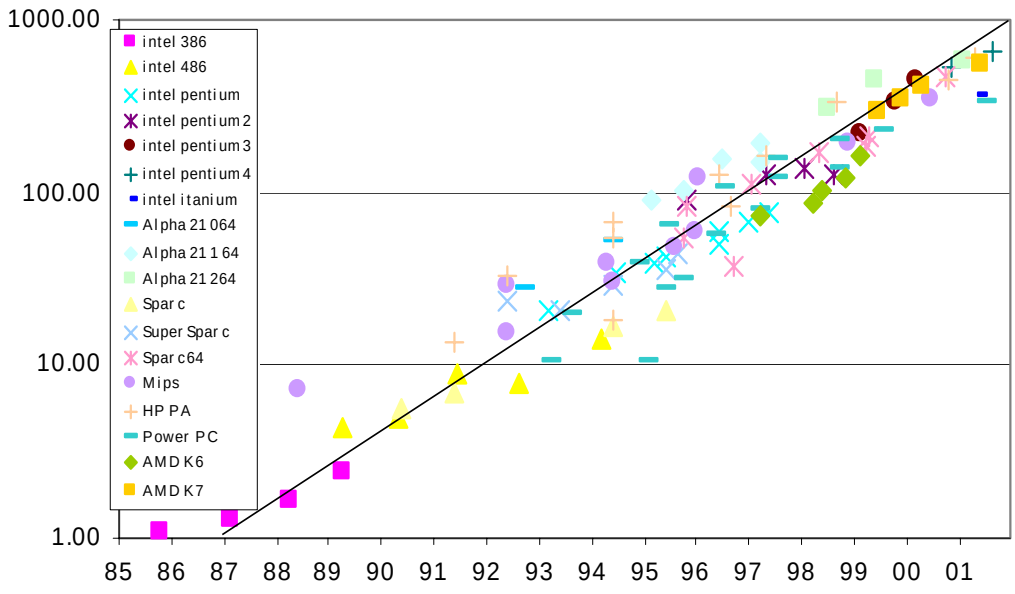
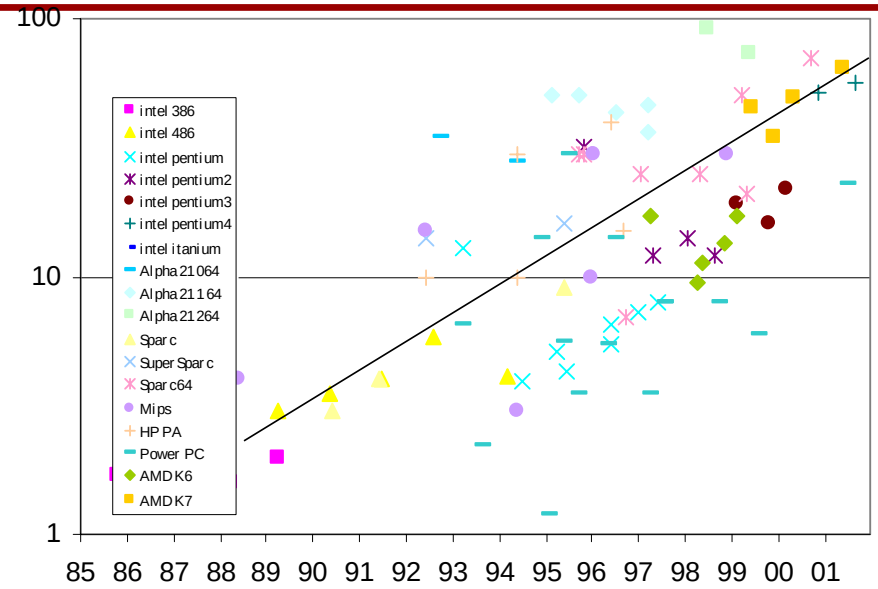
Power in Digital CMOS Circuits

Mark Horowitz
Computer Systems Laboratory
Stanford University
horowitz@stanford.edu

Fruits of Scaling – SpecInt 2000



The Darker Side of Scaling - Power



- At least it is scaling slower than performance

Power is Important

Three reasons we care about power:

