Announcements

- Virtual Front Row today, 3/16:
  - Power and Energy
  - Ben Tait
  - Khashayar Pirouzmand
  - Jennifer Zhou
  - Zltao Fang
- HW7 to be posted later this week
- Midterm grading underway - grades posted by Thursday
- Feedback welcome!

The Watt: Unit of power. A rate of energy (J/s). A gas pump hose delivers 6 MW.

1 J = 1 W * s
1 W = 1 J/s.

120 KW: The power delivered by a Tesla Supercharger. Tesla Model S has a 306 MJ battery (good for 265 miles).

120 KW: The power delivered by a Tesla Supercharger. Tesla Model S has a 306 MJ battery (good for 265 miles).

Chevy Bolt battery capacity: 66 KWhr = 237 MJ (good for 259 miles).
Energy and Power

Energy is the ability to do work (W).

Power is rate of expending energy.

Energy Efficiency: energy per operation

- **Handheld and portable** (battery operated):
  - Energy Efficiency - limits battery life
  - Power - limited by heat

- **Infrastructure and servers** (connected to power grid):
  - Energy Efficiency - dictates operation cost
  - Power - heat removal contributes to TCO

Remember: \( P = IV \)
Sad fact: Computers turn electrical energy into heat. Computation is a byproduct.

Energy and Performance

Air or water carries heat away, or chip melts.
1 Joule heats 1 gram of water 0.24 degree C

This is how electric tea pots work...

1 Joule of Heat Energy per Second

1 Watt: Unit of power. The rate at which energy dissipated in the resistor.

The Watt

20 W rating: Maximum power the package is able to transfer to the air. Exceed rating and resistor burns.

1 Ohm Resistor

The Joule: Unit of energy. Can also be expressed as Watt-Seconds. Burning 1 Watt for 100 seconds uses 100 Watt-Seconds of energy.
Old example: Cooling an iPod nano ...

Like resistor on last slide, iPod relies on passive transfer of heat from case to the air.

Why? Users don’t want fans in their pocket ...

To stay “cool to the touch” via passive cooling, power budget of 5 W.

If iPod nano used 5W all the time, its battery would last 15 minutes ...
Powering an iPod nano (2005 edition)

1.2 W-hour battery:
Can supply 1.2 watts of power for 1 hour.

1.2 W-hr / 5 W ≈ 15 minutes.

More W-hours require bigger battery and thus bigger "form factor" -- it wouldn't be "nano" anymore :-).

Real specs for iPod nano:
14 hours for music,
4 hours for slide shows.

85 mW for music.
300 mW for slides.
What’s happened since 2005?

2010 nano

0.74 ounces
(50% of 2005 Nano)

“Up to” 24 hours audio playback.
70% improvement from 2005 nano.

0.39 W Hr
(33% of 2005 Nano)
3.8 V, 0.78 Wh lithium-ion battery on 38mm model. Apple claims the 205 mAh battery should provide up to 18 hours of use (which translates to 6.5 hours of audio playback, 3 hours of talk time, or 72 hours of Power Reserve mode.)
A clever prism projects a layer over reality light.
1.76 ounces - 4X the weight of iPod Shuffle

2.1 Wh battery - 5.3x as much energy as 2010 Nano.

Battery life very usage dependent.

640 x 360 Liquid Crystal on Silicon (LCoS) prism projector.
iPhone 6

4.7 inch iPhone 6:
1,810mAh battery
@3.8V = 6.88 Wh

iPhone 5s: 1570mAh
@3.8V = 6 Wh
The A8 is manufactured on a 20 nm process by TSMC. It contains 2 billion transistors. Its physical size is 89 mm$^2$. It has 1 GB of LPDDR3 RAM included in the package. It is dual core, and has a frequency of 1.38 GHz.

