

EECS151/251A
Spring 2023
Digital Design and Integrated
Circuits

Instructors:
John Wawrzynek

Lecture 22 - 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).

$$1 \text{ J} = 1 \text{ W} * \text{s} \quad 1 \text{ W} = 1 \text{ J/s.}$$

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.

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

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.**

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

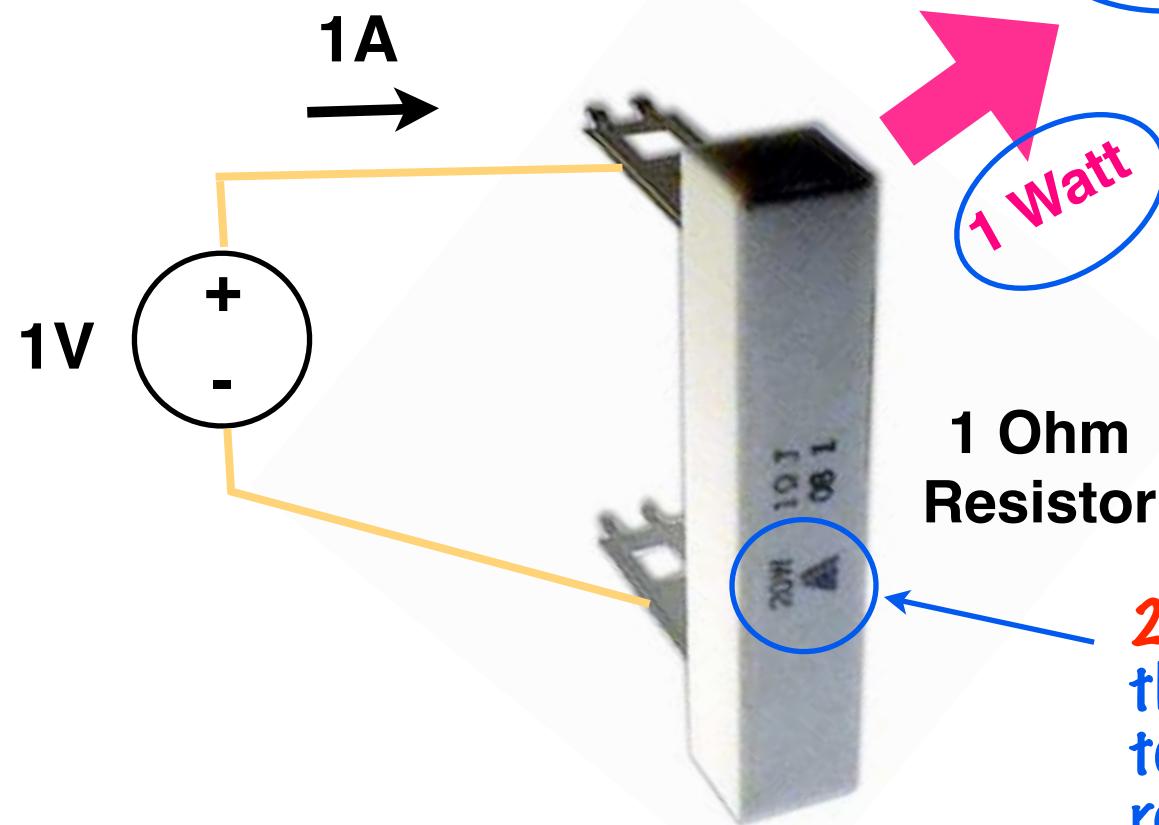
This is how electric tea pots work ...

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.



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 ...



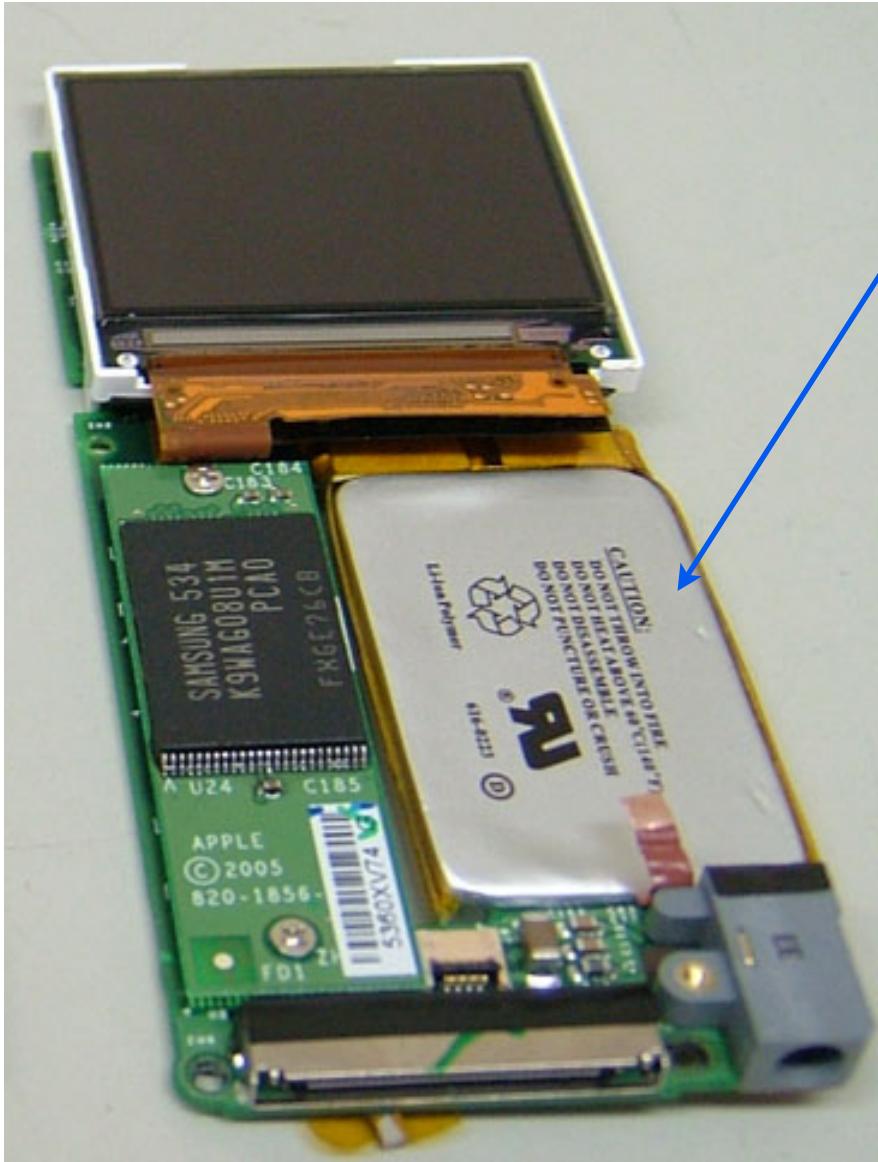
Like resistor on last slide, iPod relies on passive transfer of heat from case to the air.

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.**

$$1.2 \text{ W-hr} / 5 \text{ W} \approx 15 \text{ minutes.}$$

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.

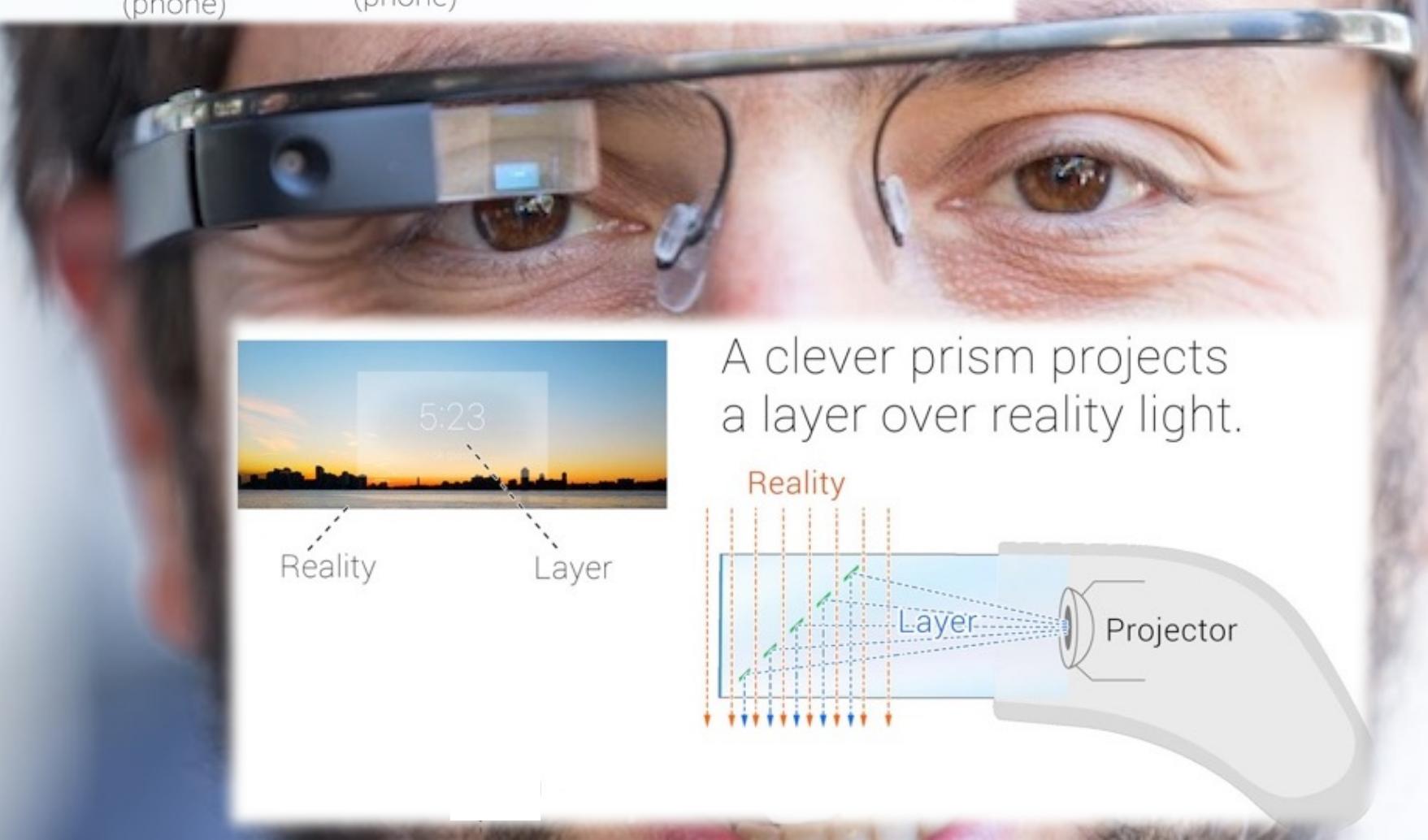
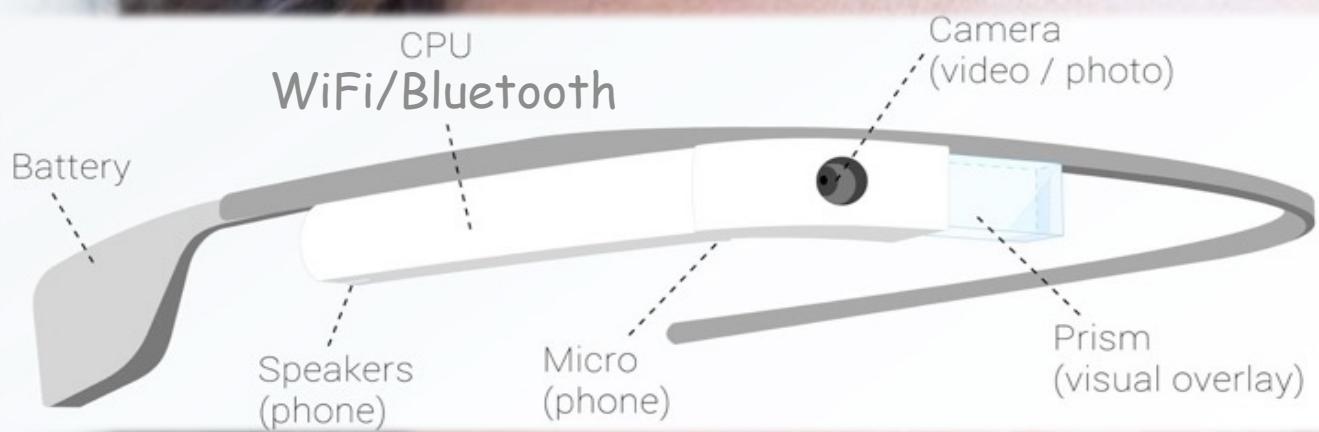
**85 mW for music.
300 mW for slides.**

2015

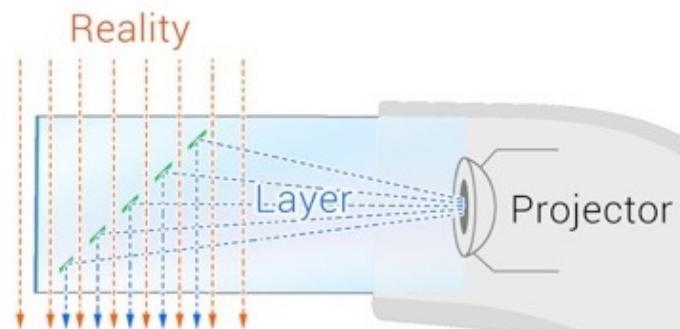
Apple Watch



3.8 V, 0.78 Whr 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.)



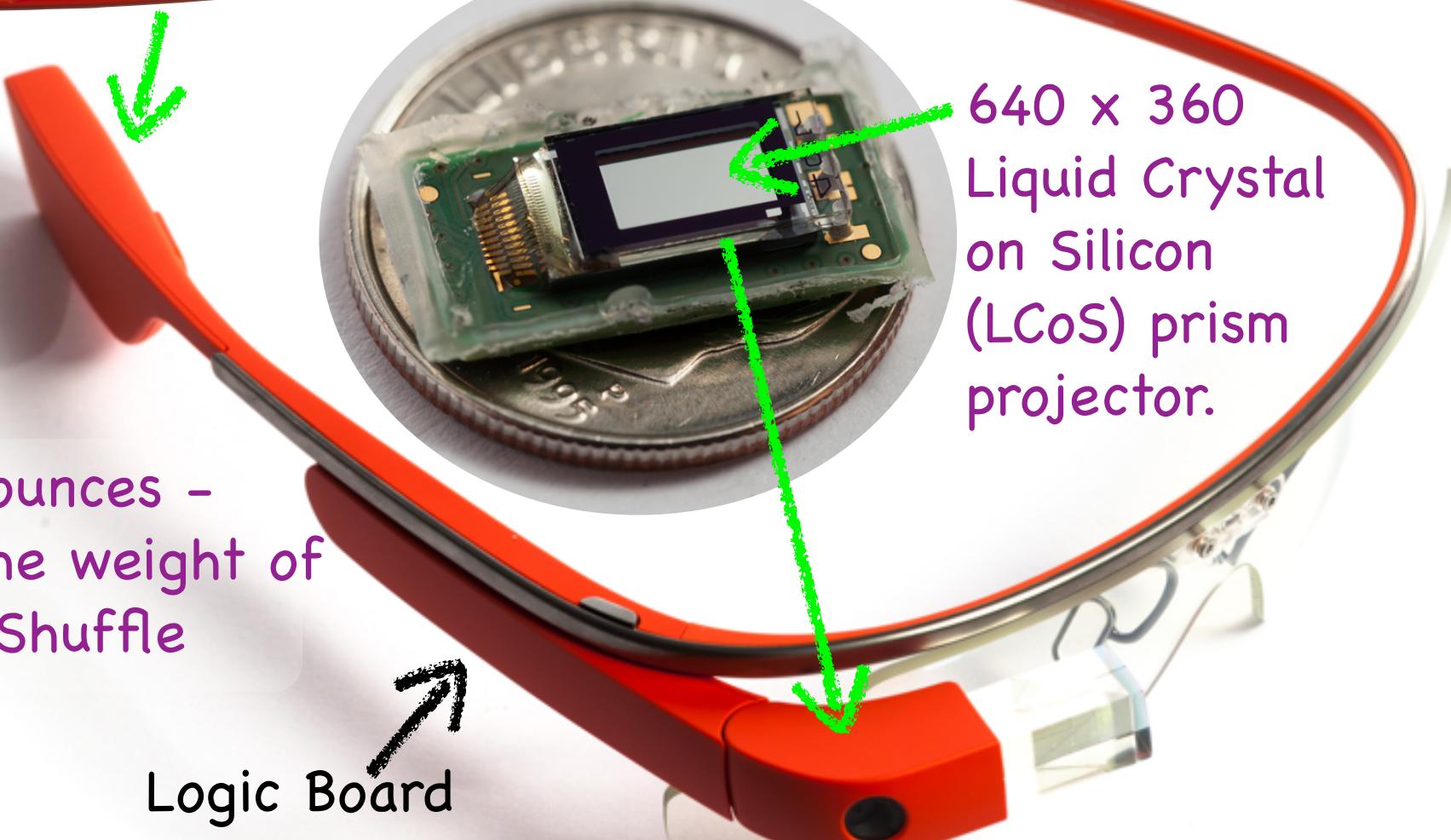
A clever prism projects a layer over reality light.



2.1 Wh battery - 2.7x as much energy as Apple watch.



Battery life very usage dependent.



1.76 ounces -
4X the weight of
iPod Shuffle

Logic Board

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

iPhone6



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

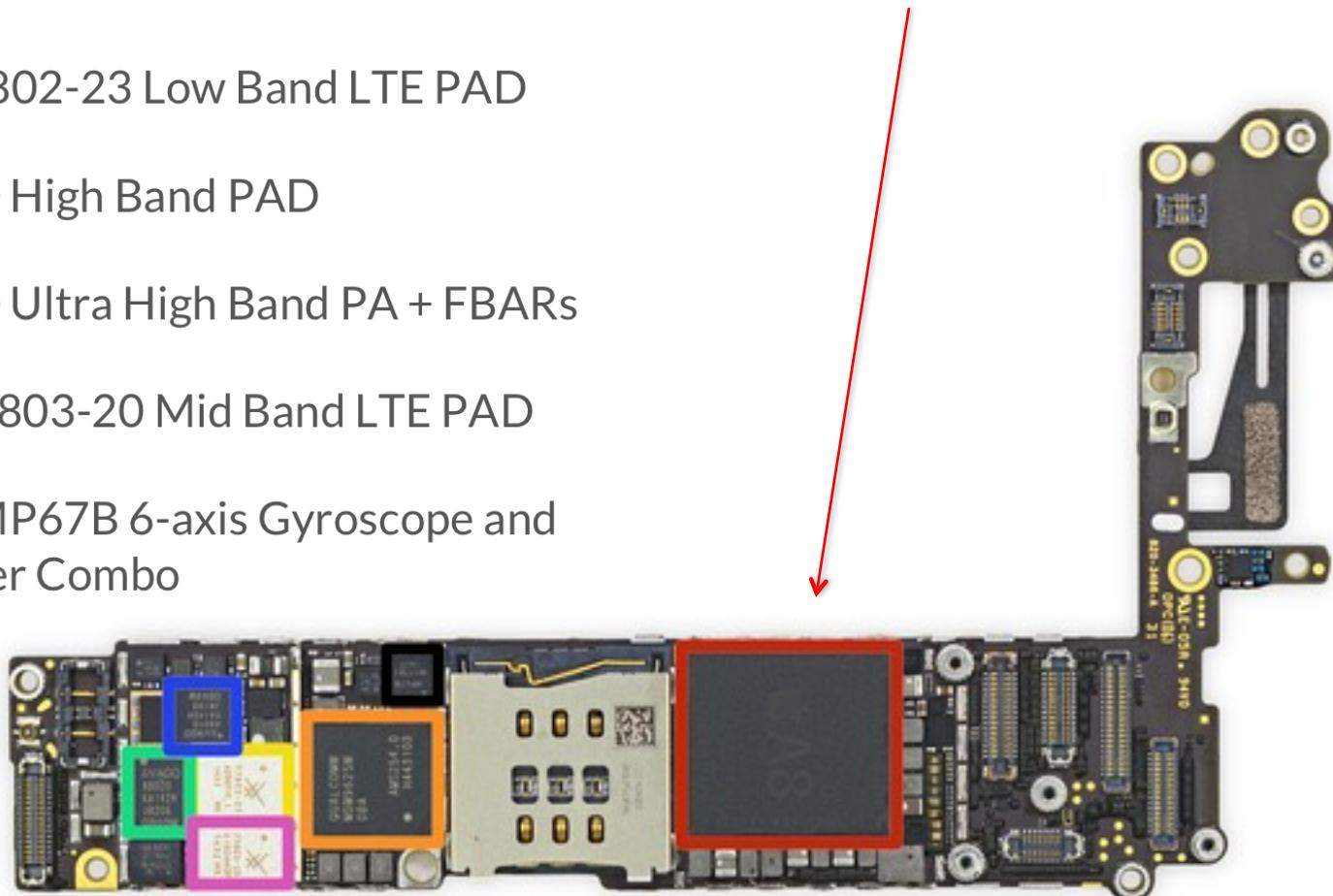


**iPhone 5s: 1570mAh
@3.8V = 6 Wh**

- The front side of the logic board:

- Apple A8 APL1011 SoC + SK Hynix RAM as denoted by the markings 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

The A8 is manufactured on a 20 nm process by TSMC. It contains 2 billion transistors. Its physical size is 89 mm². [1] It has 1 GB of LPDDR3 RAM included in the package. It is dual core, and has a frequency of 1.38 GHz.