1. Need to get the power into the chip
   - 60W@ 1.2V is many Amps
2. Need to get the power out of the chip
   - How low thermal resistance is possible?
   - Plastic packages w/o forced air
     - High thermal resistances
3. Energy is heavy
   - Need to carry the energy
   - 20Wh/lb

Important Questions

- How did we end up in this situation?
  - Power was not an issue in the mid 80s
  - Scaling theory said power would be constant
  - Energy efficiency would improve
- Is there any hope for the future?
  - More issues that need to be addressed?
  - Silver bullets to solve the power problem?
  - Techniques that will help

Power in CMOS Circuits

- Dynamic power
  - Proportional to C Vswing Vdd F
  - Dominates most circuits
- Static power
  - Idc (usually leakage current now) *Vdd
  - Has been very small, and is still small
  - Issue when circuit is idle and dynamic power is zero

Historical Power Scaling

- In current technology shrinks, X,Y, V all scale
  - Implies that C also scales
- If scale a chip to a new technology, operate at F
  - C, and V both scale, so power DECREASES by α^3
  - Power decreases by 3x for each technology generation
- If scale a chip to a new technology, operate at F/α
  - C, and V both scale down, F scales up
    - So power DECREASES by α^2
  - 1.4 times faster chip, for ½ the power
    - Every time you move to a new technology

Power Scaling – Static Power

- X,Y, V all scale.
  - Gate oxide is scaling faster than α
    - Now have leakage current through the gate
    - GIDL is another issue
- Scaling Vth
  - Performance depends on Vdd/Vth ratio
  - Transistor leakage tied to Vth
  - Leakage current increases exponentially

$$I_{leak} := I_s \cdot e^{\frac{-V_{th}}{\alpha v_T}}$$

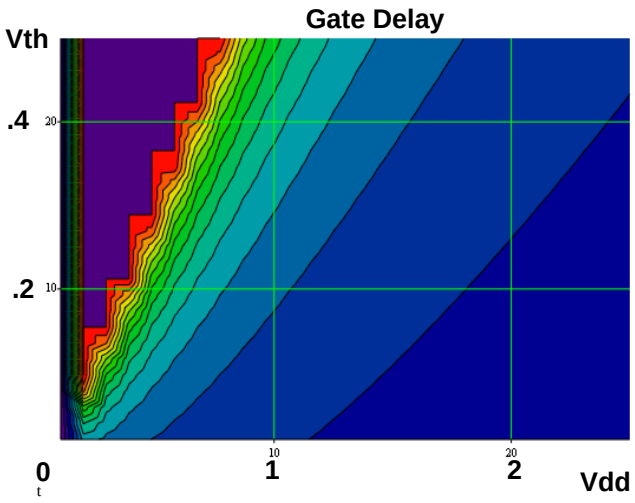
- If Vth does not scale
  - Leakage power scales as α^2

New Problem – Leakage

- Scaling Vth
  - From .5μ generation, Vdd = 10V/μ *L
    - Seems like this scaling is still on track
  - Vth was Vth = 1.4V/μ *L
    - But that would mean that 0.18μ has 250mV Vth
    - Leakage issues
  - Vdd/Vth of gates are starting to fall
    - Performance of gates will drop
      - Haven't seen this yet, since technologies are pushed
    - Variability of delay will increase
      - Subtracting two large numbers to get overdrive
- Additional leakage paths – gate tunnel current

