F04 Delays Per Cycle for Processor Designs

F04 delay per cycle is roughly proportional to the amount of computation completed per cycle.
With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004.

With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004.

With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004.

With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004.

With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004.

With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004.

With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004.

With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004.

With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004.

With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004.

With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004.

With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004.

With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004.

With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004.

With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004.

With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004.

With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004.
Sad fact: Computers turn electrical energy into heat. Computation is a byproduct.

Energy and Performance

Air or water carries heat away, or chip melts.
The Joule: Unit of energy. A 1 Gallon gas can holds 130 MJ of energy.

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.
The Joule: Unit of energy. Can also be expressed as Watt-Seconds. Burning 1 Watt for 100 seconds uses 100 Watt-Seconds of energy.

This is how electric tea pots work...

1 Joule heats 1 gram of water 0.24 degree C

1 Joule of Heat Energy per Second

The Watt: Unit of power. The amount of energy burned in the resistor in 1 second.

1 Watt

1 Ohm Resistor

20 W rating: Maximum power the package is able to transfer to the air. Exceed rating and resistor burns.
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, “hard” 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 / 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.
Finding the (2005) iPod nano CPU ...

A close relative ...

Two 80 MHz CPUs. One CPU used for audio, one for slides.

Low-power ARM roughly 1mW per MHz ... variable clock, sleep modes.

85 mW system power realistic ...
Year-to-year: continuous improvements

iPod nano
2005
14 hours
battery life
(audio playback)

What changed inside?

iPod nano
2006
24 hours
battery life
(audio playback)

Source: ifixit.com
iPod nano 2005 - a C-shaped PC board, with a battery in the “C” opening.

iPod nano 2006 - battery lies on top of PC board.

Source: ifixit.com
How? Small IC packages, fewer parts

iPod nano 2006 → iPod nano 2005

Source: arstechnica.com
Aluminum permits thinner case ... What’s happened since 2006?

Source: ilounge.com
2010 Nano:
“up to” 24 hours audio playback

2010 Shuffle:
“up to” 15 hours audio playback

0.74 ounces
0.39 W Hr
(33% of 2005 Nano)

0.44 ounces
0.19 W Hr

Sources: iFixit, Apple
0.44 ounce
0.19 W Hr

~0.6 ounce for frame
~0.3 ounce per lens
Desired screen size sets smartphone W x L

Depth? : Thin body vs. battery life
22% gain in battery energy over 5 iterations
iPhone 4S

Battery

L-shape
Main Board

Metal frame acts as antenna
In 4 years:

- 6.8x increase in transistor count
- 33% max clock speed increase

Attached DRAM:
- 128 MB → 512 MB

6.8x transistors:
- Dual CPU and GPU, and to save energy.
Notebooks ... as designed in 2006 ...

2006 Apple MacBook -- 5.2 lbs

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!

At 1 GHz, CPU consumes 13 Watts. "Energy saver" option uses this mode...

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

46x more energy than iPod nano battery. And iPod lets you listen to music for 14 hours!
MacBook Air ... design the laptop like an iPod
2011 Air: 11.8 in x 7.56 in x 0.68 in; 2.38 lbs

2006 Macbook: 12.8 in x 8.9 in x 1 in; 5.2 lbs
Mainboard: fills about 25% of the laptop

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

Top

Thunderbolt I/O

Platform Controller Hub

Core i5 CPU/GPU

Up to 4GB DRAM

Bottom

Thursday, September 6, 12
2011 Air: 35 W-h battery, 5 hour battery life*
2012 iPad: 42.5 W-h battery, 10 hour battery life*

*For a content-consumption workload.

iPad: 1.44 lbs
MacBook Air 11.6 in: 2.38 lbs

Battery-Life-Hour/W-h: 1.7x iPad advantage
iPad: iPhone++

A5X: 2 ARM Cortex9 Cores, expanded PowerVR GPU

Top

Up to 64 GB Flash

Bottom

1 GB DRAM

Cellular front-end chips on separate board.
Apple A5X

2012 iPad CPU/IGP.

45 nm, 13 x 13 mm

IGP fills about 40% of die.

IGP: 2% of Kepler (in GFLOPs).
2011 Air: $999 -- 64 GB SSD, 2 GB RAM, x86
2012 iPad: $699 -- 64 GB SSD, 1 GB RAM, ARM

iPad 2012 CPU: iPhone 4S, with 2X GPU and RAM
iPad 2011→2012, battery W-hours increased by 70% 

Weight increase of 0.11 pound, thicker by 0.03 inches.

Increase needed to double display resolution while keeping battery life @ 10 hours.
The CPU is only part of power budget!

2004-era notebook running a full workload.

"Amdahl’s Law for Power"

If our CPU took no power at all to run, that would only double battery life!

iPad 2011→2012 power++ comes from this side.
Servers: Total Cost of Ownership (TCO)

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).
Processors 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.
Scaling switching energy per gate ...

IC process scaling ("Moore's Law")

Due to reducing V and C (length and width of Cs decrease, but plate distance gets smaller).