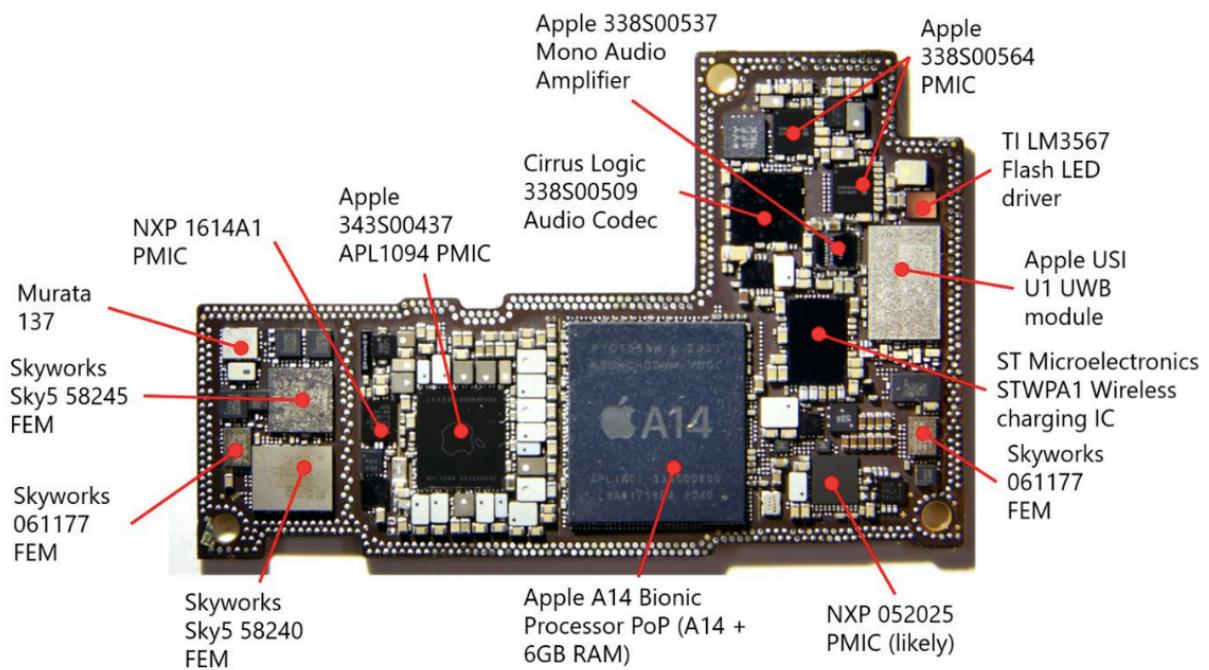
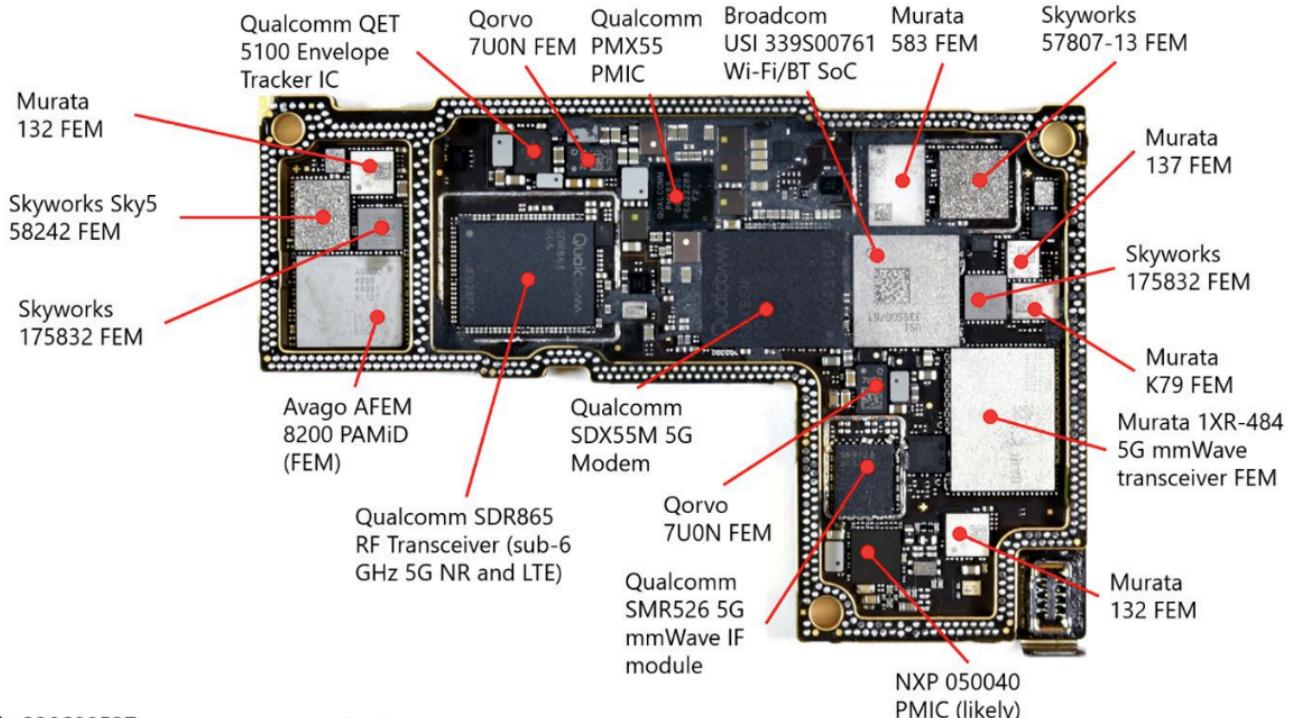
Iphone 12:

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

iPhone Model	Battery Capacity	14.13 Wh @ 3.8V (Pro Max)
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	

Accessory Identification
NFC Antenna





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!

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

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 ...

50Wh is 180,000 Joules!

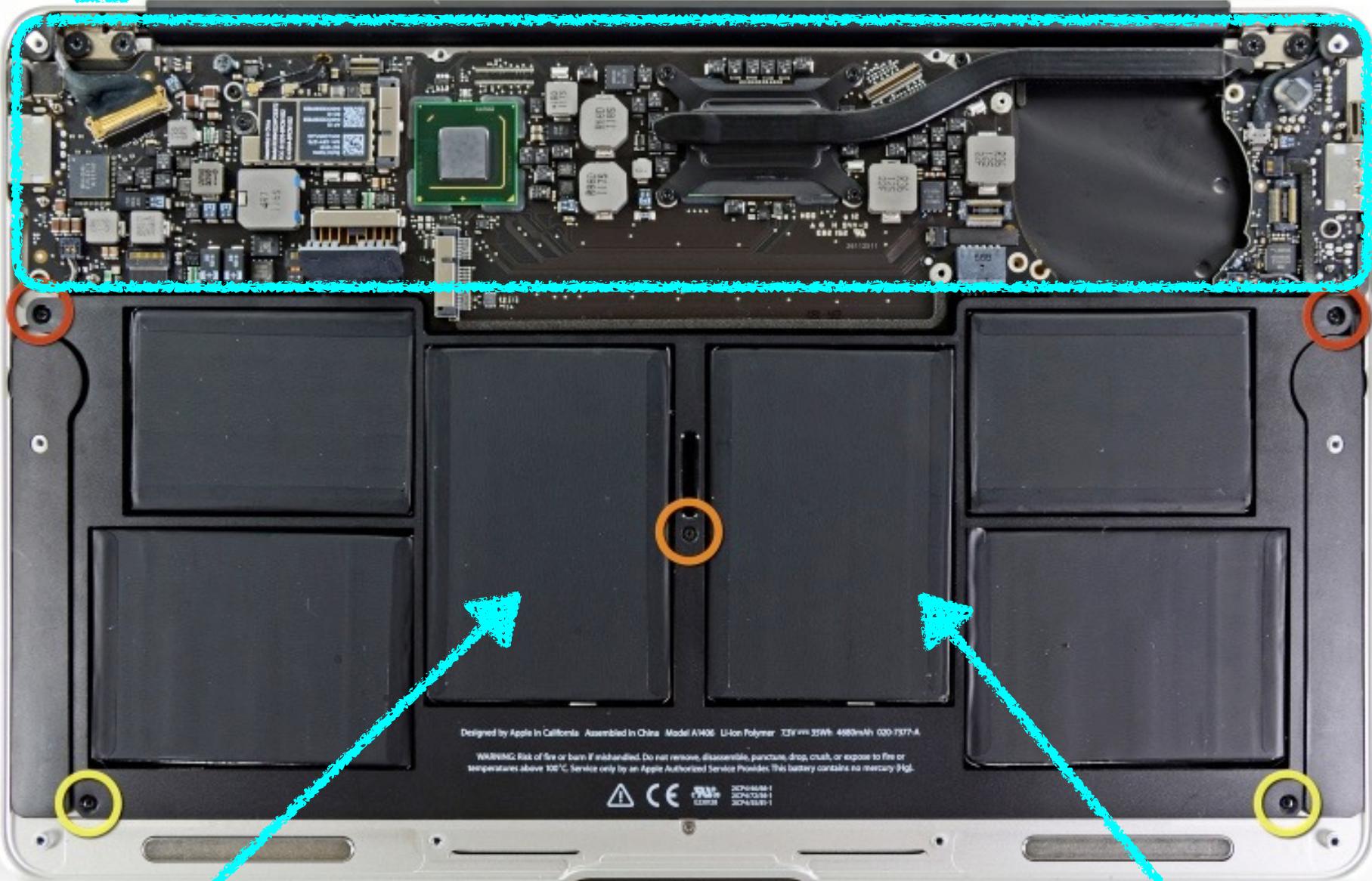




MacBook Air ... design the laptop
like an iPod/iPhone



Mainboard: fills about 25% of the laptop



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

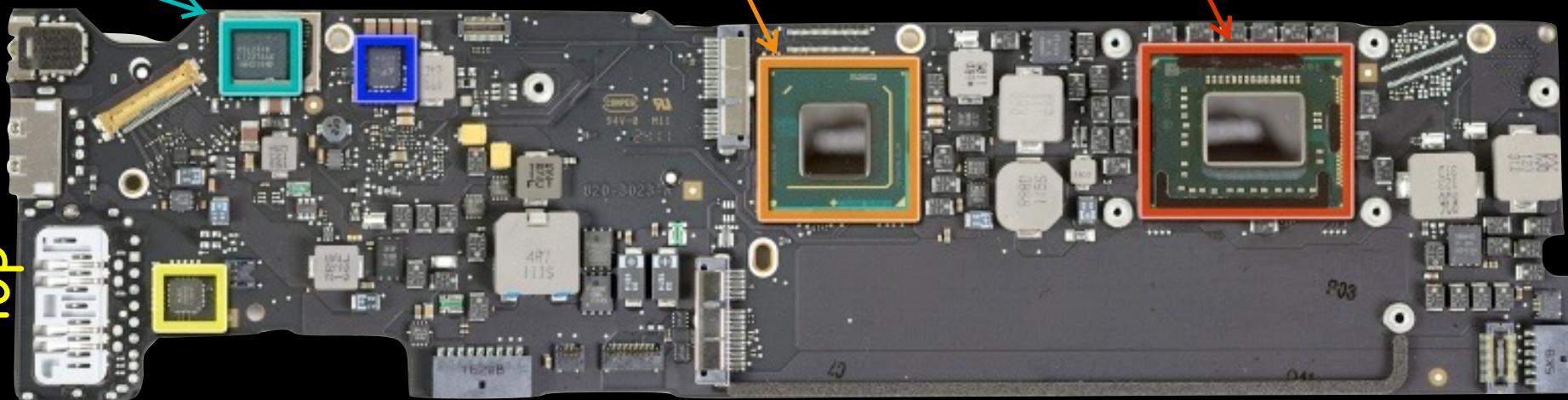
MacBook Air: Full PC



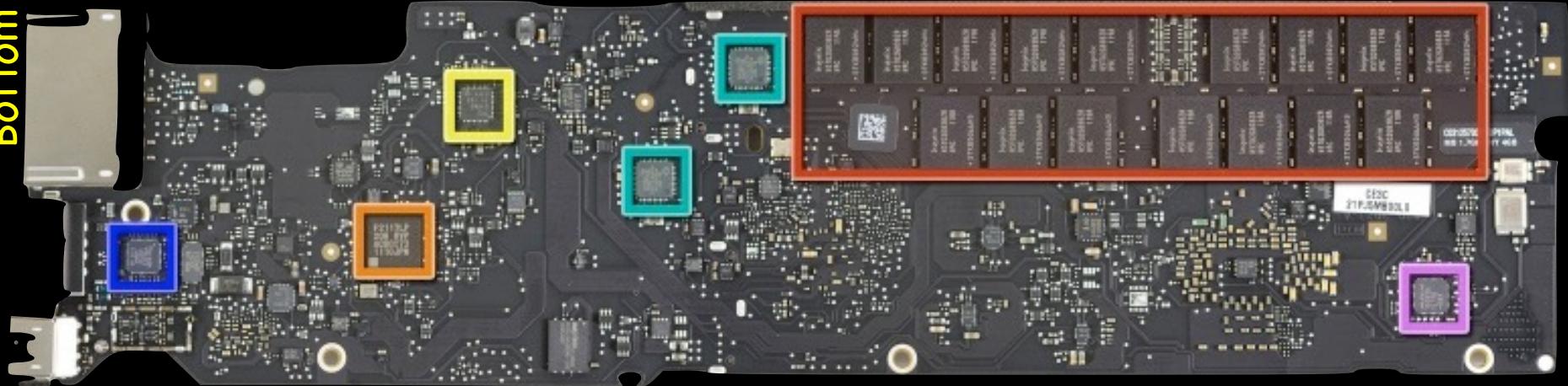
Thunderbolt I/O

Platform
Controller
Hub

Core i5
CPU/GPU



Up to
4GB DRAM



Servers: Total Cost of Ownership (TCO)

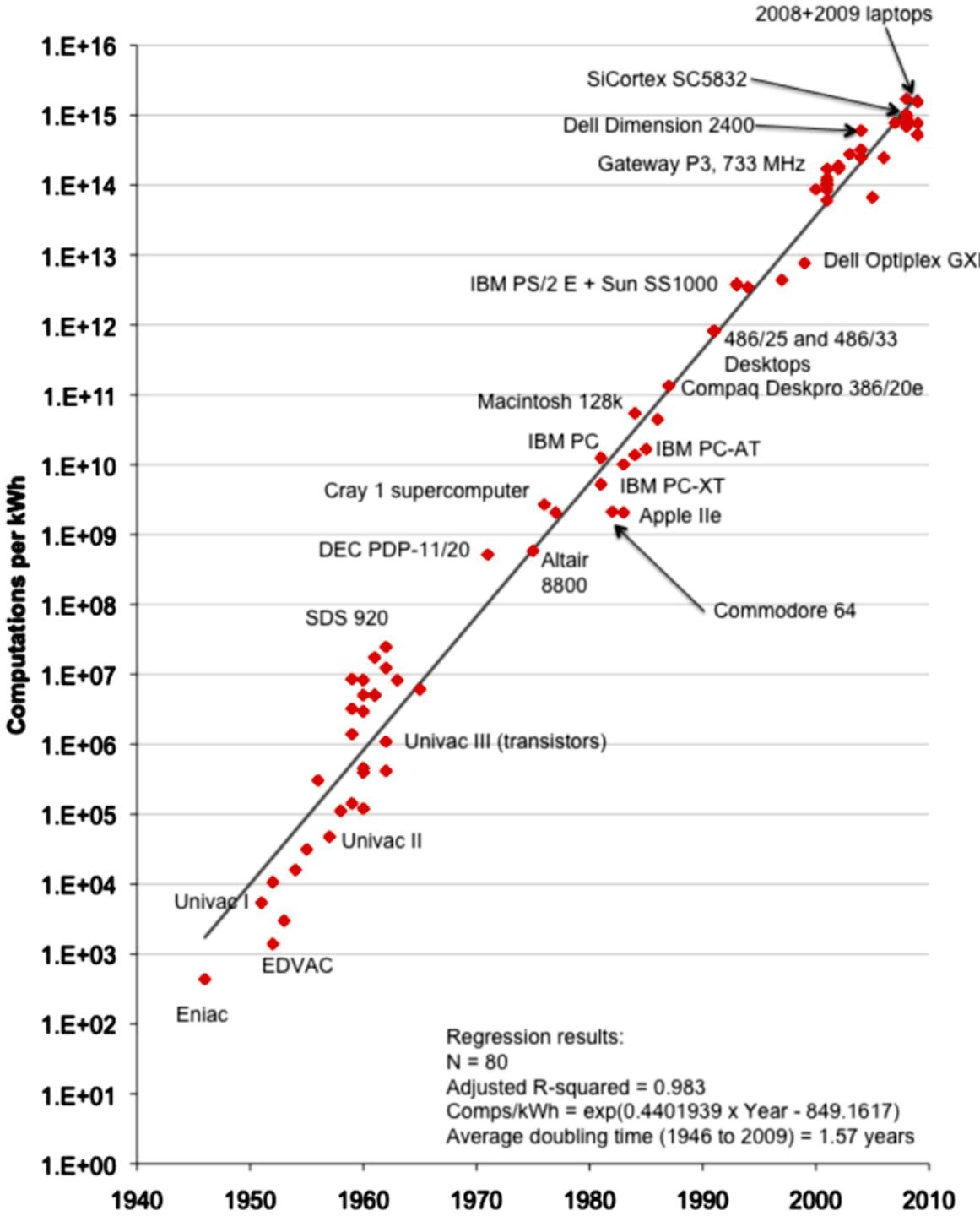


Reliability: running computers hot makes them fail more often.

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.

Computations
per W-h
doubles every
1.6 years,
going back
to the first
computer.