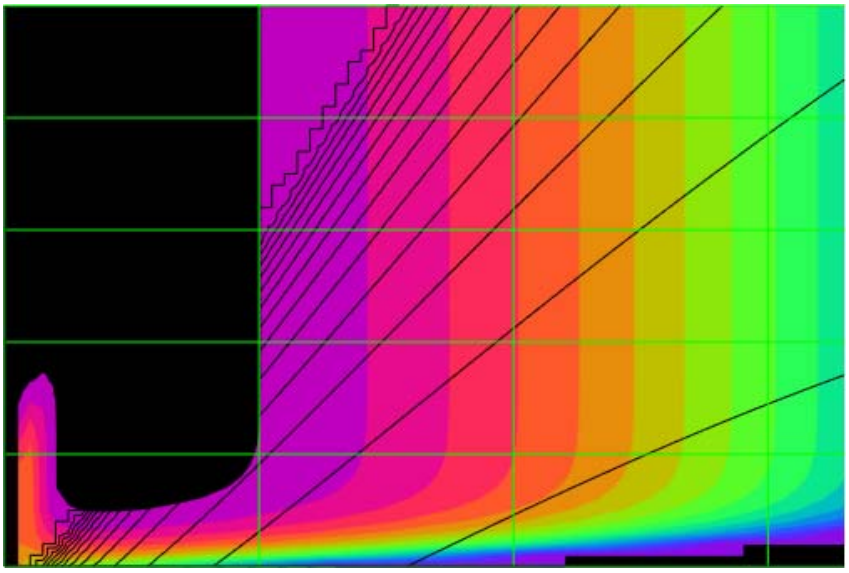
Dilemma

- What Vdd and Vth
- Lower Vth
  - Need less Vdd
  - Less dynamic power
  - More leakage current
- Correct choice depends on operation condition
- Also at low Vdd to Vth ratios the variation in delay will be larger



Energy-Delay Trade-offs

- Placed lines of constant delay on top of a contour plot of energy
- Much lower energy/op if operate at low Vdd, Vth
- If, of course, the circuit is active



E, t

Do We Have a Power Problem?

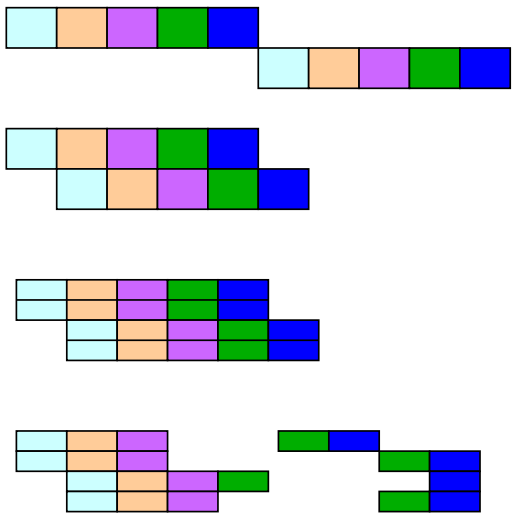
- That depends on your point of view:
  - Cost per operation is MUCH cheaper at same F
    - Look at what we do in cell phones / laptop computers
    - This cost will continue to fall
  - Cost per operation is cheaper even at higher F
    - Get this reduced cost even when you run part faster
  - But we are greedy
    - Want machine to run faster than technology scaling
    - Build the most complex machines possible
    - This combination means power has been growing.
- Power and performance are related
  - Faster often means much more power

Future Scaling Will Be Much Worse

- If Vdd does not scale
  - Energy/gate scales only as α
  - This is because C scales
- This means that:
  - Power of gate will be constant if F increases

Architecture

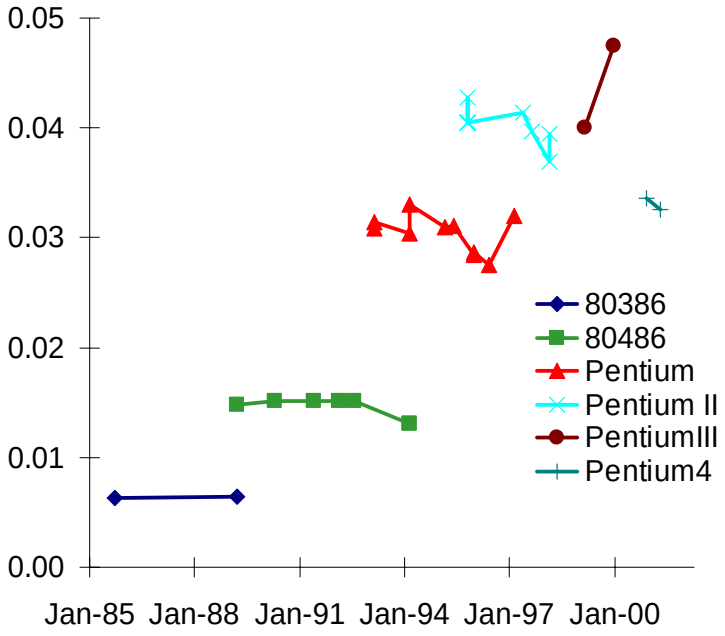
- Convert transistors to performance
- Use transistors to
  - Exploit parallelism
  - Or create it (speculate)
- Processor generations
  - Simple machine
  - Pipelined
  - Super-scalar
  - Out-of-order
  - Speculation



