Introduction to CMOS VLSI Design

Scaling

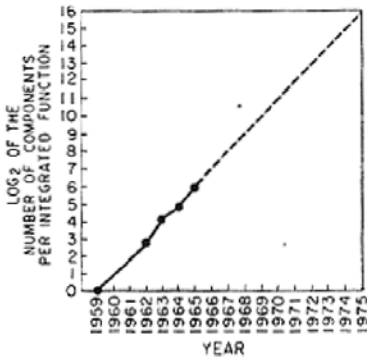
Lecture by Peter Kogge
University of Notre Dame
Fall 2011, 2015, 2018
Modified from presentation by Jay Brockman in 2008
Based on lecture slides by David Harris, Harvey Mudd College
http://www.cmosvlsi.com/coursematerials.html

Outline

- Moore's Law & ITRS Roadmap
- Ideal Scaling
- Real World Scaling
- 2004
- Scaling in the World of Multi-core

Moore's Law

- In 1965, Gordon Moore predicted the exponential growth of the number of transistors on an IC
- Transistor count doubled every year since invention
- Predicted > 65,000 transistors by 1975!
- Growth limited by power



[Moore65]

ITRS

- **ITRS**: INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS
- International group of experts from
  - Industry, Research Labs, Academia
- Run by **Semiconductor Industries of America** – **SIA**
- Every 3 years produce huge document describing projections on how future commercial technology will improve
- On non-full document years, update tables

ITRS Feature Size

- Feature Size used to be minimum gate length/width
- Now on interconnect pitch by product category: DRAM, Logic, Flash
- And based on what is in widespread use
- Also distinguish between “Drawn” & “Physical” Gate length

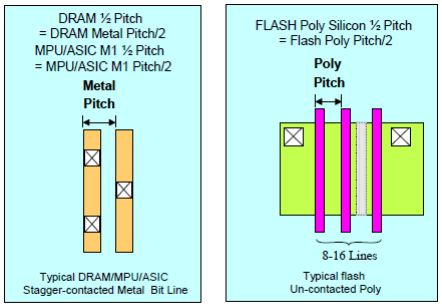
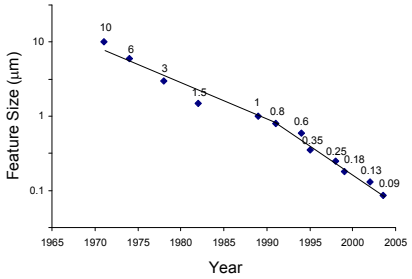


Figure 1 2009 Definition of Pitches

Scaling

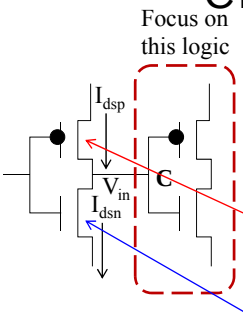
- **Feature size** used to shrink by 30% every 2-3 years
  - Transistors became smaller, faster
  - Circuits got ***smaller*** and ***faster***
  - More transistors fit on chip
  - Double gain in performance: speed & density
- Define **Scale factor** S
  - Applied to feature size
  - ~ every 0.7 shrink factor
    - Corresponds to $S = \sqrt{2}$
    - Called a ***Technology node***



Scaling Effects

- **Area**: if all dimensions shrink by 1/S, area shrinks by $1/S^2$
  - S = √2 => area shrink by 1/2
- **Speed** or **Clock** rate of a circuit
  - Function of how fast a logic gate can change input voltage of another downstream gate
  - Depends on transistor gate capacitance and saturation current
    - Smaller W & L reduces capacitance
    - W/L and V_{dd} affects saturation current
- **Power** drawn by a circuit
  - Function power lost charging and discharging capacitor
  - Depends on transistor gate capacitance, V_{dd}, & clock