- Apple A8 APL1011 SoC + SK Hynix RAM as denoted by the markings H9CKNNN8KTMWR-NTH (we presume it is 1 GB LPDDR3 RAM, the same as in the iPhone 6 Plus)

- Qualcomm MDM9625M LTE Modem

- Skyworks 77802-23 Low Band LTE PAD

- Avago A8020 High Band PAD

- Avago A8010 Ultra High Band PA + FBARs

- SkyWorks 77803-20 Mid Band LTE PAD

- InvenSense MP67B 6-axis Gyroscope and Accelerometer Combo
<table>
<thead>
<tr>
<th>iPhone Model</th>
<th>Battery Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>iPhone 12 Mini</td>
<td>2,227 mAh</td>
</tr>
<tr>
<td>iPhone 12</td>
<td>2,815 mAh</td>
</tr>
<tr>
<td>iPhone 12 Pro</td>
<td>2,815 mAh</td>
</tr>
<tr>
<td>iPhone 12 Pro Max</td>
<td>3,687 mAh</td>
</tr>
<tr>
<td>iPhone 11</td>
<td>3,110 mAh</td>
</tr>
<tr>
<td>iPhone 11 Pro</td>
<td>3,046 mAh</td>
</tr>
<tr>
<td>iPhone 11 Pro Max</td>
<td>3,969 mAh</td>
</tr>
</tbody>
</table>

14.13 Wh @ 3.8V

https://unitedlex.com/insights/apple-iphone-12-pro-max-teardown-report
Performance: Must be “close enough” to desktop performance ... most people no longer used a desktop (even in 2006).

Size and Weight. Ideal: paper notebook.

Heat: No longer “laptops” -- top may get “warm”, bottom “hot”. Quiet fans OK.
Battery: Set by size and weight limits ...

Battery rating: 55 W-hour.

At 2.3 GHz, Intel Core Duo CPU consumes 31 W running a heavy load - under 2 hours battery life! And, just for CPU!

46x more energy than iPod nano battery. And iPod lets you listen to music for 14 hours!

Almost full 1 inch depth. Width and height set by available space, weight.

At 1 GHz, CPU consumes 13 Watts. “Energy saver” option uses this mode ...
MacBook Air ... design the laptop like an iPod
Mainboard: fills about 25% of the laptop

35 W-h battery: 63% of 2006 MacBook's 55 W-h
MacBook Air: Full PC

- Thunderbolt I/O
- Platform Controller Hub
- Core i5 CPU/GPU
- Up to 4GB DRAM
50Wh is 180,000 Joules!
Machine rooms are expensive.
Removing heat dictates how many servers to put in a machine room.

Electric bill adds up! Powering the servers + powering the air conditioners is a big part of TCO.

Reliability: running computers hot makes them fail more often.
Computations per W-h doubles every 1.6 years, going back to the first computer.

(Jonathan Koomey, Stanford).
CMOS Circuits and Energy
Switching Energy: Fundamental Physics

Every logic transition dissipates energy.

How can we limit switching energy?

1. Reduce # of clock transitions. But we have work to do ...
2. Reduce Vdd. But lowering Vdd limits the clock speed ...
3. Fewer circuits. But more transistors can do more work.
4. Reduce C per node. One reason why we scale processes.

Strong result: Independent of technology.
Chip-Level “Dynamic” Power

\[ P_{sw} = \frac{1}{2} \alpha C V_{dd}^2 F \]

“activity factor”, average percentage of capacitance switching per cycle (~ number of nodes to switch)

Total chip capacitance to be switched

Clock Frequency
Additional Dynamic Power - “short circuit current”

When gate switches, brief period when both pullup network and pulldown network could be on.

Worse when input is changing slowly compared to the output.
Another Factor: Leakage Currents

Even when a logic gate isn’t switching, it burns power.

\[ OV = V_{IN} \]

I\text{Gate}: Ideal capacitors have zero DC current. But modern transistor gates are a few atoms thick, and are not ideal.

I\text{Sub}: Even when this nFet is off, it passes an I\text{off} leakage current.

We can engineer any I\text{off} we like, but a lower I\text{off} also results in a lower I\text{on}, and thus a lower maximum clock speed.

Intel’s 2006 processor designs, leakage vs switching power

A lot of work was done to get a ratio this good ... 50/50 is common.

Bill Holt, Intel, Hot Chips 17.
We can increase $I_{on}$ by raising $V_{dd}$ and/or lowering $V_t$.

$0.7 = V_{dd}$

$I_{on} = 1.2 \text{ mA}$

$I_{off} = 0$ ???

$0.25 \approx V_t$
Plot on a “Log” Scale to See “Off” Current

We can decrease $I_{off}$ by raising $V_t$ - but that lowers $I_{on}$.

$I_{off} \approx 10 \text{ nA}$

$1.2 \text{ mA} = I_{on}$

$0.25 \approx V_t$

$0.7 = V_{dd}$
Customize processes for product types ...

