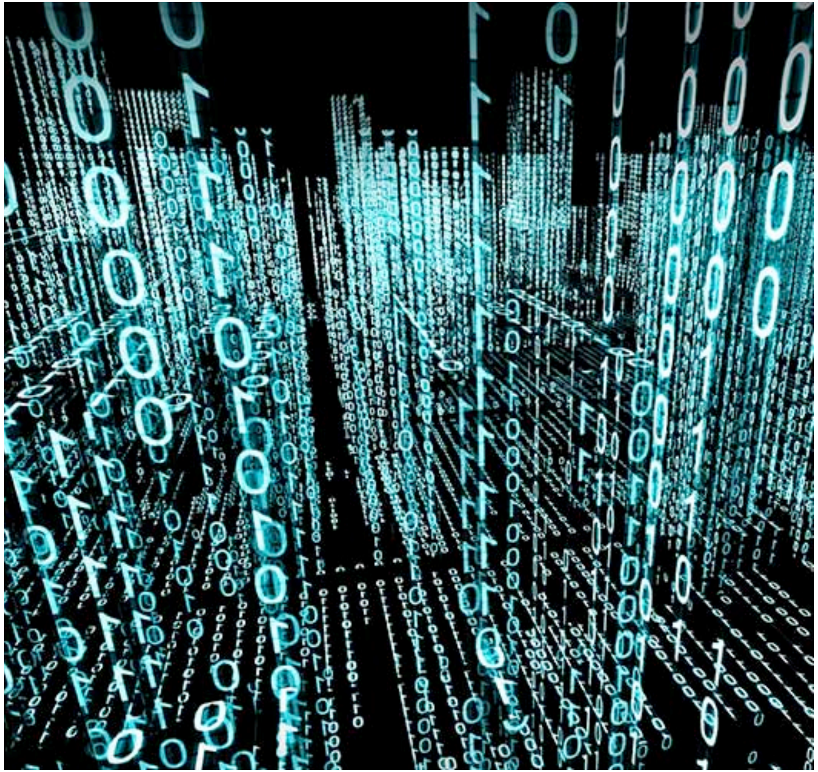


EECS151/251A
Spring 2024
Digital Design and Integrated Circuits

Instructor:
John Wawrzynek

Lecture 18 - Energy

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

The Joule: Unit of
energy. A 1 Gallon gas
container holds
130 MJ of energy.

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

$1\ J = 1\ W * s \qquad 1\ W = 1\ J/s.$

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

$$P = \frac{dW}{dt}$$

- ❑ ***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.

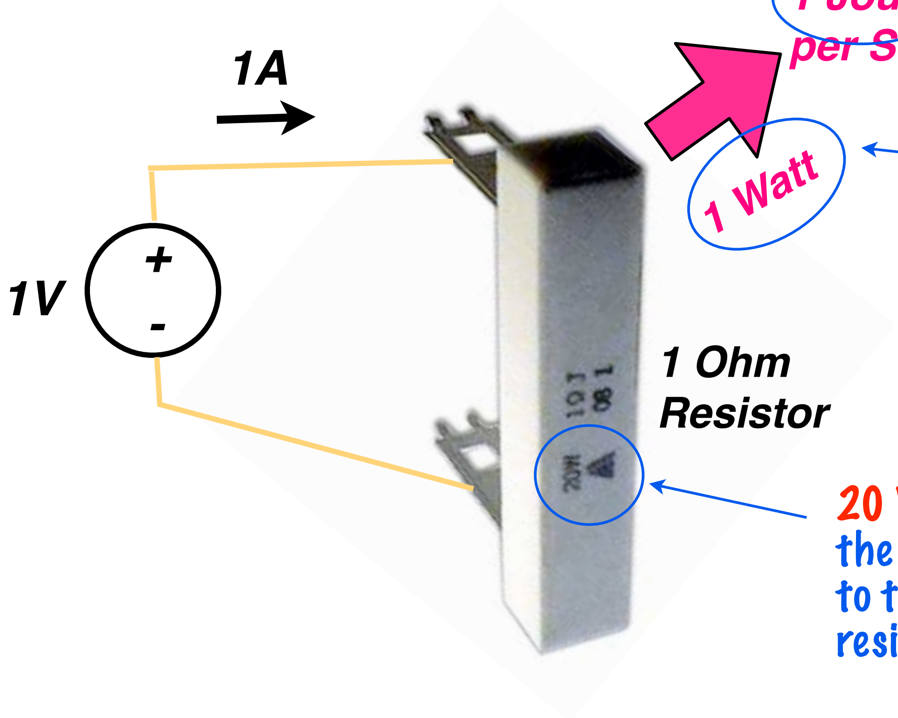
Energy and Performance

Air or water carries heat away, or chip melts.

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

1 Watt

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

1 Ohm Resistor

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

Old example: Cooling an iPod nano ...

portable media player

2005-2017

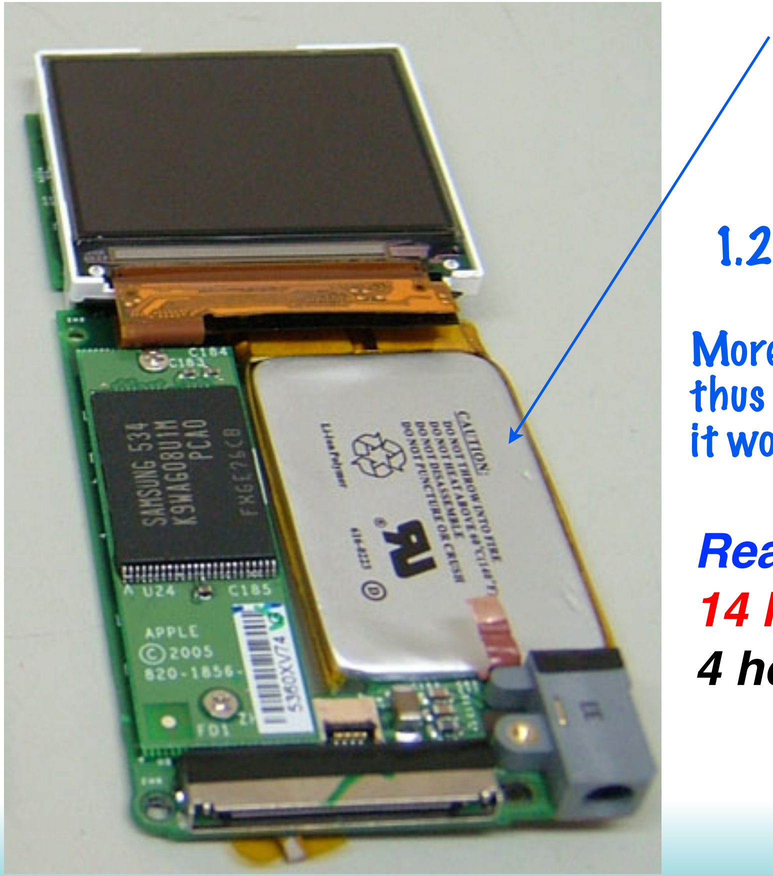
Like resistor on last slide, iPod relies on passive transfer of heat from case to the air.

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)



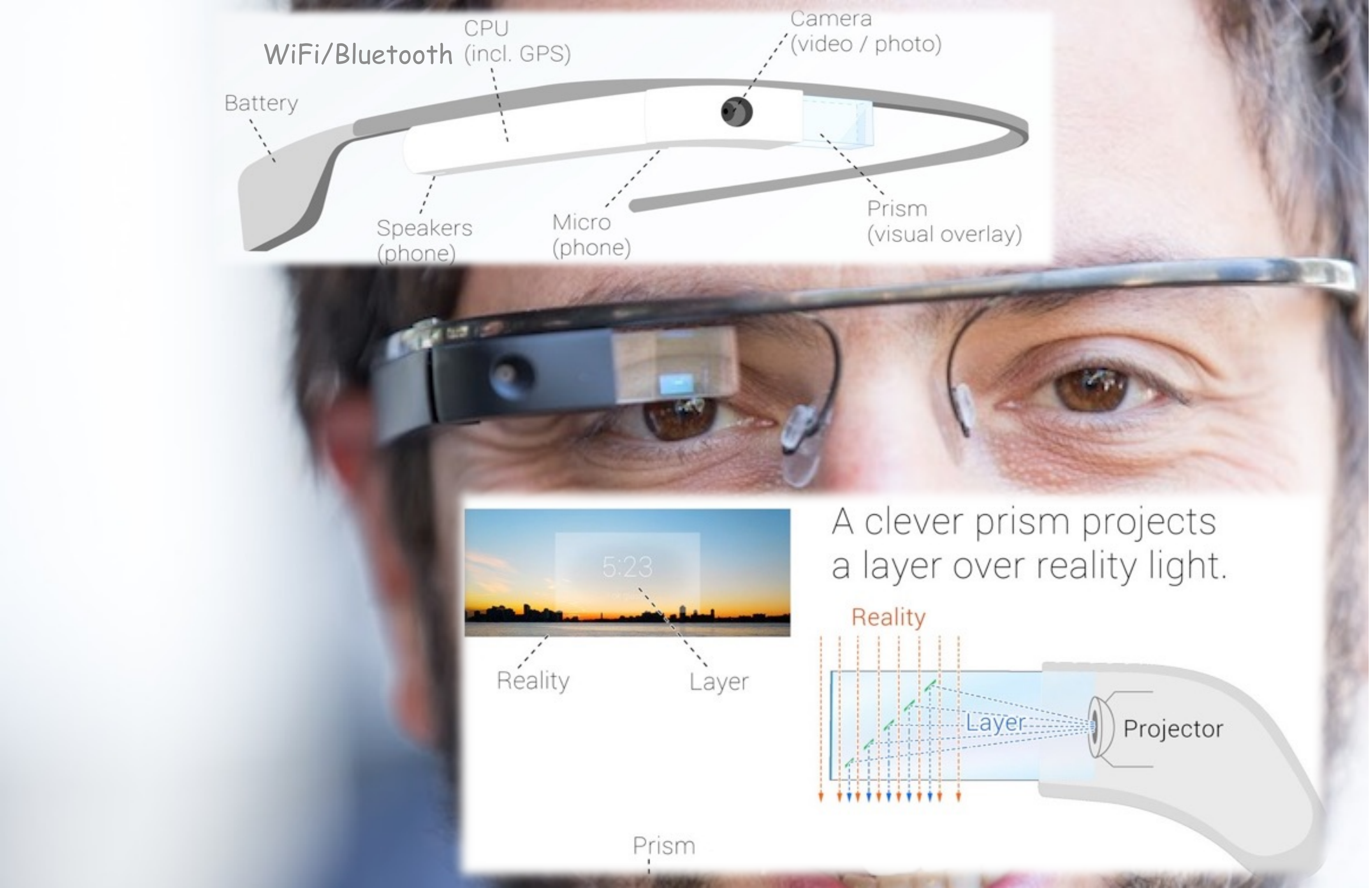
1.2 W-hour battery:
Can supply 1.2 watts
of power for 1 hour.

1.2 W-hr / 5 W ≈ 15 minutes.

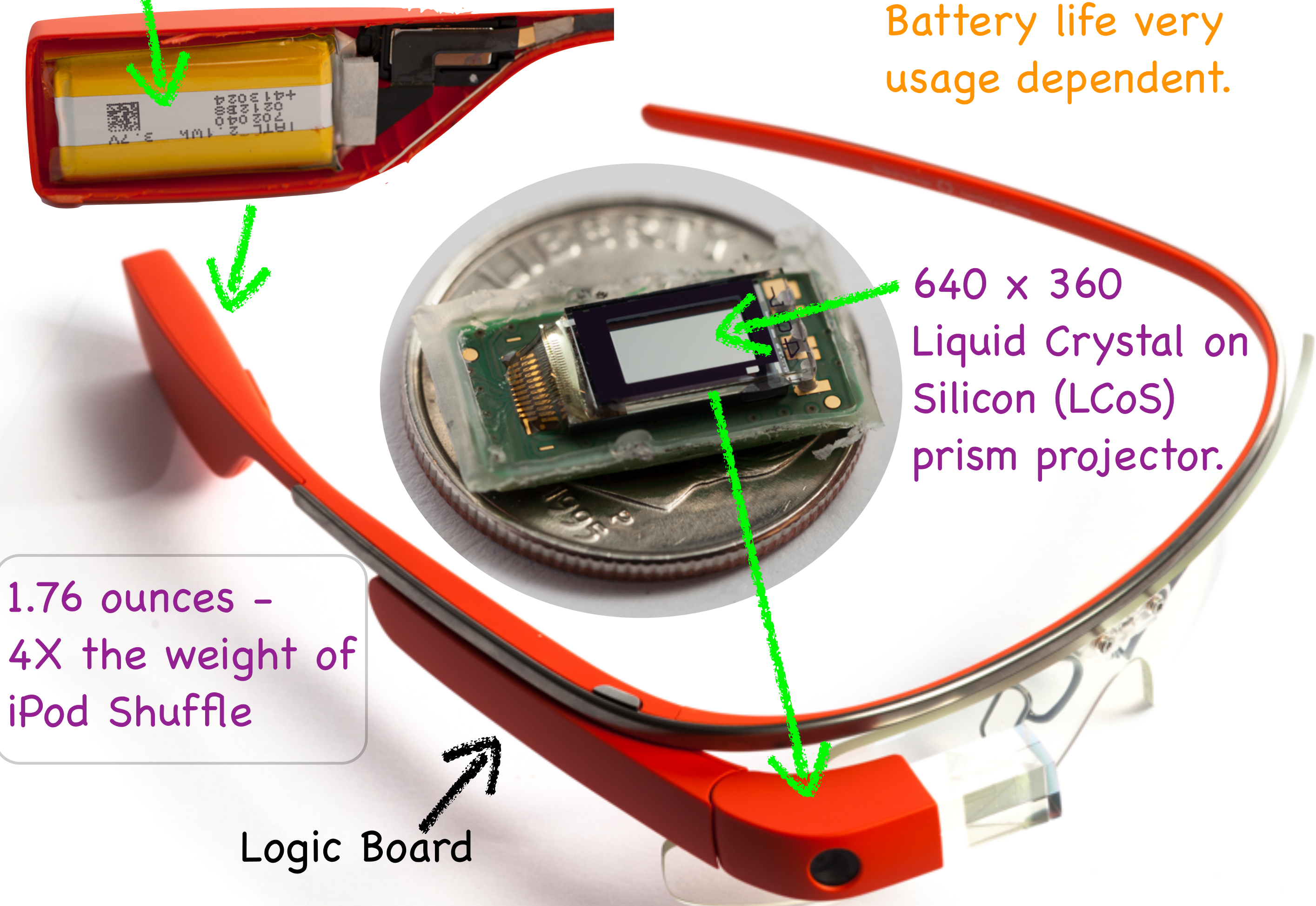
More W-hours require bigger battery and
thus bigger "form factor" --
it wouldn't be "nano" anymore :-).

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

85 mW for music.
300 mW for slides.



A clever prism projects a layer over reality light.