CMOS Energy 101



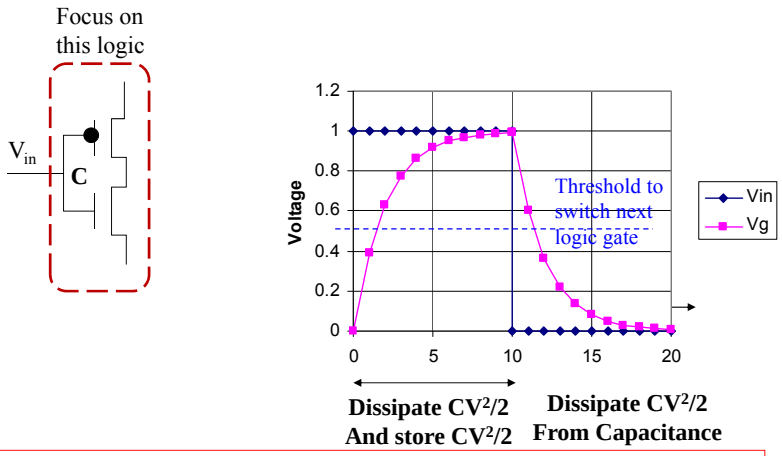
Actually there's more than just gate capacitance

Basic Equations:

- Power = ∫ Energy lost
  - = ∫ V(t)*I(t) = V_{dd}∫I(t)
- C = sum of $\varepsilon LW/t_{ox}$
- $I_{dsat} = \beta(V_{gs}-V_t)V_{dd}^2/2$
- $V_{across\_cap} = Q/C$, $Q = \int I(t)$
- When charging C:
  - Power dissipated in P-type
  - Energy stored in C
- When discharging C:
  - Power dissipated in N-type

If we define an **activity cycle** is V_{in} going from 0 to V and back to 0, then energy lost from one activity is CV^2.
If logic has clock f cycles/sec, and α activity cycle per clock period, then power dissipated by this gate is αfCV^2

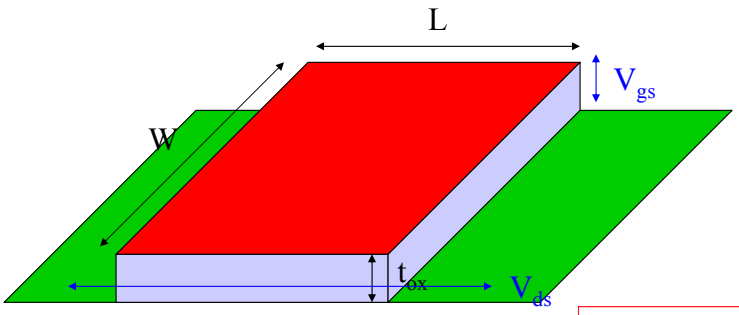
CMOS Energy 101



One clock cycle dissipates $C*V^2$
Smaller C => less power, faster transition

Constant Field
a.k.a. Dennard
Scaling

The Scaling of the Transistor



What can change?

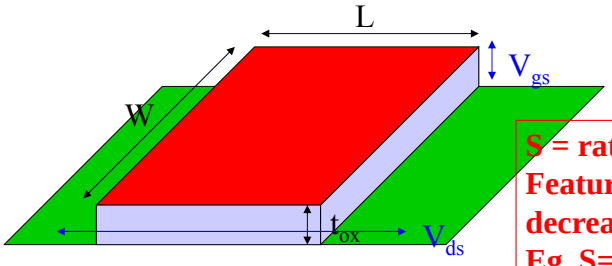
- Doping levels in diffusion
- Physical dimensions: W, L, t_{ox}
- Electrical parameters Vgs, Vds

A key parameter: The Electric Field across the Gate proportional to V_{gs}/t_{ox}

Scaling Variations

- What changes between technology nodes?
- ***Constant Field Scaling (aka Dennard Scaling)***
  - All dimensions (W, L, t_{ox}) scale by same S
  - Voltage (V_{DD}) Scales down by S
  - Doping levels change
- ***Constant Voltage Scaling*** (Today)
  - All but voltage changes
- ***Lateral Scaling***
  - Only gate length L shrinks
  - Often done as a quick **gate shrink** (S = 1.05)

Why Is It Called "Constant Field"?