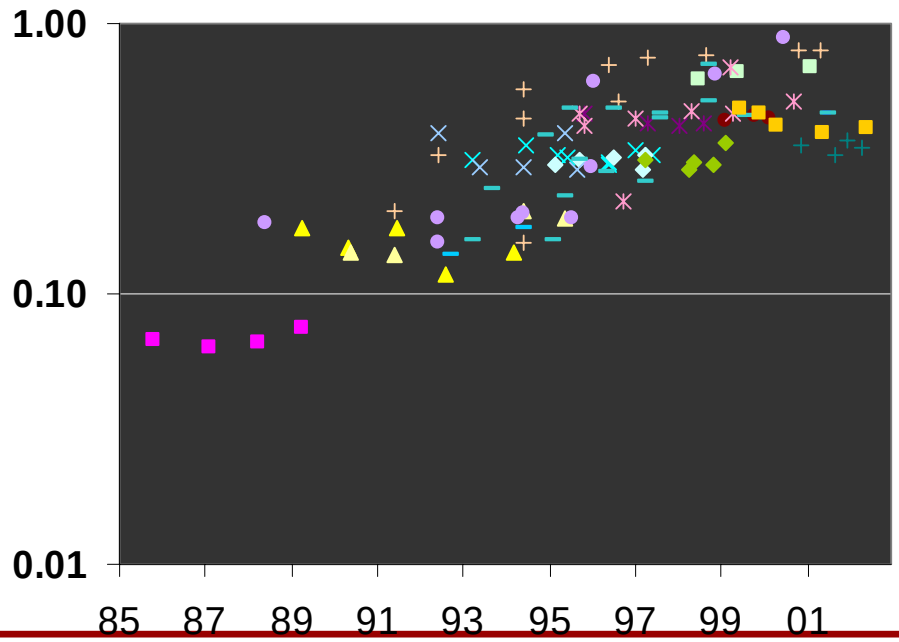
- Each design has more logic to accomplish same task

Architecture Scaling

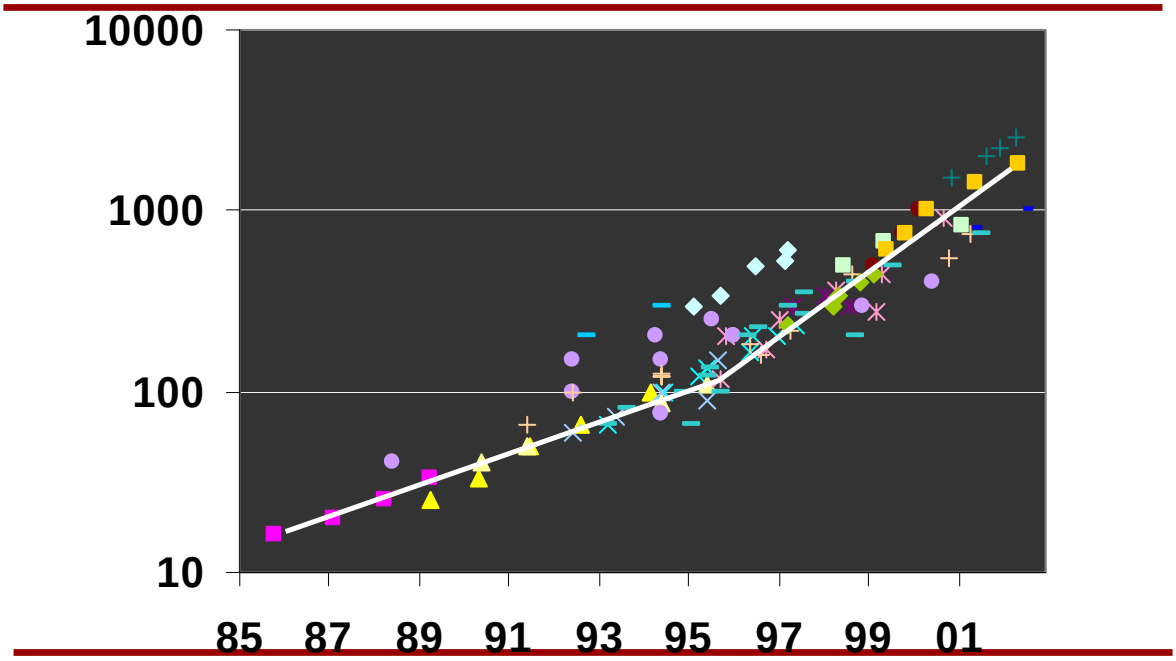
- Plot of IPC
  - Real IPC
  - Compiler
- Hardware
  - Grows rapidly
  - More FU
  - Wires don't shrink



