Why Power Matters

- Packaging costs
- Power supply rail design
- Chip and system cooling costs
- Noise immunity and system reliability
- Battery life (in portable systems)
- Environmental concerns
  - Office equipment accounted for 5% of total US commercial energy usage in 1993
  - Energy Star compliant systems
Why worry about power? -- Power Dissipation

Lead microprocessors power continues to increase

Power delivery and dissipation will be prohibitive

Source: Borkar, De Intel

Problem Illustration

What happens when the CPU cooler is removed?

www.tomshardware.de
www.tomshardware.com
Why worry about power? –
Battery Size/Weight

Expected battery lifetime increase over the next 5 years: 30 to 40%

From Rabaey, 1995

Why worry about power? -- Standby Power

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Power supply $V_{dd}$ (V)</td>
<td>1.5</td>
<td>1.2</td>
<td>0.9</td>
<td>0.7</td>
<td>0.6</td>
</tr>
<tr>
<td>Threshold $V_T$ (V)</td>
<td>0.4</td>
<td>0.4</td>
<td>0.35</td>
<td>0.3</td>
<td>0.25</td>
</tr>
</tbody>
</table>

- Drain leakage will increase as $V_T$ decreases to maintain noise margins and meet frequency demands, leading to excessive battery draining standby power consumption.

Source: Borkar, De Intel

...and phones leaky!

Source: Borkar, De Intel
Power and Energy Figures of Merit

- **Power** consumption in Watts
  - determines battery life in hours
- **Peak power**
  - determines power ground wiring designs
  - sets packaging limits
  - impacts signal noise margin and reliability analysis
- **Energy efficiency in Joules**
  - rate at which power is consumed over time
- **Energy = power * delay**
  - Joules = Watts * seconds
  - lower energy number means less power to perform a computation at the same frequency

Power versus Energy

- **Power** is height of curve
  - Lower power design could simply be slower
- **Energy** is area under curve
  - Two approaches require the same energy

Approach 1

Approach 2
PDP and EDP

- **Power-delay product** (PDP) = $P_{av} \cdot t_p = \frac{1}{2} C_L V_{DD}^2$
  - PDP is the average energy consumed per switching event (Watts * sec = Joule)
  - lower power design could simply be a slower design

- **Energy-delay product** (EDP) = $PDP \cdot t_p = P_{av} \cdot t_p^2$
  - EDP is the average energy consumed multiplied by the computation time required
  - takes into account that one can trade increased delay for lower energy/operation (e.g., via supply voltage scaling that increases delay, but decreases energy consumption)

- allows one to understand tradeoffs better

---

Understanding Tradeoffs

Which design is the “best” (fastest, coolest, both)?

```
   Energy
     b
    c    a
   d
```

```
1/Delay
   better
```

---

10/18/2005  VLSI Design I;  A. Milenkovic  |  10/18/2005  VLSI Design I;  A. Milenkovic
Understanding Tradeoffs

Which design is the “best” (fastest, coolest, both)?

![Diagram showing tradeoffs between energy and 1/Delay]

CMOS Energy & Power Equations

\[
E = C_L V_{DD}^2 P_{0\to 1} + t_{sc} V_{DD} I_{peak} P_{0\to 1} + V_{DD} I_{leakage} 
\]

\[
P = C_L V_{DD}^2 f_{0\to 1} + t_{sc} V_{DD} I_{peak} f_{0\to 1} + V_{DD} I_{leakage} 
\]

- Dynamic power
- Short-circuit power
- Leakage power
Dynamic Power Consumption

\[ \text{Energy/transition} = C_L \cdot V_{DD}^2 \cdot P_{0 \rightarrow 1} \]

\[ P_{\text{dyn}} = \text{Energy/transition} \cdot f = C_L \cdot V_{DD}^2 \cdot P_{0 \rightarrow 1} \cdot f \]

\[ P_{\text{dyn}} = C_{\text{EFF}} \cdot V_{DD}^2 \cdot f \quad \text{where} \quad C_{\text{EFF}} = \frac{P_{0 \rightarrow 1}}{C_L} \]

Not a function of transistor sizes!
Data dependent - a function of switching activity!

Pop Quiz

- Consider a 0.25 micron chip, 500 MHz clock, average load cap of 15fF/gate (fanout of 4), 2.5V supply.
  - Dynamic Power consumption per gate is ??