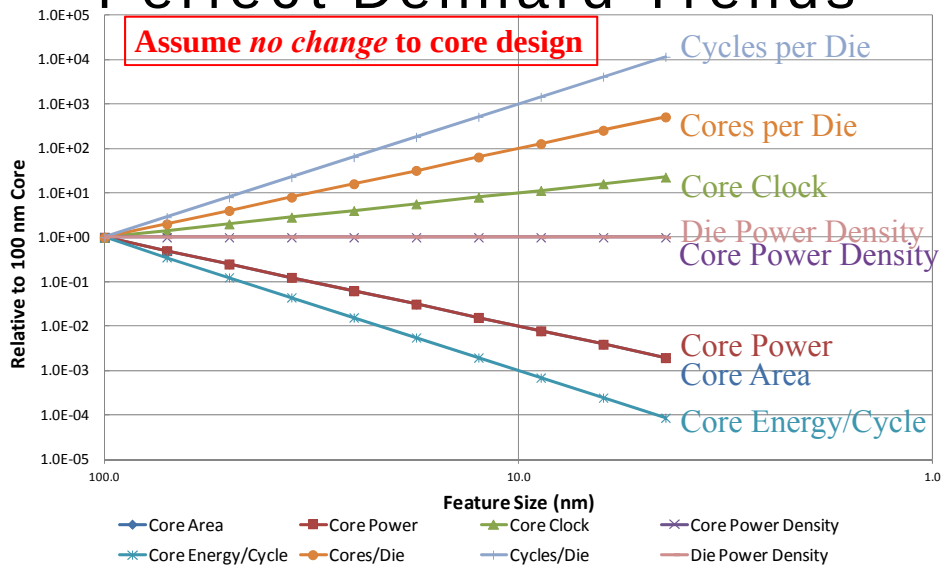
S = ratio by which Feature Size (i.e. L) decreases.
Eg. S=2 => L is ½ of previous value

- If V decreases by factor of 1/S
- And t_{ox} decreases by factor of 1/S
- *Then E field (V/t_{ox}) remains constant*!
- So mobility & the like remain constant
- And then if L gets shorter by 1/S
- Then capacitance (LW/t_{ox}) drops as (1/S)*(1/S)/(1/S) = 1/S
- Then the transistor is ***faster*** by factor of S
- And energy per cycle (CV^2) ***goes down*** as $(1/S)^3$

Translating into Core Parameters

- Core Clock ≈ S
- Core Capacitance ≈ 1/S
- Core Voltage ≈ 1/S
- Core Power ≈ Capacitance*Clock*V_{dd}^2 ≈ $1/S^2$
- Cores/Constant Area Die ≈ $1/S^2$
- **Power Density ≈ Constant**
- **Compute Cycles/Die ≈ S^3**
- **Energy/Cycle = Power/Clock ≈ $1/S^3$**

Perfect Dennard Trends



Device Scaling

Parameter	Sensitivity	Dennard Scaling	Constant Voltage	Lateral Scaling
L: Length		1/S	1/S	1/S
W: Width		1/S	1/S	1
t_{ox}: gate oxide thickness		1/S	1/S	1
V_{DD}: supply voltage		1/S	1	1
V_t: threshold voltage		1/S	1	1
NA: substrate doping		S	S	1
β	$W/(Lt_{ox})$	S	S	S
I_{on}: ON current	$\beta(V_{DD}-V_t)^2$	1/S	S	S
R: effective resistance	V_{DD}/I_{on}	1	1/S	1/S
C: gate capacitance	WL/t_{ox}	1/S	1/S	1/S
τ: gate delay	RC	1/S	$1/S^2$	$1/S^2$
f: clock frequency	$1/\tau$	S	S^2	S^2
E: switching energy / gate	CV_{DD}^2	$1/S^3$	1/S	1/S
P: switching power / gate	Ef	$1/S^2$	S	S
A: area per gate	WL	$1/S^2$	$1/S^2$	1
Switching power density	P/A	1	S^3	S
Switching current density	I_{on}/A	S	S	S