2.1 Wh battery

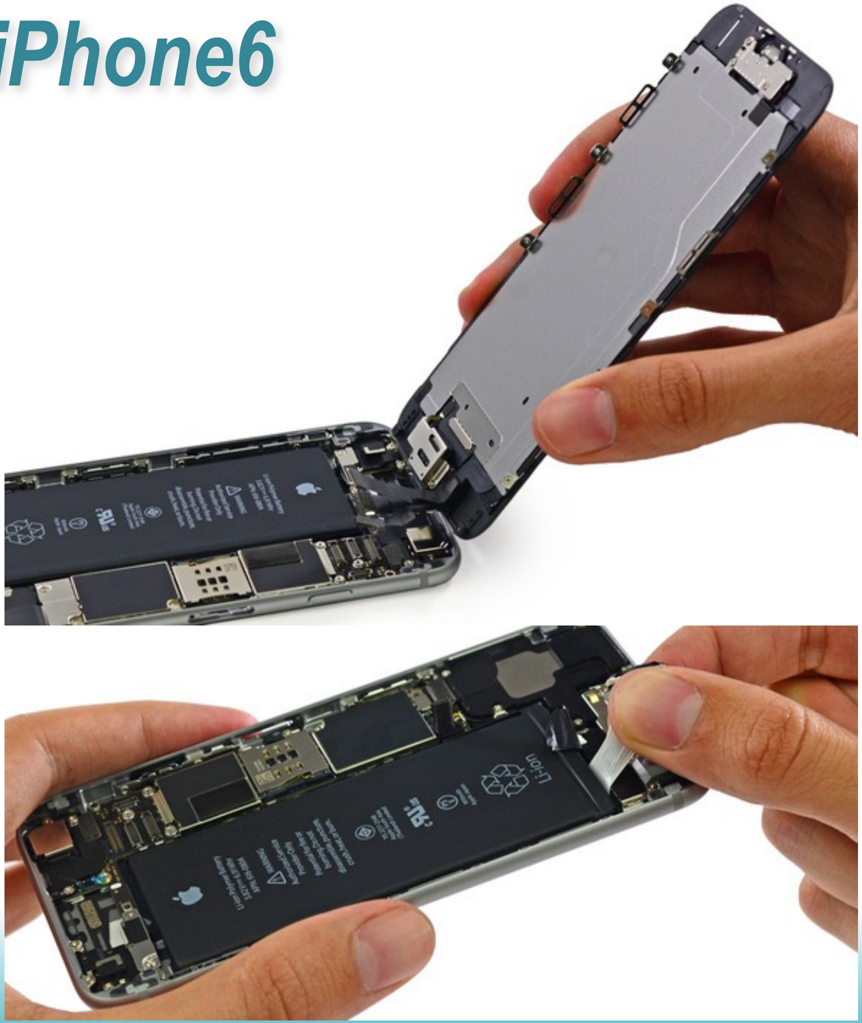
Battery life very
usage dependent.

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

1.76 ounces –
4X the weight of
iPod Shuffle

Logic Board

iPhone6



4.7 inch iPhone6:
1,810mAh battery
@3.8V = 6.88 Wh

F5D
4273
1J9H
FW5T4X

MFG
07/
2014

美国苹果公司 锂电池/二次鋰電池組
제품명: 전지(리튬폴리머전지)
执行标准: GB/T 18287-2013
警告:请勿拆解,刺破,挤压,加热或投
入火中/警告:請勿拆解,刺破,擠壓,
加熱或投入火中 1ICP4/39/96
充电限制电压 4.35V ⎓
额定电压和容量/額定電壓和
容量/정격: 3.82V ⎓ 1810mAh
工厂:惠州市德赛电池有限公司
製造商:惠州市德賽電池有限公司
제조업체: Huizhou Desay
Battery Co., Ltd. 제조국: 중국
中国制造/Assembled in China
型号/型號/모델명: 616-0805
전화: 080-333-4000

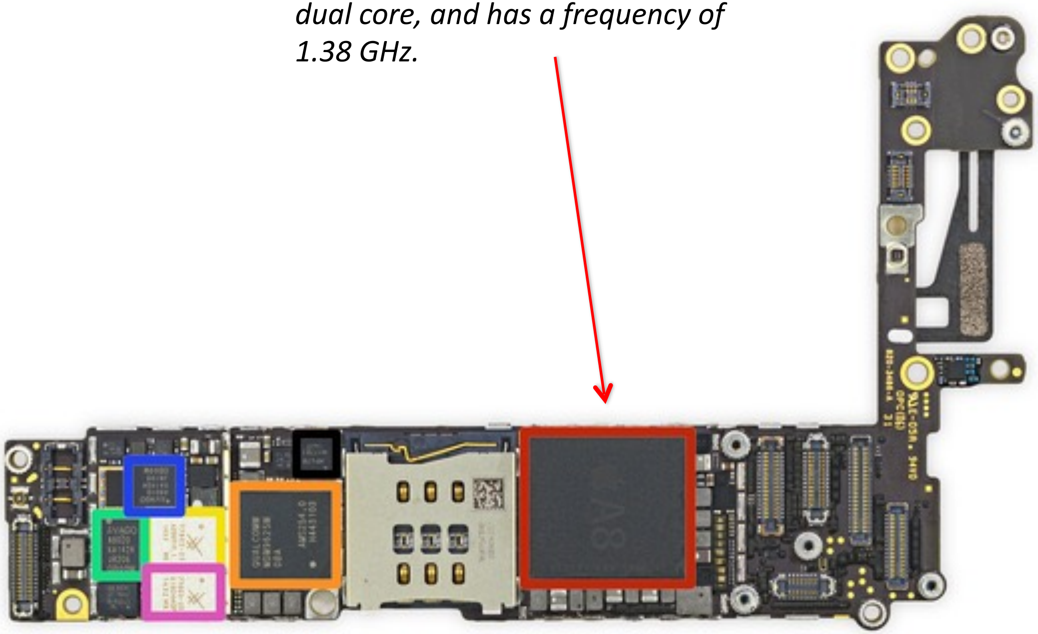
R35393

YU12246-
14002

iPhone 5s:
1570mAh
@3.8V = 6 Wh

- The front side of the logic board:
  - 🔴 Apple A8 APL1011 SoC + SK Hynix RAM as denoted by the markings H9CKNNN8KTMRWR-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

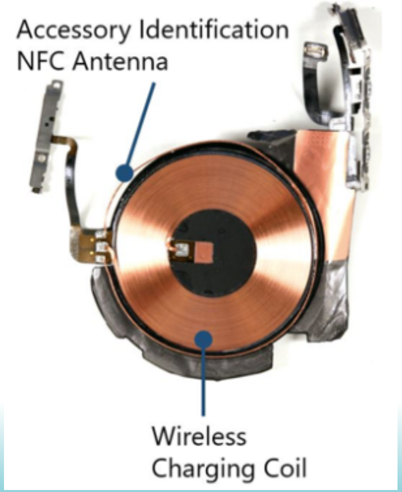
The A8 is manufactured on a 20 nm process by TSMC. It contains 2 billion transistors. Its physical size is 89 mm^2.] It has 1 GB of LPDDR3 RAM included in the package. It is dual core, and has a frequency of 1.38 GHz.



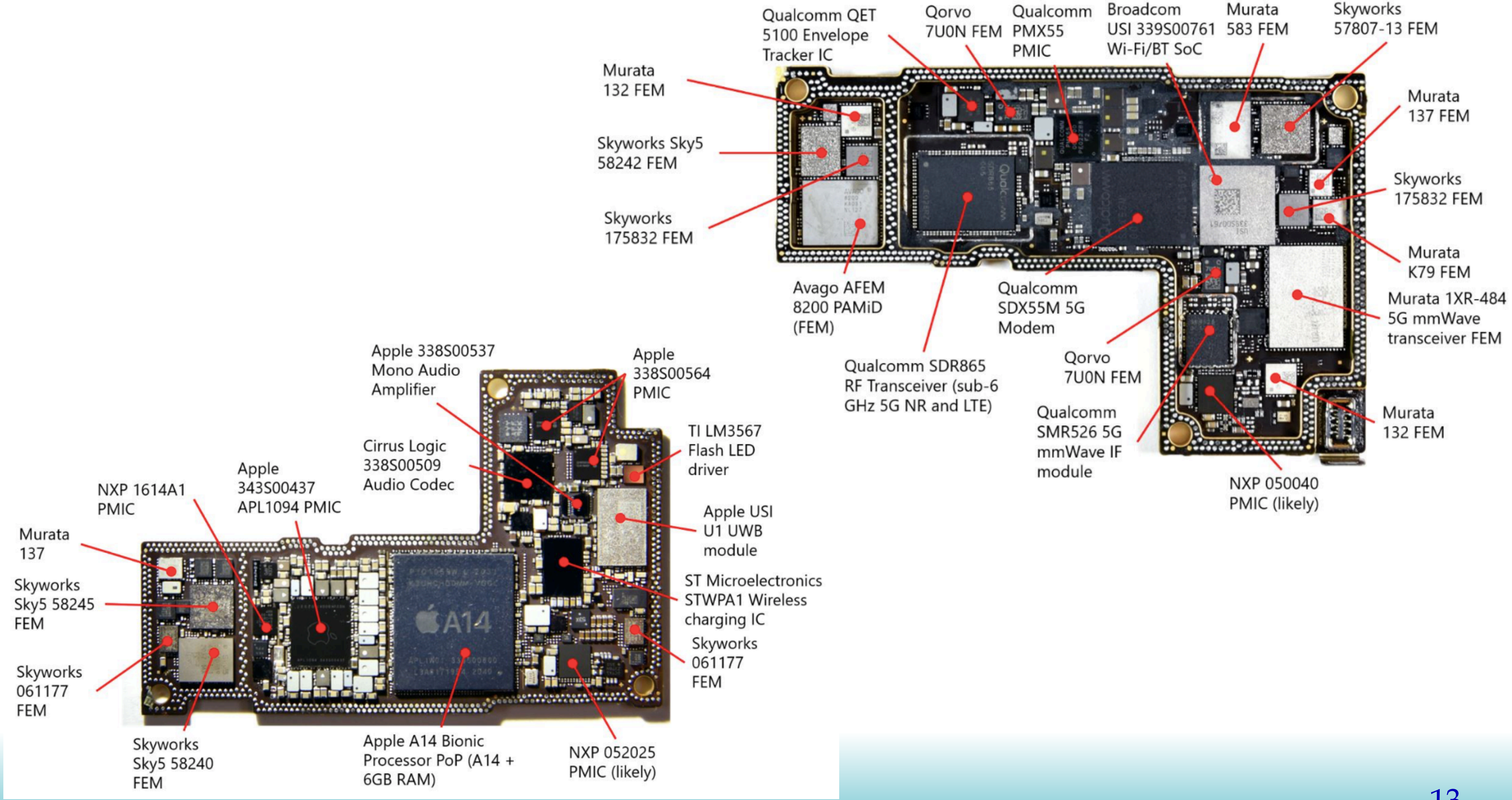
Iphone 12:

https://unitedlex.com/insights/apple-iphone-12-pro-max-teardown-report

iPhone Model	Battery Capacity
iPhone 12 Mini	2,227 mAh
iPhone 12	2,815 mAh
iPhone 12 Pro	2,815 mAh
iPhone 12 Pro Max	3,687 mAh
iPhone 11	3,110 mAh
iPhone 11 Pro	3,046 mAh
iPhone 11 Pro Max	3,969 mAh



14.13 Wh @ 3.8V (Pro Max)


Qualcomm QET 5100 Envelope Tracker IC
Qorvo 7U0N FEM
Qualcomm PMX55 PMIC
Broadcom USI 339S00761 Wi-Fi/BT SoC
Murata 583 FEM
Skyworks 57807-13 FEM
Murata 132 FEM
Murata 137 FEM
Skyworks Sky5 58242 FEM
Skyworks 175832 FEM
Skyworks 175832 FEM
Murata K79 FEM
Avago AFEM 8200 PAMiD (FEM)
Qualcomm SDX55M 5G Modem
Murata 1XR-484 5G mmWave transceiver FEM
Qualcomm SDR865 RF Transceiver (sub-6 GHz 5G NR and LTE)
Qorvo 7U0N FEM
Qualcomm SMR526 5G mmWave IF module
Murata 132 FEM
NXP 050040 PMIC (likely)
Apple 338S00537 Mono Audio Amplifier
Apple 338S00564 PMIC
TI LM3567 Flash LED driver
Cirrus Logic 338S00509 Audio Codec
Apple 343S00437 APL1094 PMIC
NXP 1614A1 PMIC
Apple USI U1 UWB module
Murata 137
Skyworks Sky5 58245 FEM
ST Microelectronics STWPA1 Wireless charging IC
Skyworks 061177 FEM
Skyworks 061177 FEM
Skyworks Sky5 58240 FEM
Apple A14 Bionic Processor PoP (A14 + 6GB RAM)
NXP 052025 PMIC (likely)

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 ...

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

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!

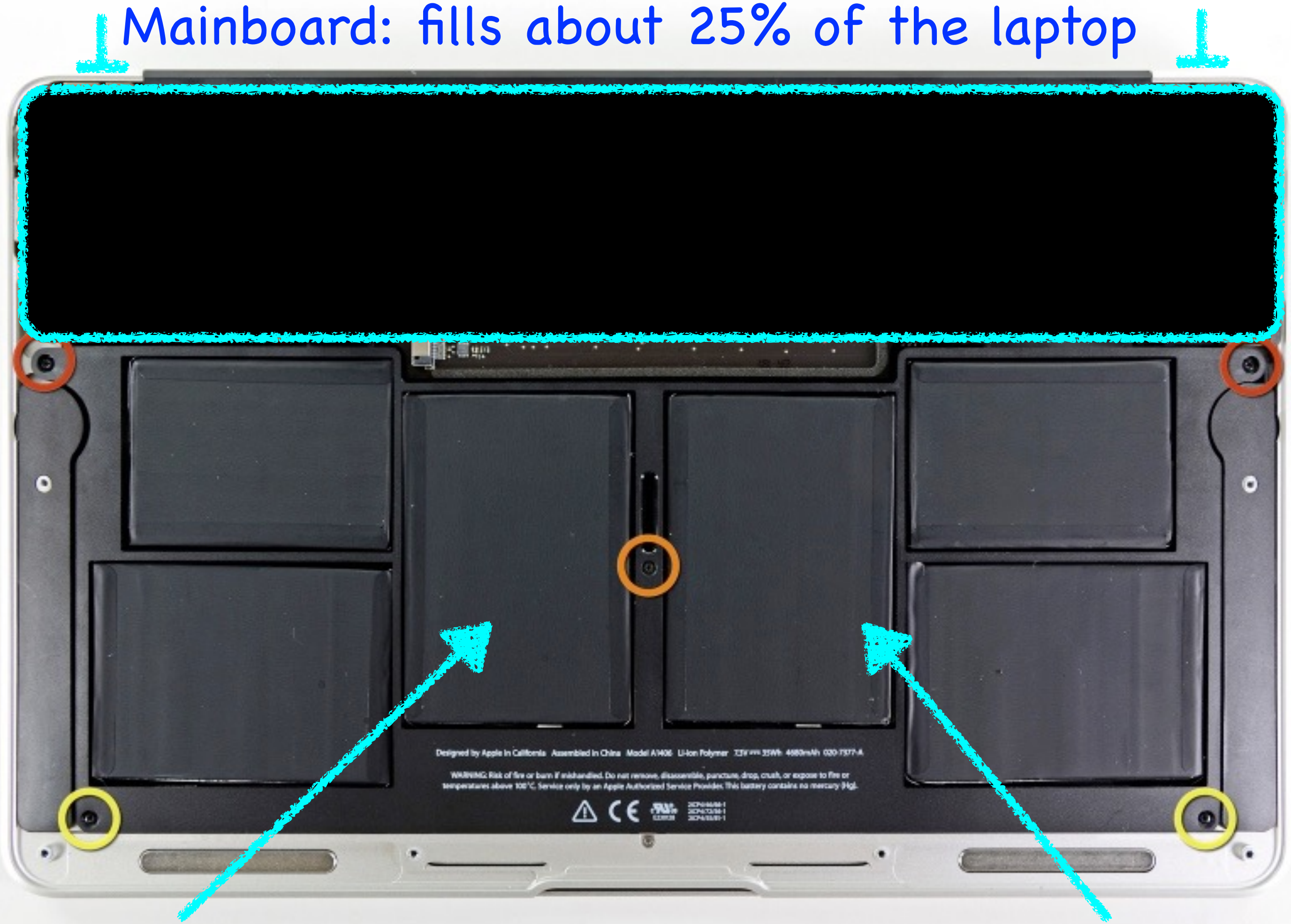
Almost full 1 inch depth. Width and height set by available space, weight.

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

50Wh is 180,000 Joules!

