun system
- Requirements of area are approximated with the number of transistors
- Estimation models for calculating delay and switching power dissipation for the circuits with three different logic styles have been proposed and their accuracies are verified with Spice and Design Analyzer in Cadence
- MOSFET parameters are used from 0.18mm process technology

Ajit Pal IIT Kharagpur
## Experimental Results

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Area (#Transistor)</th>
<th>Delay (ns)</th>
<th>Power (mW)</th>
<th>Area</th>
<th>Delay (ns)</th>
<th>Power (mW)</th>
<th>Area</th>
<th>Delay (ns)</th>
<th>Power (mW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C432</td>
<td>692</td>
<td>3.32</td>
<td>122</td>
<td>581</td>
<td>1.82</td>
<td>88</td>
<td>546</td>
<td>2.08</td>
<td>91</td>
</tr>
<tr>
<td>C499</td>
<td>1880</td>
<td>2.23</td>
<td>367</td>
<td>1506</td>
<td>1.69</td>
<td>248</td>
<td>1428</td>
<td>1.62</td>
<td>167</td>
</tr>
<tr>
<td>C880</td>
<td>1412</td>
<td>2.21</td>
<td>293</td>
<td>1249</td>
<td>1.79</td>
<td>166</td>
<td>988</td>
<td>1.18</td>
<td>267</td>
</tr>
<tr>
<td>C1355</td>
<td>1880</td>
<td>2.61</td>
<td>400</td>
<td>1603</td>
<td>1.51</td>
<td>280</td>
<td>1203</td>
<td>1.04</td>
<td>379</td>
</tr>
<tr>
<td>C1908</td>
<td>1756</td>
<td>2.91</td>
<td>367</td>
<td>1689</td>
<td>1.80</td>
<td>251</td>
<td>1088</td>
<td>1.57</td>
<td>298</td>
</tr>
<tr>
<td>C2670</td>
<td>1804</td>
<td>2.94</td>
<td>493</td>
<td>1584</td>
<td>1.74</td>
<td>395</td>
<td>1010</td>
<td>1.54</td>
<td>449</td>
</tr>
<tr>
<td>C3540</td>
<td>4214</td>
<td>4.57</td>
<td>409</td>
<td>2815</td>
<td>2.63</td>
<td>314</td>
<td>2782</td>
<td>2.58</td>
<td>294</td>
</tr>
<tr>
<td>C5315</td>
<td>7058</td>
<td>3.65</td>
<td>830</td>
<td>5970</td>
<td>2.69</td>
<td>515</td>
<td>5364</td>
<td>1.62</td>
<td>778</td>
</tr>
<tr>
<td>C6288</td>
<td>11222</td>
<td>11.84</td>
<td>409</td>
<td>8716</td>
<td>4.64</td>
<td>504</td>
<td>6060</td>
<td>4.69</td>
<td>445</td>
</tr>
<tr>
<td>C7552</td>
<td>8214</td>
<td>2.99</td>
<td>1604</td>
<td>7328</td>
<td>1.98</td>
<td>1173</td>
<td>5682</td>
<td>1.66</td>
<td>1328</td>
</tr>
</tbody>
</table>

Average % reduction compared to static CMOS circuits

-16%  -37%  -25%  -33%  -47%  -17%

Ajit Pal  IIT Kharagpur
Comparison of area in terms of the # of Transistors

Ajit Pal    IIT Kharagpur
Comparison of Energy requirement

Switching energy (fJ)

- Static CMOS
- Dynamic CMOS
- PTL

Symbols:
- C432
- C499
- C880
- C1355
- C1908
- C2670
- C3540
- C5315
- C6288
- C7552

Ajit Pal    IIT Kharagpur
Leakage Power Limits Vt Scaling

Must stop at 50%
Threshold Voltage Scaling

- Scale down the threshold voltage
  - As $V_{th}$ is reduced, the subthreshold leakage current increases leading to increase in power dissipation
Threshold Voltage (V_T) Scaling

Scale down the threshold voltage for low voltage low power circuits to increase performance

\[ V_T \downarrow = \text{Delay} \downarrow + I_{\text{leakage}} \uparrow \]