Green: "Better"; Yellow: "Neutral"; Shades of Red: "Worse"

Interconnect (aka Wires)

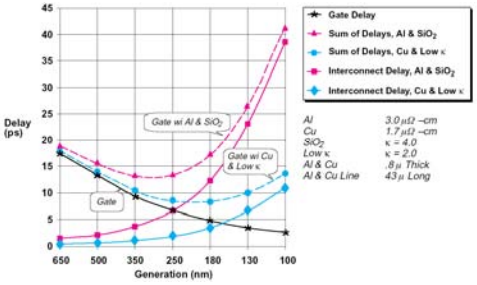
Parameter	Sensitivity	Scale Factor
w: width		1/S
s: spacing		1/S
t: thickness		1/S
h: height		1/S
D_c: die size		D_c
R_w: wire resistance/unit length	1/wt	S^2
C_{wf}: fringing capacitance / unit length	t/s	1
C_{wp}: parallel plate capacitance / unit length	w/h	1
C_w: total wire capacitance / unit length	$C_{wf} + C_{wp}$	1
t_{wu}: unrepeated RC delay / unit length	R_wC_w	S^2
t_{wr}: repeated RC delay / unit length	sqrt(RCR_wC_w)	sqrt(S)
Crosstalk noise	w/h	1
E_w: energy per bit / unit length	$C_wV_{DD}^2$	$1/S^2$

More on this in later lecture

Parameter	Sensitivity	Local / Semiglobal	Global
l: length		1/S	D_c
Unrepeated wire RC delay	l^2t_{wu}	1	$S^2D_c^2$
Repeated wire delay	lt_{wr}	sqrt(1/S)	$D_c$sqrt(S)
Energy per bit	IE_w	$1/S^3$	D^c/S^2

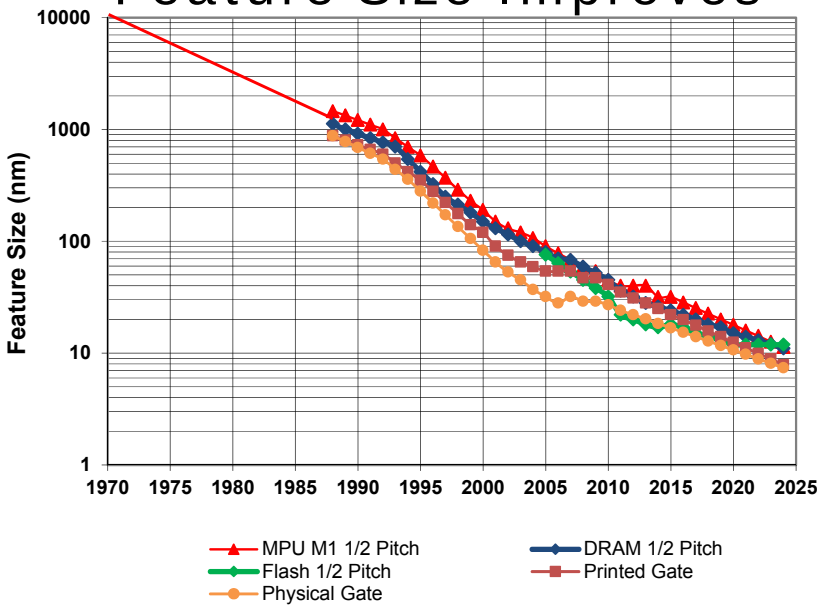
Interconnect Observations

- Capacitance per micron is remaining constant
  - About 0.2 fF/μm
  - Roughly 1/10 of gate capacitance
- Local wires are getting faster
  - Not quite tracking transistor improvement
  - But not a major problem
- Global wires are getting slower
  - No longer possible to cross chip in one cycle