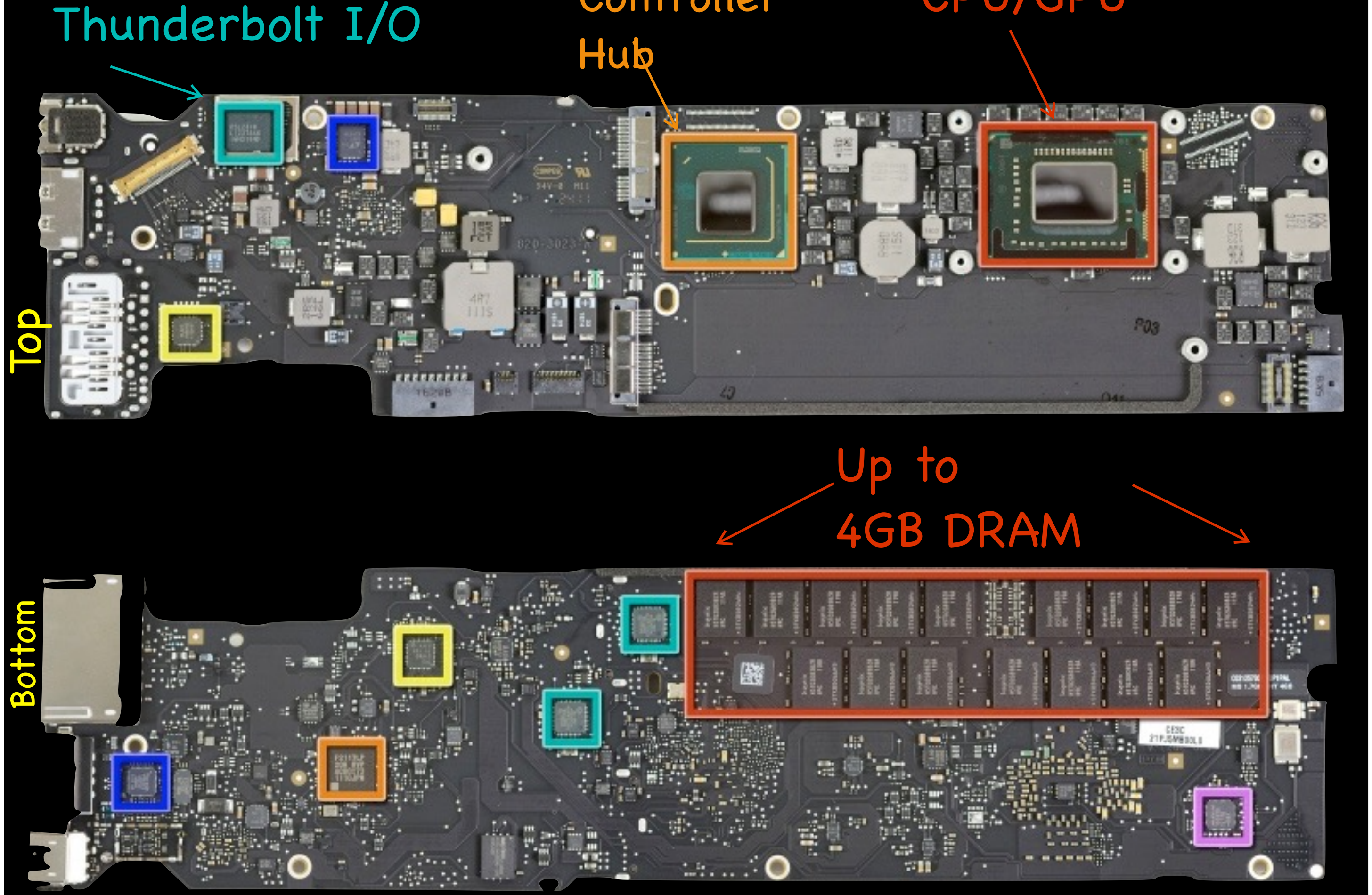
MacBook Air … design the laptop like an iPod/iPhone

Mainboard: fills about 25% of the laptop



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

MacBook Air: Full PC



Servers: Total Cost of Ownership (TCO)

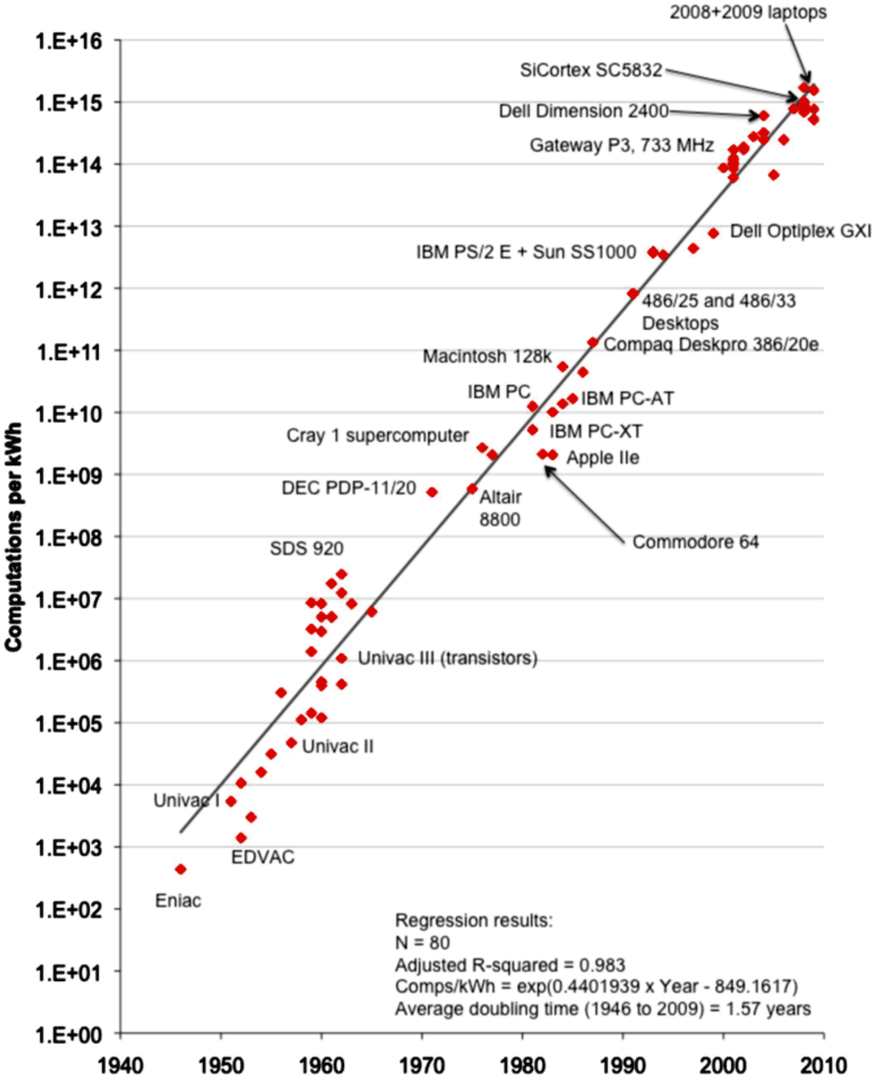
Machine rooms are expensive. Removing heat dictates how many servers to put in a machine room.

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).

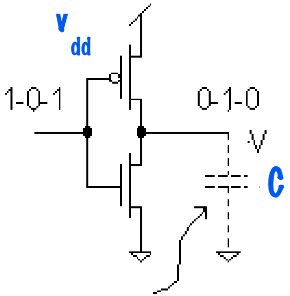


Regression results:
N = 80
Adjusted R-squared = 0.983
Comps/kWh = exp(0.4401939 x Year - 849.1617)
Average doubling time (1946 to 2009) = 1.57 years

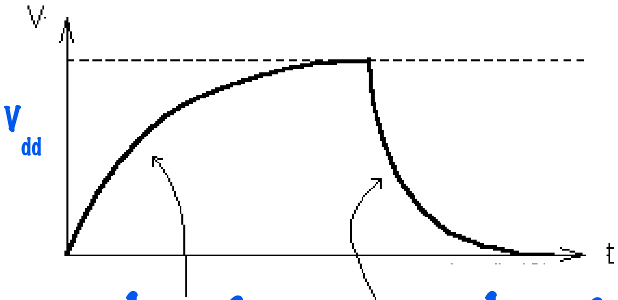
CMOS Circuits and Energy

Switching Energy: Fundamental Physics

Every logic transition dissipates energy.



Models inputs to other gates & wire capacitance



$$E_{0\text{->}1} = \frac{1}{2} C V_{dd}^{2} \qquad E_{1\text{->}0} = \frac{1}{2} C V_{dd}^{2}$$

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.

Chip-Level “Dynamic” Power

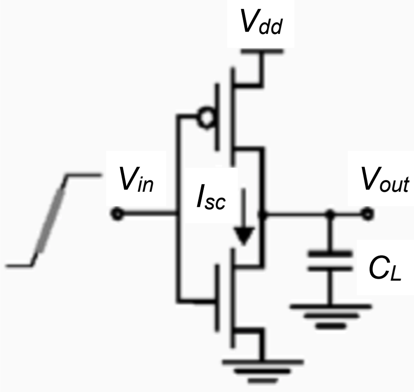
$$P_{sw} = 1/2\ \alpha\ C\ V_{dd}^{2}\ F$$

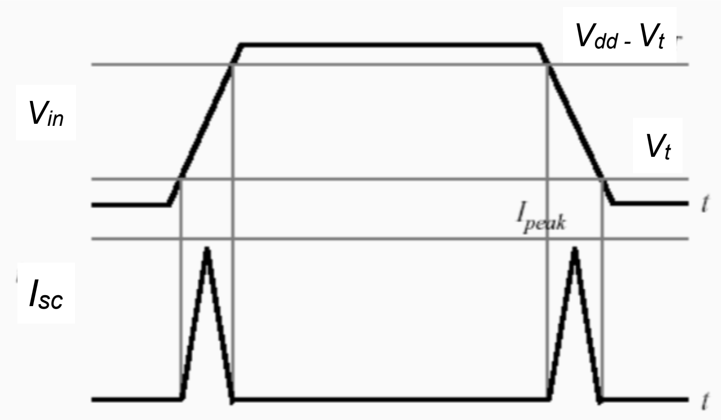
“activity factor”, average percentage of capacitance switching per cycle (~ number of nodes to switch)

Total chip capacitance to be switched

Clock Frequency

Additional Dynamic Power - “short circuit current”



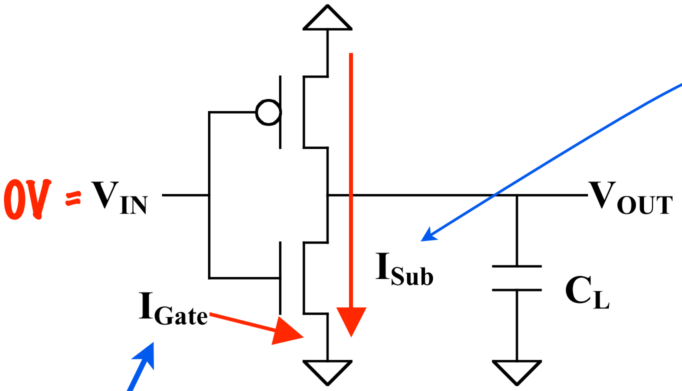


When gate switches, brief period when both pullup network and pulldown network could be on.

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.



Isub: Even when this nFet is off, it passes an Ioff leakage current.

We can engineer any Ioff we like, but a lower Ioff also results in a lower Ion, and thus a lower maximum clock speed.

Igate: 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.

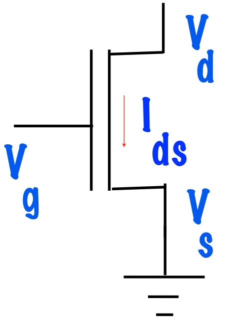
Bill Holt, Intel, Hot Chips 17.

Plot on a “Log” Scale to See “Off” Current

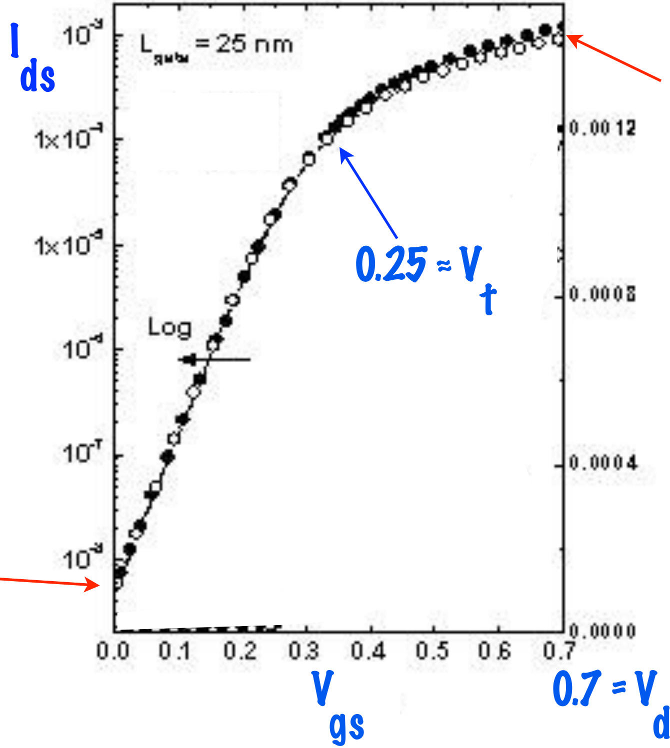
Process engineers can:

increase I_{on} by lowering V_t - but that raises I_{off}

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



$V_{ds} = V_{dd}$



$1.2\ mA = I_{on}$

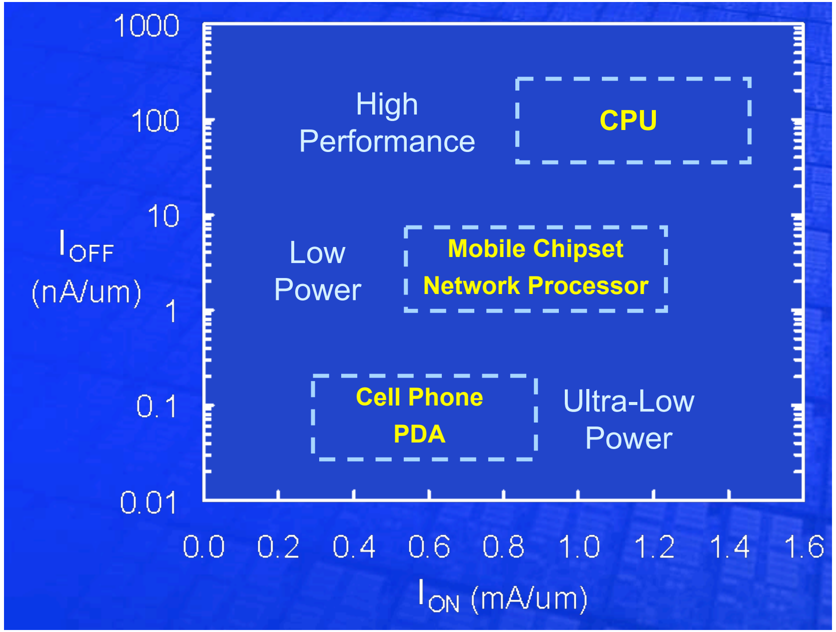
$0.25 \approx V_t$

$V_{ds} = V_{dd}$

$I_{off} \approx 10\ nA$

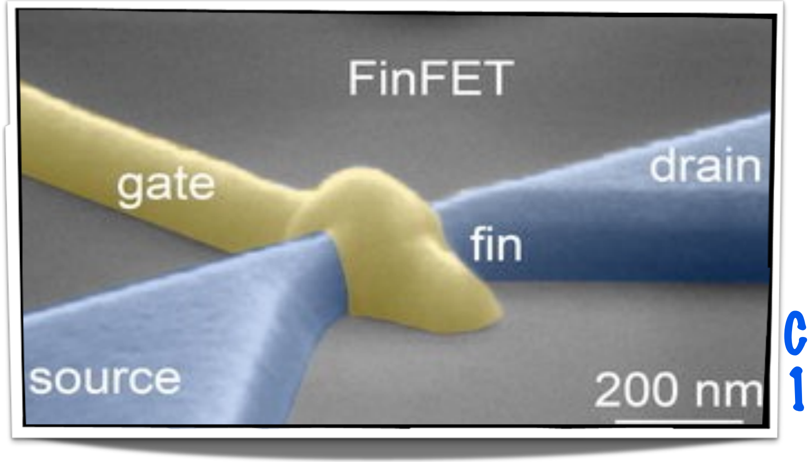
$0.7 = V_{dd}$

Customize processing for product types and different circuit paths …



From: "Facing the Hot Chips Challenge Again", Bill Holt, Intel, presented at Hot Chips 17, 2005.