- With 1 million gates (assuming each transitions every clock)
  - Dynamic Power of entire chip = ??.

10/18/2005 VLSI Design I; A. Milenkovic
**Lowering Dynamic Power**

Capacitance: Function of fan-out, wire length, transistor sizes

Supply Voltage: Has been dropping with successive generations

\[ P_{\text{dyn}} = C_L V_{DD}^2 P_{0\rightarrow 1} f \]

Activity factor: How often, on average, do wires switch?

Clock frequency: Increasing...

---

**Short Circuit Power Consumption**

Finite slope of the input signal causes a direct current path between \( V_{DD} \) and GND for a short period of time during switching when both the NMOS and PMOS transistors are conducting.
Short Circuit Currents Determinates

\[ E_{sc} = t_{sc} V_{DD} I_{peak} P_{0\rightarrow1} \]
\[ P_{sc} = t_{sc} V_{DD} I_{peak} f_{0\rightarrow1} \]

- Duration and slope of the input signal, \( t_{sc} \)
- \( I_{peak} \) determined by
  - the saturation current of the P and N transistors which depend on their sizes, process technology, temperature, etc.
  - strong function of the ratio between input and output slopes
    - a function of \( C_L \)

Impact of \( C_L \) on \( P_{sc} \)

Large capacitive load

Output fall time significantly larger than input rise time.

Small capacitive load

Output fall time substantially smaller than the input rise time.
**I\text{\textsubscript{peak}} as a Function of C\textsubscript{L}**

When load capacitance is small, $I\text{\textsubscript{peak}}$ is large.

Short circuit dissipation is minimized by matching the rise/fall times of the input and output signals - slope engineering.

500 psec input slope

**P\textsubscript{sc} as a Function of Rise/Fall Times**

When load capacitance is small ($t\text{\textsubscript{sin}}/t\text{\textsubscript{sout}} > 2$ for $V\text{\textsubscript{DD}} > 2V$) the power is dominated by $P\text{\textsubscript{sc}}$

If $V\text{\textsubscript{DD}} < V\text{\textsubscript{Tn}} + |V\text{\textsubscript{Tp}}|$ then $P\text{\textsubscript{sc}}$ is eliminated since both devices are never on at the same time.
Leakage (Static) Power Consumption

Sub-threshold current is the dominant factor.

All increase exponentially with temperature!

Leakage as a Function of $V_T$

- Continued scaling of supply voltage and the subsequent scaling of threshold voltage will make subthreshold conduction a dominant component of power dissipation.
- An 90mV/decade $V_T$ roll-off - so each 255mV increase in $V_T$ gives 3 orders of magnitude reduction in leakage (but adversely affects performance)
TSMC Processes Leakage and $V_T$

<table>
<thead>
<tr>
<th></th>
<th>CL018 G</th>
<th>CL018 LP</th>
<th>CL018 ULP</th>
<th>CL018 HS</th>
<th>CL015 HS</th>
<th>CL013 HS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{dd}$</td>
<td>1.8 V</td>
<td>1.8 V</td>
<td>1.8 V</td>
<td>2 V</td>
<td>1.5 V</td>
<td>1.2 V</td>
</tr>
<tr>
<td>$T_{ox}$ (effective)</td>
<td>42 Å</td>
<td>42 Å</td>
<td>42 Å</td>
<td>42 Å</td>
<td>29 Å</td>
<td>24 Å</td>
</tr>
<tr>
<td>$L_{gate}$</td>
<td>0.16 µm</td>
<td>0.16 µm</td>
<td>0.18 µm</td>
<td>0.13 µm</td>
<td>0.11 µm</td>
<td>0.08 µm</td>
</tr>
<tr>
<td>$I_{Dsat(n/p)}$</td>
<td>600/260</td>
<td>500/180</td>
<td>320/130</td>
<td>780/360</td>
<td>860/370</td>
<td>920/400</td>
</tr>
<tr>
<td>$I_{off (leakage)}$</td>
<td>20</td>
<td>1.60</td>
<td>0.15</td>
<td>300</td>
<td>1,800</td>
<td>13,000</td>
</tr>
<tr>
<td>$V_{Th}$</td>
<td>0.42 V</td>
<td>0.63 V</td>
<td>0.73 V</td>
<td>0.40 V</td>
<td>0.29 V</td>
<td>0.25 V</td>
</tr>
<tr>
<td>FET Perf. (GHz)</td>
<td>30</td>
<td>22</td>
<td>14</td>
<td>43</td>
<td>52</td>
<td>80</td>
</tr>
</tbody>
</table>