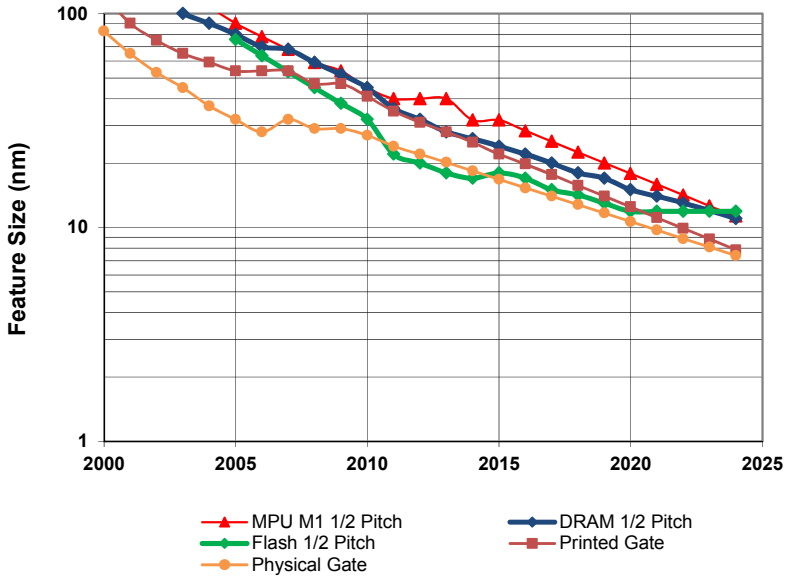


Scaling in the "Old" Real World

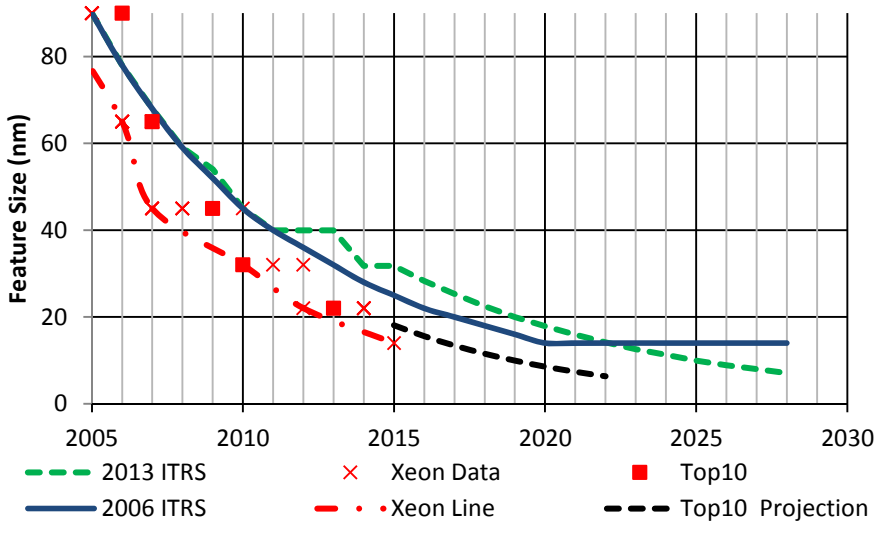
Feature Size Improves



Feature Size In Detail

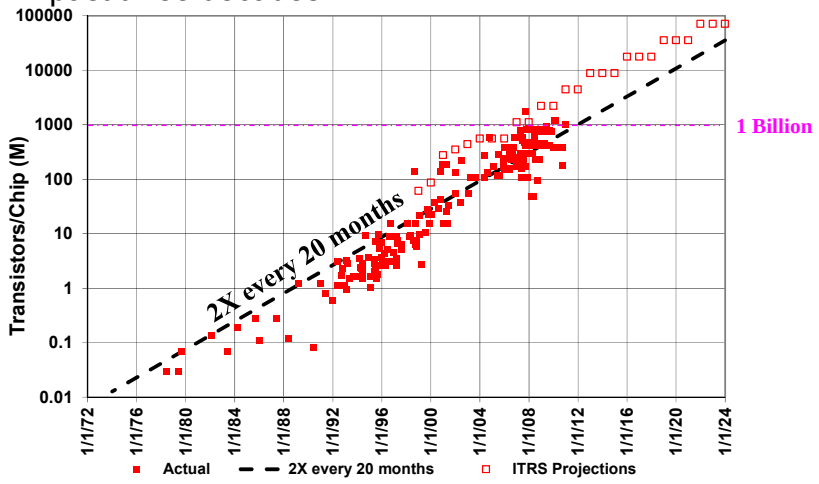