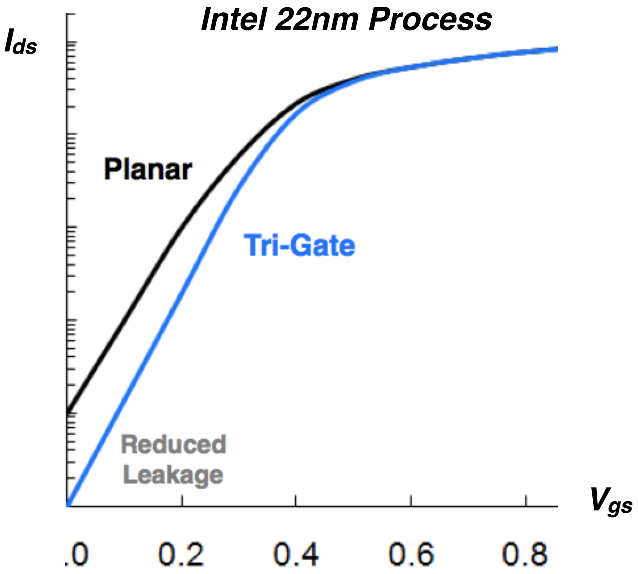
- *Vt is controlled by channel doping.*
- *Modern IC processes have 2 or 3 different Vt values available.*
- *Standard cell libraries offer low Vt and high Vt versions of cells so that the tools can optimize on a per instance basis.*
- *(If high performance not needed then use high Vt to reduce leakage).*

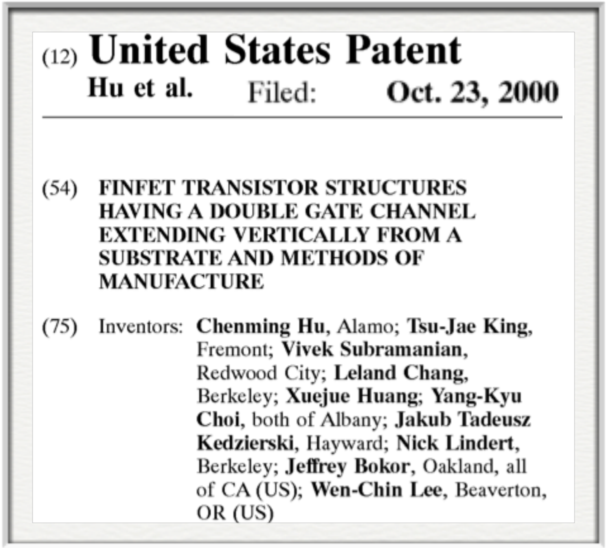


Transistor channel is a raised fin.

Gate controls channel from sides and top.

Channel depth is fin width. 12-15nm for L=22nm.

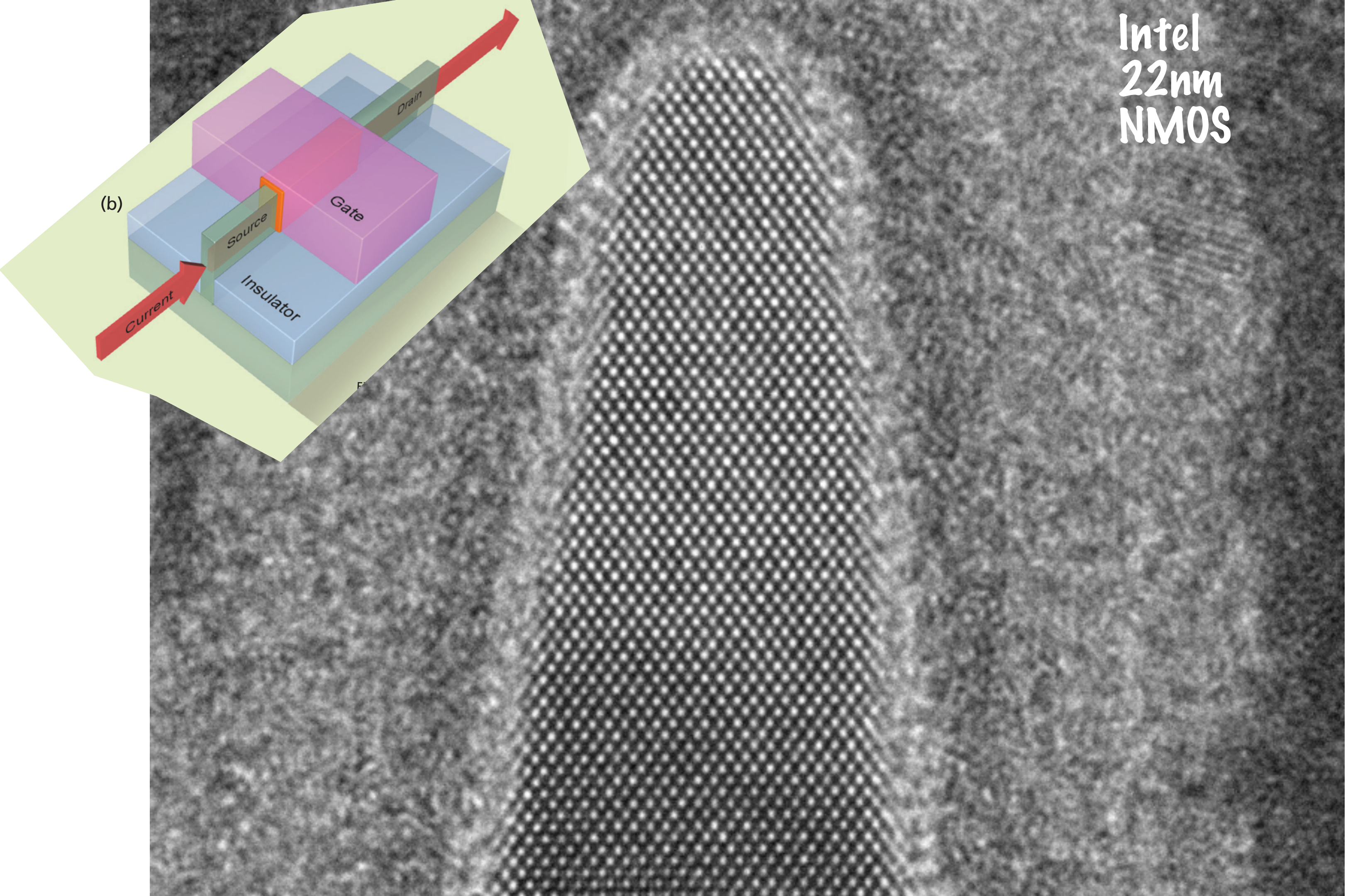



(12) **United States Patent**

Hu et al. Filed: **Oct. 23, 2000**

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

(75) 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
22nm
NMOS

Dynamic versus Leakage Power

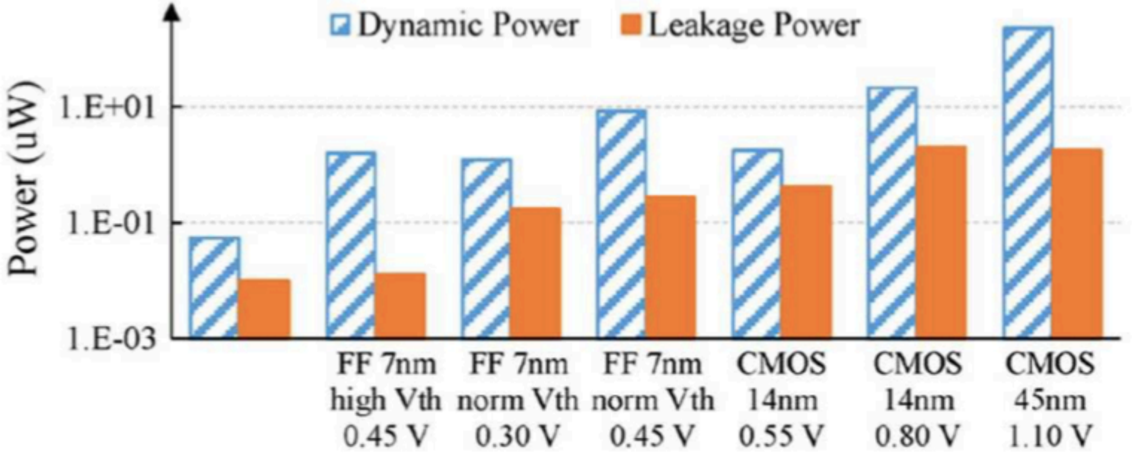
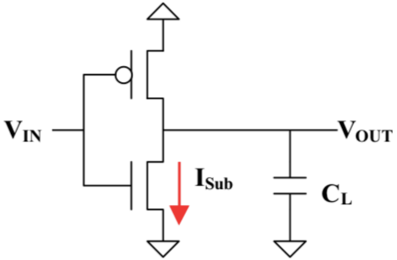


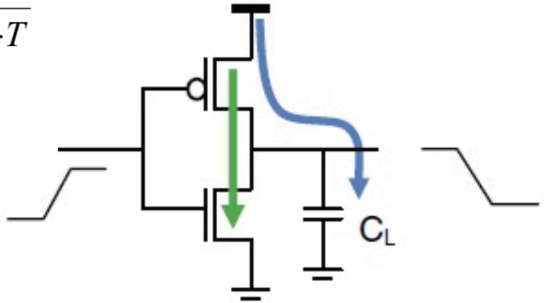
Figure 1: The reduction of feature sizes from 45 to 7nm may induce drastic gains in power consumption and leakage power [Xie2015]

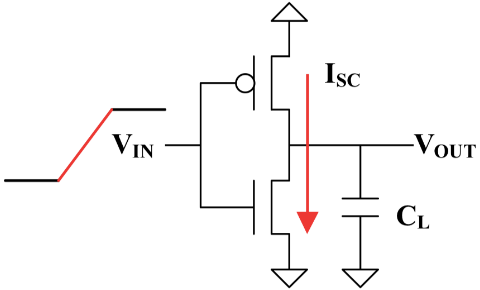
Xie, Q. (2015). Performance Comparisons between 7-nm FinFET and Conventional Bulk CMOS Standard Cell Libraries. IEEE Transactions on Circuits and Systems II: Express Briefs, 62(8), 761-765.

Total Power = $P_{switching} + P_{short\text{-}circuit} + P_{leakage}$



$$I_{DSub} = k \cdot e^{\frac{-q \cdot V_T}{a \cdot k_a \cdot T}}$$





Some low-power design techniques

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

Design Technique #1

Trading Hardware for Power

via Parallelism and Pipelining ...

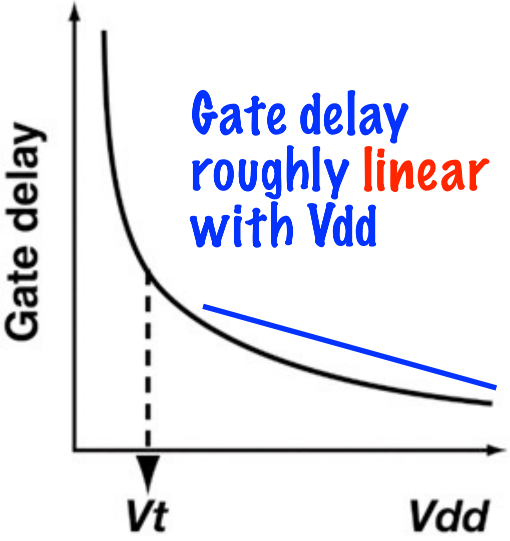
Voltage Scaling

$$P_{sw} = 1/2\ \alpha\ C\ V_{dd}^{2}\ F$$

Reducing F, reduces power, but our computation now takes longer, and total energy does not change.

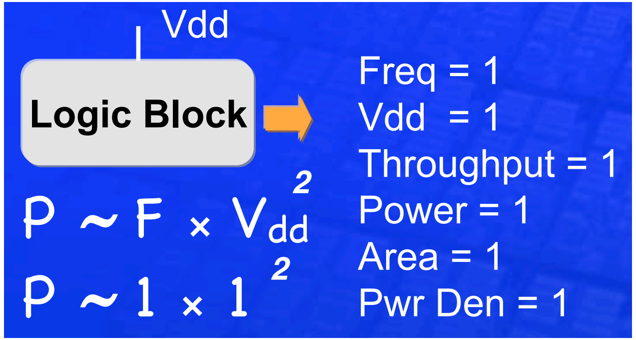
Reducing both F and Vdd, reduces power but also improves energy efficiency (total energy for computation is less).

Parallelism gives us a way to make up for lower performance from voltage scaling.



Gate delay roughly **linear** with Vdd

And so, we can transform this:



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

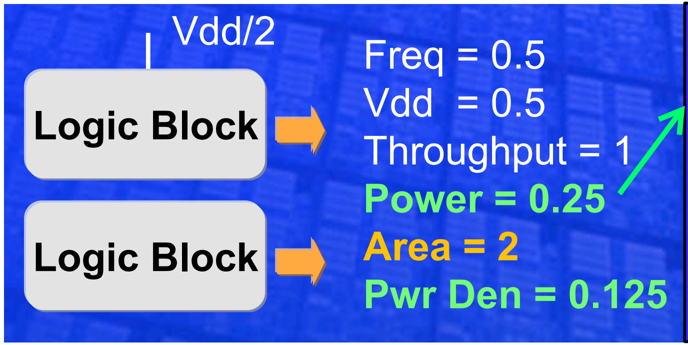
$P \sim 1 \times 1^{2}$

Freq = 1
Vdd = 1
Throughput = 1
Power = 1
Area = 1
Pwr Den = 1

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

Into this:

Top block processes "left", bottom "right".



Freq = 0.5
Vdd = 0.5
Throughput = 1
Power = 0.25
Area = 2
Pwr Den = 0.125

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

$P \sim 2 \times 1/2 \times 1/4 = 1/4$

CV^2 power only

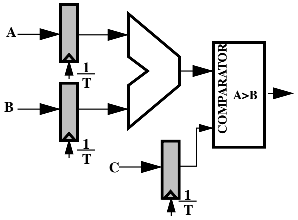
THIS MAGIC TRICK BROUGHT TO YOU BY CORY HALL ...