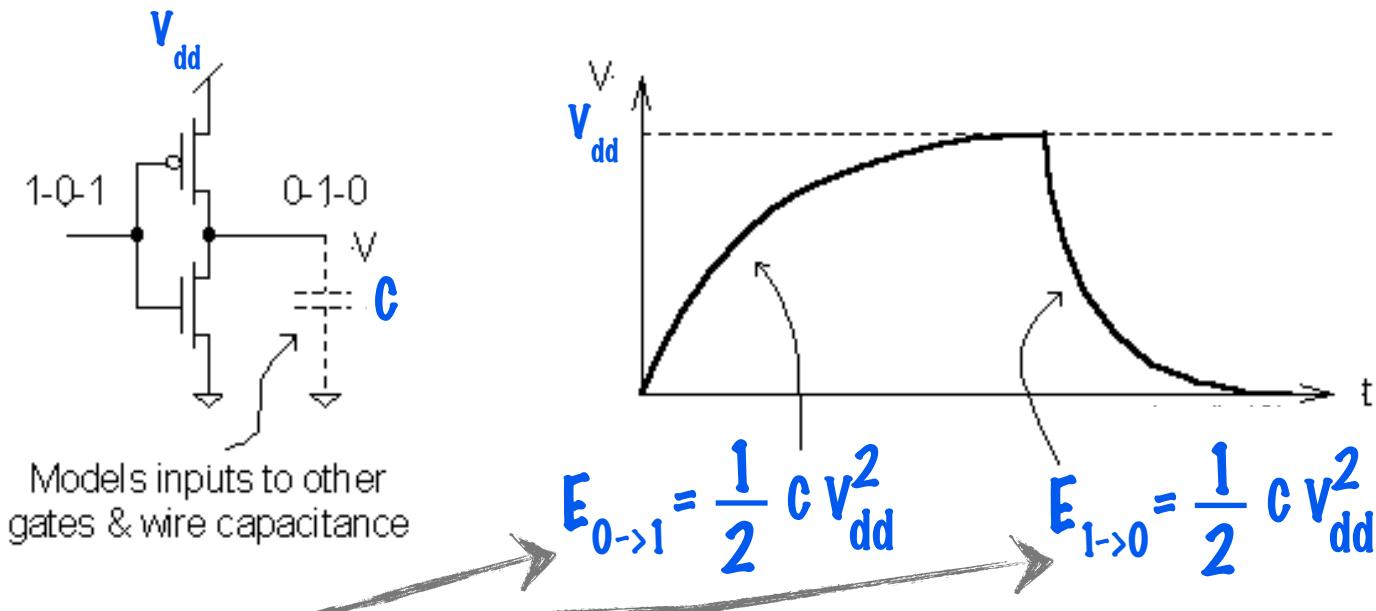


(Jonathan
Koomey,
Stanford).

CMOS Circuits and Energy

Switching Energy: Fundamental Physics

Every logic transition dissipates energy.



Strong result: Independent of technology.

How can we limit switching energy?

- (1) Reduce # of clock transitions. But we have work to do ...
- (2) Reduce V_{dd} . But lowering V_{dd} 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

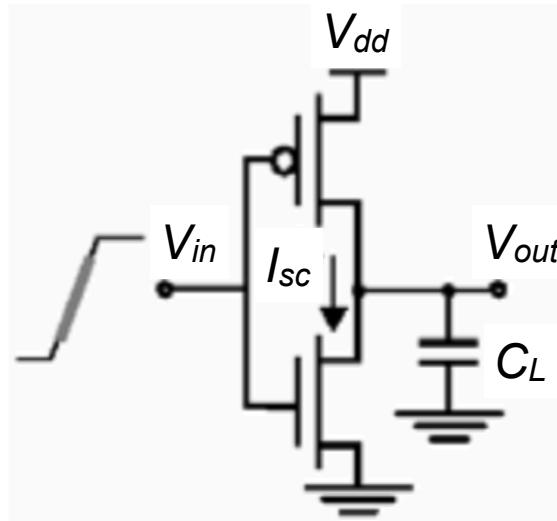
$$P_{sw} = 1/2 \alpha C V_{dd}^2 F$$

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

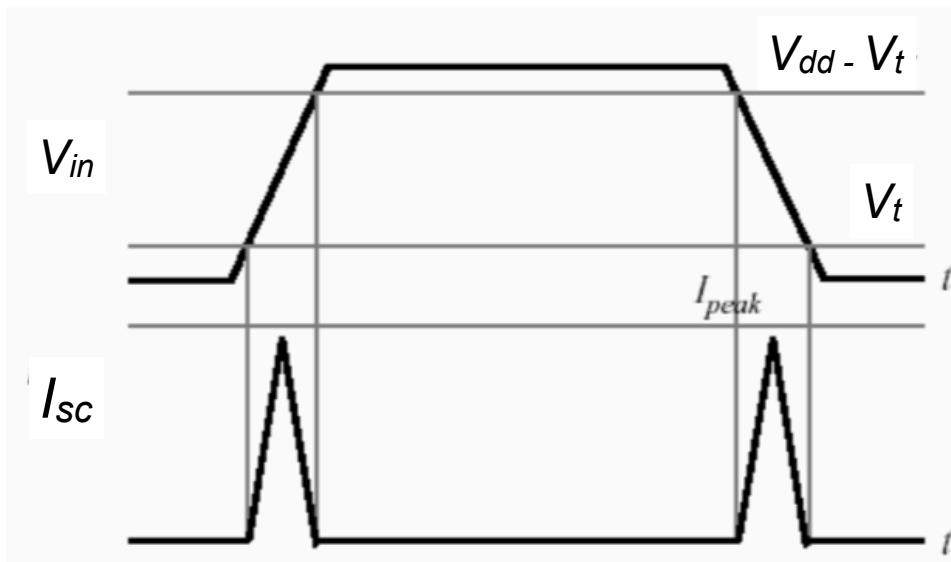
Total chip capacitance to be switched

Clock Frequency

Additional Dynamic Power - “short circuit current”



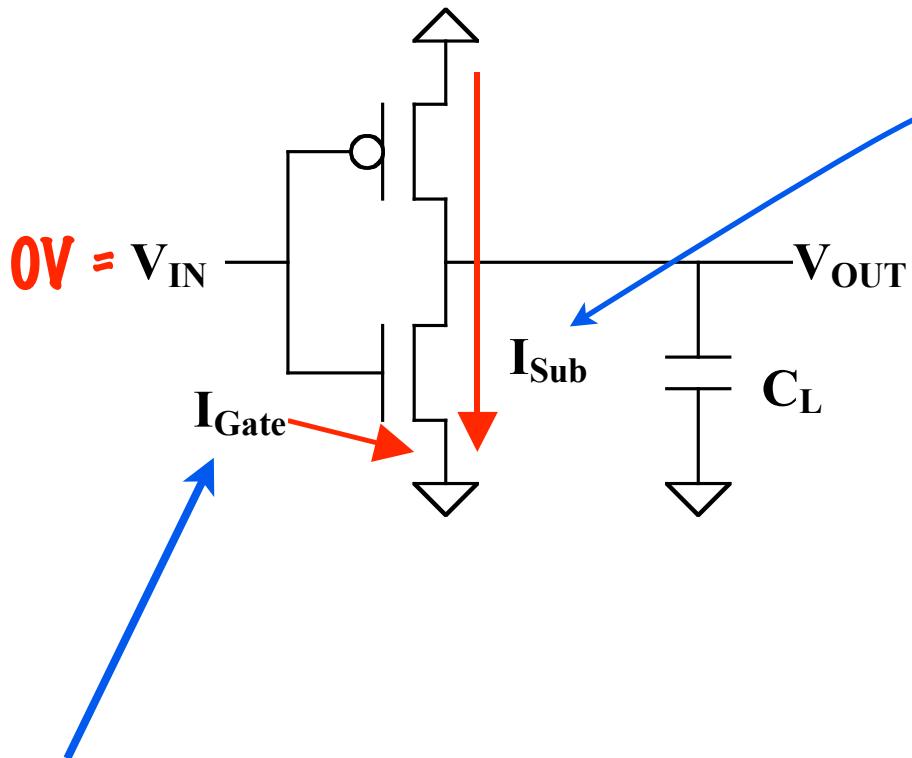
When gate switches, brief period when both pullup network and pulldown network could be on.



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.

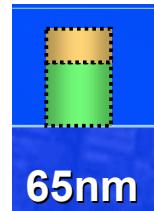


I_{Gate} : Ideal capacitors have zero DC current. But modern transistor gates are a few atoms thick, and are not ideal.

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.

Intel's 2006 processor designs, leakage vs switching power

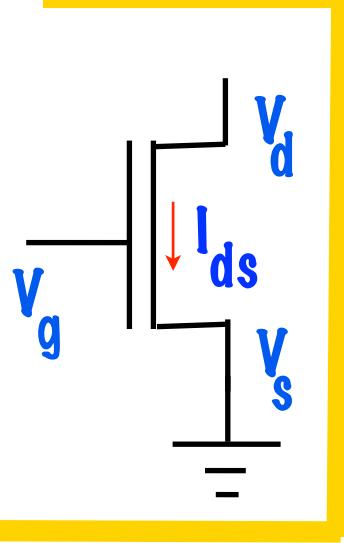


- Leakage
- Dynamic

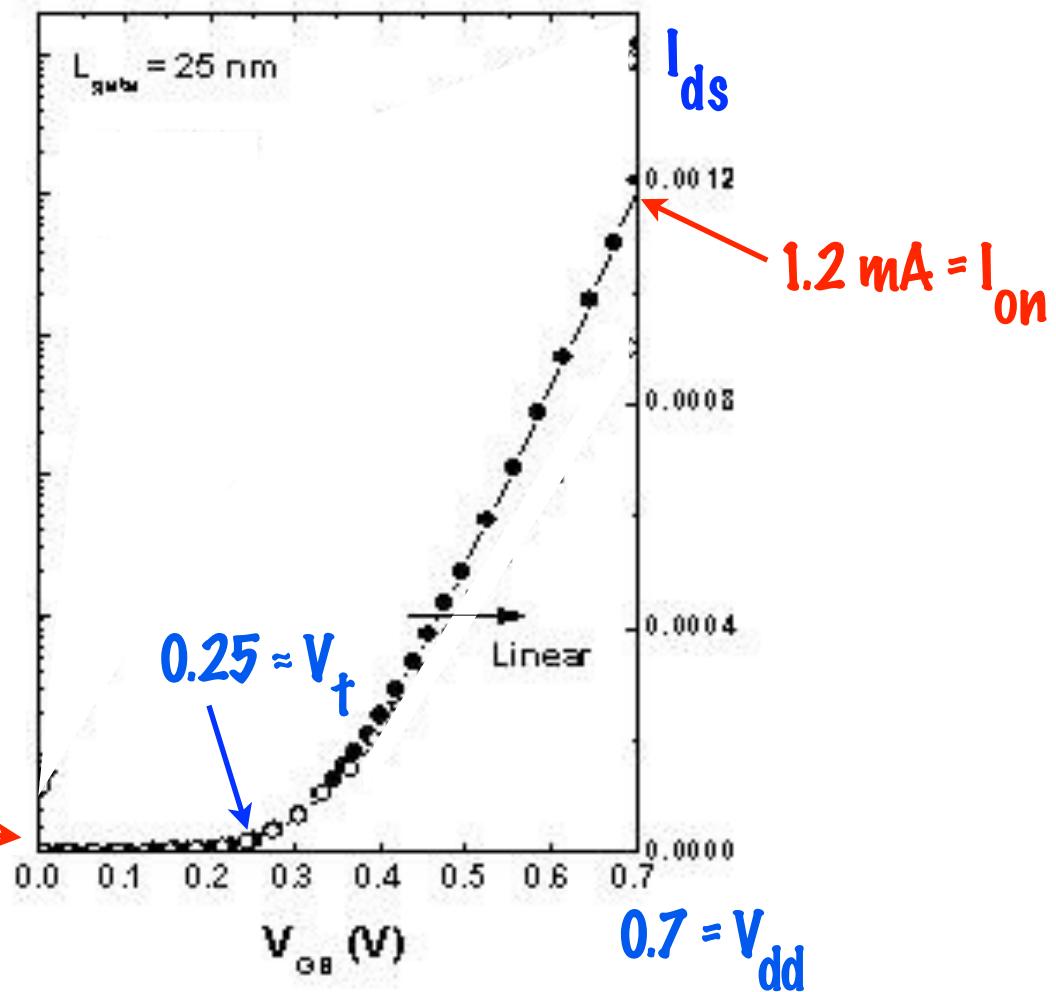
Bill Holt, Intel, Hot Chips 17.

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

Engineering “On” Current at 25 nm ...

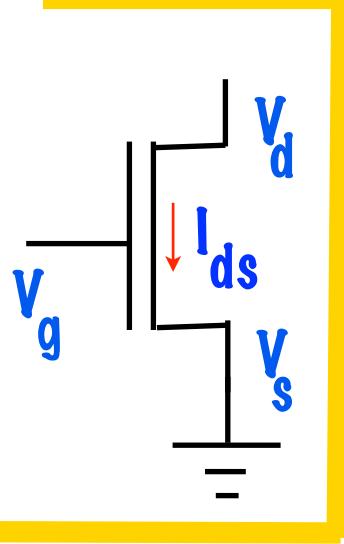


We can increase I_{on} by raising V_{dd} and/or lowering V_t .

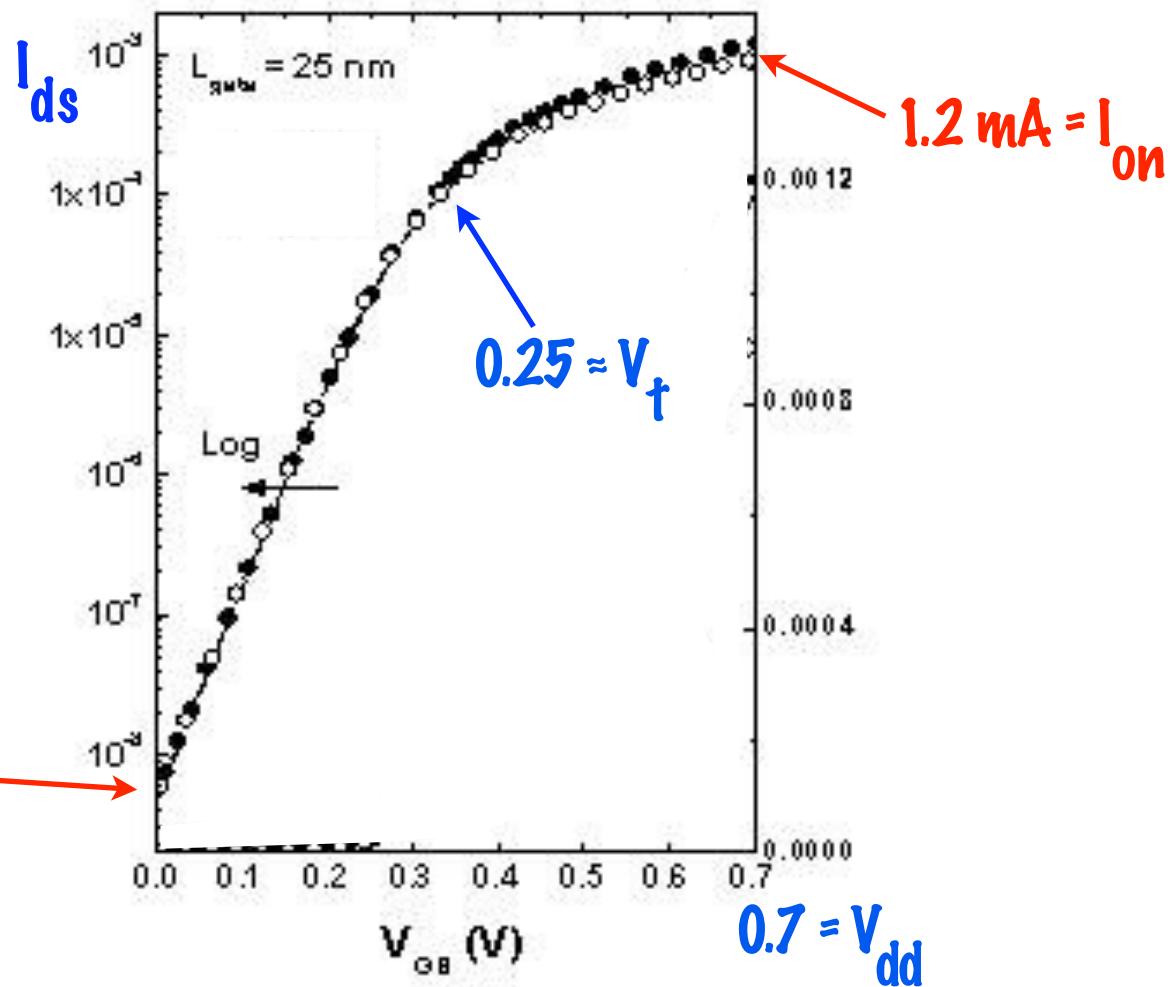


$I_{off} = 0 ???$

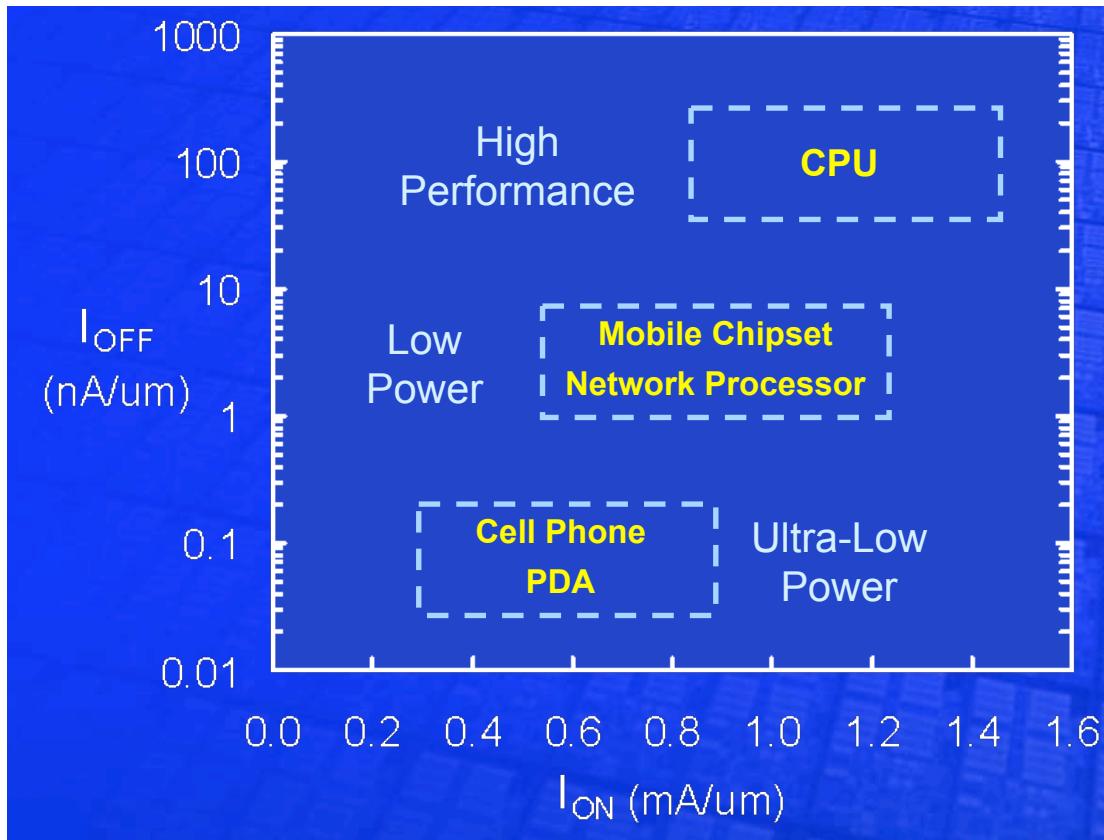
Plot on a “Log” Scale to See “Off” Current



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

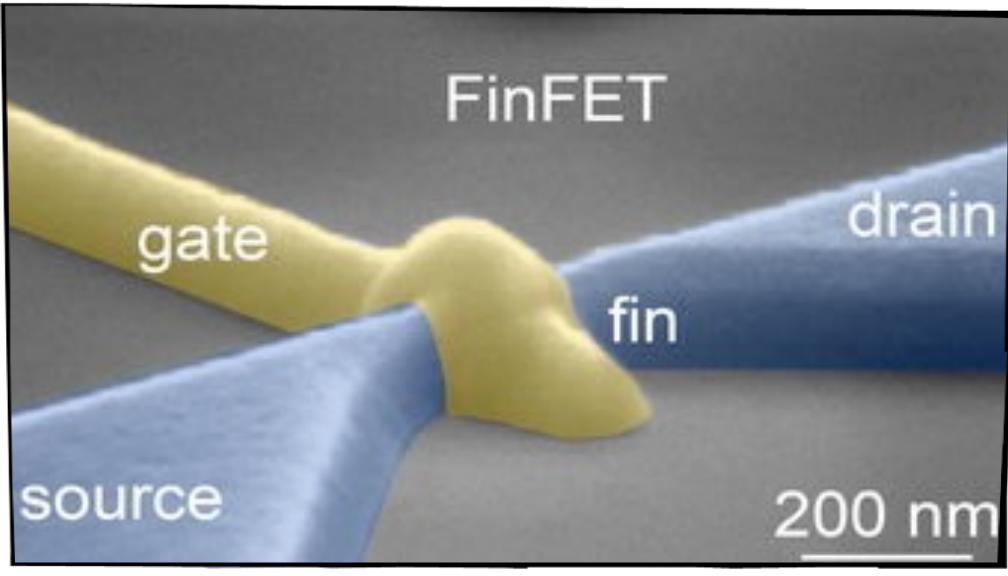


Customize processes for product types ...