Leading Edge Fabs Are Doing Better

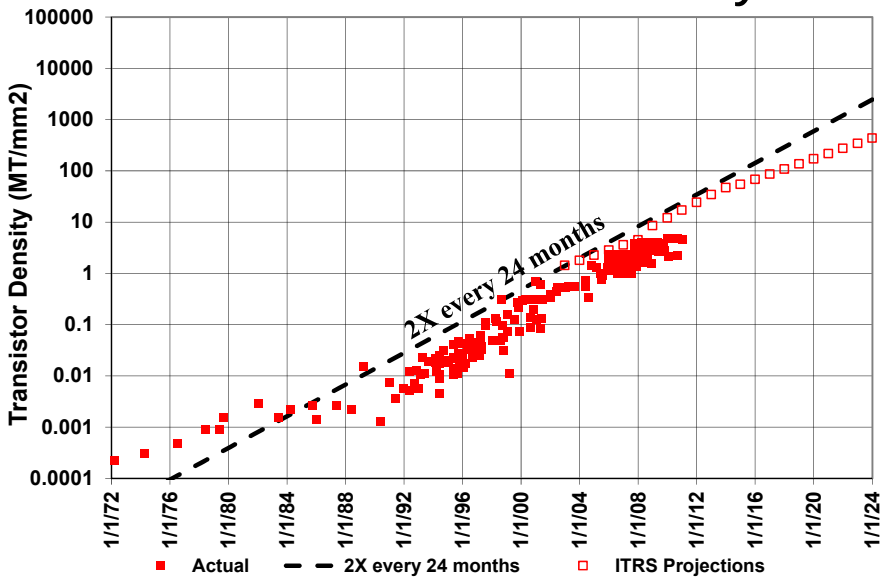


More Moore

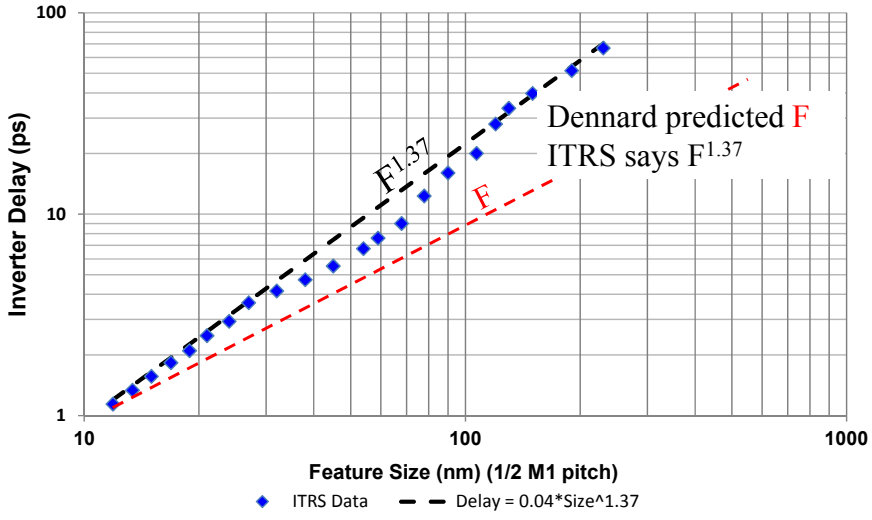
- Transistor counts have doubled periodically for the past three decades