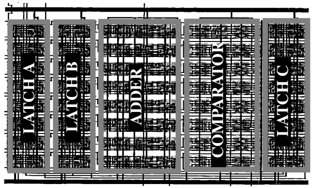
Chandrakasan & Brodersen (UCB, 1992)

Architecture	Power (normalized)
Simple	1
Parallel	0.36
Pipelined	0.39
Pipelined-Parallel	0.2

Architecture	Area (normalized)
Simple	1
Parallel	3.4
Pipelined	1.3
Pipelined-Parallel	3.7

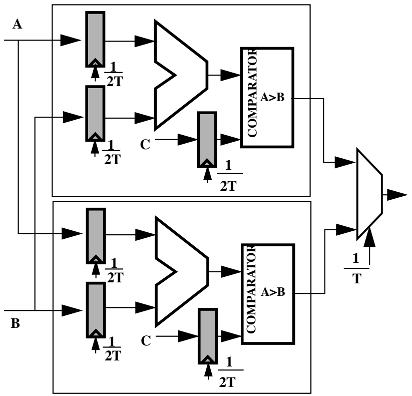
Architecture	Voltage
Simple	5V
Parallel	2.9V
Pipelined	2.9V
Pipelined-Parallel	2.0

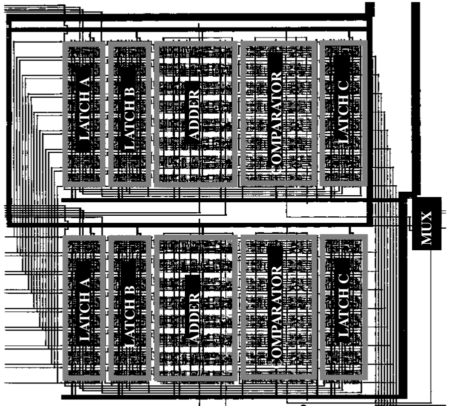




Area = 636 x 833 μ^2

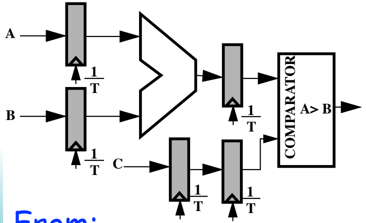
Simple

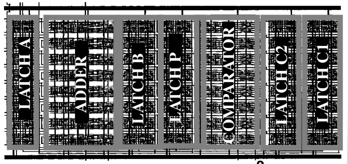




Area = 1476 x 1219 μ^2

Parallel





Area = 640 x 1081 μ^2

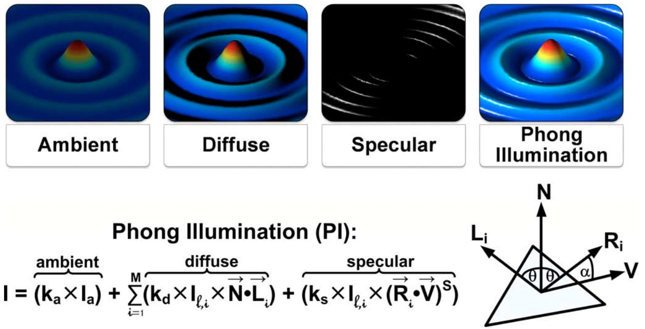
Pipelined

From:

Minimizing Power Consumption in CMOS Circuits

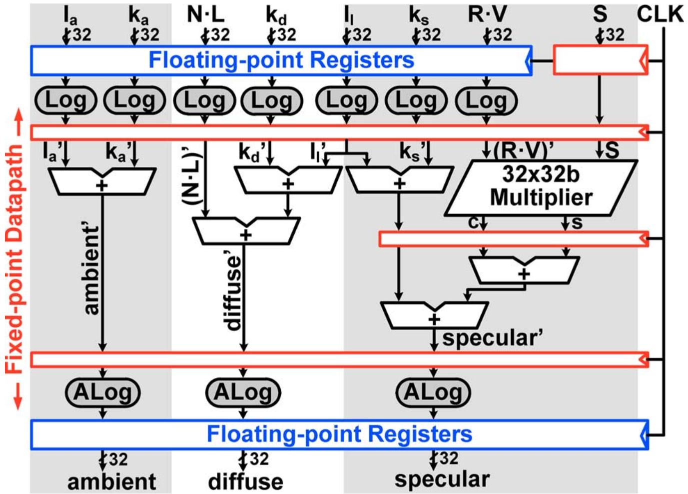
Anantha P. Chandrakasan
Robert W. Brodersen

Example: Intel Graphics Pipeline IP

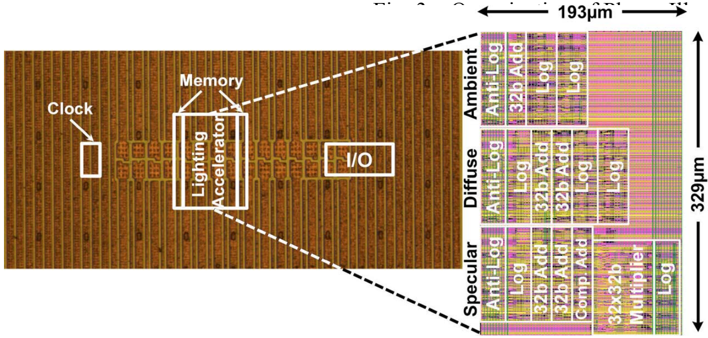


$$I = \underbrace{(k_a \times I_a)}_{\text{ambient}} + \sum_{i=1}^{M} \underbrace{(k_d \times I_{\ell,i} \times \vec{N} \cdot \vec{L}_i)}_{\text{diffuse}} + \underbrace{(k_s \times I_{\ell,i} \times (\vec{R}_i \cdot \vec{V})^S)}_{\text{specular}}$$

Fig. 1. Phong Illumination for vertex and pixel shading.



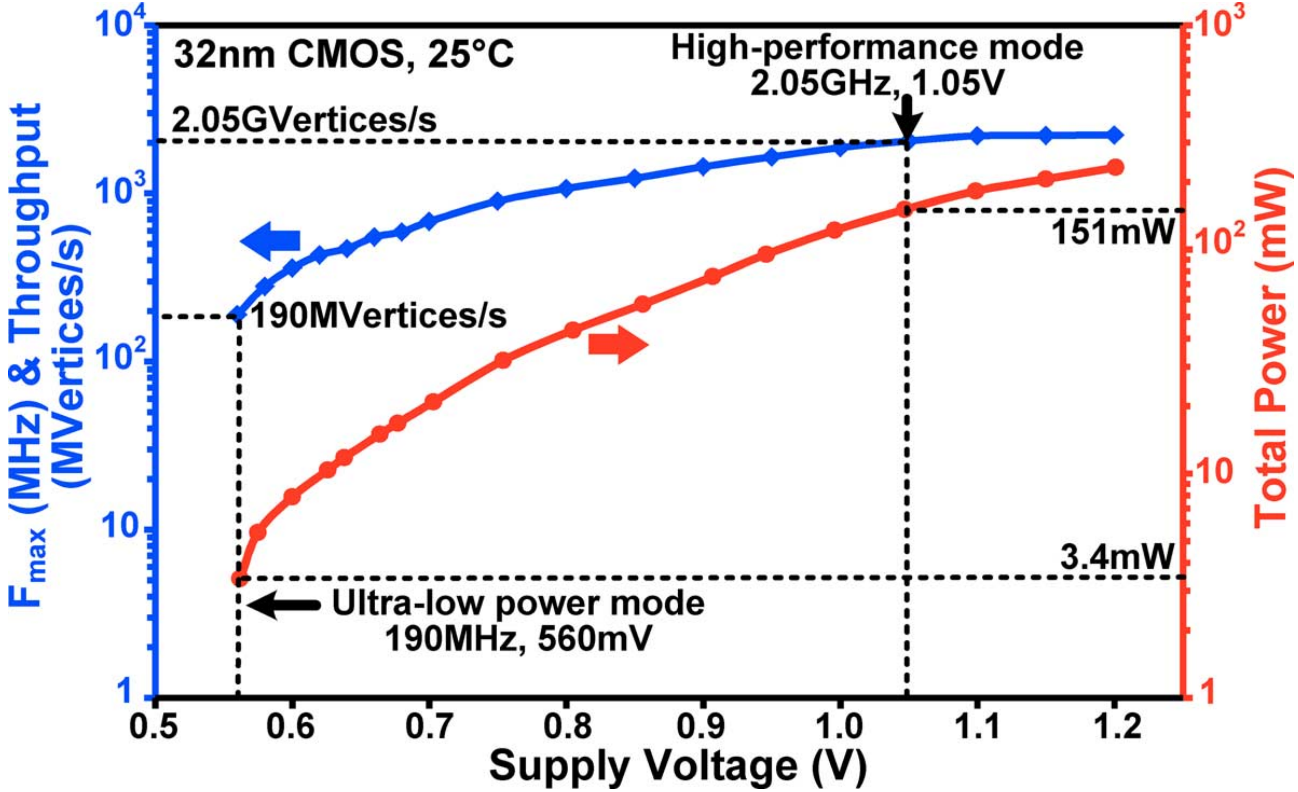
nination lighting accelerator.



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*

Clock Rate and Power vs Voltage



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

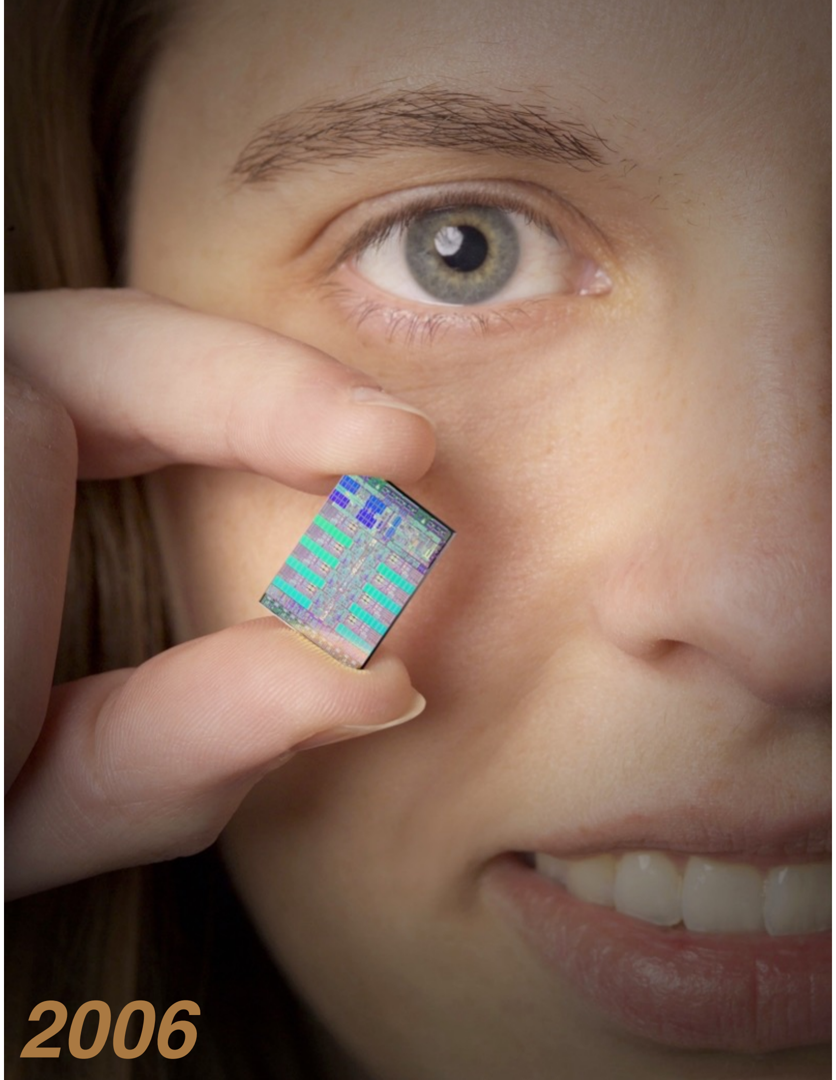
Trade hardware for power, on a large scale ...

Cell:
The PS3 chip

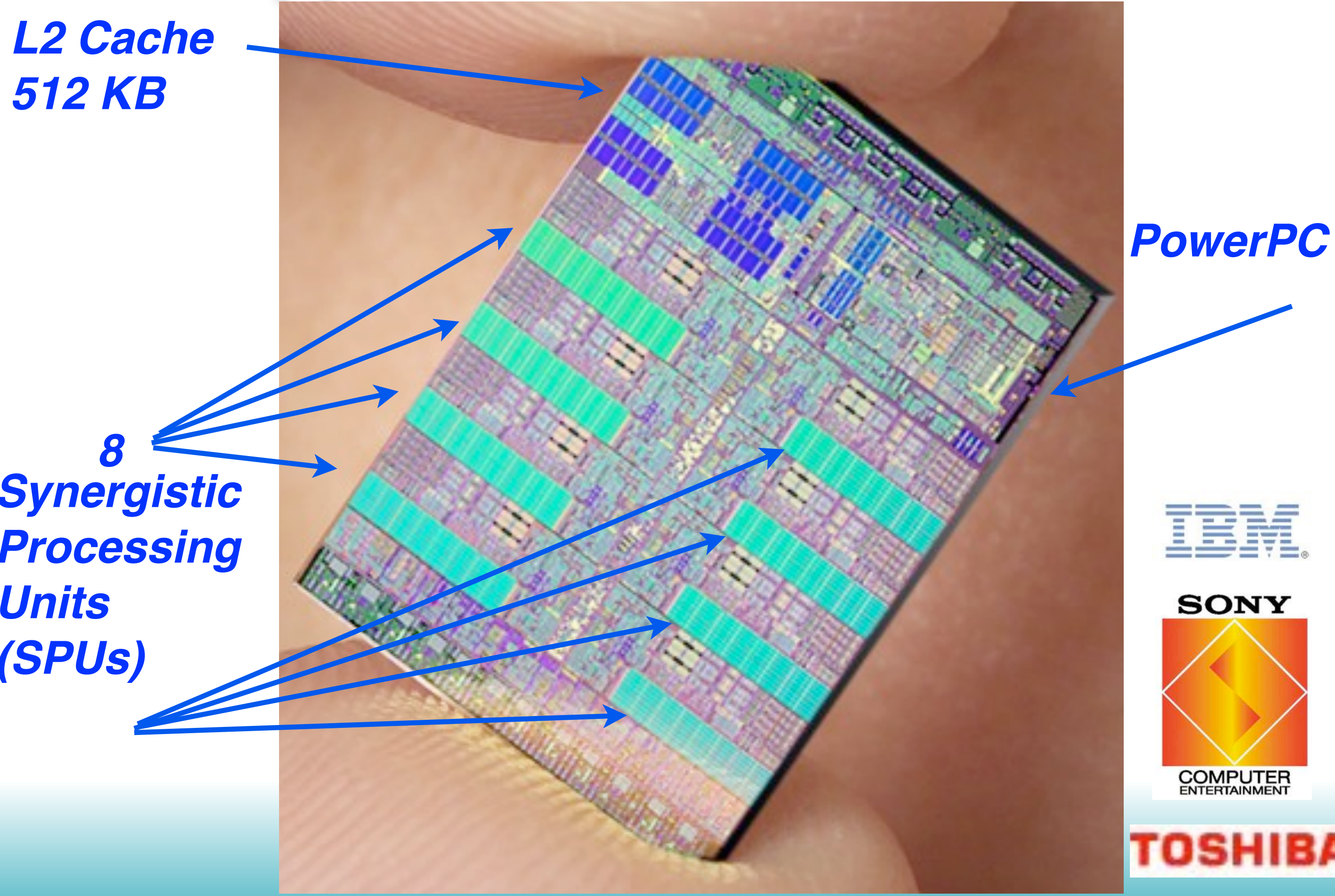
IBM

SONY
COMPUTER
ENTERTAINMENT

TOSHIBA


2006

Cell (PS3 Chip): 1 CPU + 8 "SPUs"



IBM

SONY COMPUTER ENTERTAINMENT

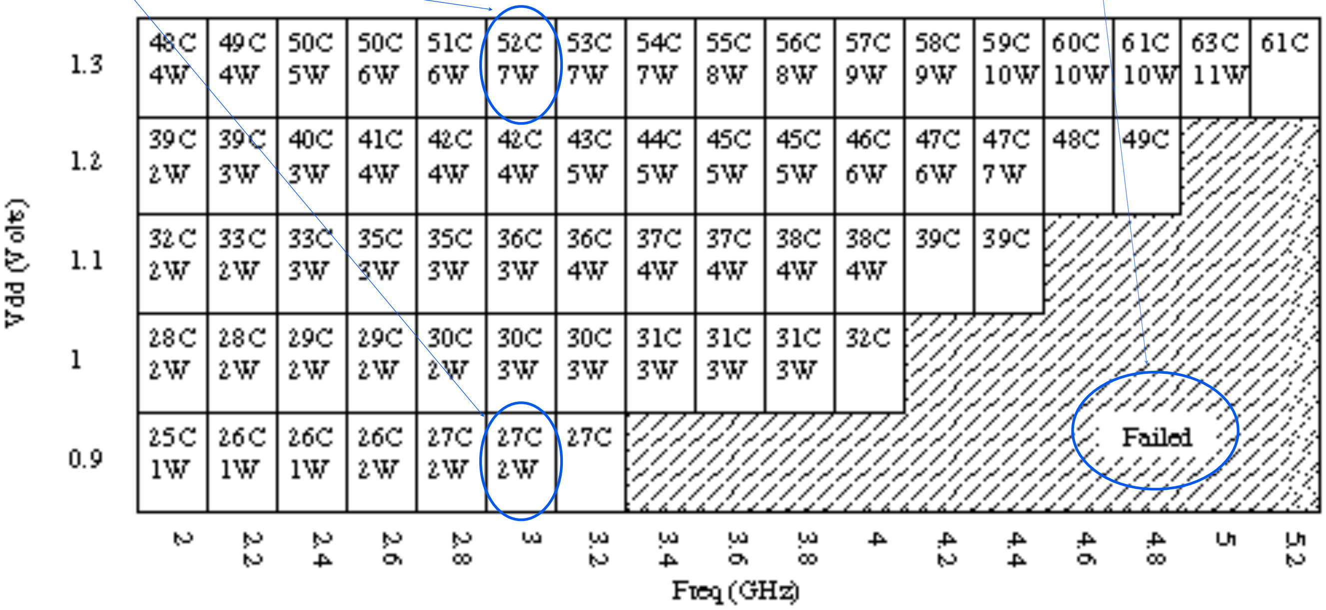
TOSHIBA

A “Schmoo” plot for a Cell SPU ...