Recent slope more shallow because V is being scaled less aggressively.


CS 250 L5: Power and Energy

Thursday, September 6, 12
With this open database, you can mine microprocessor trends over the past 40 years.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

**Voltage Versus Feature Size**

- **$V_{dd}$** vs. feature size (um)
- Slope = 0.63

Not surprisingly, specification data gets harder to find the older the processor becomes, especially for those that are no longer made, or worse, whose manufacturers no longer exist. We have been collecting this type of data for three decades and are now releasing it in the form of an open repository of processor specifications. The goal of CPU DB is to aggregate detailed processor specifications into a convenient form and to encourage community participation, both to leverage this information and to keep it accurate and current.

CPU DB (cpudb.stanford.edu) is populated with desktop, laptop, and server processors, for which we use SPEC CoreMark benchmark for performance. With this open database, you can mine microprocessor trends over the past 40 years.

In addition, we provide a methodology to separate the effect of technology scaling from improvements on other frontiers (e.g., architecture and software), allowing the comparison of machines built in different technologies. To demonstrate the utility of this data and analysis, we use it to decompose processor improvements into contributions from the physical scaling of devices, and where did these incredible gains come from? This article sheds some light on this question by extending the coverage of embedded processors in the database.

In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004. It had 2,300 transistors, ran at a clock speed of up to 740 KHz, and delivered 3 instructions per second while dissipating 0.5 watts. The following four decades witnessed exponential growth in compute power, a trend that has enabled applications as diverse as climate modeling, protein folding, and computing real-time ballistic trajectories of angry birds. Today's microprocessor chips employ billions of transistors, include multiple processor cores on a single silicon die, run at clock speeds measured in gigahertz, and deliver more than 4 million times the performance of the original 4004.

In addition, the database contains limited data on embedded cores, for which we are using our performance-measuring tool. In November 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004. It had 2,300 transistors, ran at a clock speed of up to 740 KHz, and delivered 3 instructions per second while dissipating 0.5 watts. The following four decades witnessed exponential growth in compute power, a trend that has enabled applications as diverse as climate modeling, protein folding, and computing real-time ballistic trajectories of angry birds. Today's microprocessor chips employ billions of transistors, include multiple processor cores on a single silicon die, run at clock speeds measured in gigahertz, and deliver more than 4 million times the performance of the original 4004.

With the CoreMark benchmark for performance.
**Second Factor: Leakage Currents**

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

\[ O V = V_{IN} \]

\[ I_{Gate} \]

\[ I_{Sub} \]

\[ V_{OUT} \]

\[ C_L \]

**I_{sub}:** Even when this nFet is off, it passes an \( I_{off} \) leakage current.

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

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

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.
Engineering “On” Current at 25 nm ...

We can increase $I_{on}$ by raising $V_{dd}$ and/or lowering $V_t$.

$0.7 = V_{dd}$

$0.25 = V_t$

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

$I_{off} = 0$ ???
Plot on a “Log” Scale to See “Off” Current

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

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

$0.25 \approx V_t$

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

$0.7 = V_{dd}$

$I_{off} \approx 10 \text{ nA}$
Device engineers trade speed and power

We can reduce \( CV^2 (P_{\text{active}}) \) by lowering \( V_{dd} \).

We can increase speed by raising \( V_{dd} \) and lowering \( V_t \).

We can reduce leakage \( (P_{\text{standby}}) \) by raising \( V_t \).

---

From: Silicon Device Scaling to the Sub-10-nm Regime
Meikei Ieong, Bruce Doris, Jakub Kedzierski, Ken Rim, Min Yang
Customize processes for product types ...

(H, S, L) == High Vt, Standard Vt, Low Vt

(40, 45, 50) are channel lengths (in nm)
Transistor physics revisited ...

The drain junction is also a capacitor, and puts - charges in the substrate.

Away from the surface, the drain-induced charges remain even when the gate is off!

As we make $L$ smaller, source and drain come closer, and $I_{off}$ gets larger!
Solution concept: Fully-depleted channel

On:
\[ V_s = 0V \quad V_g = 1V \quad V_d = 1V \]
\[ I \approx \mu A \]

Off:
\[ V_s = 0V \quad V_g = 0V \quad V_d = 1V \]
\[ I \approx nA \]

We limit the depth of the channel so that the gate voltage “wins” over the drain voltage.

Done as shown, 5 to 7 nm depth for a 20 nm transistor. Requires expensive wafers.

“FD-SOI” -- Fully-Depleted Silicon-On-Insulator
Transistor channel is a raised fin.
Gate controls channel from sides and top.
Channel depth is fin width. 12-15nm for L=22nm.