Low -V_T : Provides high performance

\[ V_T \uparrow = \text{Delay} \uparrow + I_{\text{leakage}} \downarrow \]

High -V_T : Reduces subthreshold leakage

\[ 0.2V_{DD} \leq V_T \leq 0.5V_{DD} \]
Threshold Voltage Scaling

- Fabrication of multiple threshold voltages:
  - Multiple channel doping
  - Multiple Oxide thickness
  - Multiple channel length
  - Multiple body bias
- Various Approaches:
  - Variable-threshold-voltage CMOS (VTCMOS) approach
  - Multi-threshold-voltage CMOS (MTCMOS) approach
  - Dual-$V_t$ assignment approach
VTCMOS Approach

Typical n-well CMOS

\[ V_{Tp} = \begin{cases} -0.2 \text{ V} & \text{in active mode} \\ -0.6 \text{ V} & \text{in stand-by mode} \end{cases} \]

\[ V_{Tn} = \begin{cases} 0.2 \text{ V} & \text{in active mode} \\ 0.6 \text{ V} & \text{in stand-by mode} \end{cases} \]

\[ V_{Bp} = \begin{cases} 2 \text{ V} & \text{in active mode} \\ 4 \text{ V} & \text{in stand-by mode} \end{cases} \]

\[ V_{Bn} = \begin{cases} 0 \text{ V} & \text{in active mode} \\ -2 \text{ V} & \text{in stand-by mode} \end{cases} \]
Multi-threshold-voltage CMOS (MTCMOS)

- MTCMOS (Multi-threshold CMOS) (S. Mutoh et al. 1996)

Q1, Q2 = Sleep control transistors

MTCMOS Circuit Scheme
MTCMOS Performance

- Simulation results

![Graph showing simulation results for MTCMOS and Conv. CMOS with normalized delay and energy against supply voltage.](image-url)
Advantages and Limitations of MTCMOS

- MTCMOS can be easily implemented using existing circuits
- MTCMOS reduces only the standby power
- Large inserted MOSFETs will increase area and delay
- Extra $V_{th}$ memory circuit is needed to maintain the data in the standby mode
Dual-$V_{th}$ Assignment

Dual Threshold CMOS Technology

- Low-\(V_{th}\) transistors in critical path for high performance
- Some high-\(V_{th}\) transistors in non-critical paths to reduce leakage
Darker gates on the critical path

Ajit Pal  IIT Kharagpur
HighVt = 0.25 assigned to all gates in the off-critical path
HighVt = 0.396 assigned to some gates in the off-critical path

Ajit Pal  IIT Kharagpur
HighVt = 0.46 assigned to some gates in the off-critical path

Ajit Pal    IIT Kharagpur
Dual-$V_{th}$ Assignment Problem

• However, not all the transistors in non-critical paths can be assigned a high-$V_{th}$
• How to selectively assign dual $V_{th}$ to achieve the best leakage saving under performance constraint?

Ajit Pal   IIT Kharagpur
Represents the circuits as a DAG where each node represents a gate and each edge represents a connection.

1. Initialize all nodes with low-$V_{th}$.
2. Compute the critical path(s).
3. Using BFS traversal, assign high-$V_{th}$ to a node such that it does not alter the critical path.
4. Optimal high-$V_{th}$ calculation.
   Repeat the assignment with different high-$V_{th}$ ($0.2V_{dd}<\text{high}-V_{th}<0.5V_{dd}, V_{dd}=1V$) for which maximum number of node assignment and hence minimum leakage power is possible.

Ajit Pal   IIT Kharagpur
Optimal Dual-$V_{th}$ Assignment
another Approach

Delay-Constrained Dual-$V_T$
Static CMOS Circuits

Critical path

Assigned with high-$V_T$ transistors

Ajit Pal    IIT Kharagpur
Delay-Constrained Dual-$V_T$ Assignment

Assume the circuit as a DAG where each node represents a gate and each edge represents a connection.