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

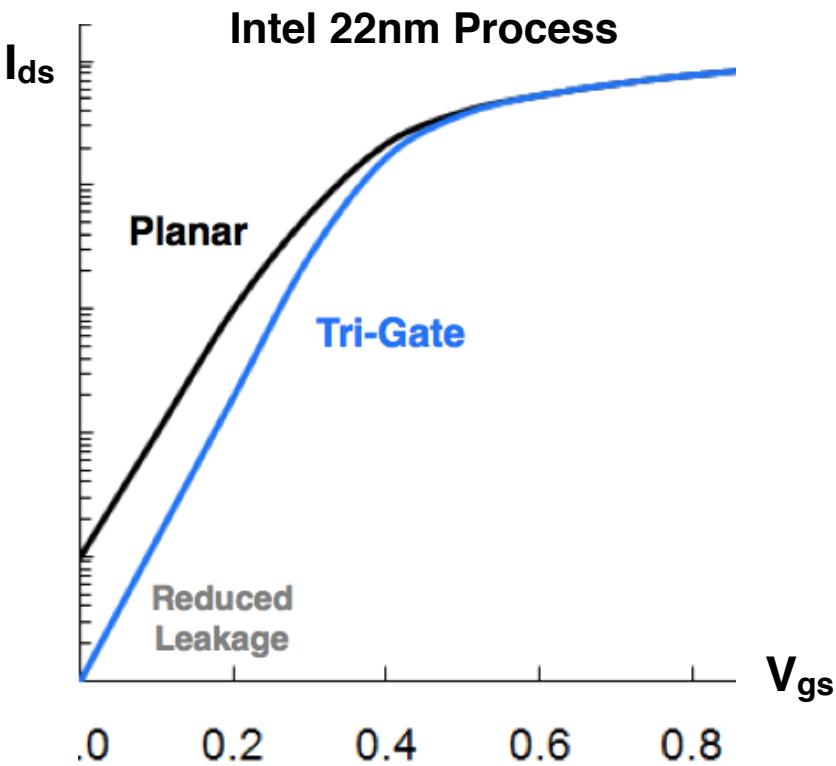
- *V_t is controlled by channel doping.*
- *Modern IC processes have 2 or 3 different V_t values available.*
- *Standard cell libraries offer low V_t and high V_t versions of cells so that the tools can optimize on a per instance basis.*
- *(If high performance not needed then use high V_t to reduce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



Transistor channel is a raised fin.

Gate controls channel from sides and top.

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

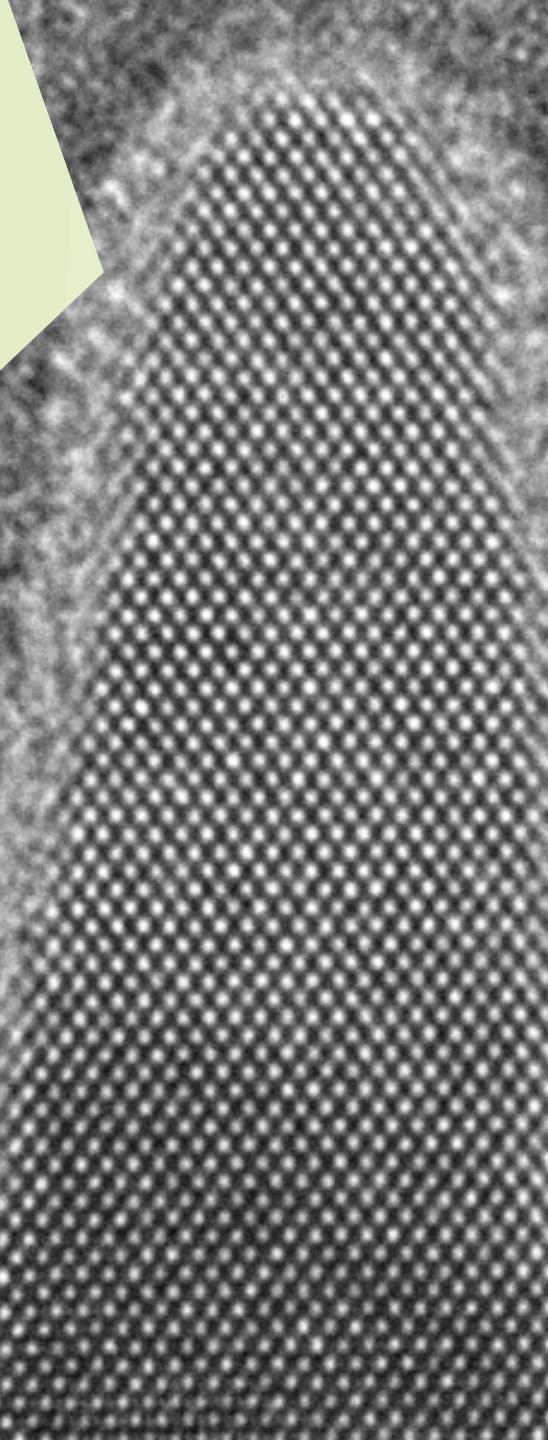
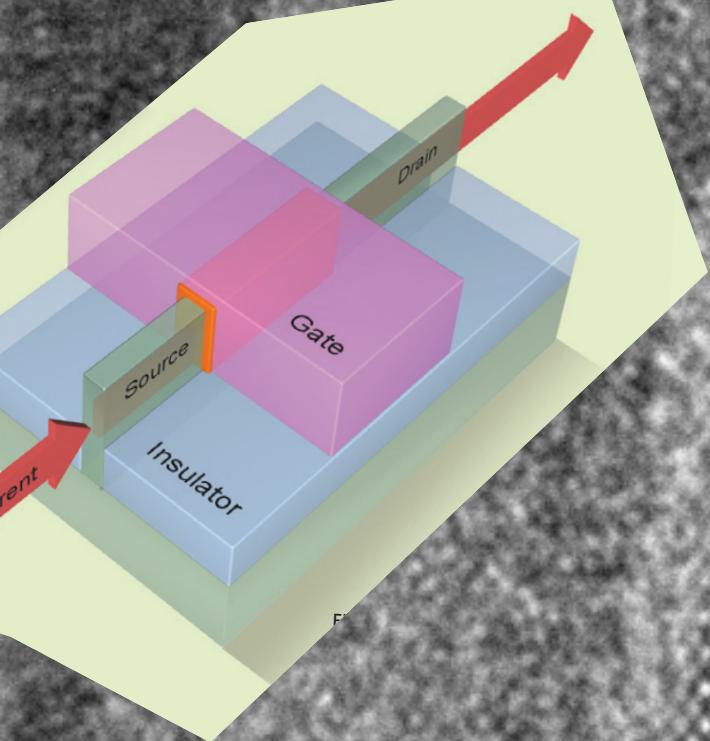


(12) **United States Patent**
Hu et al. Filed: Oct. 23, 2000

(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

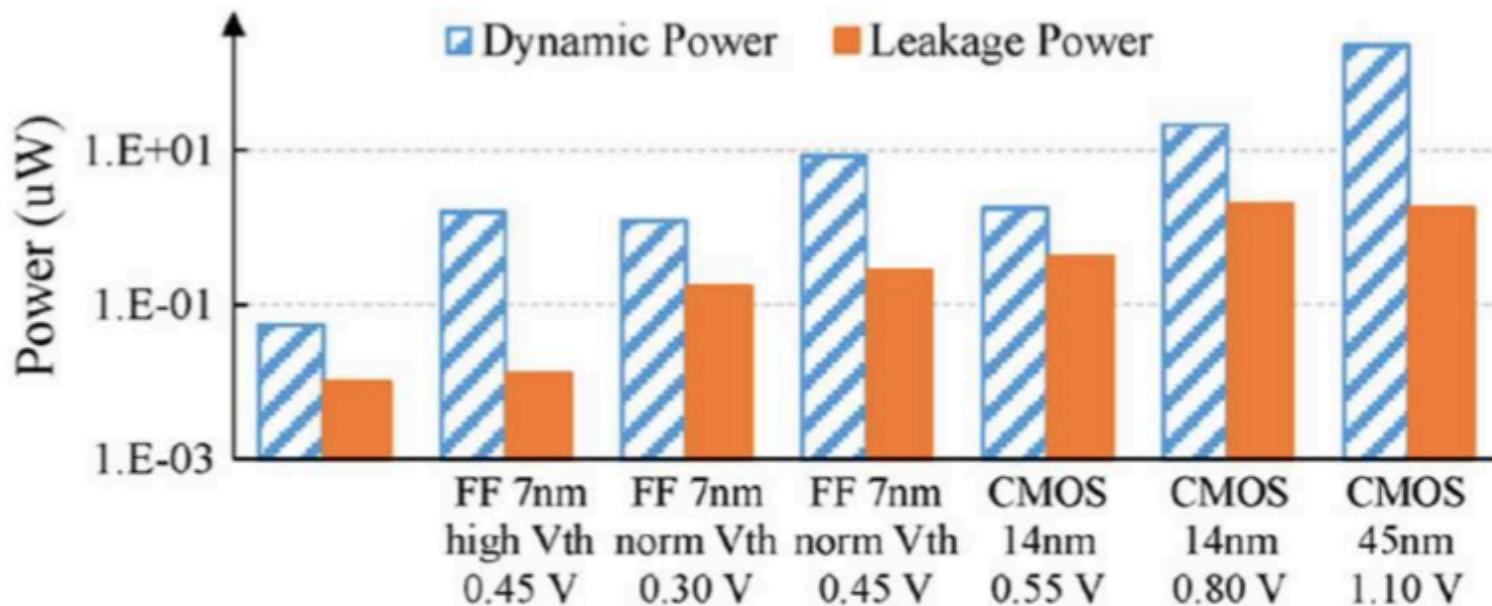
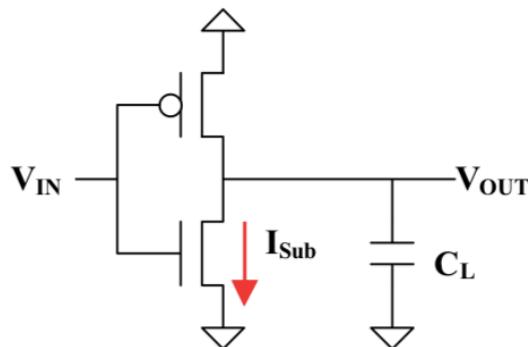


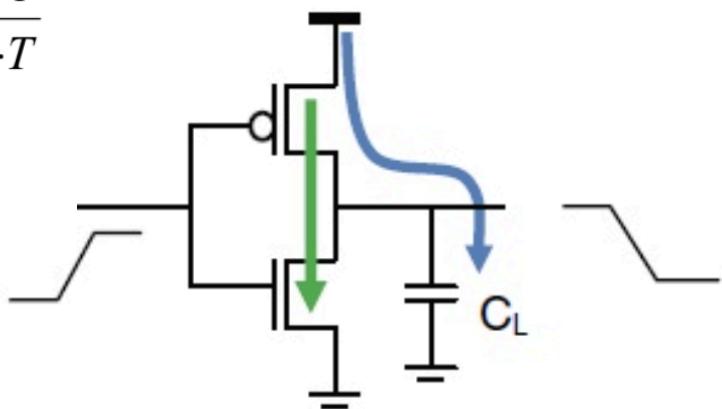
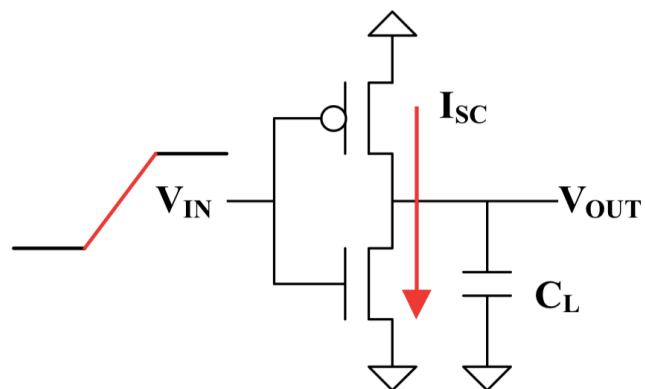
Figure 1: The reduction of feature sizes from 45 to 7nm may induce drastic gains in power consumption and leakage power [Xie2015]

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



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

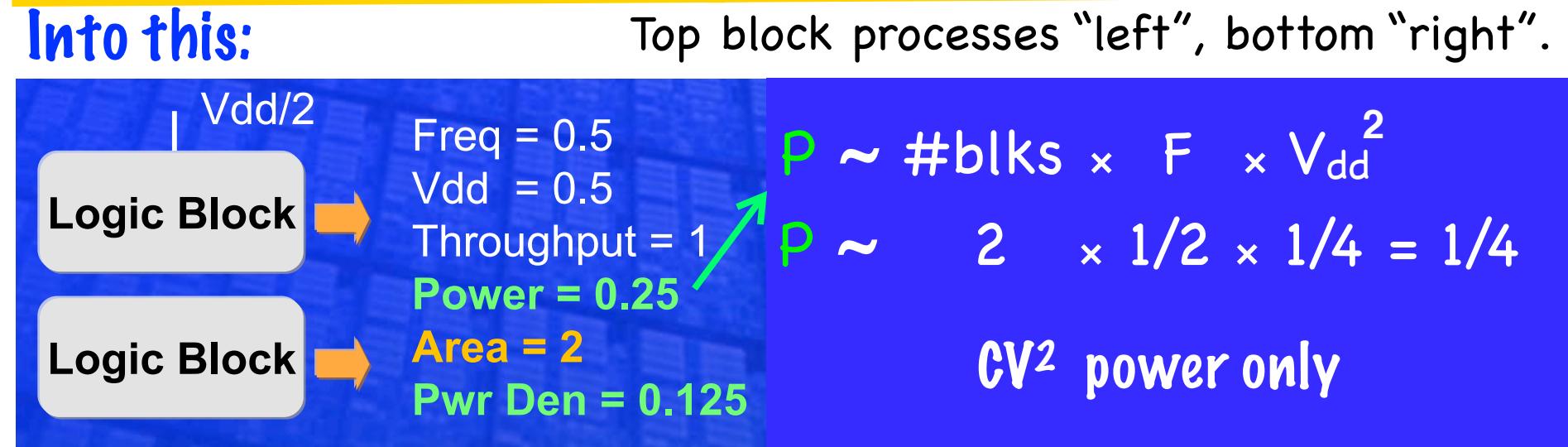
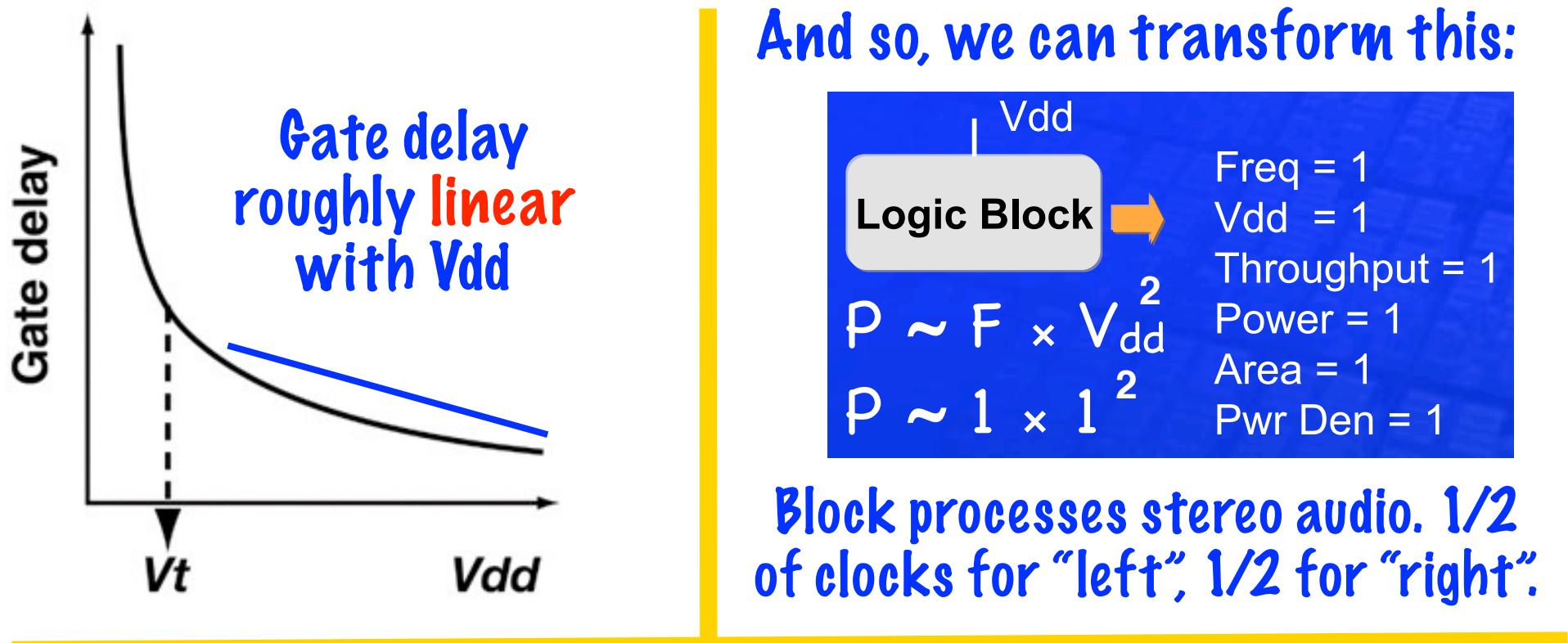


Some low-power design techniques

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

Trading Hardware for Power

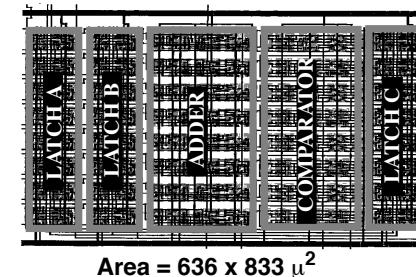
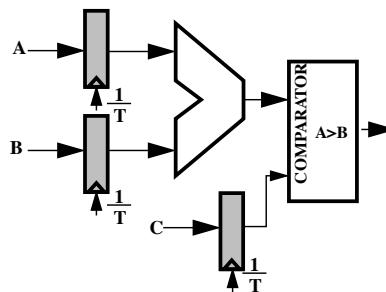
via Parallelism and Pipelining ...