- $V_t$ is controlled by channel doping.
- Modern IC processes have 2 or 3 different $V_t$ values available.
- Standard cell libraries offer low $V_t$ and high $V_t$ versions of cells so that the tools can optimize on a per instance basis.
- (If high performance not needed then use high $V_t$ to reduce leakage).

Transistor channel is a raised fin. Gate controls channel from sides and top. Channel depth is fin width. 12-15nm for L=22nm.
Thanks in large part to remarkable research started by Cal Berkeley professor Chenming Hu under a DARPA contract, the 20-nm process will likely be the last hurrah for the planar transistor (at least as we know it today), as the industry moves to FETs built with fins.

**INS AND OUTS OF FINS**

In a planar transistor of today, electrical current flows from source to drain through a flat, 2D horizontal channel underneath the gate. The gate voltage controls current flow through the channel. As transistor size shrinks with the introduction of each new silicon process, the planar transistor cannot adequately stop the flow of current when it is in an “off” state, which results in leakage and heat. In a FinFET MOSFET transistor, the gate wraps around the channel on three sides, giving the gate much better electrostatic control to stop the current when the transistor is in the “off” state. Superior gate control in turn allows designers to increase the current and switching speed and, thus, the performance of the IC.

Because the gate wraps around three sides of the fin-shaped channel, the FinFET is often called a 3D transistor (not to be confused with 3D ICs, like the Virtex-7 2000T, which Xilinx pioneered with its stacked-silicon technology).

In a three-dimensional transistor (see Figure 1b), gate control of the channel is on three sides rather than just one, as in conventional two-dimensional planar transistors (see Figure 1a). Even better channel control can be achieved with a thinner fin, or in the future with a gate-all-around structure where the channel will be enclosed by a gate on all sides.

The industry believes the 16-nm/14-nm FinFET process will enable a 50 percent performance increase at the same power as a device built at 28 nm. Alternatively, the FinFET device will consume 50 percent less power at the same performance. The performance-per-watt benefits added to the continued increases in capacity make FinFET processes extremely promising for devices at 16 or 14 nm and beyond.

That said, the cost and complexity of designing and manufacturing 3D transistors is going to be higher at least for the short term, as EDA companies figure out ways to adequately model the device characteristics of these new processes and augment their tools and flows to account for signal integrity, electromigration, width quantization, resistance and capacitance. This complexity makes designing ASICs and ASSPs even riskier and more expensive than before.

Xilinx, however, shields users from the manufacturing details. Customers can reap the benefits of increased performance per watt and Xilinx’s Generation Ahead design flows to bring innovations based on the new UltraScale architecture to market faster.

**Figure 1** – The position of the gate differs in the two-dimensional traditional planar transistor (a) vs. the three-dimensional FinFET transistor (b).
Total Power = $P_{\text{switching}} + P_{\text{short-circuit}} + P_{\text{leakage}}$

$I_{DS_{\text{Sub}}} = k \cdot e^{\frac{-q \cdot V_T}{a \cdot k_B \cdot T}}$
Some low-power design techniques

- Parallelism and pipelining
- Power-down idle transistors
- Slow down non-critical paths
- Clock gating
- Thermal management
Trading Hardware for Power via Parallelism and Pipelining ...
Gate delay roughly linear with $V_{dd}$

And so, we can transform this:

- **Logic Block**
  - $V_{dd}$
  - $F = 1$
  - $V_{dd} = 1$
  - Throughput = 1
  - Power = 1
  - Area = 1
  - Pwr Den = 1

$
\text{P} \sim F \times V_{dd}^2
$

Block processes stereo audio. 1/2 of clocks for "left", 1/2 for "right".

Into this:

- **Logic Block**
  - $V_{dd}/2$
  - $F = 0.5$
  - $V_{dd} = 0.5$
  - Throughput = 1
  - Power = 0.25
  - Area = 2
  - Pwr Den = 0.125

$\text{P} \sim \#\text{blks} \times F \times V_{dd}^2$

$\text{P} \sim 2 \times 1/2 \times 1/4 = 1/4$

$CV^2$ power only

THIS MAGIC TRICK BROUGHT TO YOU BY CORY HALL ...
Chandrakasan & Brodersen (UCB, 1992)