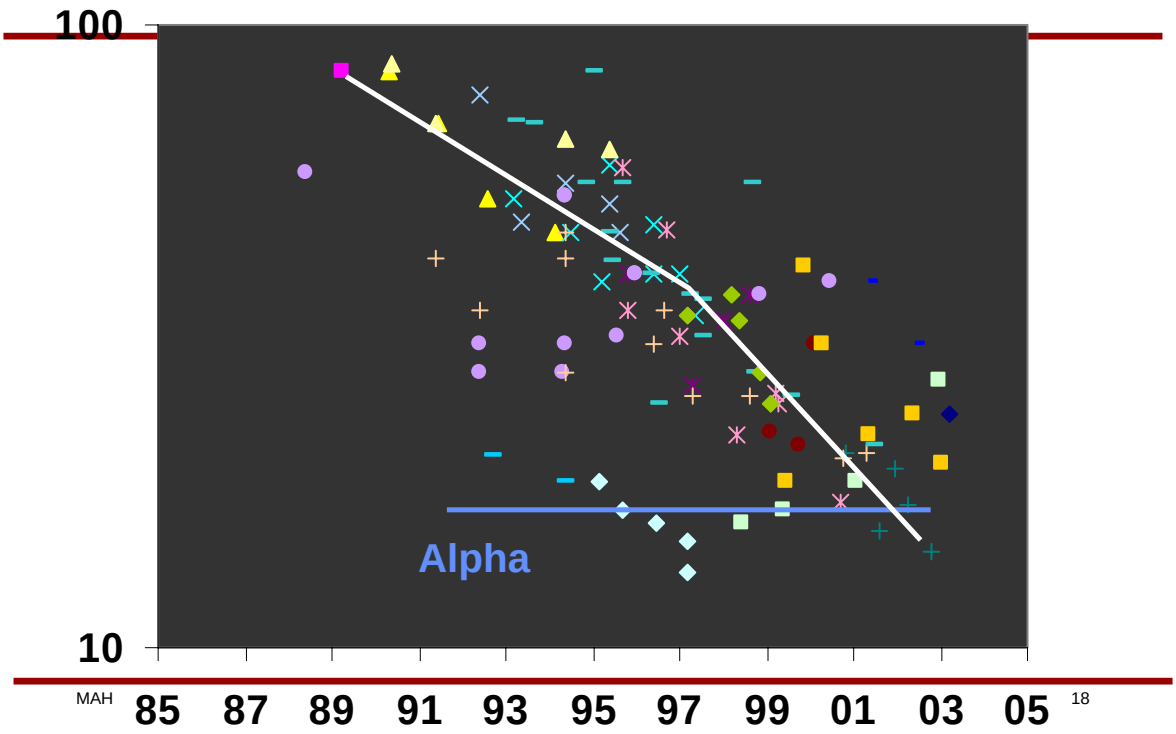
SpecInt/MHz



Clock Frequency Scaling



Clock Cycle in ‘FO4’



Power Scaling

- Complexity of chip is scaling as α^2
- Freq is scaling as $1/\alpha^2$
  - Means we have more flops/gate
  - Higher average power per gate
- Power should scale as faster than $1/\alpha$
  - For the biggest / fastest chips
- This scaling has changed the rules of design
  - It is not sustainable
  - Processors have a limited power budget
  - Fastest processor for a given power budget

Now What?