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

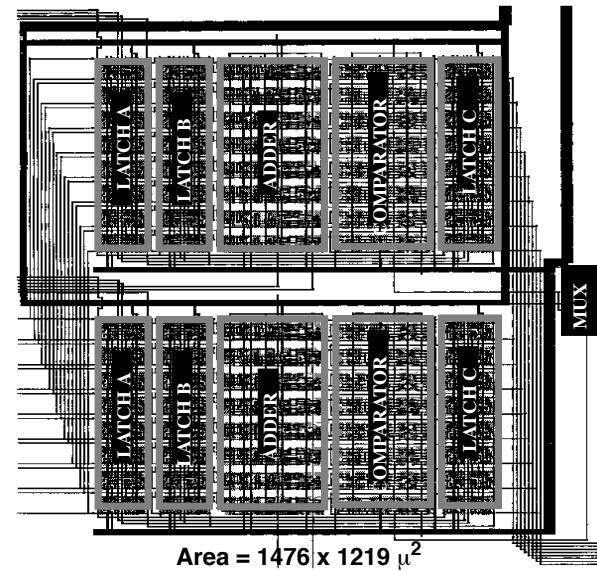
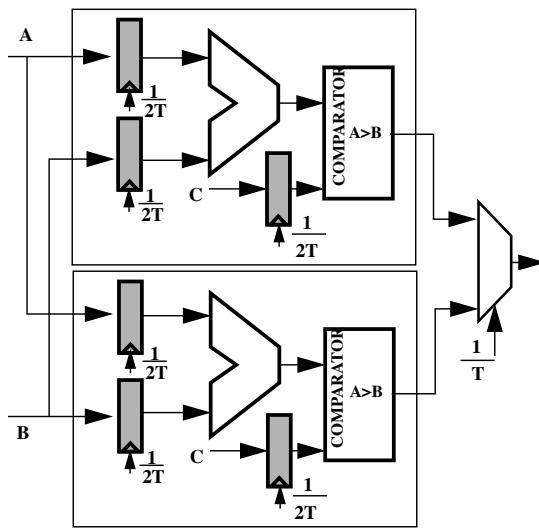
Chandrakasan & Brodersen (UCB, 1992)

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



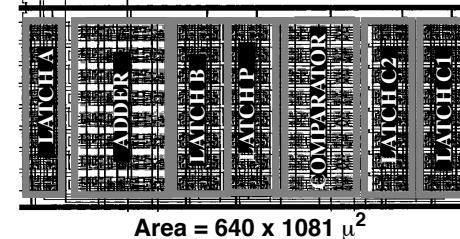
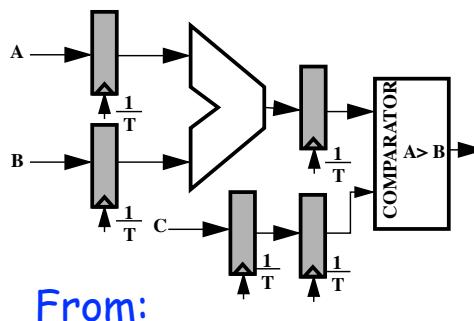
Simple

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



Parallel

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



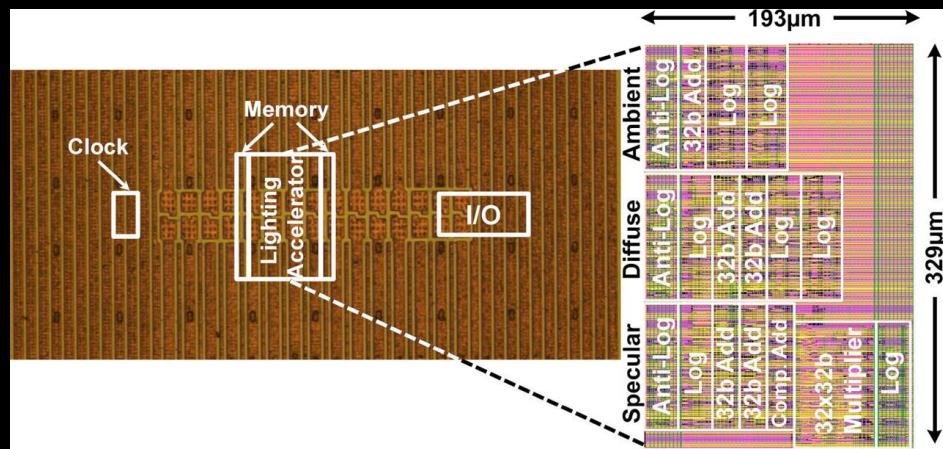
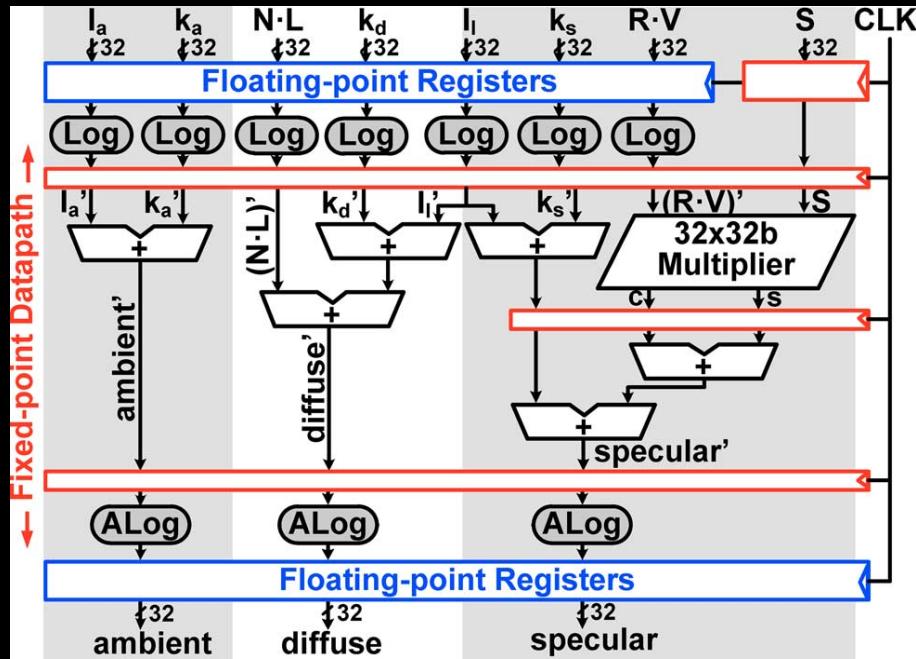
Pipelined

Example: Intel Graphics Pipeline IP



Phong Illumination (PI):

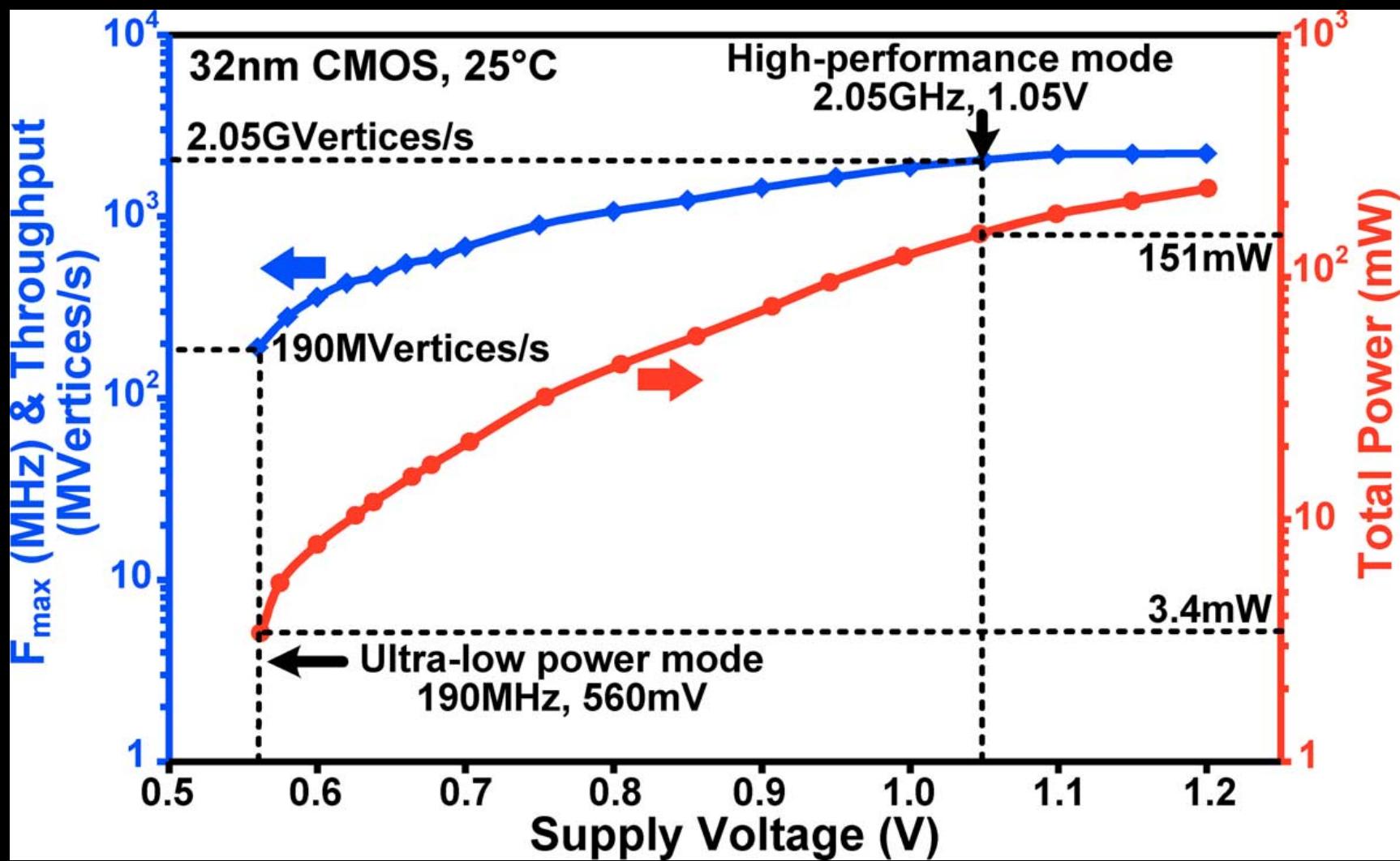
$$I = \underbrace{(k_a \times I_a)}_{\text{ambient}} + \sum_{i=1}^M \underbrace{(k_d \times I_{l,i} \times N \cdot L_i)}_{\text{diffuse}} + \underbrace{(k_s \times I_{l,i} \times (R_i \cdot V)^S)}_{\text{specular}}$$



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*

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.

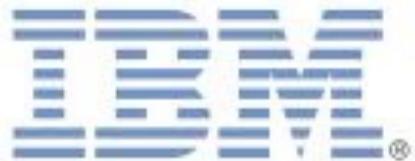
Reducing both F and V_{dd} , reduces power but also improves energy efficiency (total energy for computation is less).

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

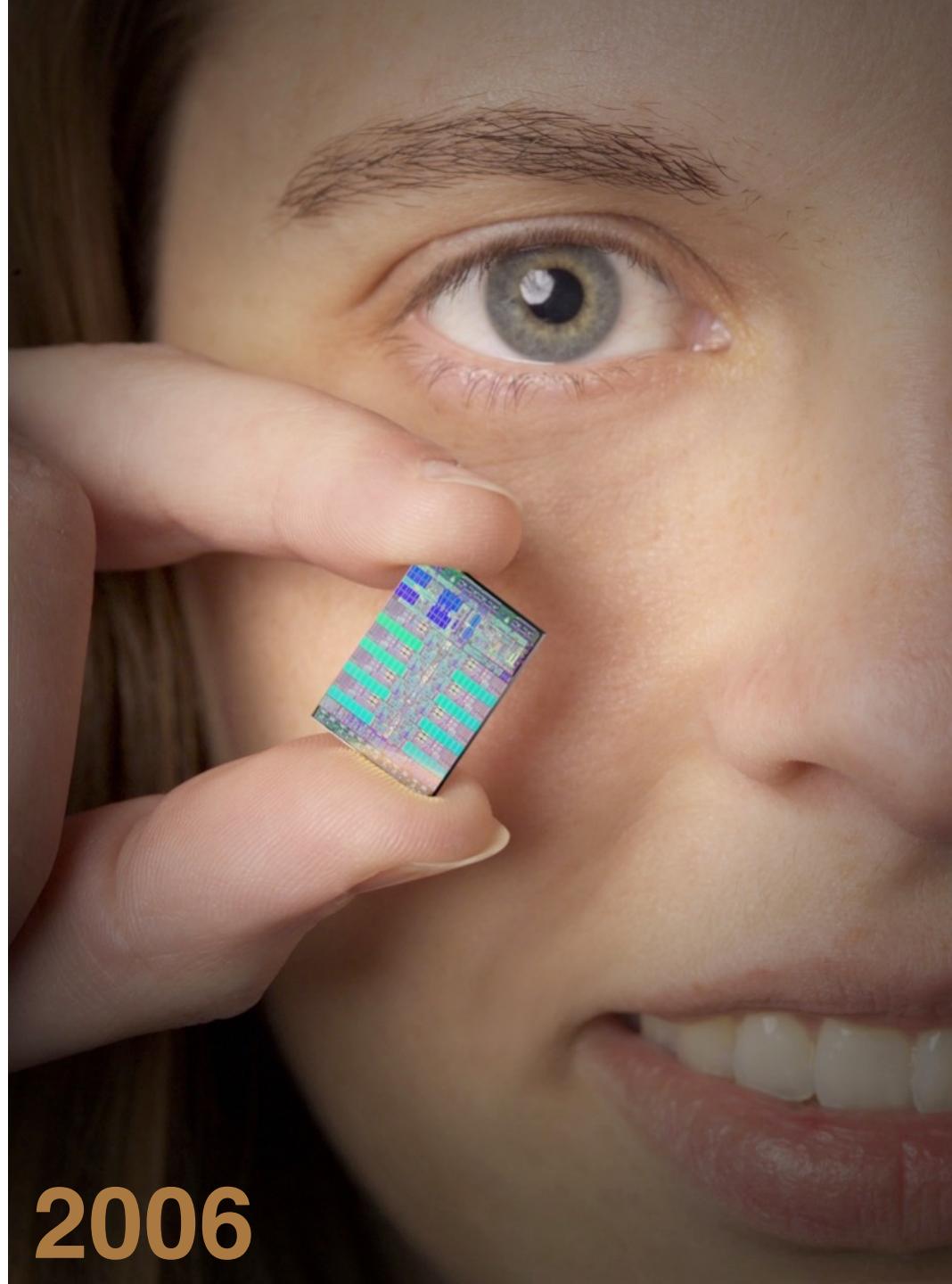
Multiple Cores for Low Power

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

Cell: The PS3 chip



TOSHIBA



2006

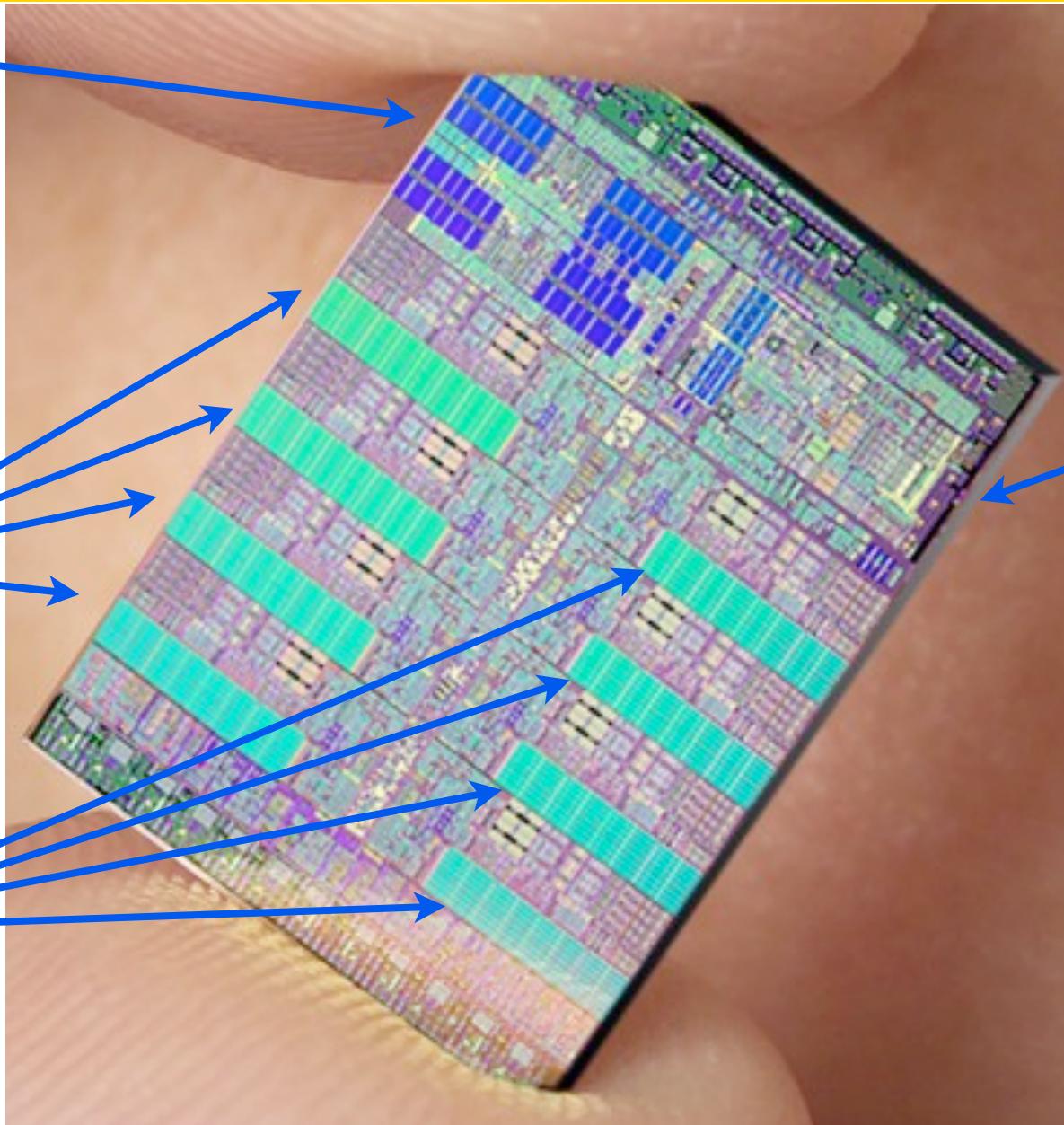
Cell (PS3 Chip): 1 CPU + 8 “SPUs”

L2 Cache

512 KB

8

Synergistic
Processing
Units
(SPUs)



PowerPC

IBM

SONY



COMPUTER
ENTERTAINMENT

TOSHIBA

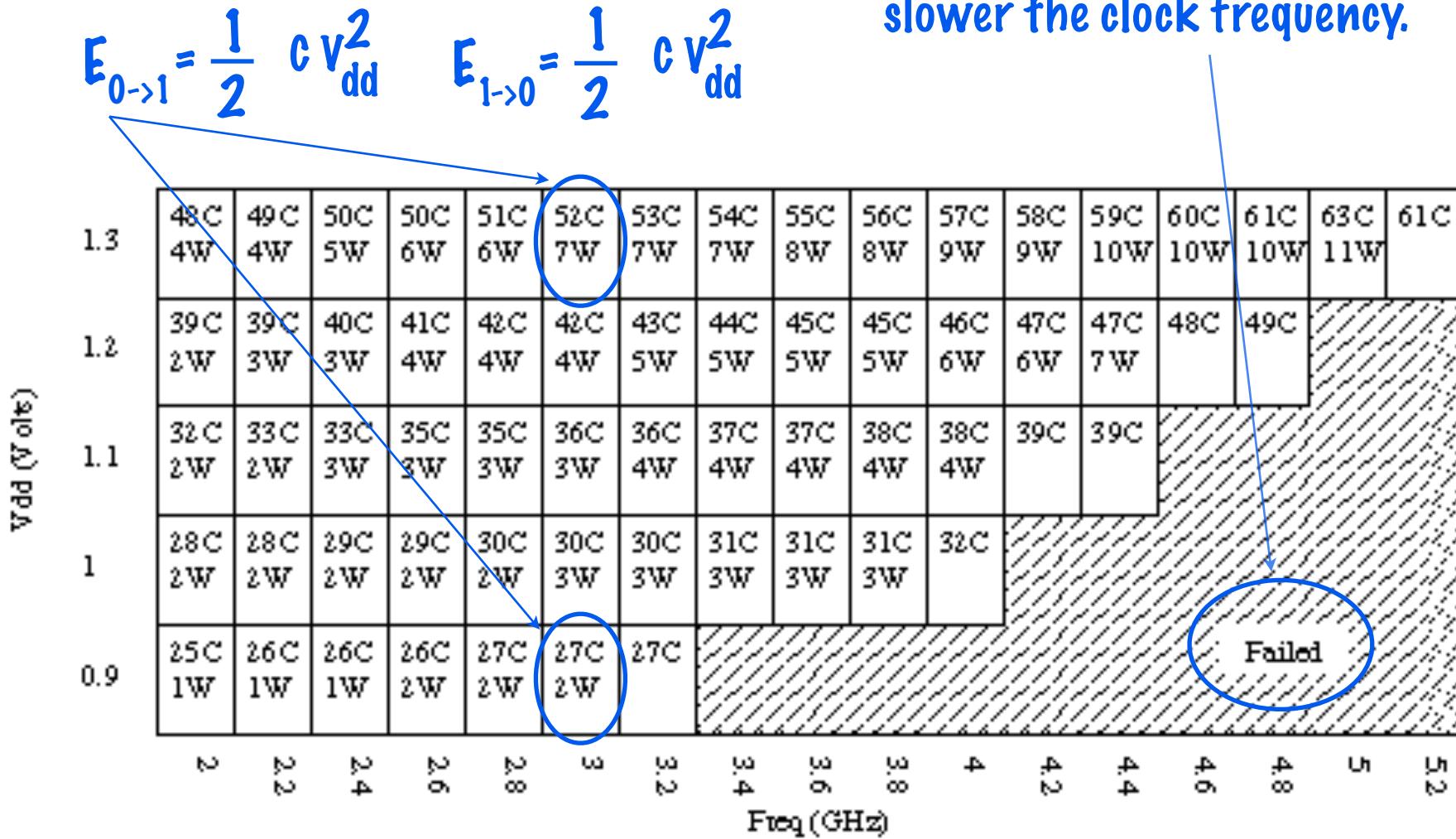
A “Schmoo” plot for a Cell SPU ...