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Power (normalized)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>1</td>
</tr>
<tr>
<td>Parallel</td>
<td>0.36</td>
</tr>
<tr>
<td>Pipelined</td>
<td>0.39</td>
</tr>
<tr>
<td>Pipelined-Parallel</td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Area (normalized)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>1</td>
</tr>
<tr>
<td>Parallel</td>
<td>3.4</td>
</tr>
<tr>
<td>Pipelined</td>
<td>1.3</td>
</tr>
<tr>
<td>Pipelined-Parallel</td>
<td>3.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Voltage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>5V</td>
</tr>
<tr>
<td>Parallel</td>
<td>2.9V</td>
</tr>
<tr>
<td>Pipelined</td>
<td>2.9V</td>
</tr>
<tr>
<td>Pipelined-Parallel</td>
<td>2.0</td>
</tr>
</tbody>
</table>

From: Minimizing Power Consumption in CMOS Circuits

Anantha P. Chandrakasan
Robert W. Brodersen
A 2.05 GVertices/s 151 mW Lighting Accelerator for 3D Graphics Vertex and Pixel Shading in 32 nm CMOS
A 2.05 GVertices/s 151 mW Lighting Accelerator for 3D Graphics Vertex and Pixel Shading in 32 nm CMOS

Farhana Sheikh, Member, IEEE, Sanu K. Mathew, Member, IEEE, Mark A. Anders, Member, IEEE, Himanshu Kaul, Member, IEEE, Steven K. Hsu, Member, IEEE, Amit Agarwal, Member, IEEE, Ram K. Krishnamurthy, Fellow, IEEE, and Shekhar Borkar, Fellow, IEEE
Multiple Cores for Low Power

Trade hardware for power, on a large scale ...
Cell: The PS3 chip
Cell (PS3 Chip): 1 CPU + 8 “SPUs”

- L2 Cache: 512 KB
- PowerPC
- 8 Synergistic Processing Units (SPUs)
A “Schmoo” plot for a Cell SPU ...

The lower Vdd, the less dynamic energy consumption.

\[ E_{0\rightarrow1} = \frac{1}{2} C V_{dd}^2 \]
\[ E_{1\rightarrow0} = \frac{1}{2} C V_{dd}^2 \]

The lower Vdd, the longer the maximum clock period, the slower the clock frequency.

<table>
<thead>
<tr>
<th>Vdd (Volts)</th>
<th>1.3</th>
<th>1.2</th>
<th>1.1</th>
<th>1.0</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>48C 4W</td>
<td>49C 4W</td>
<td>50C 5W</td>
<td>50C 6W</td>
<td>51C 7W</td>
<td>52C 7W</td>
</tr>
<tr>
<td>39C 2W</td>
<td>39C 3W</td>
<td>40C 3W</td>
<td>41C 4W</td>
<td>42C 4W</td>
<td>43C 4W</td>
</tr>
<tr>
<td>32C 2W</td>
<td>33C 2W</td>
<td>33C 3W</td>
<td>35C 3W</td>
<td>36C 3W</td>
<td>36C 4W</td>
</tr>
<tr>
<td>25C 1W</td>
<td>26C 1W</td>
<td>26C 2W</td>
<td>27C 2W</td>
<td>27C 2W</td>
<td>27C 2W</td>
</tr>
</tbody>
</table>

Freq (GHz) 2 2.5 3 3.5 4 4.5 5 5.5

Failed
Clock speed alone doesn’t help $E/\text{op}$...

But, lowering clock frequency while keeping voltage constant spreads the same amount of work over a longer time, so chip stays cooler...

$$E_{0\rightarrow1} = \frac{1}{2} CV_{dd}^2$$
$$E_{1\rightarrow0} = \frac{1}{2} CV_{dd}^2$$
Scaling V and f does lower energy/op

1 W to get 2.2 GHz performance. 26 C die temp. 7 W to reliably get 4.4 GHz performance. 47 C die temp.

If a program that needs a 4.4 GHz CPU can be recoded to use two 2.2 GHz CPUs ... big win.
Dynamic Voltage/Frequency Scaling (DVFS)

Many modern processors have controls for dynamically changing operating frequency and voltage.

Intel power states

- BIO/OS software can adjust frequency to reduce heat and/or improve power efficiency with high performance not needed.
- Adjusting both voltage and frequency helps improve energy efficiency and allows higher frequency for a given power level.
Powering down idle circuits
Add “sleep” transistors to logic ...

Example: Floating point unit logic.