*From MPR, 2000*

Exponential Increase in Leakage Currents

*From De, 1999*
Review: Energy & Power Equations

\[ E = C_L V_{DD}^2 I_{0\to1}^{\text{leakage}} + t_{sc} V_{DD} I_{\text{peak}} P_{0\to1} + V_{DD} \]

\[ f_{0\to1} = P_{0\to1} \cdot f_{\text{clock}} \]

\[ P = C_L V_{DD}^2 f_{0\to1}^{1} + t_{sc} V_{DD} I_{\text{peak}} f_{0\to1}^{1} + V_{DD} I_{\text{leakage}} \]

- **Dynamic power** (~90% today and decreasing relatively)
- **Short-circuit power** (~8% today and decreasing absolutely)
- **Leakage power** (~2% today and increasing)

Power and Energy Design Space

<table>
<thead>
<tr>
<th>Constant Throughput/Latency</th>
<th>Variable Throughput/Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Energy</strong></td>
<td><strong>Design Time</strong></td>
</tr>
<tr>
<td><strong>Active</strong></td>
<td>Logic Design</td>
</tr>
<tr>
<td>Reduced ( V_{dd} )</td>
<td>Reduced ( V_{dd} )</td>
</tr>
<tr>
<td>Sizing</td>
<td>Multi-( V_{dd} )</td>
</tr>
<tr>
<td><strong>Leakage</strong></td>
<td>+ Multi-( V_T )</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

10/18/2005 VLSI Design I; A. Milenkovic
Dynamic Power as a Function of Device Size

- Device sizing affects dynamic energy consumption
  - gain is largest for networks with large overall effective fan-outs ($F = C_L/C_{g,1}$)

- The optimal gate sizing factor ($f$) for dynamic energy is smaller than the one for performance, especially for large $F$'s
  - e.g., for $F=20$, $f_{opt}(\text{energy}) = 3.53$ while $f_{opt}(\text{performance}) = 4.47$

- If energy is a concern avoid oversizing beyond the optimal

Dynamic Power Consumption is Data Dependent

- Switching activity, $P_{0\to1}$, has two components
  - A static component – function of the logic topology
  - A dynamic component – function of the timing behavior (glitching)

Static transition probability

\[ P_{0\to1} = P_{\text{out}=0} \times P_{\text{out}=1} = P_0 \times (1-P_0) \]

With input signal probabilities

\[ P_{A=1} = 1/2 \]
\[ P_{B=1} = 1/2 \]

NOR static transition probability

\[ = 3/4 \times 1/4 = 3/16 \]

2-input NOR Gate

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
NOR Gate Transition Probabilities

- Switching activity is a strong function of the input signal statistics
  - \( P_A \) and \( P_B \) are the probabilities that inputs A and B are one

\[
P_{0\rightarrow1} = P_0 \times P_1 = (1-(1-P_A)(1-P_B)) \times (1-P_A)(1-P_B)
\]

Transition Probabilities for Some Basic Gates

<table>
<thead>
<tr>
<th>Gate</th>
<th>( P_{0\rightarrow1} = P_{out=0} \times P_{out=1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOR</td>
<td>((1 - (1 - P_A)(1 - P_B)) \times (1 - P_A)(1 - P_B))</td>
</tr>
<tr>
<td>OR</td>
<td>((1 - P_A)(1 - P_B) \times (1 - (1 - P_A)(1 - P_B)))</td>
</tr>
<tr>
<td>NAND</td>
<td>(P_A P_B \times (1 - P_A P_B))</td>
</tr>
<tr>
<td>AND</td>
<td>((1 - P_A P_B) \times P_A P_B)</td>
</tr>
<tr>
<td>XOR</td>
<td>((1 - (P_A + P_B - 2P_A P_B)) \times (P_A + P_B - 2P_A P_B))</td>
</tr>
</tbody>
</table>

For X: \( P_{0\rightarrow1} = \)