The lower Vdd, the less dynamic energy consumption.

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

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

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

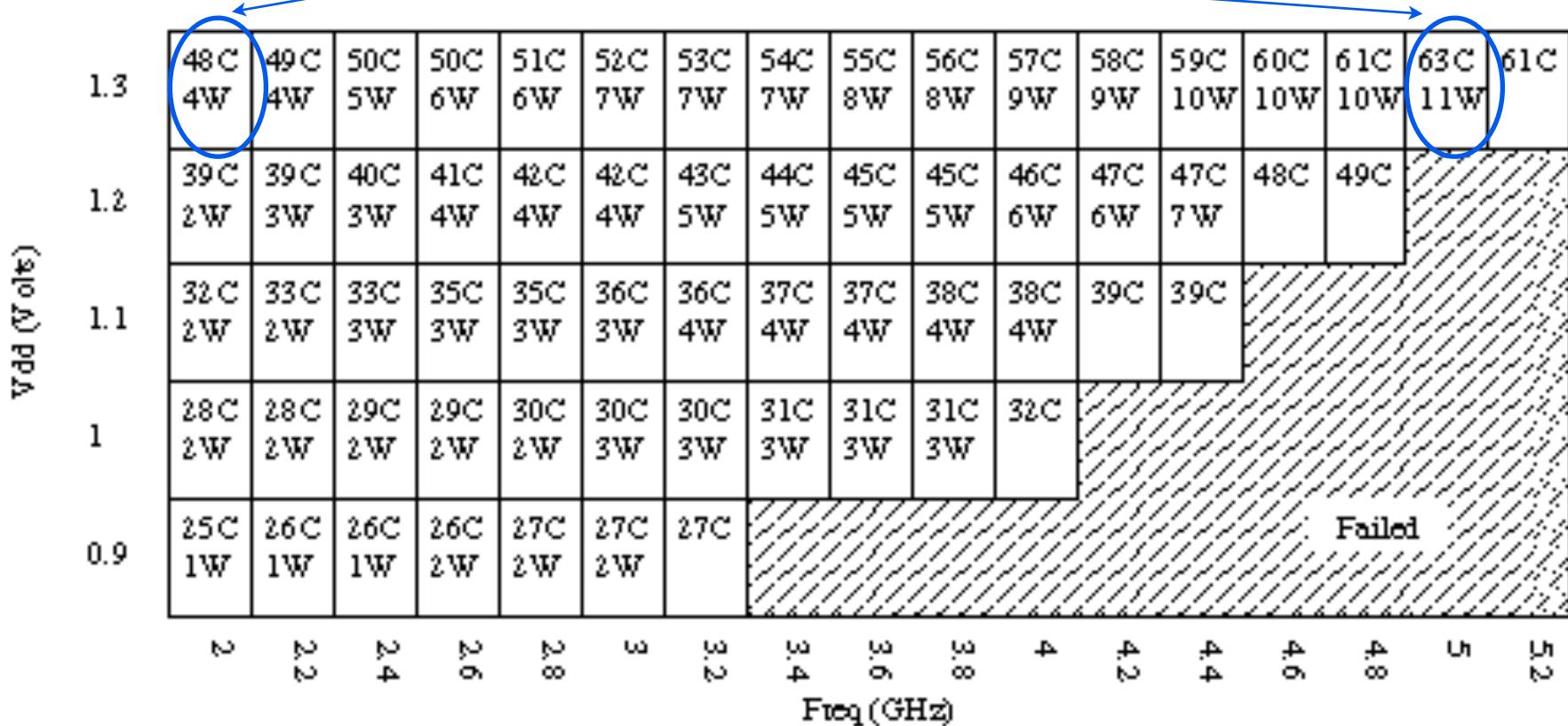


Clock speed alone doesn't help E/op ...

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

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

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

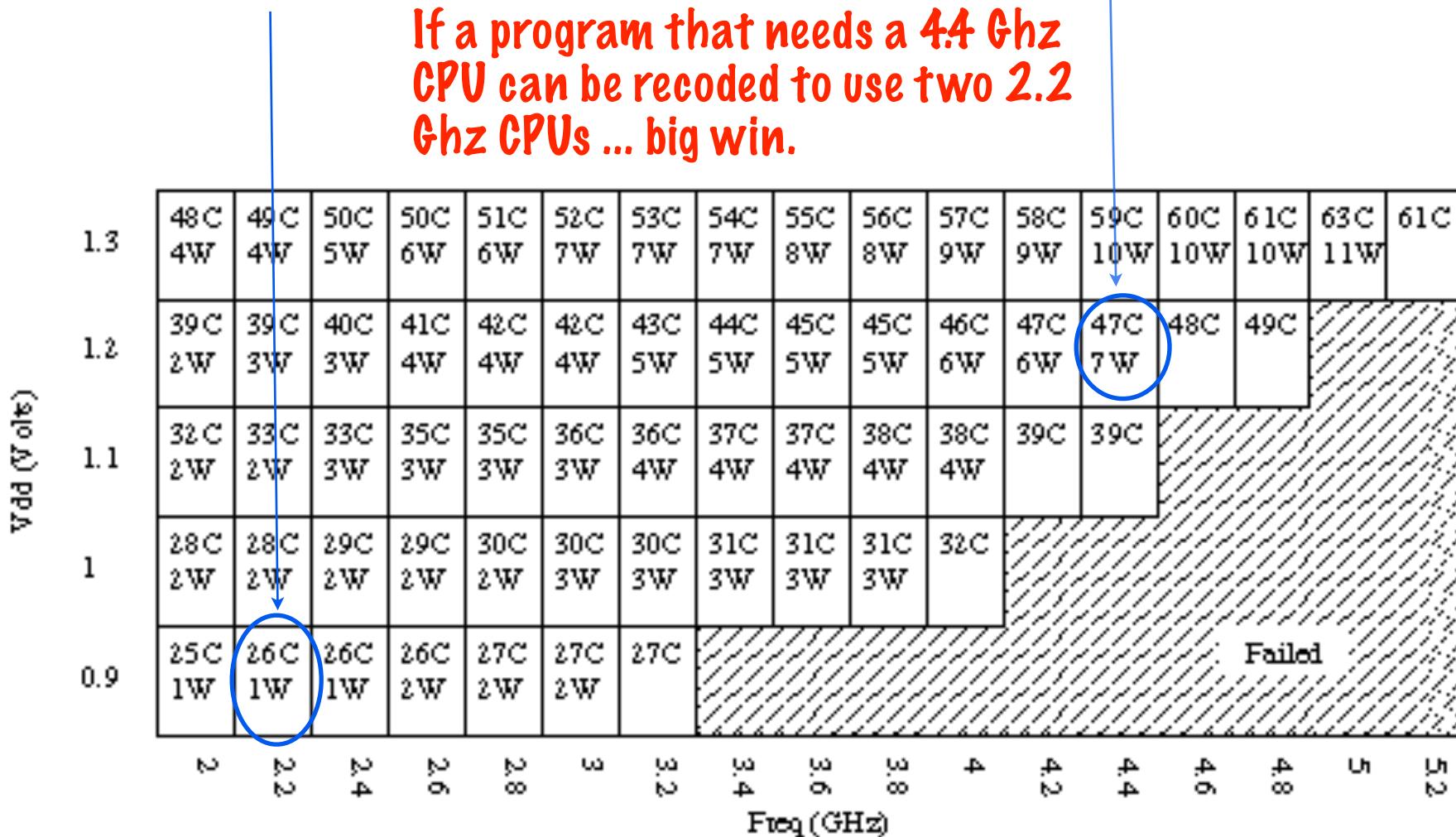


Scaling V and f does lower energy/op

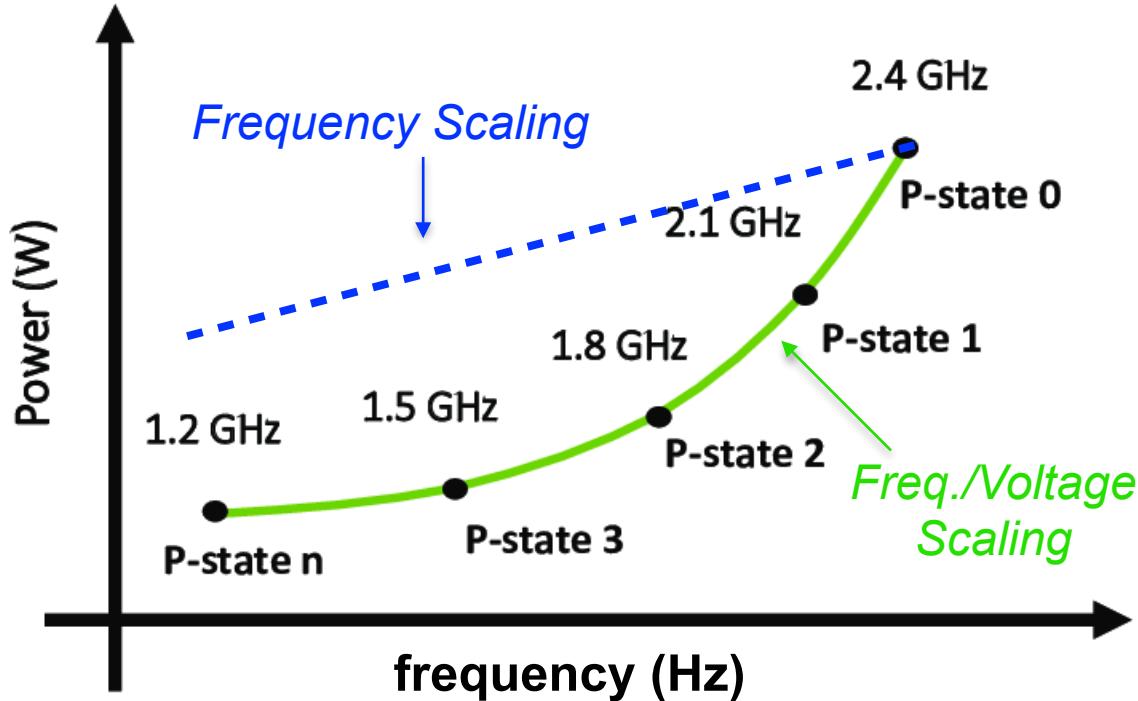
1 W to get 2.2 GHz performance. 26 C die temp.

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.



Dynamic Voltage/Frequency Scaling (DVFS)



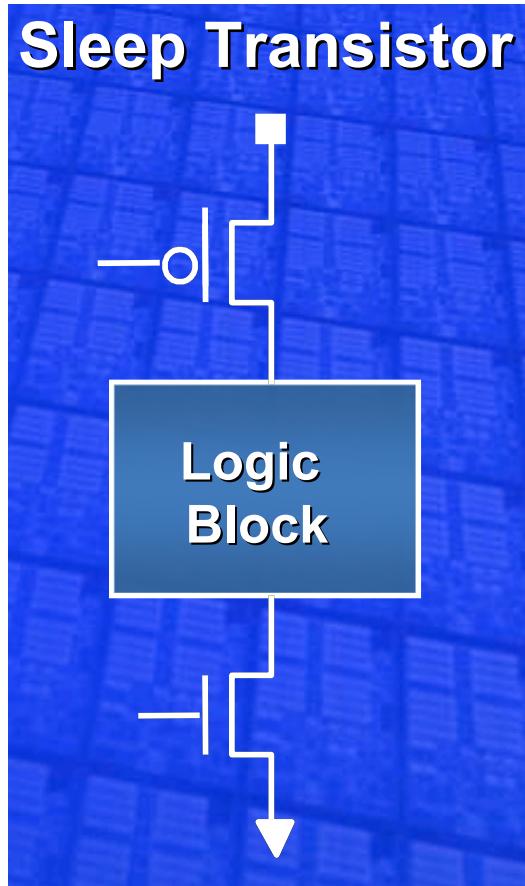
Many modern processors have controls for dynamically changing operating frequency and voltage.

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 ...



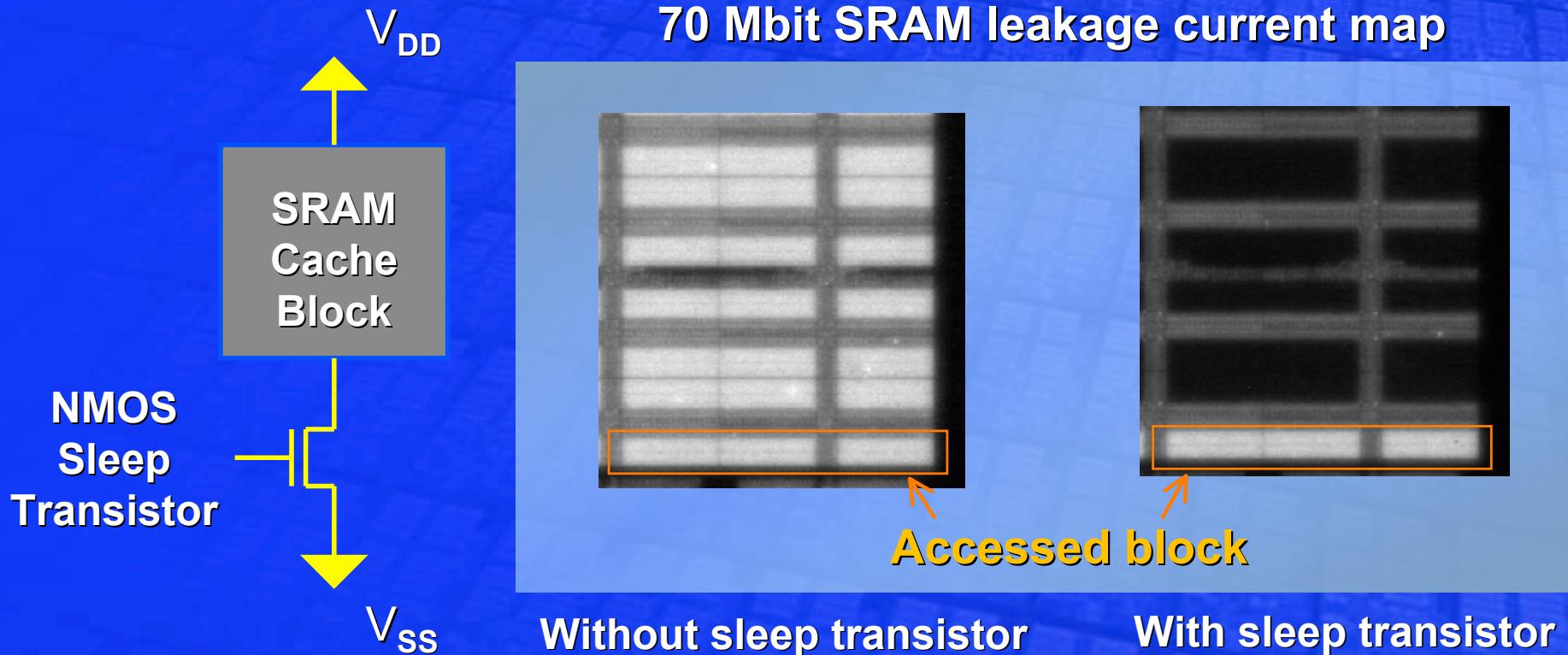
Example: Floating point unit logic.