The lower Vdd, the less dynamic energy consumption.

$$E_{0\rightarrow 1} = \frac{1}{2} C V_{dd}^{2} \qquad E_{1\rightarrow 0} = \frac{1}{2} C V_{dd}^{2}$$

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



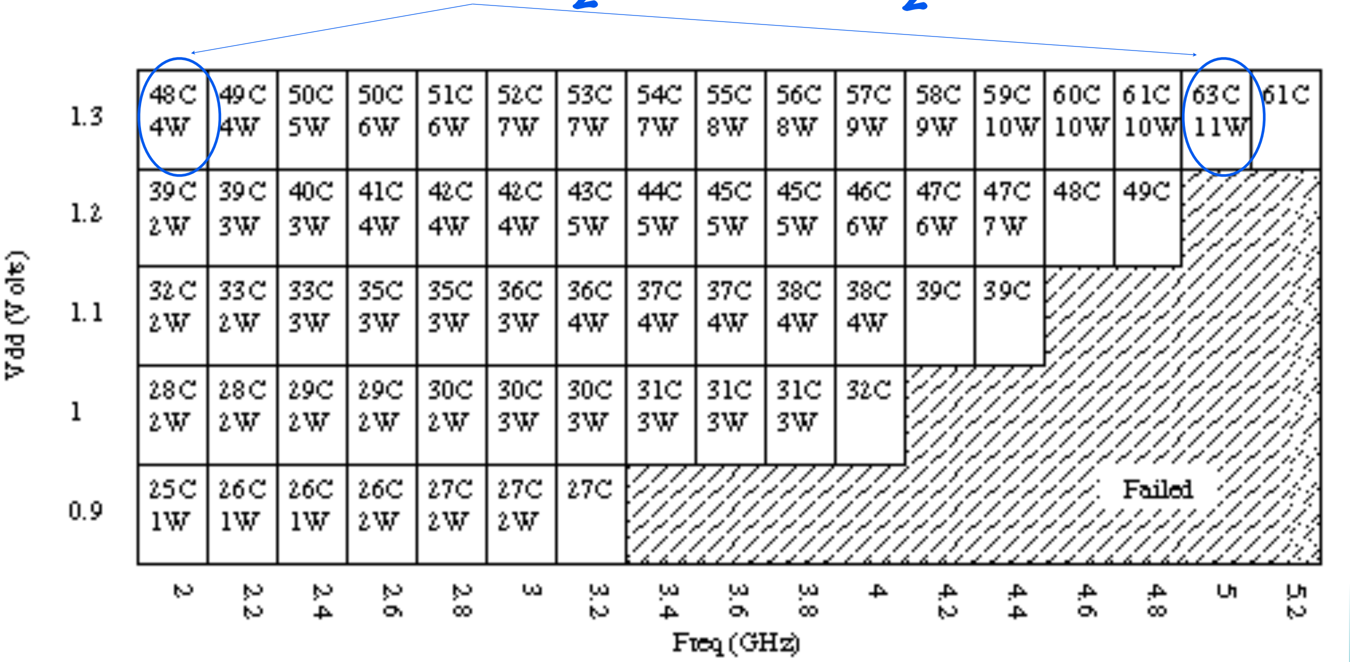
Vdd (Volts)	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2
1.3	48C 4W	49C 4W	50C 5W	50C 6W	51C 6W	52C 7W	53C 7W	54C 7W	55C 8W	56C 8W	57C 9W	58C 9W	59C 10W	60C 10W	61C 10W	63C 11W	61C
1.2	39C 2W	39C 3W	40C 3W	41C 4W	42C 4W	42C 4W	43C 5W	44C 5W	45C 5W	45C 5W	46C 6W	47C 6W	47C 7W	48C	49C		
1.1	32C 2W	33C 2W	33C 3W	35C 3W	35C 3W	36C 3W	36C 4W	37C 4W	37C 4W	38C 4W	38C 4W	39C	39C				
1	28C 2W	28C 2W	29C 2W	29C 2W	30C 2W	30C 3W	30C 3W	31C 3W	31C 3W	31C 3W	32C						
0.9	25C 1W	26C 1W	26C 1W	26C 2W	27C 2W	27C 2W	27C								Failed		

Freq (GHz)

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 ...

$$E_{0\text{->}1} = \frac{1}{2} C V_{dd}^2 \qquad E_{1\text{->}0} = \frac{1}{2} C V_{dd}^2$$



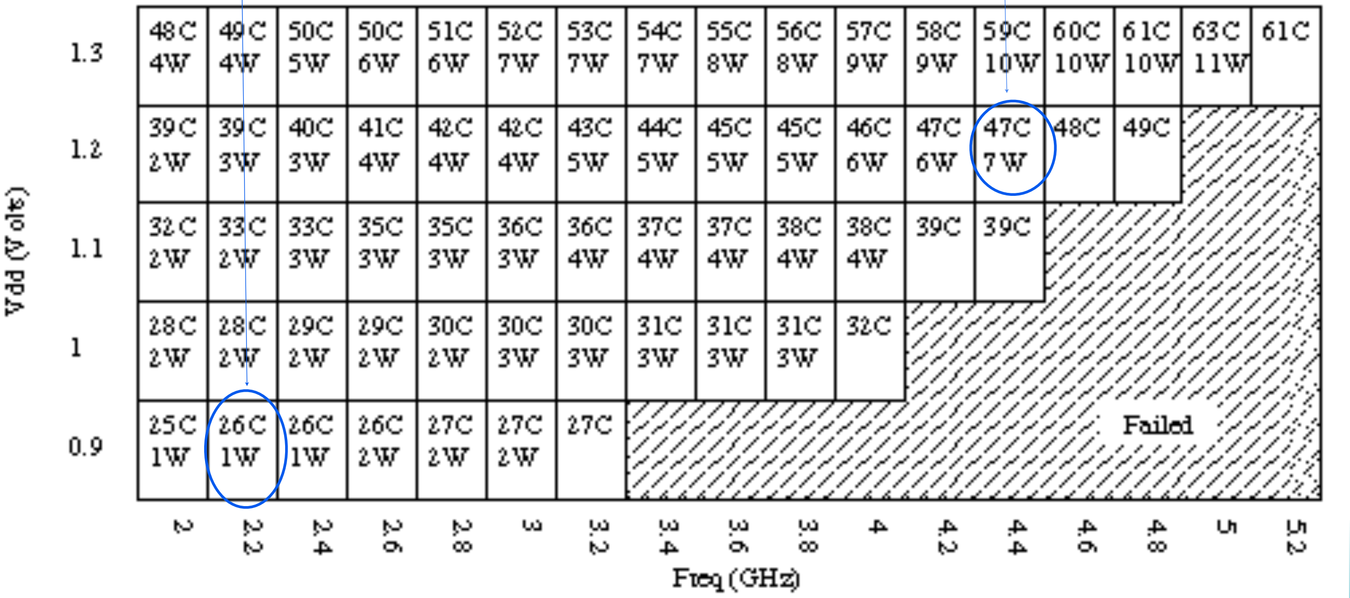
Vdd (Volts) \ Freq (GHz)	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2
1.3	48C 4W	49C 4W	50C 5W	50C 6W	51C 6W	52C 7W	53C 7W	54C 7W	55C 8W	56C 8W	57C 9W	58C 9W	59C 10W	60C 10W	61C 10W	63C 11W	61C
1.2	39C 2W	39C 3W	40C 3W	41C 4W	42C 4W	42C 4W	43C 5W	44C 5W	45C 5W	45C 5W	46C 6W	47C 6W	47C 7W	48C	49C	▨	▨
1.1	32C 2W	33C 2W	33C 3W	35C 3W	35C 3W	36C 3W	36C 4W	37C 4W	37C 4W	38C 4W	38C 4W	39C	39C	▨	▨	▨	▨
1	28C 2W	28C 2W	29C 2W	29C 2W	30C 2W	30C 3W	30C 3W	31C 3W	31C 3W	31C 3W	32C	▨	▨	▨	▨	▨	▨
0.9	25C 1W	26C 1W	26C 1W	26C 2W	27C 2W	27C 2W	27C	▨	▨	▨	▨	▨	▨	▨	Failed	▨	▨

Scaling V *and* f does lower energy/op

1 W to get 2.2 GHz performance. 26 C die temp.

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

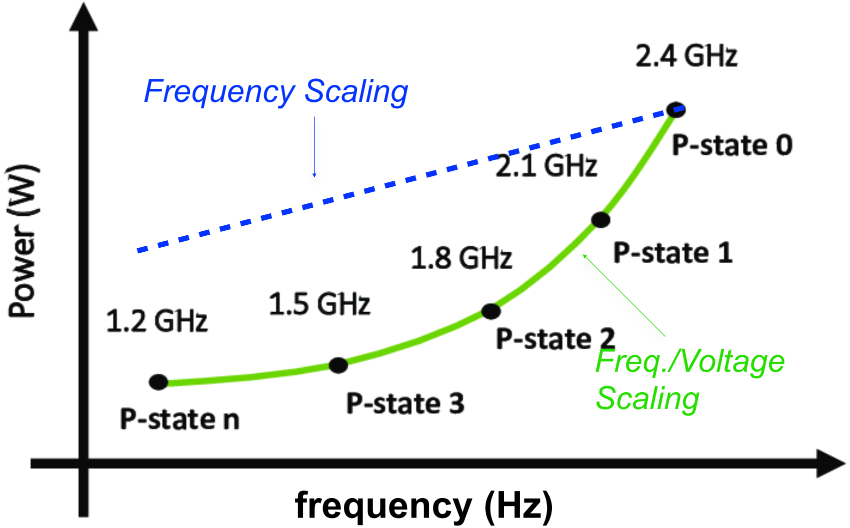
If a program that needs a 4.4 Ghz CPU can be recoded to use two 2.2 Ghz CPUs ... big win.



Vdd (Volts)	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2
1.3	48C 4W	49C 4W	50C 5W	50C 6W	51C 6W	52C 7W	53C 7W	54C 7W	55C 8W	56C 8W	57C 9W	58C 9W	59C 10W	60C 10W	61C 10W	63C 11W	61C
1.2	39C 2W	39C 3W	40C 3W	41C 4W	42C 4W	42C 4W	43C 5W	44C 5W	45C 5W	45C 5W	46C 6W	47C 6W	47C 7W	48C	49C		
1.1	32C 2W	33C 2W	33C 3W	35C 3W	35C 3W	36C 3W	36C 4W	37C 4W	37C 4W	38C 4W	38C 4W	39C	39C				
1	28C 2W	28C 2W	29C 2W	29C 2W	30C 2W	30C 3W	30C 3W	31C 3W	31C 3W	31C 3W	32C						
0.9	25C 1W	26C 1W	26C 1W	26C 2W	27C 2W	27C 2W	27C								Failed		

Freq (GHz)

Dynamic Voltage/Frequency Scaling (DVFS)



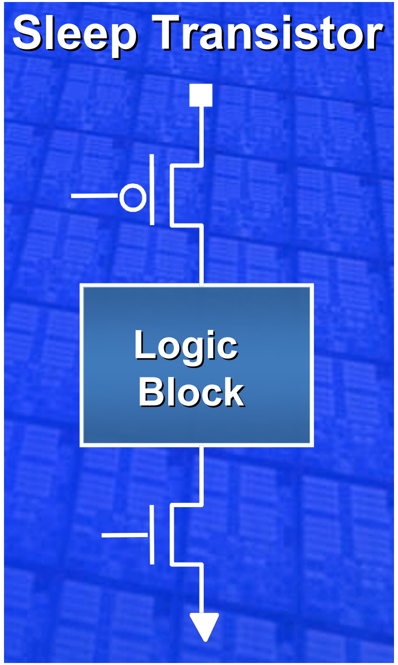
Intel power states

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

- 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 ...



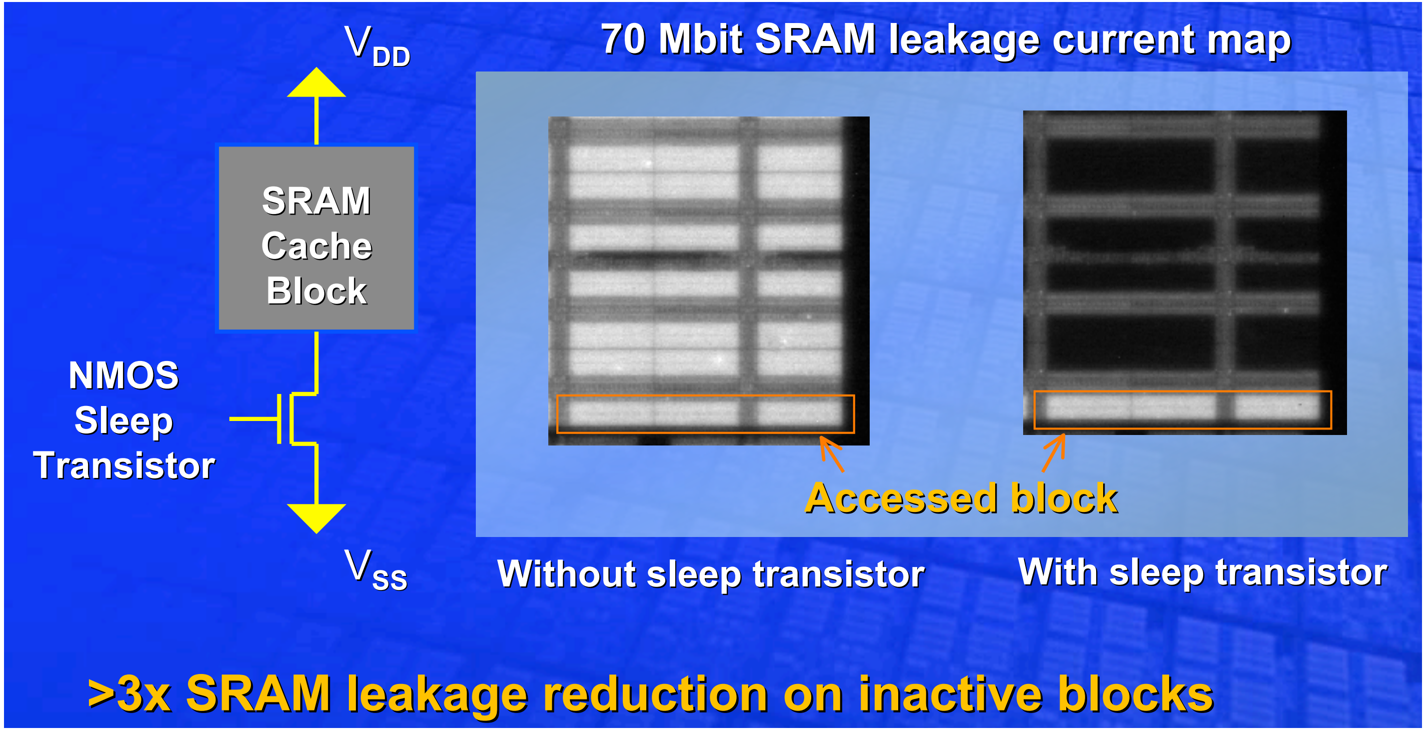
Example: Floating point unit logic.

