Review: 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

Review: CMOS Energy & Power Equations

\[ E = C_L V_{DD}^2 P_{0-1} + \tau_{LC} V_{DD} I_{peak} P_{0-1} + V_{DD} I_{leak} \]

\[ P = C_L V_{DD}^2 f_{0-1} + \tau_{SC} V_{DD} I_{peak} f_{0-1} + V_{DD} I_{leak} \]

Dynamic power
Short-circuit power
Leakage power

Dynamic Power Consumption

\[ P_{dyn} = C_L V_{DD}^2 f_{0-1} \]

Energy/transition = \( C_L \cdot V_{DD}^2 \cdot f_{0-1} \)

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

Lowering Dynamic Power

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

Supply Voltage:
Has been dropping with successive generations

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

\[ E_{sc} = t_{sc} V_{DD} I_{peak} \quad P_{sc} = t_{sc} V_{DD} I_{peak} f_{0 \rightarrow 1} \]

- 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 \( C_L \)

\[ E_{sc} = t_{sc} V_{DD} I_{peak} P_{0 \rightarrow 1} \]
\[ P_{sc} = t_{sc} V_{DD} I_{peak} f_{0 \rightarrow 1} \]

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

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

\[ \text{Large capacitive load} \]
\[ \text{Small capacitive load} \]

Output fall time significantly larger than input rise time.

\[ \text{Output fall time substantially smaller than the input rise time.} \]

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

\[ \text{vin} \]
\[ \text{vout} \]
\[ \text{CL} \]
\[ \text{Isc} \approx 0 \]
\[ \text{Isc} \approx I_{\text{max}} \]

Leakage (Static) Power Consumption

- Sub-threshold current is the dominant factor.
- All increase exponentially with temperature!

\[ \text{Leakage (Static) Power Consumption} \]

\[ V_{DD} I_{\text{leakage}} \]

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)

\[ \text{Leakage as a Function of } V_T \]

\[ \text{VDD} \]

\[ \text{leakage} \]

\[ \text{Drain junction leakage} \]

\[ \text{Sub-threshold current} \]

\[ \text{Gate leakage} \]
TSMC Processes Leakage and $V_T$

### Table

<table>
<thead>
<tr>
<th>Process</th>
<th>$V_T$ (effective)</th>
<th>$t_{ns}$ (ns)</th>
<th>$I_{leak}(nA/\mu m)$</th>
<th>$I_{DSat}(\mu A/\mu m)$</th>
<th>$L_{gate}$</th>
<th>$V_{DD}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL018G</td>
<td>1.8 V</td>
<td>0.16 µm</td>
<td>60/260</td>
<td>750/130</td>
<td>0.13 µm</td>
<td>1.2 V</td>
</tr>
<tr>
<td>CL018LP</td>
<td>1.8 V</td>
<td>0.16 µm</td>
<td>500/180</td>
<td>780/200</td>
<td>0.13 µm</td>
<td>1.2 V</td>
</tr>
<tr>
<td>CL018ULP</td>
<td>1.8 V</td>
<td>0.16 µm</td>
<td>320/130</td>
<td>900/380</td>
<td>0.13 µm</td>
<td>1.2 V</td>
</tr>
<tr>
<td>CL015HS</td>
<td>2.0 V</td>
<td>0.11 µm</td>
<td>860/400</td>
<td>920/400</td>
<td>0.11 µm</td>
<td>1.5 V</td>
</tr>
<tr>
<td>CL013HS</td>
<td>2.0 V</td>
<td>0.11 µm</td>
<td>13,000</td>
<td>24,000</td>
<td>0.11 µm</td>
<td>1.5 V</td>
</tr>
</tbody>
</table>

For $V_T$: 0.42 V, $t_{ns}$: 0.63 V, $I_{leak}$: 0.73 V, $I_{DSat}$: 0.40 V

### Equations

#### Dynamic Power Consumption is Data Dependent

- Switching activity, $P_{0\rightarrow1}$, has two components
  - A static component – function of the logic topology
  - A dynamic component – function of the timing behavior (glitching)