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

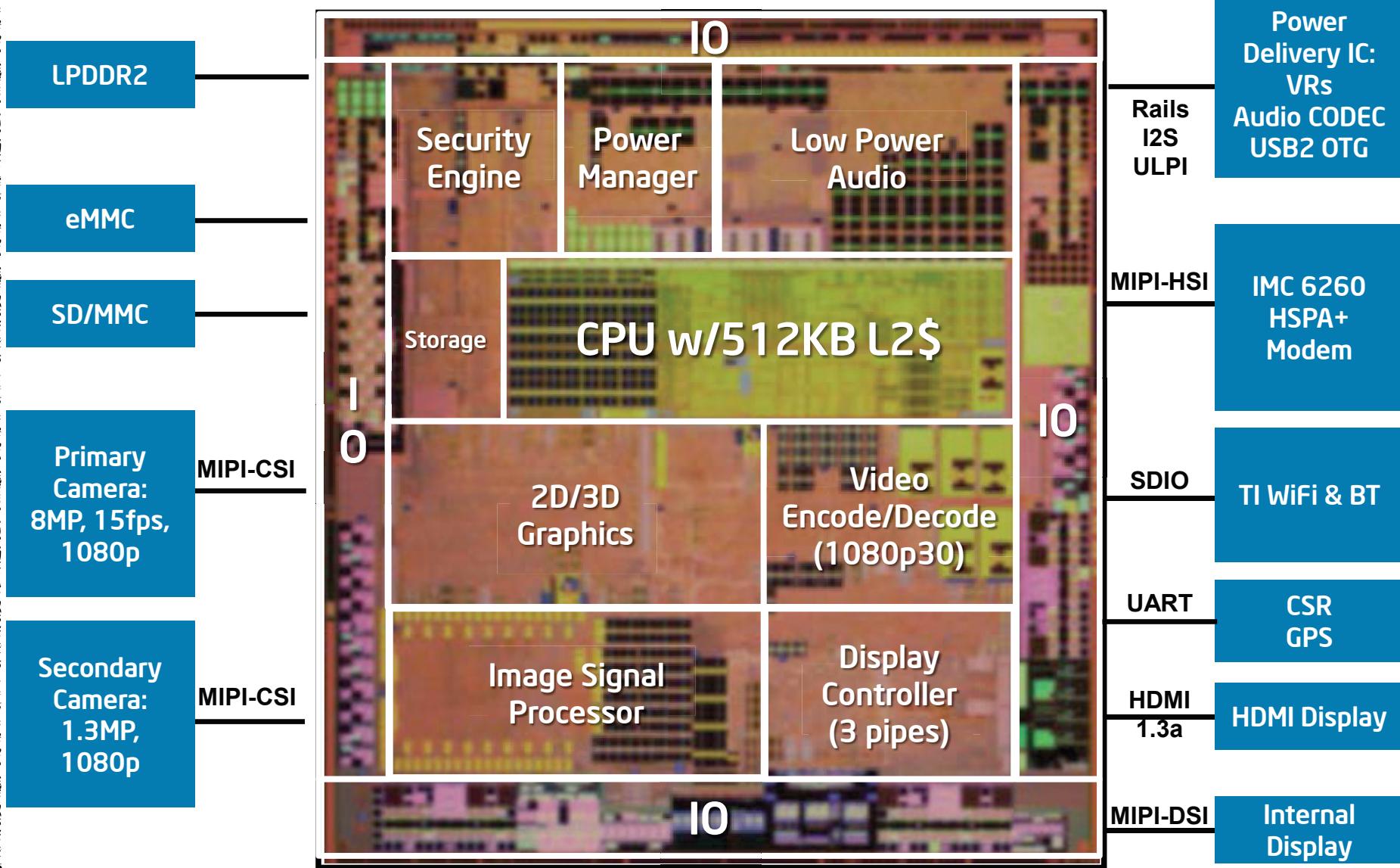


>3x SRAM leakage reduction on inactive blocks

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

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

Intel Medfield



Intel Medfield

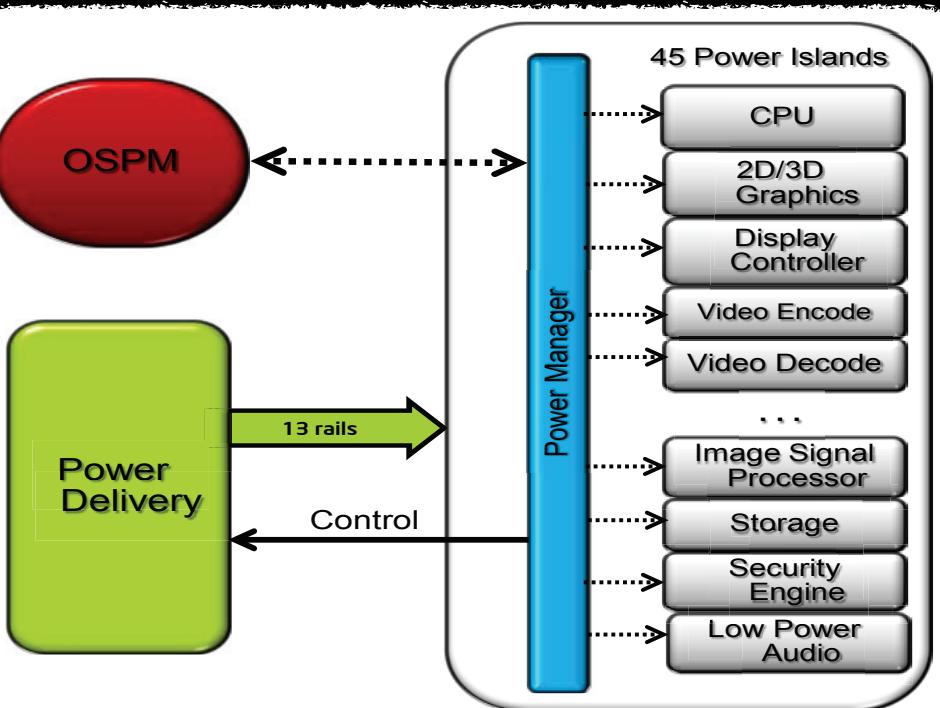
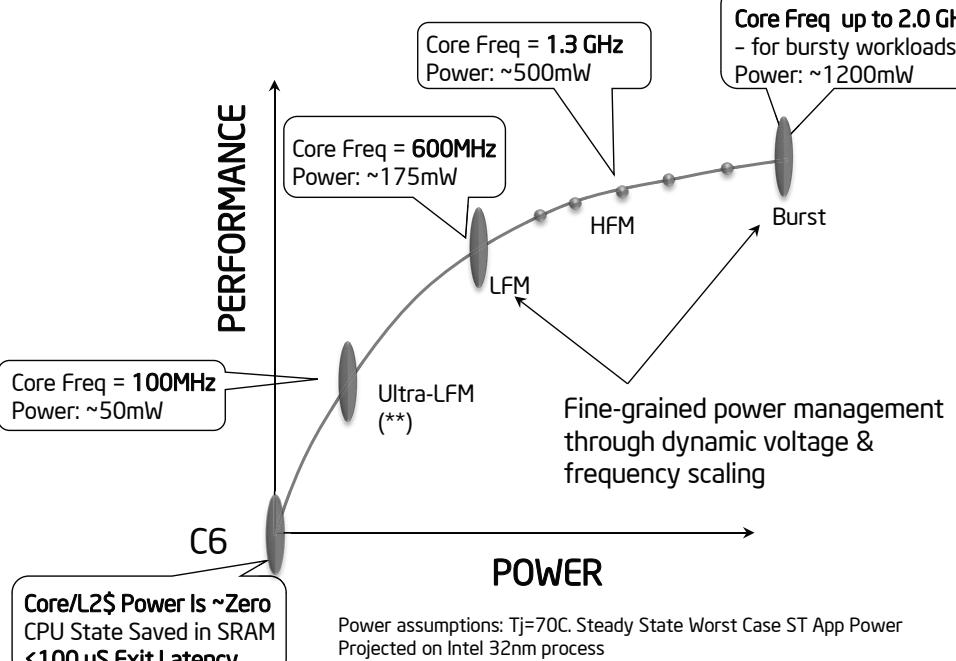
Switches 45 power "islands."

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

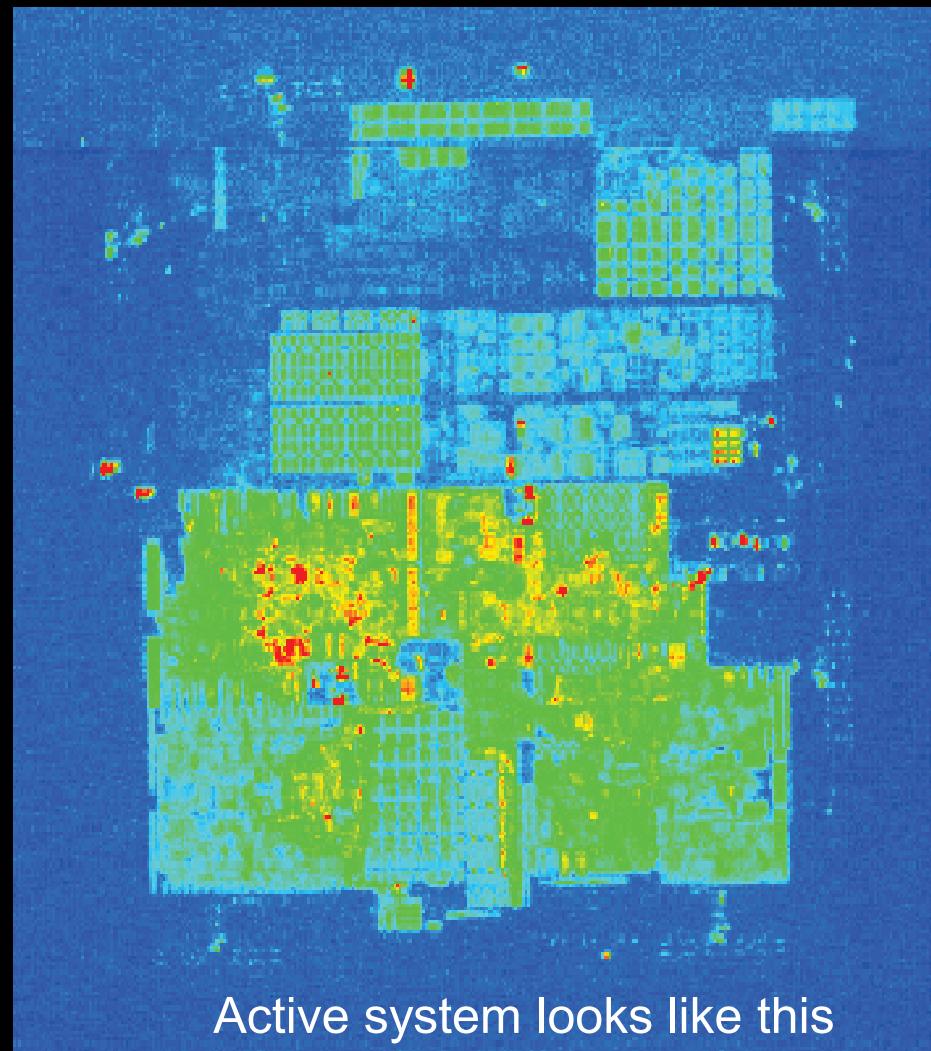
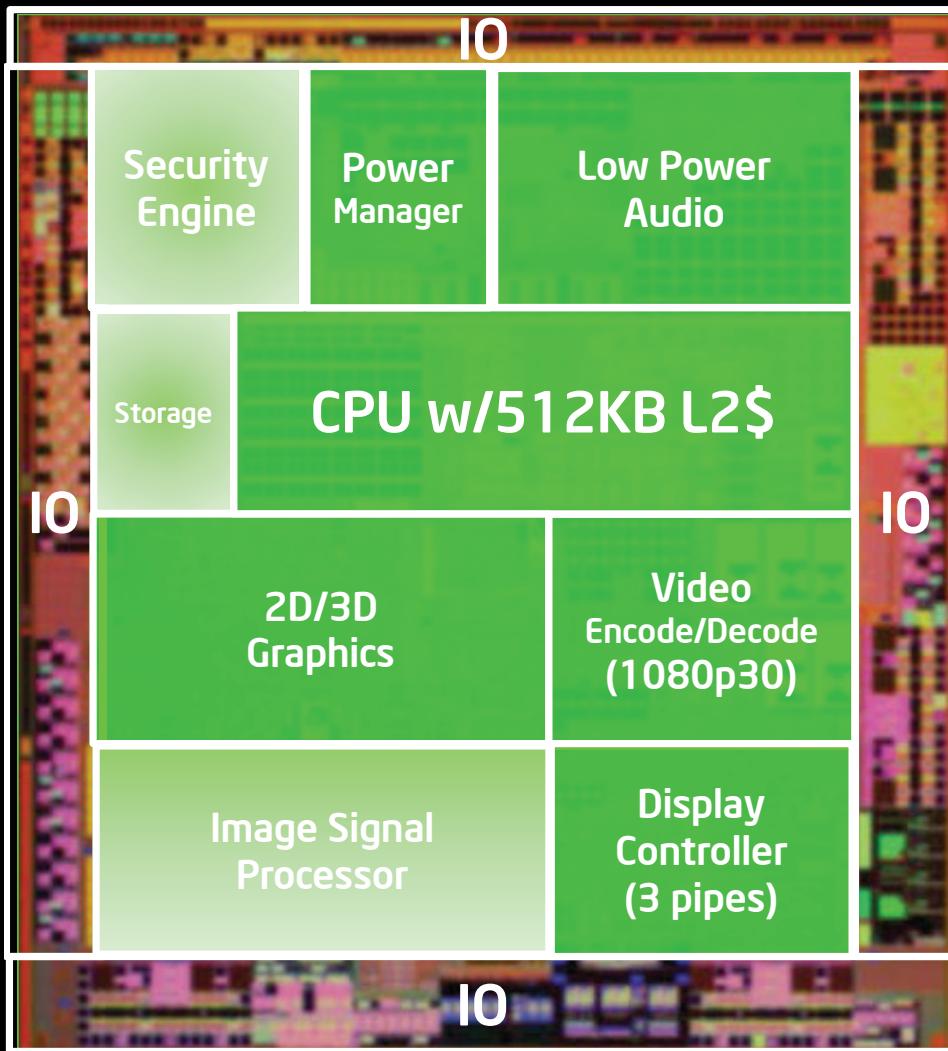
"Race to idle" strategy -- finish tasks quickly, to get to power down.



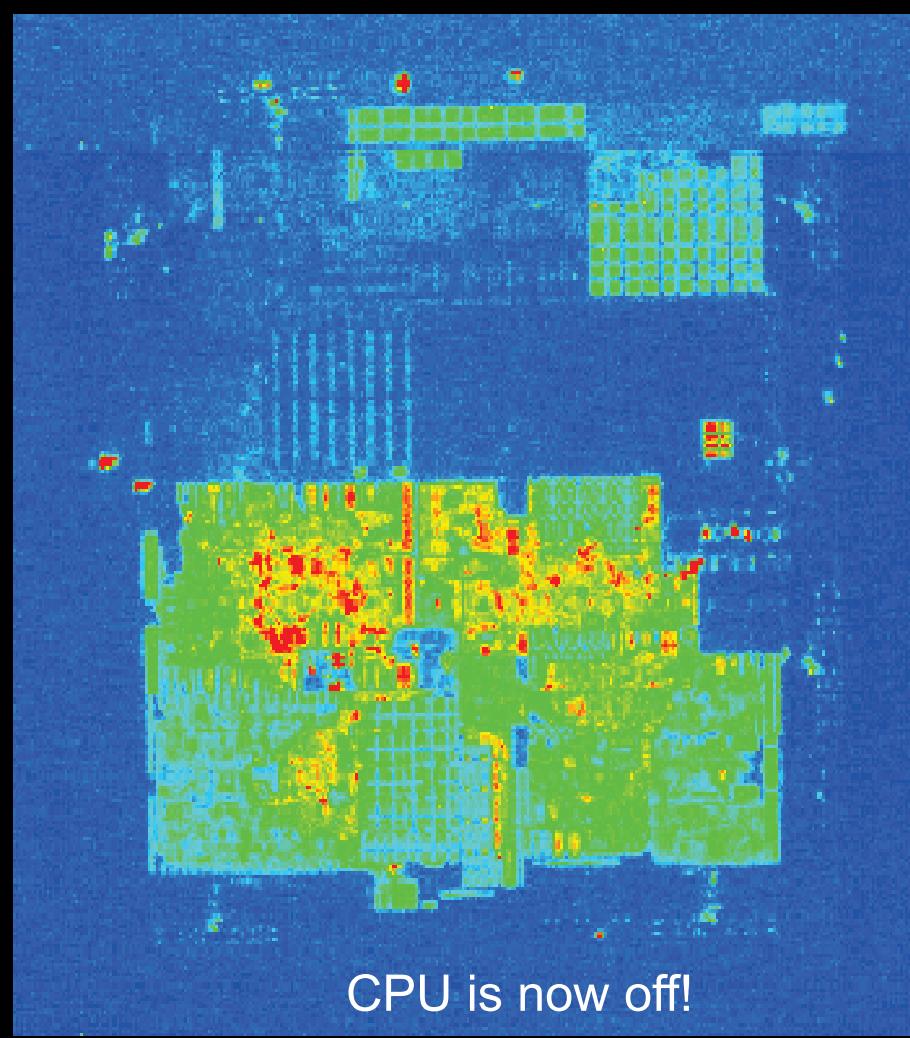
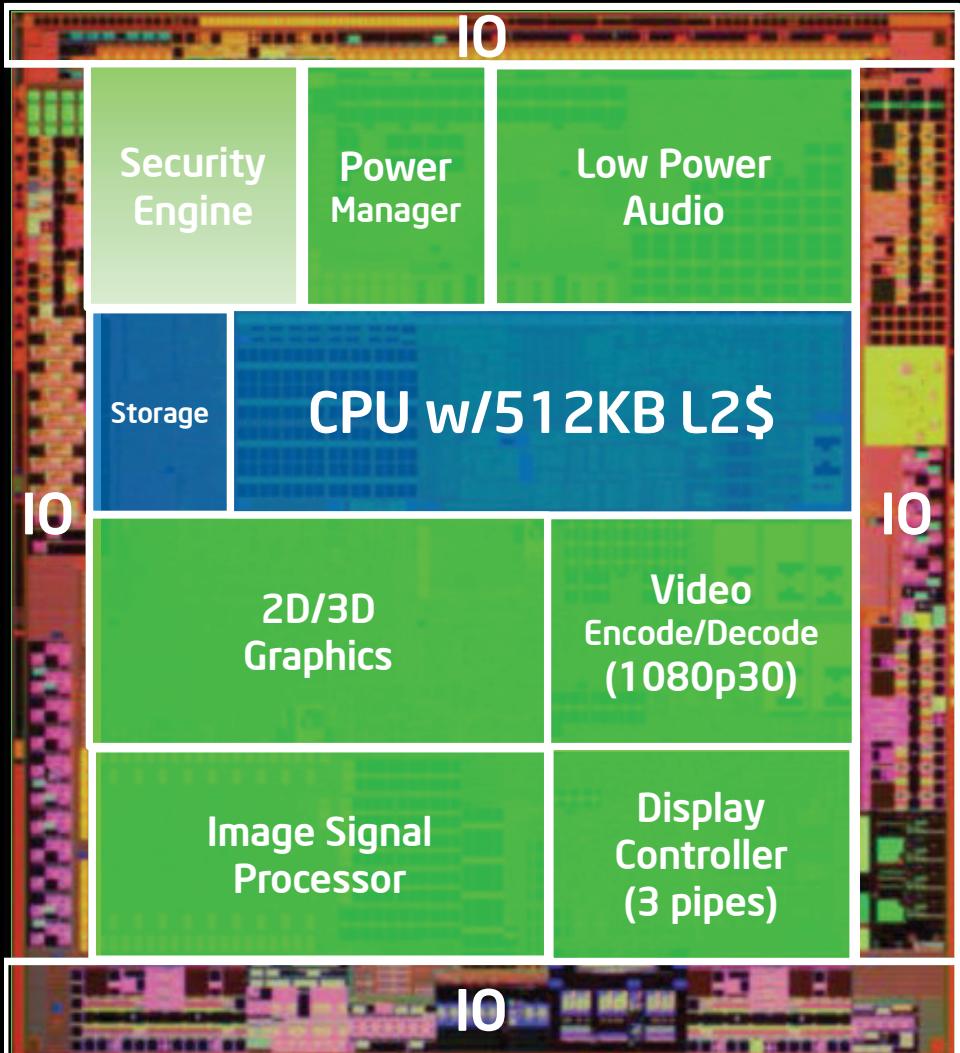
Intel® Smartphone Reference Design