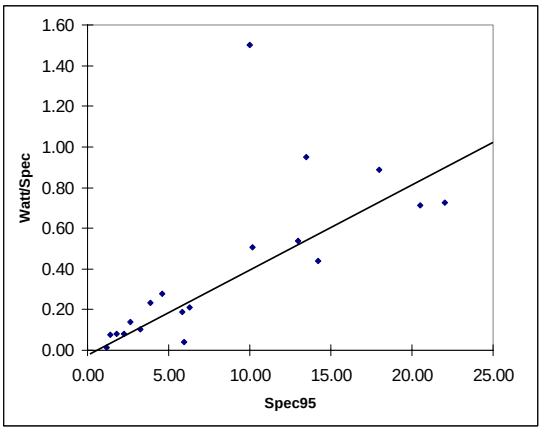
- The power issue is more complex
  - Power efficiency of digital hardware is improving
    - Lots of hardware now can run on small batteries
  - But demands are growing faster
    - Power on top-end processors is growing
- Technology scaling is becoming more interesting
  - Scaling now has its own set of trade-offs
  - Interesting techniques to deal with these issues

Evaluating Power Efficiency

- Want to select the most energy efficient solution
  - Can't use power, $P=CV^2F$
    - Lowering the operating frequency lowers the power
  - Many people use Mips/W
    - This is really an energy metric
    - Mips/W = Reciprocal of Joules/million instructions
    - Energy metrics are kCV^2
    - Lower Voltage will increase this metric
      - Look at the Xscale processor
      - Highest Mips/Watt for lowest supply voltage
- Need to plot performance and power

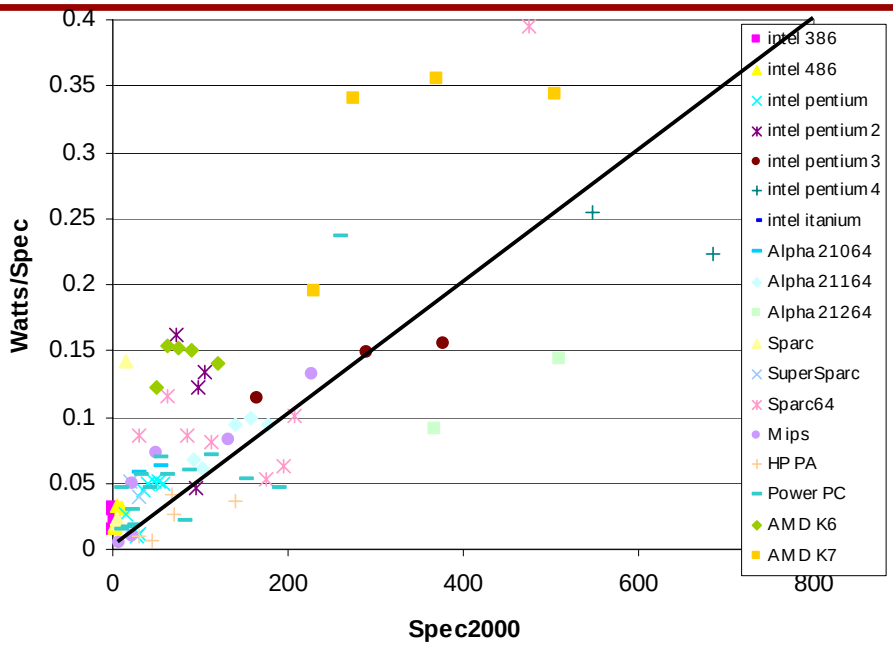
Energy –Performance Graphs

- Use two axis
  - Power or Energy/Op
  - Performance
- Normalize for technology
  - Energy scales as α^3
- Makes tradeoff clearer
  - At performance level
    - Choose lowest energy
  - At power level
    - Choose highest performance

