Why Worry About Power?

- Mobile and Embedded systems:
  - Battery life
- High-end processors:
  - Cooling (costs $1 per chip per Watt if operating @ >40W)
  - Packaging
  - Reliability

Power Consumption in CMOS

- Dynamic Power Consumption
  - Charging and discharging capacitors

Sources of power consumption in CMOS:

- Dynamic or active power (due to the switching of transistors)
- Short-circuit power
- Leakage power

Peak Power in Today’s CPUs*

- Intel 486 5W
- AMD Athlon XP 67W
- HP PA-8700 75W
- Alpha 21264 95W
- IBM Power 4 135W
- Intel Itanium 130W

*source: Microprocessor Report, August 2002
Power Consumption in CMOS

- Dynamic Power Consumption
  - Charging and discharging capacitors

\[ P = \frac{E \times f}{C} = C \times V^2 \times f \]
**Dynamic Power Consumption**

- **Capacitance**: function of wire length, transistor size
- **Clock frequency**: increasing
- **Activity factor**: how often do wires switch
- **Supply voltage**: has been dropping with successive process generations

\[ \text{Power} = \alpha \times C \times V^2 \times f \]

**Power Consumption in CMOS**

- **Dynamic Power**
  - Short-circuit power
    - Both PMOS and NMOS are conducting
  - About 2% of the overall power.

- **Leakage power** – transistors are not perfect switches and they leak

- **Voltage Scaling**
  - Transistor switches slower at lower voltage. Need to also scale threshold voltage to maintain performance
  - Leakage current grows exponentially with decreases in threshold voltage
  - Leakage power goes to the roof

- **Leakage power**

  - 7% now, expect 20% in next technology, 50% in next one
Technology Scaling: Ideal Shrink

- New process generation every 2-3 years
- $V_{dd}$ scales down by 30%
- Gate delays are shortened by 30%
  - $\sim 50\%$ frequency gain
  - ($500\text{ps cycle} = 2\text{GHz clock}, 333\text{ps cycle} = 3\text{GHz clock}$)
- Transistor density increases by 2X
  - 0.7X shrink on a side, 2X area reduction
- Cap/transistor reduced by 30% ($W/L_{tx}=0.7x0.7/0.7=0.7X$)

*Source: “Power – the Next Frontier: a Microarchitecture Perspective”, Ronny Ronen, Keynote speech at PACS’02 Workshop.

- 66% reduction in energy/transition!
  - $(CV^2 \rightarrow 0.7x0.7^2 = 0.34X)$
- 50% reduction in power!
  - $(CV^2f \rightarrow 0.7x0.7^2 x 1.5 = 0.5X)$
- No change in power density!
  - $(CV^2f / 0.7x0.7 \rightarrow 0.7x0.7^2 x 1.5/0.7x0.7 = 0.5 / 0.5 = 1X)$

Process Technology – the Reality*

- Performance does not scale w/ frequency
  - New designs increase frequency by 2X
  - New designs use 2X-3X more transistors to get 1.4X-1.8X performance*
- So, every new process and March generation:
  - Power goes up by about 2X (3X transistors * 2X switches * 1/3 energy)
  - Power density goes up 30%-80% (2X power / 1.X area)
  - Leakage power goes to the roof

*Source: “Power – the Next Frontier: a Microarchitecture Perspective”, Ronny Ronen, Keynote speech at PACS’02 Workshop.

Mind-Boggling Numbers*

<table>
<thead>
<tr>
<th></th>
<th>486 (0.8\mu)</th>
<th>Pentium 4 (0.18\mu)</th>
<th>Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transistors</td>
<td>1.2M</td>
<td>42M</td>
<td>35x</td>
</tr>
<tr>
<td>Frequency</td>
<td>50 MHz</td>
<td>2000 MHz</td>
<td>40x</td>
</tr>
<tr>
<td>Voltage</td>
<td>5 V</td>
<td>1.65 V</td>
<td>1/3x</td>
</tr>
<tr>
<td>Peak Power</td>
<td>5 W</td>
<td>100 W</td>
<td>20x</td>
</tr>
<tr>
<td>Die size</td>
<td>0.73 cm$^2$</td>
<td>2.17 cm$^2$</td>
<td>3x</td>
</tr>
<tr>
<td>Power density</td>
<td>6.8 W/cm$^2$</td>
<td>46 W/cm$^2$</td>
<td>7x</td>
</tr>
</tbody>
</table>

*Source: “Power – the Next Frontier: a Microarchitecture Perspective”, Ronny Ronen, Keynote speech at PACS’02 Workshop.

The Problem with High Power Dissipation

- Processor Lifetime Decreases
  - Heat puts mechanical stresses on the silicon die
  - interconnections on chip can break
  - transistors can malfunction
  - the leakage component of power goes up, increasing heat dissipation further
- Cooling and power supply costs go up, driving up overall system costs
- Prevents us using faster clocks for better performance

Areal Power Density !

- Hot Spots on chip

Another Performance limiter
Circuits and process scaling alone can no longer solve all power problems.

SYSTEMS must also be power-aware:
- OS
- Compilers
- Architecture

The challenge is to reduce power WITH MINIMAL performance loss.

This is in contrast to embedded systems, where performance loss can sometimes be tolerated.

Power = α*C*V^2*f

Dynamic Power Consumption:
- Capacitance: function of wire length, transistor size
- Activity factor: how often do wires switch
- Clock frequency: increasing
- Supply voltage: has been dropping with successive process generations