#### NOR Gate

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>$P_{out}$</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>

### Exponential Increase in Leakage Currents

- From MPR, 2000

### Review: Energy & Power Equations

- $E = C_L V_{DD}^2 P_{0\rightarrow1} + t_{ns} V_{DD} I_{peak} P_{0\rightarrow1} + V_{DD}$
- $P = C_L V_{DD}^2 f_{0\rightarrow1} + t_{ns} V_{DD} I_{peak} f_{0\rightarrow1} + V_{DD} I_{leakage}$

### Power and Energy Design Space

- Constant Throughput/Latency
  - Energy Design Time
  - Non-active Modules
  - Variable Throughput/Latency
  - Run Time

- Active
  - Logic Design
  - Reduced $V_{dd}$
  - Clock Gating

- Leakage
  - Multi-V_{T}
  - Sleep Transistors

- Dynamic
  - Multi-V_{T}
  - Variable $V_{dd}$

### 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 ($F = C_L/C_{g,eff}$)
- The optimal gate sizing factor ($f$) for dynamic energy is smaller than the one for performance, especially for large $F$s
  - e.g., for $F=20$, $f_{opt}(energy) = 3.53$ while $f_{opt}(performance) = 4.47$
- If energy is a concern avoid oversizing beyond the optimal

- From Nikolic, UCB, 2005
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

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

Transition Probabilities for Some Basic Gates

<table>
<thead>
<tr>
<th>Gate</th>
<th>Transition Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOR</td>
<td>[ (1 - (1 - P_A)(1 - P_B)) \times (1 - (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 \times P_B \times (1 - (1 - P_A)(1 - P_B)) ]</td>
</tr>
<tr>
<td>AND</td>
<td>[ (1 - P_A)(1 - P_B) \times P_A \times 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>

Inter-signal Correlations

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

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

For X: \( P_{0 \rightarrow 1} = 0.5 \times 0.5 = 0.25 \)

For Z: \( P_{0 \rightarrow 1} = 0.5 \times 0.5 = 0.25 \)

Logic Restructuring

- 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)

Beneficial to postpone the introduction of signals with a high transition rate (signals with signal probability close to 0.5)

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

Balanced Delay Paths to Reduce Glitching

Gating 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>Energy</td>
<td>Design Time</td>
<td>Non-active Modules</td>
</tr>
<tr>
<td>Active</td>
<td>Logic Design</td>
<td>Reduced V_{dd}</td>
</tr>
<tr>
<td></td>
<td>Sizing</td>
<td>Multi-V_{dd}</td>
</tr>
<tr>
<td>Leakage</td>
<td>+ Multi-V_{t}</td>
<td>Sleep Transistors</td>
</tr>
<tr>
<td></td>
<td></td>
<td></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_{dd} Considerations

- How many V_{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_{ddL} drives a gate at V_{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_{ddH} and switches to V_{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>Energy</td>
<td>Design Time</td>
<td>Non-active Modules</td>
</tr>
<tr>
<td>Active</td>
<td>Logic Design</td>
<td>Reduced V_{dd}</td>
</tr>
<tr>
<td></td>
<td>Sizing</td>
<td>Multi-V_{dd}</td>
</tr>
<tr>
<td>Leakage</td>
<td>+ Multi-V_{t}</td>
<td>Sleep Transistors</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Stack Effect

- Leakage is a function of the circuit topology and the value of the inputs
  \[ V_I = V_{TH} + \gamma (\sqrt{|-2\phi F + V_{SB}|} - \sqrt{|-2\phi F|}) \]
  where V_{TH} 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
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 150 mV per voltage change in $V_{DS}$ so the stack effect is even more significant for short-channel devices.
  - $V_T$ 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 ~100 mV 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 subthreshold leakage current (exponentially)
  - 90mV reduction in $V_T$ increases leakage by an order of magnitude
- But, reducing $V_T$ decreases gate delay (increases performance)

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