Apple Announces M1 Pro & M1 Max: Giant New SoCs with All-Out Performance

October 18, 2021, AnandTech - The M1 Pro and Max both follow-up on last year’s M1, Apple’s first generation silicon that ushered in the beginning of Apple’s journey to replace x86-based chips with their own in-house designs. The M1 had been widely successful for Apple, demonstrating fantastic performance at never-before-seen power efficiency in the laptop market. Although the M1 was fast, it was still a somewhat smaller SoC - still powering devices such as the iPad Pro line-up, and a corresponding lower TDP, naturally still losing out to larger more power-hungry ships from the competition.

Wire Resistance

- $p = \text{resistivity (\Omega \cdot \text{m})}$
- $R = \frac{l}{t} \cdot \frac{1}{W} = R \frac{1}{W}$
- $R_{qs} = \text{sheet resistance (\Omega/\square)}$
- $\square$ is a dimensionless unit!
- Count number of squares
  - $R = R_{qs} \cdot (# \text{ of squares})$

Wire Capacitance

- Wire has capacitance per unit length
  - To neighbors
  - To layers above and below
- $C_{\text{total}} = C_{\text{top}} + C_{\text{bot}} + 2C_{\text{adj}}$

Wire RC Model

- Wires are a distributed system
  - Approximate with lumped element models
- 3-segment pi-model is accurate to 3% in simulation
Elmore Delay for RC Tree

\[ t_{pd} \approx \sum_{i=1}^{\text{nodes}} R_i C_i \]
\[ = R_1 C_1 + (R_1 + R_2) C_2 + \ldots + (R_1 + R_2 + \ldots + R_N) C_N \]

Example: RC Delay with Wire and Gate

Logical Effort with Wires

Administrivia

• Homework 6 due this week
• Homework 7 next week
• All labs need to be checked off by next week!
• Projects (ASIC and FPGA) start this week
• Midterm 2 is on November 4 at 7pm
• Courses:
  • EECS251B will be offered in Spring (pending Campus approval)
  • EE194/290C SoC Design

Midterm 1 Scores

EECS151: Average: 69.7/84 (83%)
EECS251B: Average 86.4/88 (83%)

Energy

Power and Energy

• Power is drawn from a voltage source attached to the \( V_{DD} \) pin(s) of a chip.

  * Instantaneous Power: \( P(t) = I(t)V(t) \)

  * Energy: \( E = \int_0^T P(t)dt \)

  * Average Power: \( P_{\text{avg}} = \frac{E}{T} = \frac{1}{T} \int_0^T P(t)dt \)

Processor Frequency Scaling
Power in a Circuit Element

\[ P_{\text{DC}}(t) = I_{\text{DD}}(t)V_{\text{DD}} \]
\[ P_{\text{ac}}(t) = \frac{I_{\text{L}}^2(t)}{R} \]
\[ E_{\text{ac}} = \frac{1}{2} C V(t)^2 \frac{dV(t)}{dt} \]

Sources of Power Dissipation

\[ P_{\text{total}} = P_{\text{dynamic}} + P_{\text{static}} \]
- Dynamic power: \( P_{\text{dynamic}} = P_{\text{switching}} + P_{\text{shortcircuit}} \)
  - Switching load capacitances
  - Short-circuit current
- Static power: \( P_{\text{static}} = (I_{\text{sub}} + I_{\text{gate}} + I_{\text{junct}} + I_{\text{contention}})V_{\text{DD}} \)
  - Subthreshold leakage
  - Gate leakage
  - Junction leakage
  - Contention current

Charging and Discharging a Capacitor

- When the gate output rises
  - Energy stored in capacitor is \( E_{\text{ac}} = \frac{1}{2} C V(t)^2 \frac{dV(t)}{dt} \)
  - But energy drawn from the supply is \( E_{\text{draw}} = \frac{1}{2} C V_{\text{DD}}^2 \frac{dV_{\text{DD}}}{dt} \)
  - Half the energy from \( V_{\text{DD}} \) is dissipated in the pMOS transistor as heat, other half stored in capacitor
- When the gate output transitions HL
  - Energy in capacitor is dumped to GND
  - Dissipated as heat in the NMOS transistor

Dynamic Power Reduction

How can we limit switching power?
- Try to minimize: Activity factor, Capacitance, Supply voltage, Frequency

Reduce Activity Factor

- Clock gating
  - The best way to reduce the activity is to turn off the clock to registers in unused blocks
  - Saves clock activity (a = 1)
  - Eliminates all switching activity in the block
  - Requires determining if block will be used

Reduce Capacitance

- Gate capacitance
  - Fewer stages of logic
  - Smaller gate sizes
- Wire capacitance
  - Good floorplanning to keep communicating blocks close to each other

Reduce Voltage/Frequency

- Voltage domains
  - Provide separate supplies to different blocks
- Dynamic voltage/frequency scaling
  - Adjust \( V_{\text{DD}} \) and \( f \) according to workload
**Subthreshold Leakage**

$$G = \frac{1}{1 + \frac{V_{DS}}{V_{LS}}^D}$$

- Nearly linear
- $$I_{DS} \approx K(V_{GS} - V_{Th})$$

**Power Gating**

- Turn OFF power to blocks when they are idle to save leakage
  - Use virtual $$V_{DD}$$ (V_{VDD})
  - Gate outputs to prevent invalid logic levels to next block
- Voltage drop across sleep transistor degrades performance during normal operation
  - Size the transistor wide enough to minimize impact
- Switching wide sleep transistor costs dynamic power
  - Only justified when circuit sleeps long enough

**Example: Power Management**

**Power states**

<table>
<thead>
<tr>
<th></th>
<th>C0 HF</th>
<th>C0 LF</th>
<th>C1 V2</th>
<th>C4</th>
<th>CE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core Voltage</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td>Core Clock</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>OFF</td>
<td>OFF</td>
</tr>
<tr>
<td>PLL</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>OFF</td>
<td>OFF</td>
</tr>
<tr>
<td>L1 Caches</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td>L2 Caches</td>
<td>[ ]</td>
<td>[ ]</td>
<td>Flushed</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td>Wake-Up Time</td>
<td>active</td>
<td>active</td>
<td>Partial Flush</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td>Power</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
</tbody>
</table>

**Summary**

- Wire contributes to delay, especially in modern technology
- We can use RC model to capture wire delays
- Energy becomes an increasingly important optimization goal
  - Dynamic energy
  - Static energy