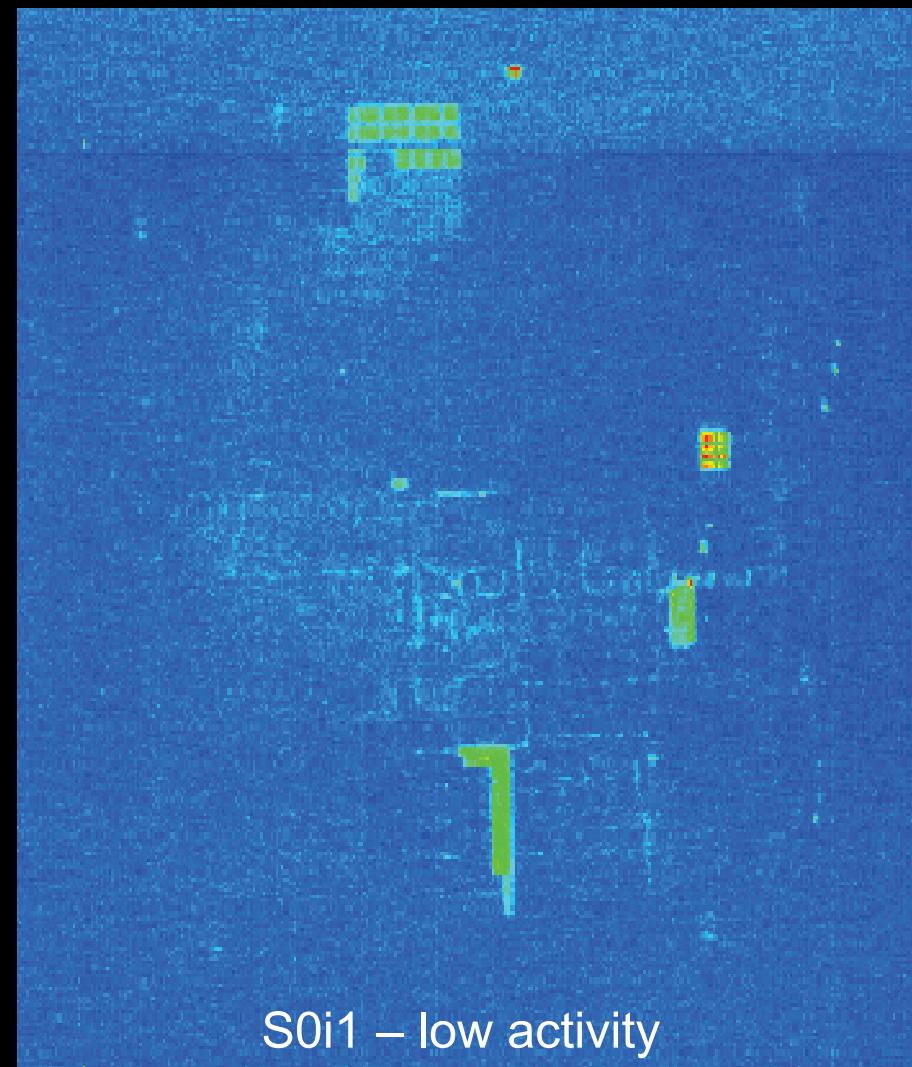
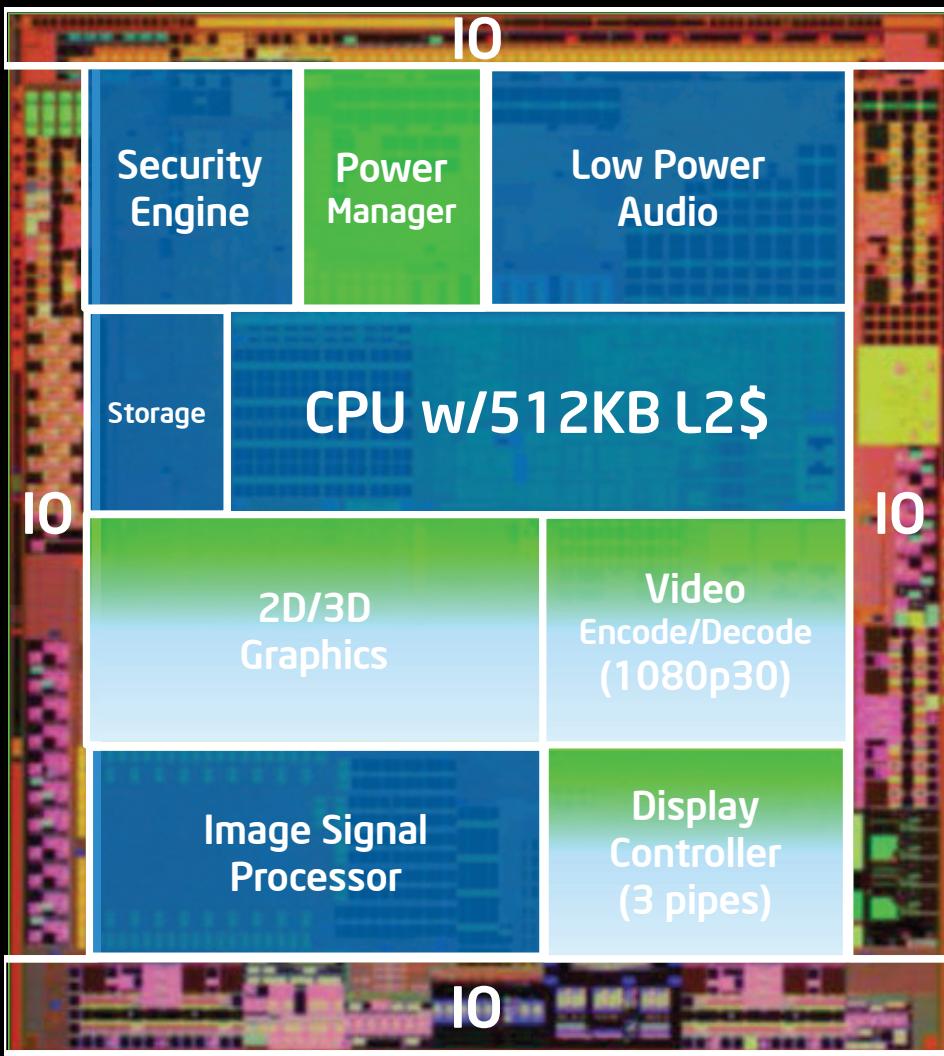
Playing a game ...



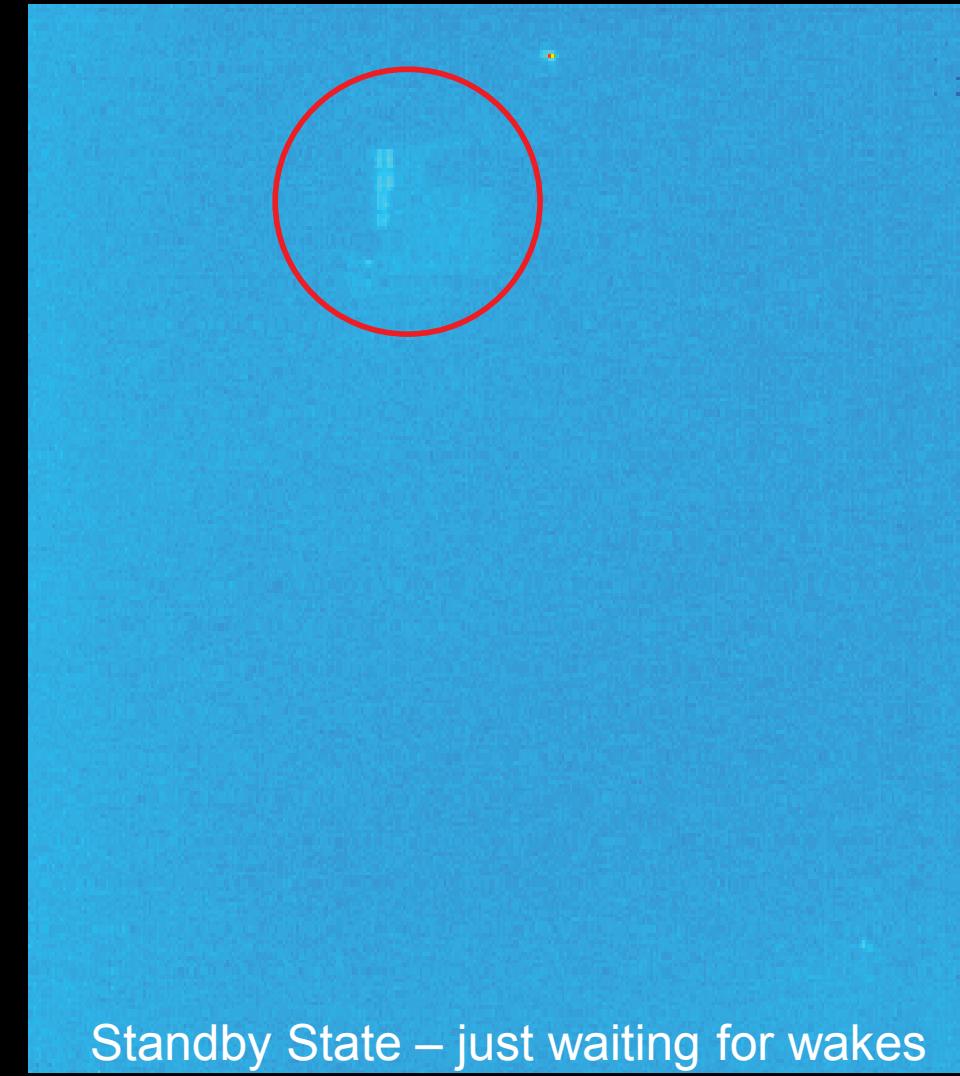
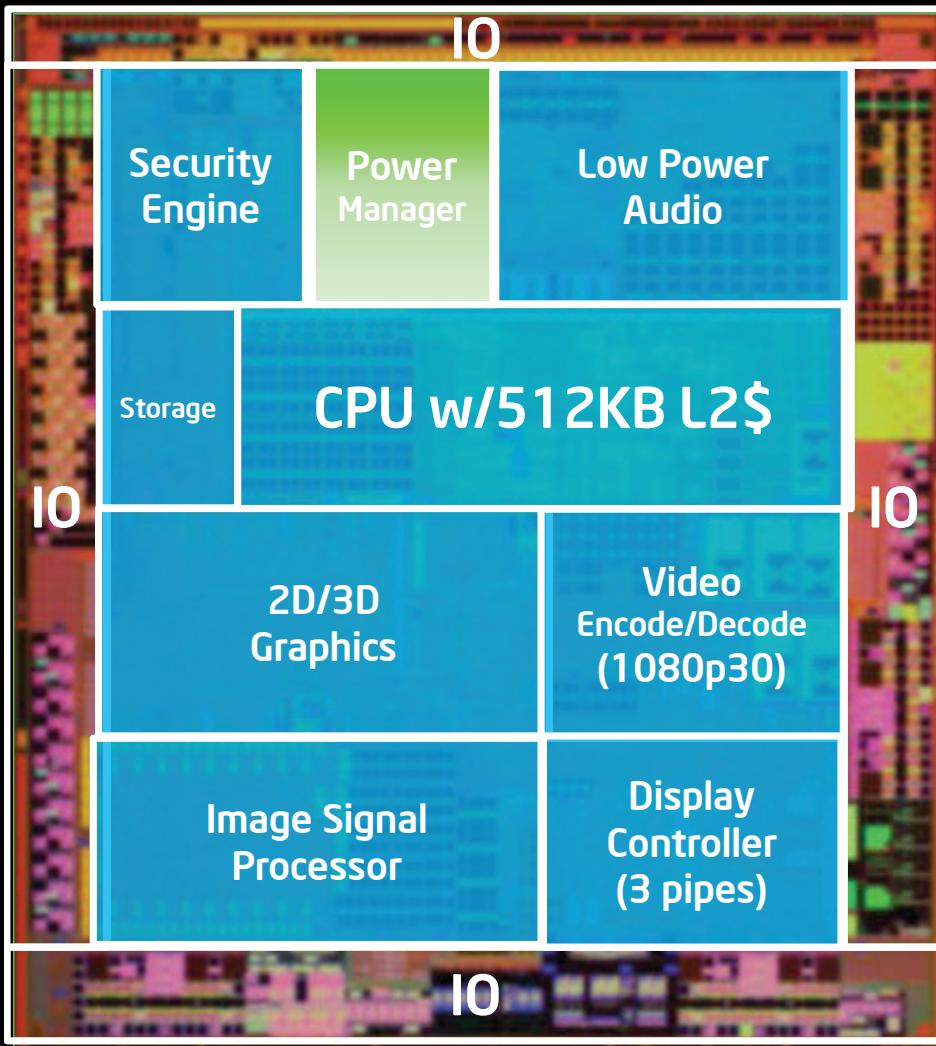
Watching a video ...



Looking at phone screen, not doing anything ...



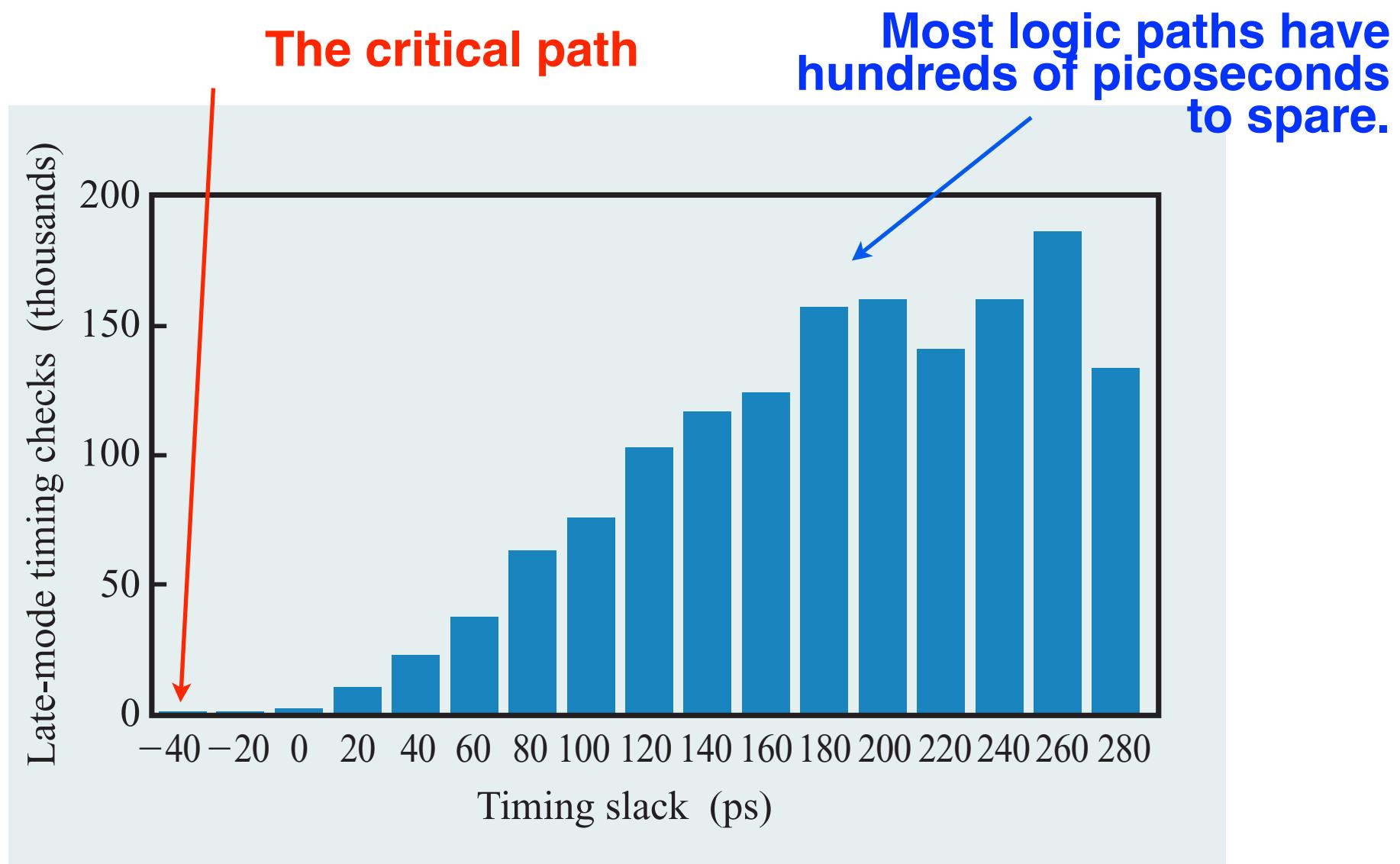
Phone in your pocket, waiting for a call ...



Standby State – just waiting for wakes

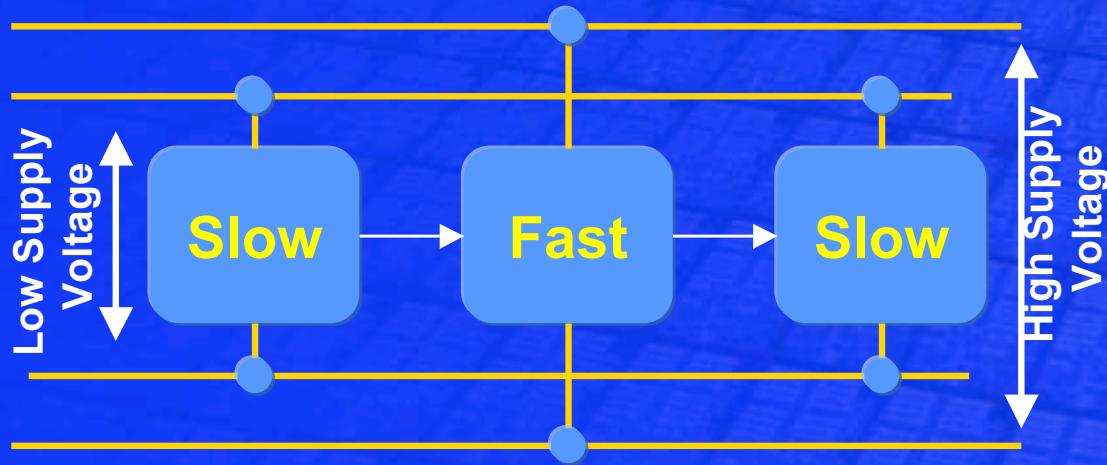
Slow down “slack paths”

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



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.

Thermal Management

Keep chip cool to minimize leakage power

Figure 3

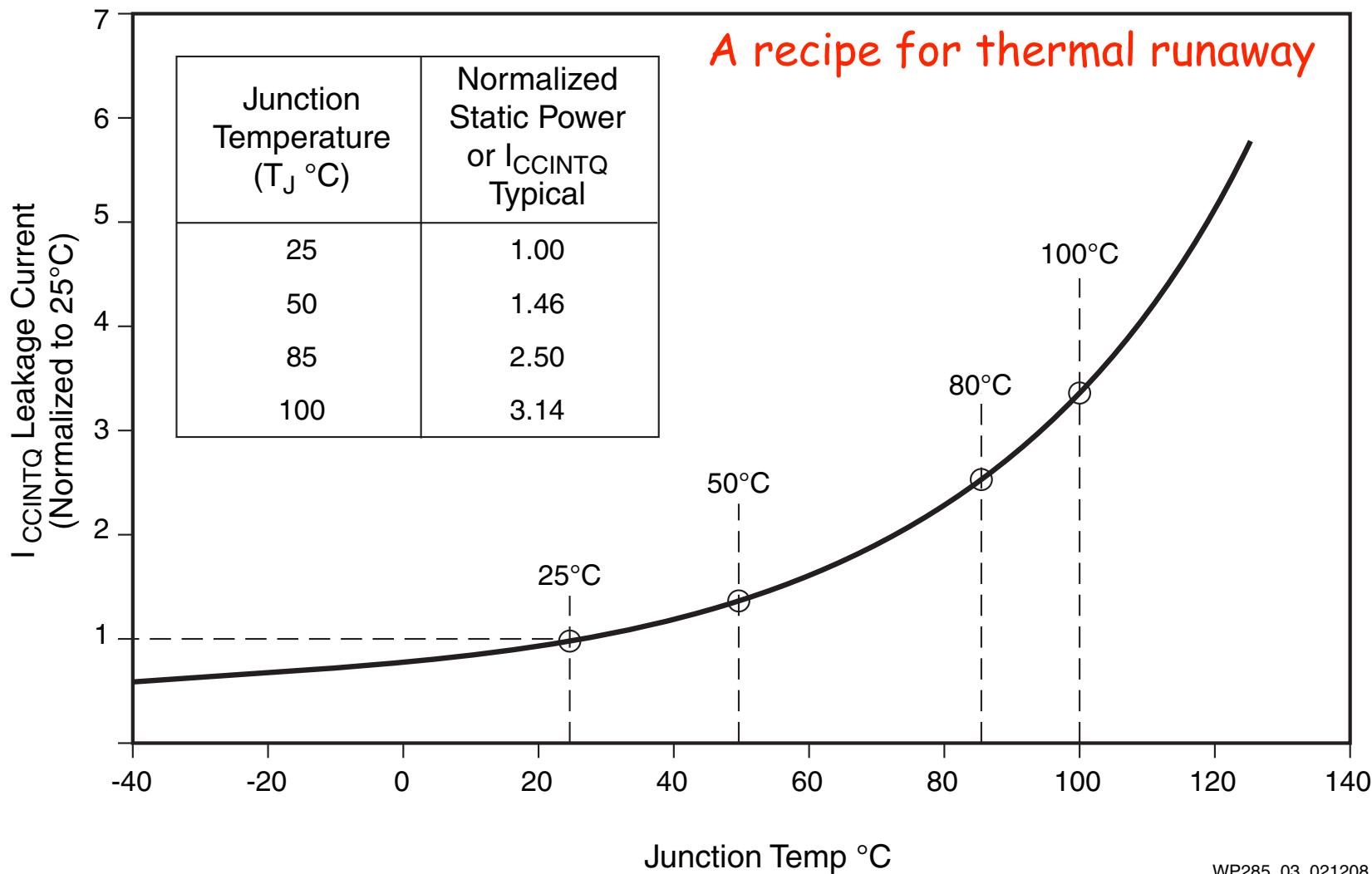


Figure 3: I_{CCINTQ} vs. Junction Temperature with Increase Relative to 25°C

WP285_03_021208

Intel realtime temp monitoring