Intel 22nm Process

United States Patent
Hu et al. Filed: Oct. 23, 2000

FINFET TRANSISTOR STRUCTURES HAVING A DOUBLE GATE CHANNEL EXTENDING VERTICALLY FROM A SUBSTRATE AND METHODS OF MANUFACTURE

Inventors: Chenming Hu, Alamo; Tsu-Jae King, Fremont; Vivek Subramanian, Redwood City; Leland Chang, Berkeley; Xuejue Huang; Yang-Kyu Choi, both of Albany; Jakub Tadeusz Kedzierski, Hayward; Nick Lindert, Berkeley; Jeffrey Bokor, Oakland, all of CA (US); Wen-Chin Lee, Beaverton, OR (US)
Intel "Ivy Bridge" 22nm CPUs, first production parts
Sandy Bridge

32nm planar

1.16B transistors

Ivy Bridge

22nm FinFet

1.4B transistors

“Less than half the power @ same performance”
Too many low-power techniques to cover!

Interpret as: 17% of respondents having “Personal Computing & Peripherals” as primary application use “Low Vdd Standby” low power design technique.
Five 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 ...
Power and Area Minimization of Reconfigurable FFT Processors: A 3GPP-LTE Example

Chia-Hsiang Yang, Member, IEEE, Tsung-Han Yu, Student Member, IEEE, and Dejan Marković, Member, IEEE
Gate delay roughly **linear** with Vdd

**Into this:**

- Logic Block
  - Vdd/2
  - Freq = 0.5
  - Vdd = 0.5
  - Throughput = 1
  - Power = 0.25
  - Area = 2
  - Pwr Den = 0.125

- Logic Block

**Top block processes “left”, bottom “right”.

**Block processes stereo audio. 1/2 of clocks for “left”, 1/2 for “right”.

- Logic Block
  - Vdd
  - Freq = 1
  - Vdd = 1
  - Throughput = 1
  - Power = 1
  - Area = 1
  - Pwr Den = 1

**And so, we can transform this:**

- P ∼ F × Vdd
- P ∼ 1 × 1²

**Minimizing Power Consumption in CMOS Circuits**

Anantha P. Chandrakasan
Robert W. Brodersen

CS 250 L5: Power and Energy

Thursday, September 6, 12
Minimizing Power Consumption in CMOS Circuits

From:

Chandrakasan & Brodersen (UCB EECS)

<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>

Area = 636 x 833 \( \mu^2 \)

Area = 1476 x 1219 \( \mu^2 \)

Area = 640 x 1081 \( \mu^2 \)
Multiple Cores for Low Power

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

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

The lower Vdd, the less dynamic energy consumption.

\[ E_{0\rightarrow1} = \frac{1}{2} CV_{dd}^2 \quad E_{1\rightarrow0} = \frac{1}{2} CV_{dd}^2 \]

The lower Vdd, the longer the maximum clock period, the slower the clock frequency.
Clock speed alone doesn’t help E/op ...

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

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

<table>
<thead>
<tr>
<th>V_{dd} (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>Freq (GHz)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Failed
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.
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

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 slow logic by using high-Vth transistors.

Logical partition into 0.8V and 1.0V nets done manually to meet 350 MHz spec (90nm).

Level-shifter insertion and placement done automatically.

Dynamic power in 0.8V section cut 50% below baseline.

Leakage power in 1.0V section cut 70% below baseline.

From a chapter from new book on ASIC design by Chinnery and Keutzer (UCB).
Synopsis Global User Survey, 2011

Number of Voltage Domains

- 1-3
- 4-5
- 6-10

2011 N = 758
Median Voltage Domains Current Design = 3

Thanks to Mary Ann White, Synopsis
28nm cell library for power/speed tradeoff

<table>
<thead>
<tr>
<th>28nm Architecture</th>
<th>High-Density</th>
<th>High-Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>9.5 grids (95nm)</td>
<td>12.5 grids (125nm)</td>
</tr>
<tr>
<td>Raw gate density**</td>
<td>3120K gates/mm²</td>
<td>2520K gates/mm²</td>
</tr>
</tbody>
</table>

** Post-shrink on silicon, 60% NAND2, 40% DFF

Thanks to Mary Ann White, Synopsis
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 (gasp) gated clocks are a big win.
But, done with CAD tools in a disciplined way.

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

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

Synopsis Data Sheet
Synopsis Global User Survey, 2011 ...

Number of Clock Domains

- 21+
- 11-20
- 6-10
- 1-5

2007 2008 2009 2010 2011

2011 N = 827
Median Clock Domains Current Design = 6

Thanks to Mary Ann White, Synopsis
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

CS 250  L5: Power and Energy

WP285 (v1.0) February 14, 2008  UC Regents Fall 2012 © UCB
IBM Power 4: How does die heat up?

4 dies on a multi-chip module

2 CPUs per die
115 Watts: Concentrated in “hot spots”

66.8 °C == 152 °F
82 °C == 179.6 °F
Idea: Monitor temperature, servo clock speed

Power consumption varies greatly by workload

Time (as we run a benchmark suite)

Power Headroom that can be utilized per core.

* Based on internal AMD modeling using benchmark simulations

TDP = Thermal Design Point
Upcoming: Project Lectures

<table>
<thead>
<tr>
<th>Tue</th>
<th>Sep 11</th>
<th>JL</th>
<th>Lecture 6: Image Processing : Theory and Practice</th>
</tr>
</thead>
</table>