For Z: \( P_{0\rightarrow1} = \)
Transition Probabilities for Some Basic Gates

<table>
<thead>
<tr>
<th>Gate</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOR</td>
<td>((1 - (1 - P_A)(1 - P_B)) \times (1 - P_A)(1 - P_B))</td>
</tr>
<tr>
<td>OR</td>
<td>((1 - P_A)(1 - P_B) \times (1 - (1 - P_A)(1 - P_B)))</td>
</tr>
<tr>
<td>NAND</td>
<td>(P_A P_B \times (1 - P_A P_B))</td>
</tr>
<tr>
<td>AND</td>
<td>((1 - P_A P_B) \times P_A P_B)</td>
</tr>
<tr>
<td>XOR</td>
<td>((1 - (P_A + P_B - 2P_A P_B)) \times (P_A + P_B - 2P_A P_B))</td>
</tr>
</tbody>
</table>

For X: \(P_{0 \rightarrow 1} = P_0 \times P_1 = (1 - P_A P_B) P_A P_B\)
\[= 0.5 \times 0.5 = 0.25\]

For Z: \(P_{0 \rightarrow 1} = P_0 \times P_1 = (1 - P_A P_B) P_A P_B\)
\[= (1 - (0.5 \times 0.5)) \times (0.5 \times 0.5) = 3/16\]

Inter-signal Correlations

- Determining switching activity is complicated by the fact that signals exhibit correlation in space and time
  - reconvergent fan-out

  \[P(Z=1) = P(B=1) \& P(X=1 | B=1)\]

- Have to use conditional probabilities
Inter-signal Correlations

- Determining switching activity is complicated by the fact that signals exhibit correlation in space and time
  - reconvergent fan-out
    \[(1-0.5)(1-0.5)x(1-(1-0.5)(1-0.5)) = 3/16\]

- Have to use conditional probabilities

Logic Restructuring

Logic restructuring: changing the topology of a logic network to reduce transitions

- Chain implementation has a lower overall switching activity than the tree implementation for random inputs
  - Ignores glitching effects
Beneficial to postpone the introduction of signals with a high transition rate (signals with signal probability close to 0.5)

(1-0.5x0.2)x(0.5x0.2)=0.09
(1-0.2x0.1)x(0.2x0.1)=0.0196

Beneficial to postpone the introduction of signals with a high transition rate (signals with signal probability close to 0.5)
Glitching in Static CMOS Networks

- Gates have a nonzero propagation delay resulting in spurious transitions or glitches (dynamic hazards)
  - glitch: node exhibits multiple transitions in a single cycle before settling to the correct logic value

![Diagram of logic gates with ABC, X, Z, A, B, C, and unit delay]

<table>
<thead>
<tr>
<th>ABC</th>
<th>101</th>
<th>000</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unit Delay
Glitching in an RCA

Balanced Delay Paths to Reduce Glitching

Glitching is due to a mismatch in the path lengths in the logic network; if all input signals of a gate change simultaneously, no glitching occurs.

So equalize the lengths of timing paths through logic.
### Power and Energy Design Space

<table>
<thead>
<tr>
<th></th>
<th>Constant Throughput/Latency</th>
<th>Variable Throughput/Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Energy</strong></td>
<td>Design Time</td>
<td>Non-active Modules</td>
</tr>
<tr>
<td><strong>Active</strong></td>
<td>Logic Design</td>
<td>Clock Gating</td>
</tr>
<tr>
<td></td>
<td>Reduced $V_{dd}$ Sizing</td>
<td>Multi-$V_{dd}$</td>
</tr>
<tr>
<td><strong>Leakage</strong></td>
<td>+ Multi-$V_T$</td>
<td>Sleep Transistors</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multi-$V_{dd}$ Variable $V_T$</td>
</tr>
</tbody>
</table>

### Dynamic Power as a Function of $V_{DD}$

- Decreasing the $V_{DD}$ decreases dynamic energy consumption (quadratically)
- But, increases gate delay (decreases performance)

- Determine the critical path(s) at design time and use high $V_{DD}$ for the transistors on those paths for speed. Use a lower $V_{DD}$ on the other gates, especially those that drive large capacitances (as this yields the largest energy benefits).
**Multiple V\textsubscript{DD} Considerations**

- How many V\textsubscript{DD}? — Two is becoming common
  - Many chips already have two supplies (one for core and one for I/O)
- When combining multiple supplies, level converters are required whenever a module at the lower supply drives a gate at the higher supply (step-up)
  - If a gate supplied with V\textsubscript{DDL} drives a gate at V\textsubscript{DDH}, the PMOS never turns off
    - The cross-coupled PMOS transistors do the level conversion
    - The NMOS transistor operate on a reduced supply
  - Level converters are not needed for a step-down change in voltage
  - Overhead of level converters can be mitigated by doing conversions at register boundaries and embedding the level conversion inside the flipflop