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



70 Mbit SRAM leakage current map

V_{DD}

SRAM Cache Block

NMOS Sleep Transistor

V_{SS}

Accessed block

Without sleep transistor

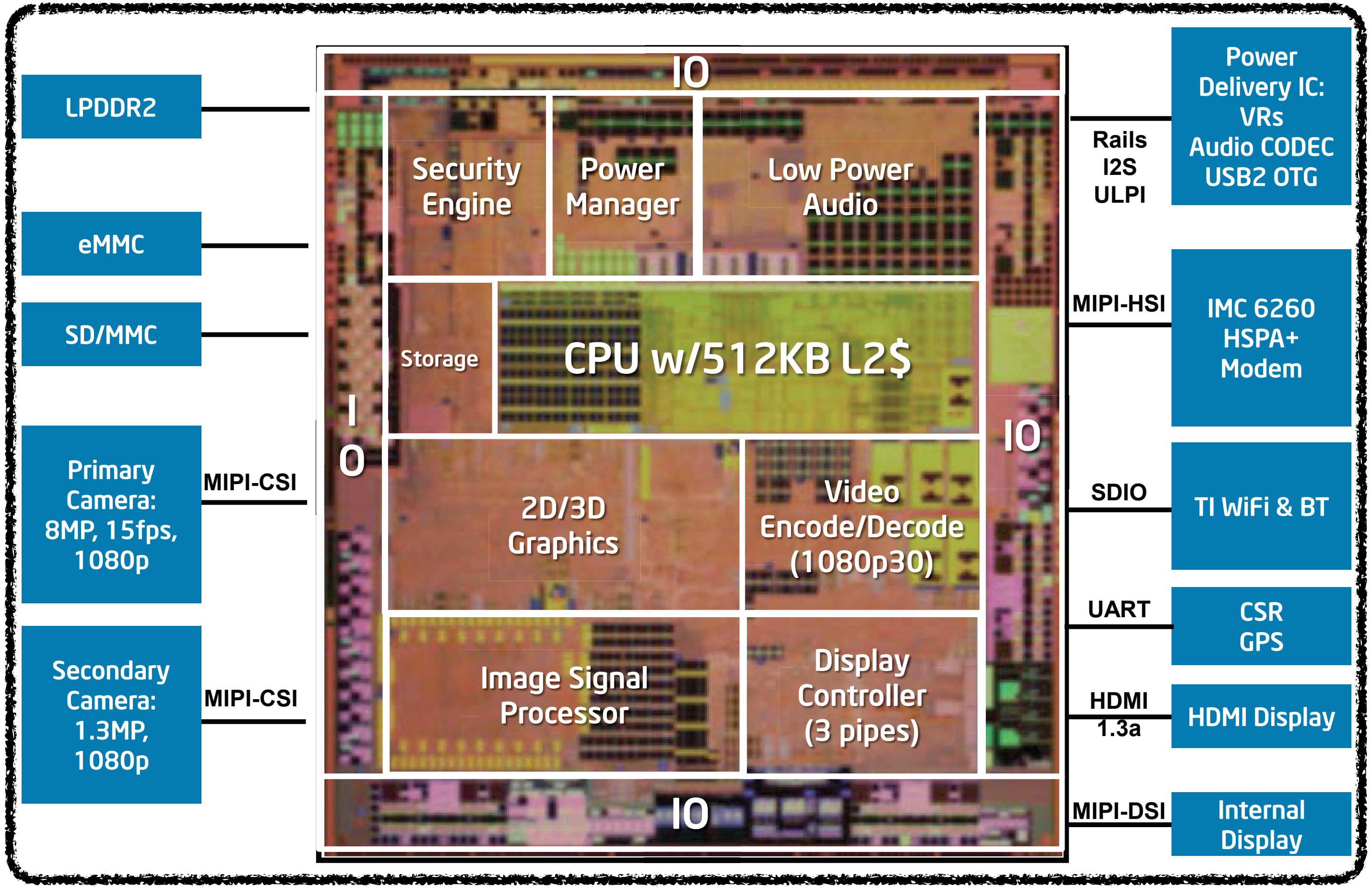
With sleep transistor

>3x SRAM leakage reduction on inactive blocks

A tiny current supplied in "sleep" maintains SRAM state.

From: "Facing the Hot Chips Challenge Again", Bill Holt, Intel, presented at Hot Chips 17, 2005.

Intel Medfield



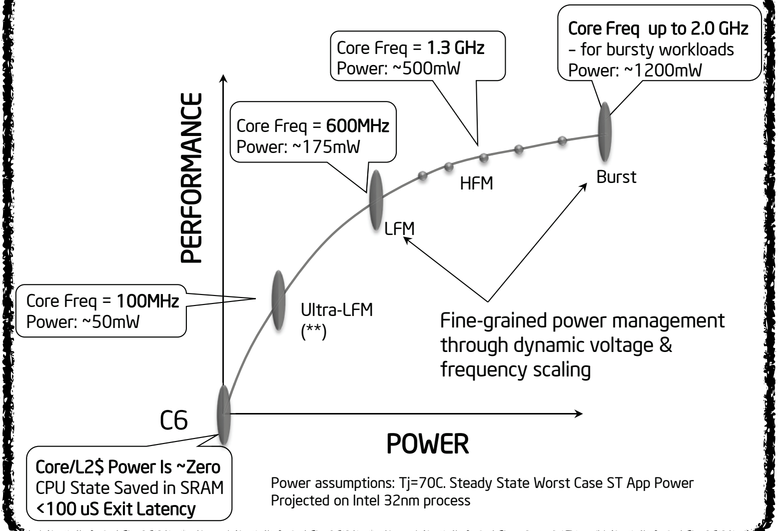
Intel Medfield

Switches 45 power "islands."

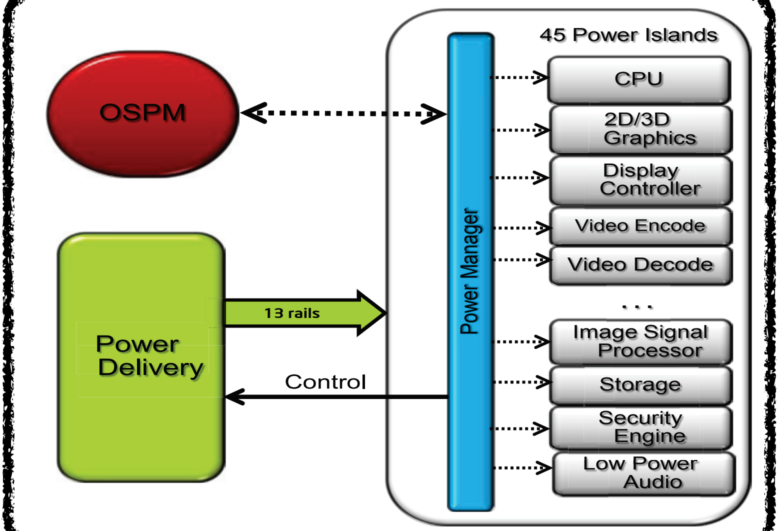
Fine-grained control of leakage power, to track user activity.

"Race to idle" strategy -- finish tasks quickly, to get to power down.

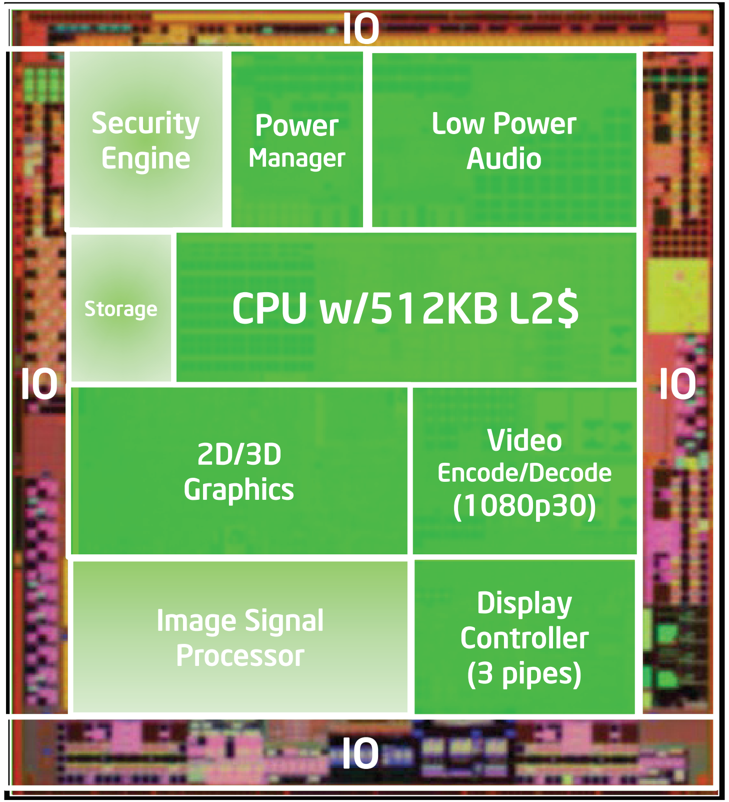
Intel® Smartphone Reference Design

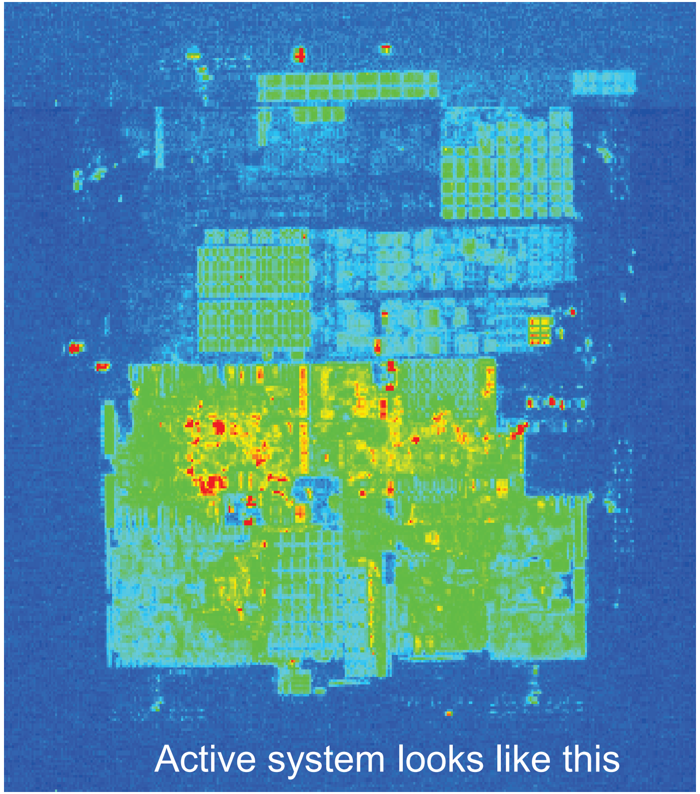


Power assumptions: Tj=70C. Steady State Worst Case ST App Power
Projected on Intel 32nm process