**Algorithm**

1. Assume $\text{low-}V_T < \text{high-}V_T < 0.5V_{DD}$
2. Initialize all nodes with high-$V_T$
3. Compute the critical path(s)
4. Using DFS traversal, assign low-$V_T$ to a node on the critical path
5. Go to Step 3 until all the nodes on the critical path are assigned with low-$V_T$

Ajit Pal IIT Kharagpur
Delay-Constrained Dual-$V_T$ Assignment

Repeat the assignment with different high-$V_T$ ($0.2V_{DD}<$high-$V_T<$0.5$V_{DD}$) for which maximum number of nodes assignment and hence minimum leakage power is possible.
Experimental Results

Comparison of our results with [Wei+99]

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>#Transistor</th>
<th>%Redn in standby leakage power</th>
<th>%Redn in total power</th>
<th>CPU time (s)</th>
<th>#Transistor</th>
<th>%Redn in standby leakage power</th>
<th>%Redn in total power</th>
<th>CPU time (S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C432</td>
<td>278</td>
<td>59.65</td>
<td>16.52</td>
<td>20</td>
<td>348</td>
<td>87.35</td>
<td>26.17</td>
<td>36</td>
</tr>
<tr>
<td>C499</td>
<td>604</td>
<td>51.09</td>
<td>7.18</td>
<td>118</td>
<td>796</td>
<td>64.45</td>
<td>12.68</td>
<td>174</td>
</tr>
<tr>
<td>C880</td>
<td>1126</td>
<td>84.87</td>
<td>14.07</td>
<td>55</td>
<td>1208</td>
<td>88.65</td>
<td>19.41</td>
<td>89</td>
</tr>
<tr>
<td>C1355</td>
<td>1232</td>
<td>49.36</td>
<td>8.51</td>
<td>198</td>
<td>1346</td>
<td>59.95</td>
<td>13.15</td>
<td>346</td>
</tr>
<tr>
<td>C1908</td>
<td>1430</td>
<td>76.21</td>
<td>15.46</td>
<td>225</td>
<td>1684</td>
<td>83.45</td>
<td>21.75</td>
<td>412</td>
</tr>
<tr>
<td>C2670</td>
<td>2736</td>
<td>81.24</td>
<td>19.27</td>
<td>269</td>
<td>3092</td>
<td>92.96</td>
<td>24.80</td>
<td>485</td>
</tr>
<tr>
<td>C3540</td>
<td>3430</td>
<td>85.60</td>
<td>21.43</td>
<td>301</td>
<td>3698</td>
<td>90.42</td>
<td>32.29</td>
<td>541</td>
</tr>
<tr>
<td>C5315</td>
<td>5432</td>
<td>83.12</td>
<td>18.44</td>
<td>342</td>
<td>5516</td>
<td>89.69</td>
<td>31.05</td>
<td>619</td>
</tr>
<tr>
<td>C6288</td>
<td>5768</td>
<td>43.38</td>
<td>19.89</td>
<td>564</td>
<td>8950</td>
<td>83.69</td>
<td>45.42</td>
<td>890</td>
</tr>
<tr>
<td>C7552</td>
<td>7102</td>
<td>76.41</td>
<td>20.35</td>
<td>387</td>
<td>7786</td>
<td>87.65</td>
<td>22.36</td>
<td>609</td>
</tr>
</tbody>
</table>

**69.01%  16.11%**  **82.82%  24.92%**
ISCAS Benchmarks Results

Average 85.84% savings in leakage power compared to 60.91% of the earlier result.
Comparison

More reduction in leakage power
(Average 25% more reduction in leakage power)
Comparison

More number of transistor assignment

No. of transistors

<table>
<thead>
<tr>
<th>Old approach</th>
<th>New approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2000</td>
<td>8000</td>
</tr>
<tr>
<td>4000</td>
<td>6000</td>
</tr>
<tr>
<td>6000</td>
<td>4000</td>
</tr>
<tr>
<td>8000</td>
<td>2000</td>
</tr>
<tr>
<td>10000</td>
<td>0</td>
</tr>
</tbody>
</table>

Ajit Pal  IIT Kharagpur
Comparison

Higher time complexity

Graph showing comparison between old and new approaches in terms of CPU time in seconds.
List of publications


Thanks!

Ajit Pal    IIT Kharagpur