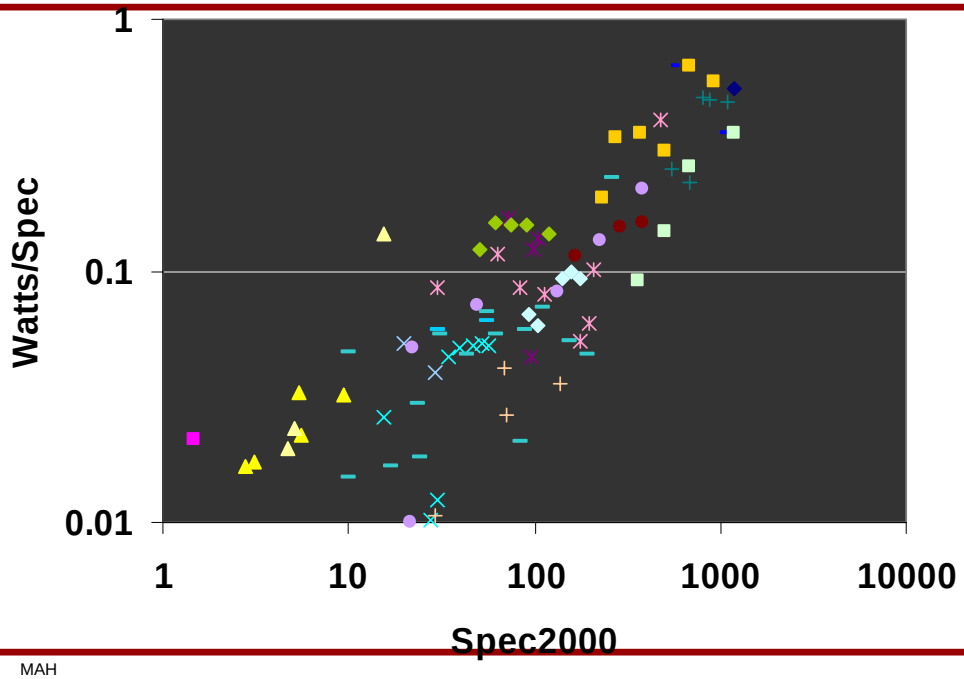


Most processors are roughly on a line of 25 $Spec^2/W$ in 98

Energy/Instruction vs. Performance



Energy/Instruction Performance



Truths About Power

- The power is super linear on required performance
  - Lower performance is lower Energy/Op
- Means you can trade excess performance for power
  - Many low power solutions are really high-performance algorithms run slowly
  - Key enabling of low power signal processing
    - Create parallel solution, and then voltage scale
    - Create new algorithms that use less computing
- Best way to save power is to do less ops
  - System level power management

Low Power Design

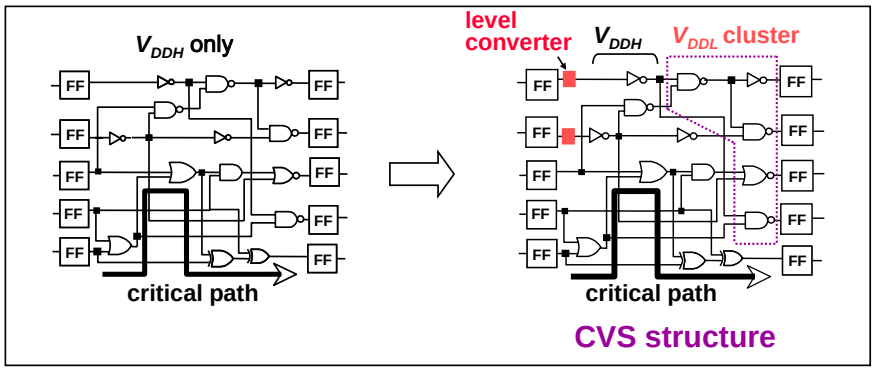
- Define your problem at the correct level
  - Architectural changes make the most difference
    - Turning the RF section off most of the time is easier than building a low power RF design
    - Doing less OPs is easier than building low power OPs
    - Doing OPs in parallel is lower power than sequentially
- Technology also makes a huge difference
  - Power scales by about 3 to 4x each generation
  - Mobile parts almost always use the ‘best’ process
- Circuits is mostly about not wasting power
  - Use the right supply, Vth, etc for the job

Solution - Multiple Flavors of Transistors

- New technologies have many different transistors
  - Vth, Gate Oxide
- Get to select which technology to use
  - Highest performance, low leakage current, etc
  - Often have a couple of transistors types
    - Dual gate oxides, multiple Vths for the transistors
- Designer
  - Needs to choose the transistor type
  - Can choose the supply voltage too

Solution – Multiple Levels for Vdd

- Basic Concept
  - Gates off the critical path ➔ run at VDDL (reduced voltage)
  - Gates on the critical path ➔ run at VDDH (higher voltage)
- Minimize # of level-converters