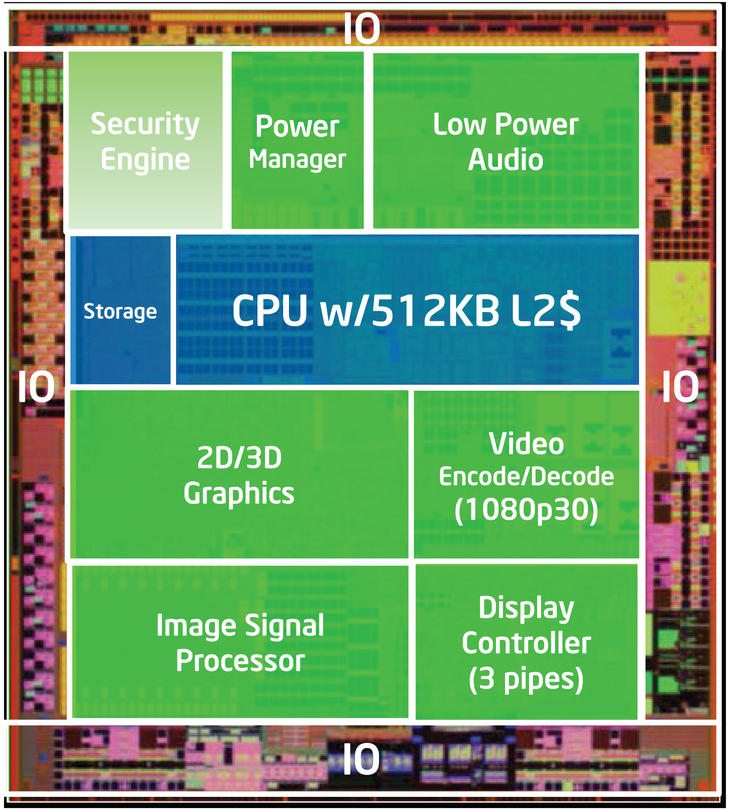


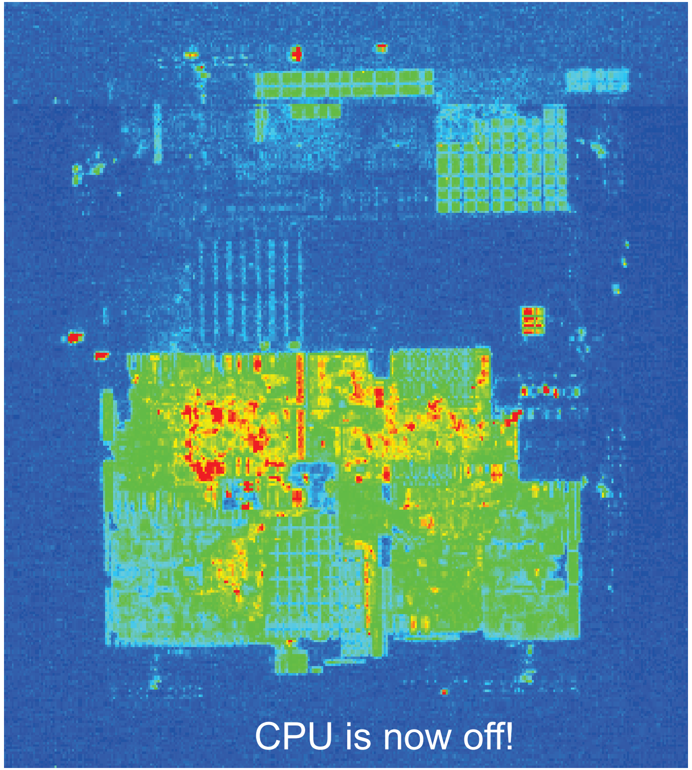
Playing a game …



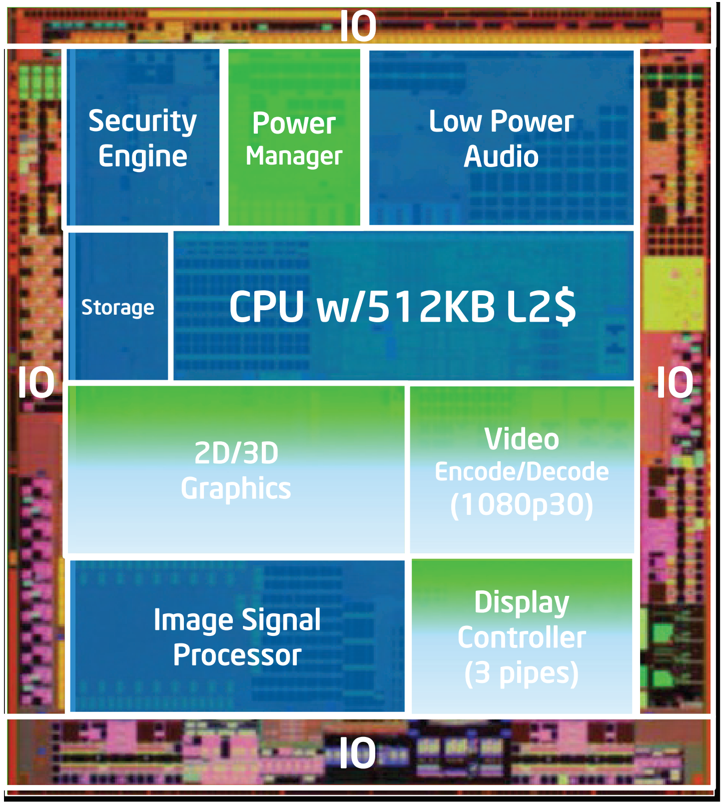


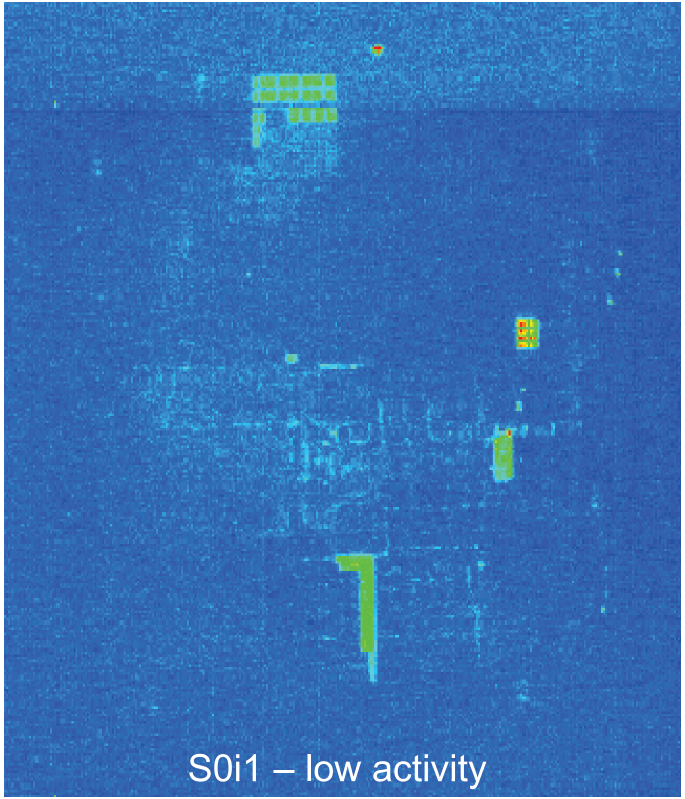
Watching a video ...



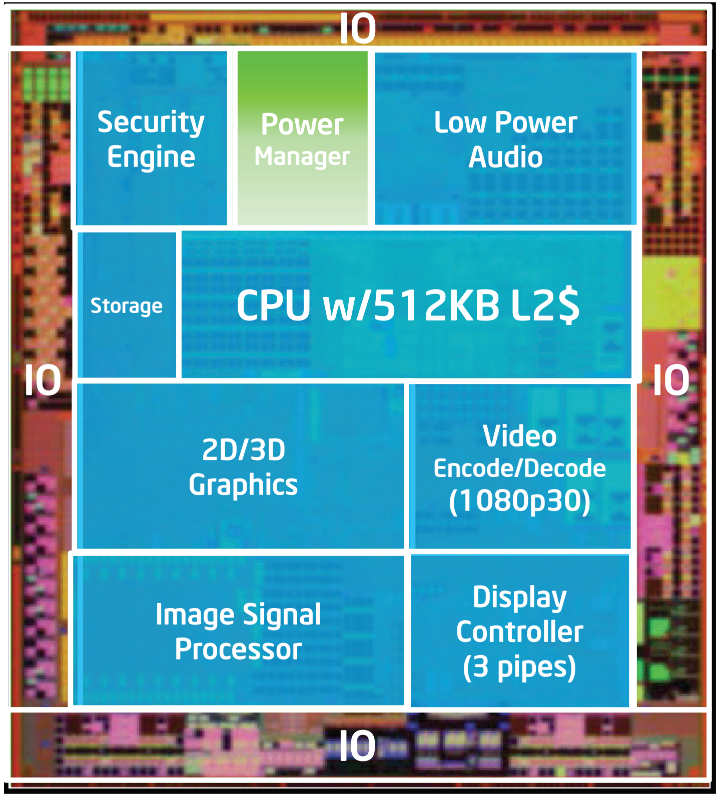


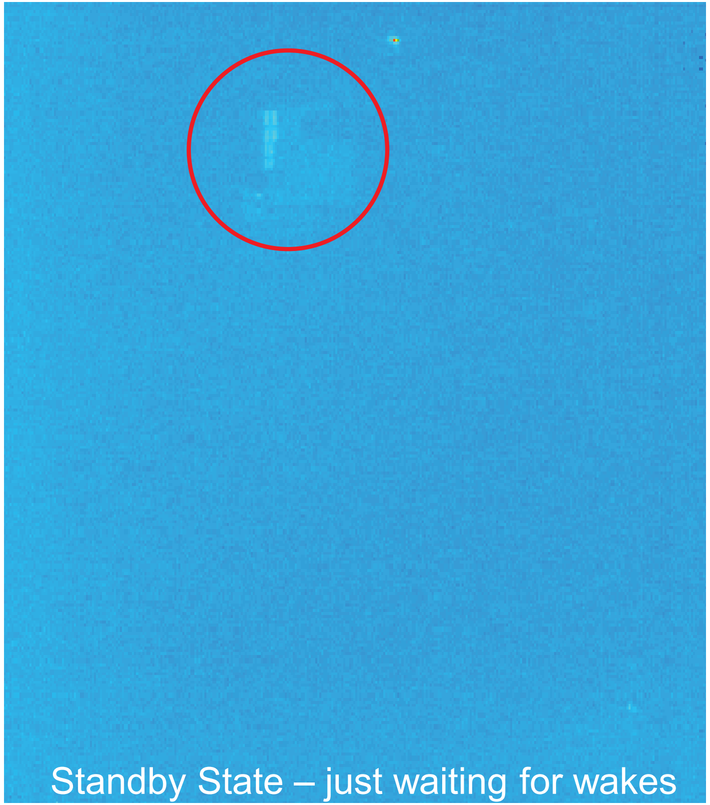
Looking at phone screen, not doing anything …





Phone in your pocket, waiting for a call …





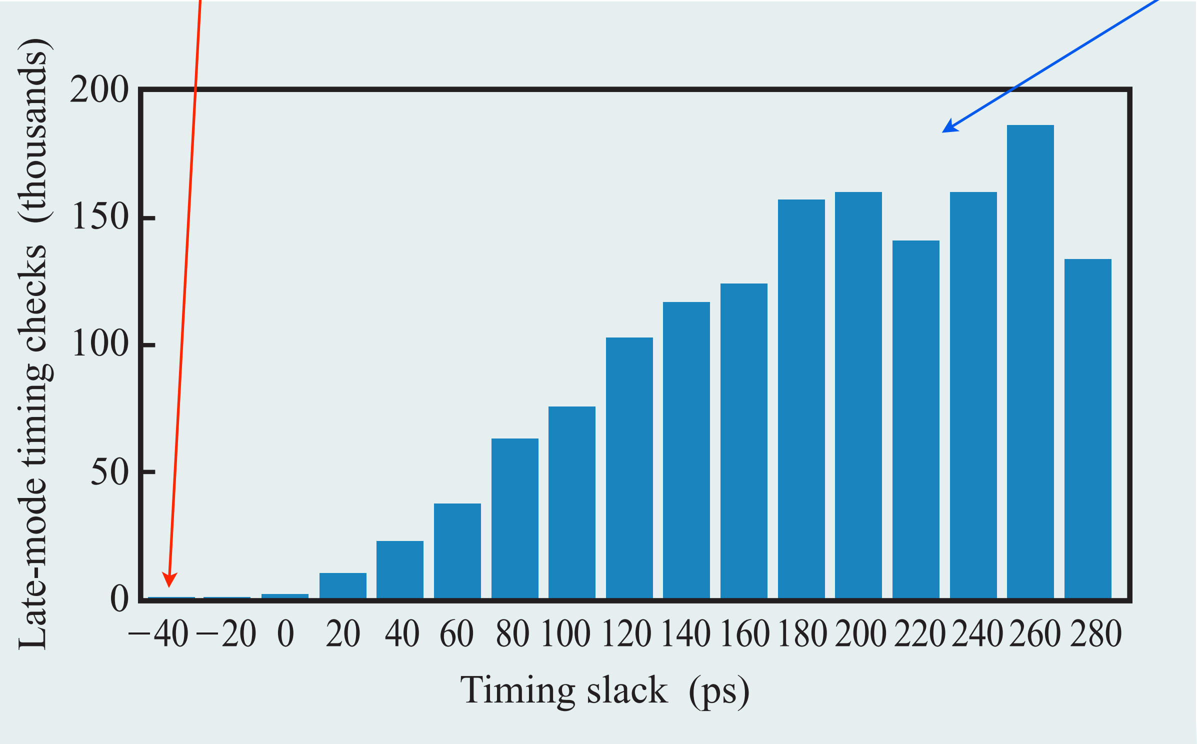
Design Technique #3 (of 5)

Slow down “slack paths”

Fact: Most logic on a chip is “too fast”

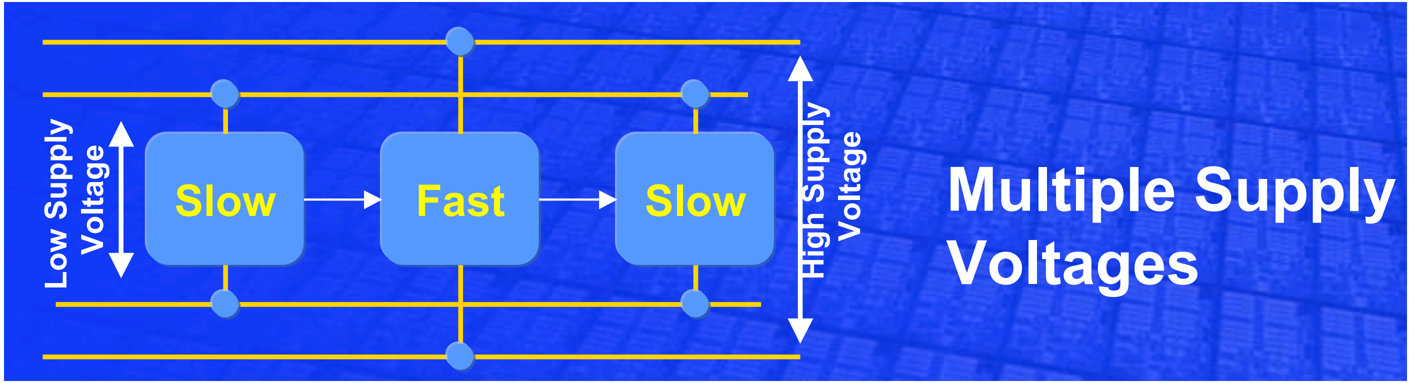
The critical path

Most logic paths have hundreds of picoseconds to spare.



From “The circuit and physical design of the POWER4 microprocessor”, IBM J Res and Dev, 46:1, Jan 2002, J.D. Warnock et al.

Use several supply voltages on a chip ...



Multiple Supply Voltages

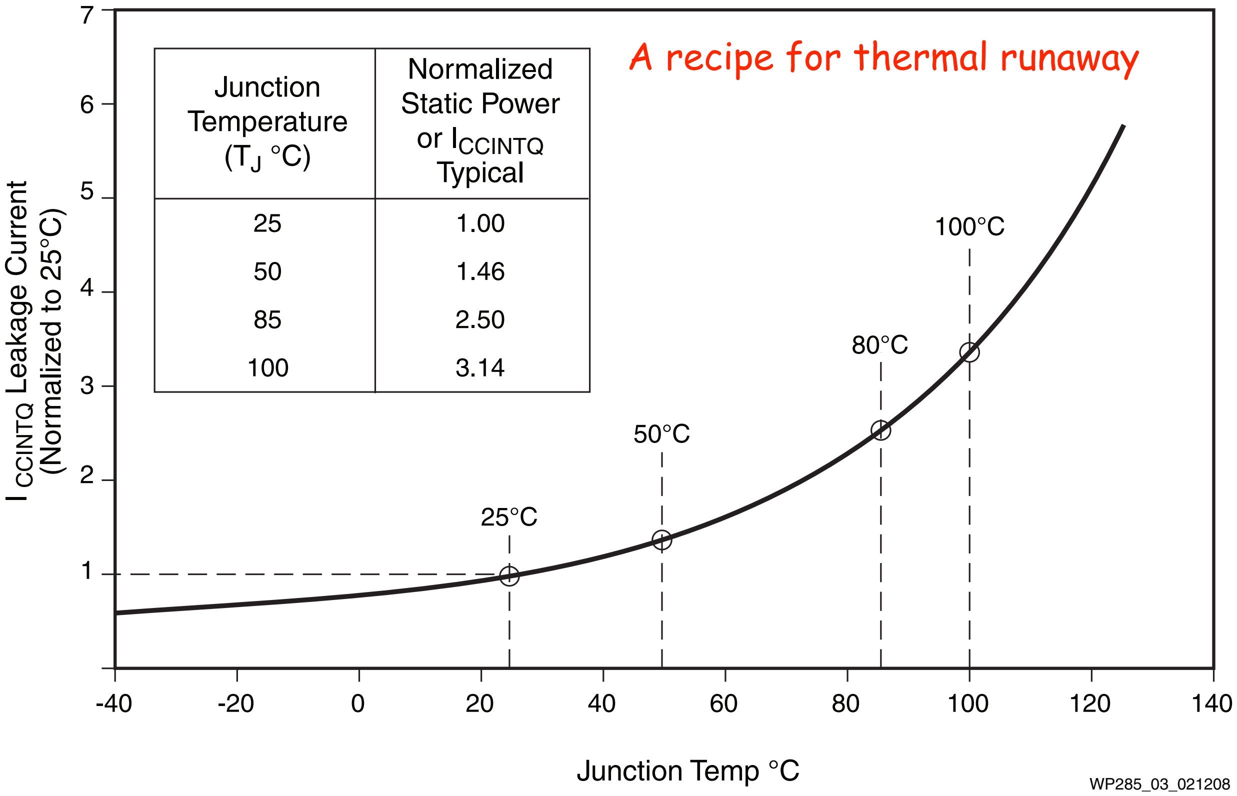
Why use multi-Vdd? We can reduce dynamic power by using low-power Vdd for logic off the critical path.

In practice, instead of multi-Vdd design ...
In a multi-Vt process, we can reduce leakage power on the off critical path logic by using high-Vth transistors.

Design Technique #5 (of 5)

Thermal Management

Keep chip cool to minimize leakage power



Junction Temperature (T_J °C)	Normalized Static Power or I_{CCINTQ} Typical
25	1.00
50	1.46
85	2.50
100	3.14

A recipe for thermal runaway

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