When running fixed-point instructions, put logic “to sleep”.

+++ When “asleep”, leakage power is dramatically reduced.

--- Presence of sleep transistors slows down the clock rate when the logic block is in use.
Intel example: Sleeping cache blocks

>3x SRAM leakage reduction on inactive blocks

A tiny current supplied in “sleep” maintains SRAM state.

The 32nm CPU in the Intel® Atom™ Processor Z2480 is a process-shrink of the original 45nm Atom™ micro-architecture [5]. The 32nm CPU doubled the size of the Gshare branch predictor to 8K entries and optimized memory copy performance. Additional low-power enhancements include operating the CPU at lower minimum voltage, reducing active power of the CPU PLL, separating voltage rails for CPU and the L2 cache, and enabling full power-gating of the CPU in the C6 standby mode. This section highlights the results of these low-power optimizations, and how they apply to smartphone use cases.

As shown in Figure 5, the Intel® Atom™ CPU provides a wide dynamic performance/power operating range. On Medfield, fine-grained active CPU power management is accomplished through dynamic frequency voltage scaling that is controlled by Enhanced Intel® SpeedStep® Technology [7], also known as CPU P-states. The dynamic range of the CPU ranges from 600MHz @ ~175mW to a sustained high frequency mode (HFM) of 1.3GHz @ ~500mW. For bursty workloads (e.g., interactive use of a web browser) the CPU can burst up to 2.0GHz @ ~1,200mW for short periods of time.

As long as thermal headroom allows, the CPU can run in burst mode until thermal monitors in the platform, the SoC, or the CPU indicate that thermal design power limits have been reached. When this occurs, a combination of firmware and software throttle the CPU back into a lower P-state. Additionally, CPU w/512KB L2$
Intel Medfield

Swatches 45 power “islands.”

Fine-grained control of leakage power, to track user activity.

“Race to idle” strategy -- finish tasks quickly, to get to power down.
Playing a game ...

Active system looks like this
Watching a video ...

CPU w/512KB L2$

IO

Security Engine

Power Manager

Low Power Audio

Storage

CPU w/512KB L2$

IO

2D/3D Graphics

Video Encode/Decode (1080p30)

Image Signal Processor

Display Controller (3 pipes)

IO

CPU is now off!
Figure 8 – S0i1 System State with power consumption in the mW range

Figure 9 – S0i3 System State with power consumption in the uW range

Looking at phone screen, not doing anything ...
Phone in your pocket, waiting for a call ...

Figure 8 – S0i1 System State with power consumption in the mW range

Figure 9 – S0i3 System State with power consumption in the uW range

Standby State – just waiting for wakes
Slow down “slack paths”
Fact: Most logic on a chip is “too fast”

- The critical path
- Most wires have hundreds of picoseconds to spare.

Use several supply voltages on a chip ...

Why use multi-Vdd? We can reduce dynamic power by using low-power Vdd for logic off the critical path.

What if we can’t do a multi-Vdd design?
In a multi-Vt process, we can reduce leakage power on the off critical path logic by using high-Vth transistors.

Gating clocks to save power
On a CPU, where does the power go?

Half of the power goes to latches (Flip-Flops).

Most of the time, the latches don’t change state.

So “gated” clocks are a big win.
But, done with CAD tools in a disciplined way.

From: Bose, Martonosi, Brooks: Sigmetrics-2001 Tutorial
Synopsis Design Compiler can do this ...

"Up to 70% power savings at the block level, for applicable circuits"

Synopsis Data Sheet
Thermal Management
Keep chip cool to minimize leakage power

A recipe for thermal runaway

<table>
<thead>
<tr>
<th>Junction Temperature (T_J °C)</th>
<th>Normalized Static Power or ( I_{CCINTQ} ) Typical</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>1.00</td>
</tr>
<tr>
<td>50</td>
<td>1.46</td>
</tr>
<tr>
<td>85</td>
<td>2.50</td>
</tr>
<tr>
<td>100</td>
<td>3.14</td>
</tr>
</tbody>
</table>

Figure 3: \( I_{CCINTQ} \) vs. Junction Temperature with Increase Relative to 25°C

Optimizing Designs for Power Consumption through Changes to the FPGA Environment
Five low-power design techniques

- Parallelism and pipelining
- Power-down idle transistors
- Slow down non-critical paths
- Clock gating
- Thermal management