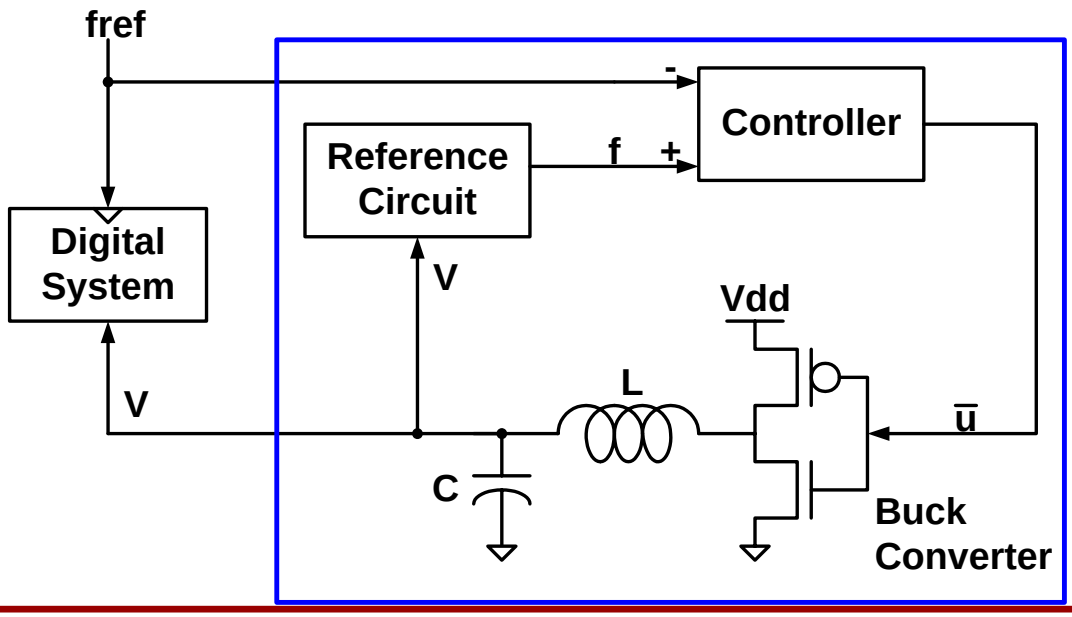
[Usami98a]

Problem With Static Approach

- Two issues with setting Vdd and Vth at design time
  - Optimal point changes with operating task
  - Variability of devices changes what you create
- Might want to have periods of high-performance
  - Adaptive supplies
- Might have periods of low activity
  - Adaptive threshold

- But how do you do this when you have variability?
  - This is still an open question

Adaptive Power-Supply Regulator



Adaptive Threshold Control

- People have used substrate voltage
  - Both forward and reverse bias
- Impedance control is the big issue
  - Have strap transistors to tie to Vdd, Gnd
  - Only drive the lines when in standby
- Other groups have tied to control Vth more directly
  - Adjust Ion to Ioff ratios to optimize performance
  - Problem is how to measure leakage
    - If devices don't match, your control device might not be correct, and the optimal point depends on activity anyhow.

Crazy Idea – Measure Power Directly

- Assume you change change Vdd and Vth
  - Why not use simple adaptive algorithm
- Build a system that has a tracking Vdd control loop
  - Vdd adapts to the value needed to run at F
  - Add a small modulation to Vthn Vthp
  - Measure power of the system as it is running
    - Measure the change in power w.r.t. Vthn, Vthp using a synchronous detector. This will tell whether lowering Vthn will increase power or decrease power for this chip.
  - Adapt Vthn Vthp in a slow loop
- Naturally will track activity ratio, temp, etc.

Summary

- Power efficiency of CMOS logic scales
  - Power per function is scaling about 3-4x per gen
  - This scaling might slow down in a couple gen
    - Trouble with scaling Vdd if Vth does not scale
    - Use adaptation to become more efficient with voltages
- Power is a large problem for current IC designers
  - We now have the capability to build very hot chips
    - Since we can build very complex circuits in small spaces
  - Need to balance performance and power
  - Best way to lower power is to find solution that needs less stuff done, or can use parallelism