**Dual-Supply Inside a Logic Block**

- Minimum energy consumption is achieved if all logic paths are critical (have the same delay)
- Clustered voltage-scaling
  - Each path starts with V\textsubscript{DDH} and switches to V\textsubscript{DDL} (gray logic gates) when delay slack is available
  - Level conversion is done in the flipflops at the end of the paths
### Power and Energy Design Space

<table>
<thead>
<tr>
<th></th>
<th>Constant Throughput/Latency</th>
<th>Variable Throughput/Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Energy</strong></td>
<td>Design Time</td>
<td>Non-active Modules</td>
</tr>
<tr>
<td><strong>Active</strong></td>
<td>Logic Design</td>
<td>Clock Gating</td>
</tr>
<tr>
<td></td>
<td>Reduced $V_{dd}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sizing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multi-$V_{dd}$</td>
<td></td>
</tr>
<tr>
<td><strong>Leakage</strong></td>
<td>+ Multi-$V_T$</td>
<td>Sleep Transistors</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multi-$V_{dd}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Variable $V_T$</td>
</tr>
</tbody>
</table>

### Stack Effect

- Leakage is a function of the circuit topology and the value of the inputs
  
  $$V_T = V_{T0} + \gamma (\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|-2\phi_F|})$$

  where $V_{T0}$ is the threshold voltage at $V_{SB} = 0$; $V_{SB}$ is the source-bulk (substrate) voltage; $\gamma$ is the body-effect coefficient.

- Leakage is least when $A = B = 0$
- Leakage reduction due to stacked transistors is called the **stack effect**

---

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>$V_X$</th>
<th>$I_{SUB}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>$V_T \ln(1+n)$</td>
<td>$V_{GS}=V_{BS}=-V_X$</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>$V_{GS}=V_{BS}=0$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>$V_{DD}-V_T$</td>
<td>$V_{GS}=V_{BS}=0$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>$V_{SG}=V_{SB}=0$</td>
</tr>
</tbody>
</table>
Short Channel Factors and Stack Effect

- In short-channel devices, the subthreshold leakage current depends on $V_{GS}$, $V_{BS}$ and $V_{DS}$. The $V_T$ of a short-channel device decreases with increasing $V_{DS}$ due to DIBL (drain-induced barrier loading).
  - Typical values for DIBL are 20 to 150mV change in $V_T$ per voltage change in $V_{DS}$ so the stack effect is even more significant for short-channel devices.
  - $V_X$ reduces the drain-source voltage of the top nFET, increasing its $V_T$ and lowering its leakage.

- For our 0.25 micron technology, $V_X$ settles to ~100mV in steady state so $V_{BS} = -100$mV and $V_{DS} = V_{DD} - 100$mV which is 20 times smaller than the leakage of a device with $V_{BS} = 0$mV and $V_{DS} = V_{DD}$.

Leakage as a Function of Design Time $V_T$

- Reducing the $V_T$ increases the sub-threshold leakage current (exponentially)
  - 90mV reduction in $V_T$ increases leakage by an order of magnitude.
- But, reducing $V_T$ decreases gate delay (increases performance).
- Determine the critical path(s) at design time and use low $V_T$ devices on the transistors on those paths for speed. Use a high $V_T$ on the other logic for leakage control.
  - A careful assignment of $V_T$'s can reduce the leakage by as much as 80%.
Dual-Thresholds Inside a Logic Block

- Minimum energy consumption is achieved if all logic paths are critical (have the same delay)
- Use lower threshold on timing-critical paths
  - Assignment can be done on a per gate or transistor basis; no clustering of the logic is needed
  - No level converters are needed

Variable $V_T$ (ABB) at Run Time

- $V_T = V_{T0} + \gamma(\sqrt{|-2\phi_F| + V_{SB}} - \sqrt{|-2\phi_F|})$
  - For an n-channel device, the substrate is normally tied to ground ($V_{SB} = 0$)
  - A negative bias on $V_{SB}$ causes $V_T$ to increase
  - Adjusting the substrate bias at run time is called adaptive body-biasing (ABB)
    - Requires a dual well fab process