Transistor Density



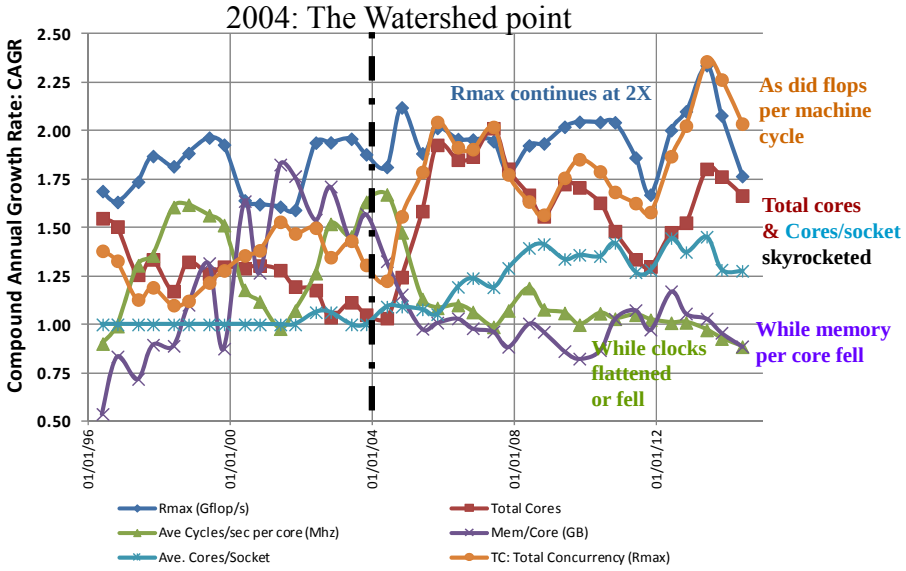
Inherent Delay Better Than Dennard



2004: The End of the World as We Knew It

The World Changed in 2004

(Data taken from top 10 supercomputers over last 20 years)



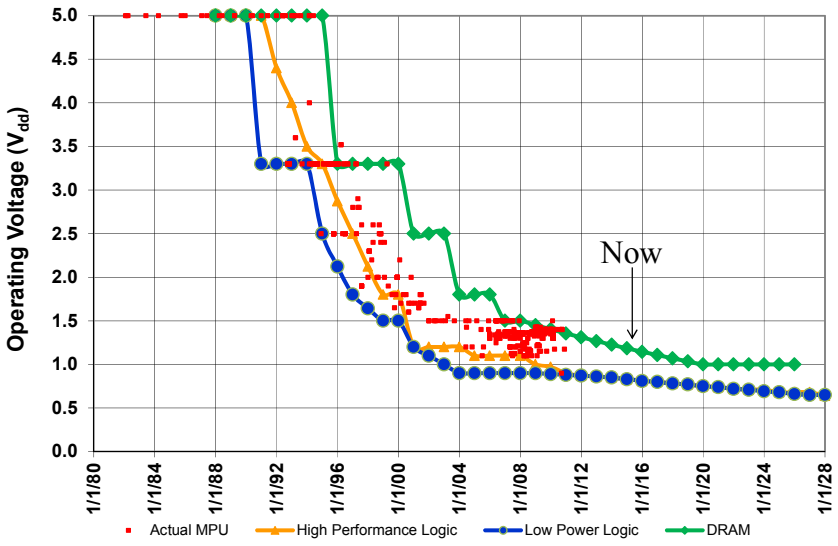
The 2004 Event

- S = scale factor from one technology node to next
- Assume we port same design unchanged
- In constant field scaling
  - Area goes down by $1/S^2$
  - V_{dd} goes down by $1/S^2$
  - Capacitance goes down by 1/S
  - Clock goes up by S
  - So power goes down by $(1/S)*S*(1/S^2) = 1/S^2$
  - And power per unit area = $(1/S^2)/(1/S^2)$ = constant
  - But if area increases, chip power increases
- In 2004 V_{dd} stopped decreasing and we maxed out our ability to cool chips

The Origins of 2004

- Entering “Constant Voltage” Scaling
- With minimal change in t_{ox}
- If Clock continued to go up by S
  - Power/core goes down by (1/S)*S*(1) = constant
  - And power per unit area = (1)/(1/S^2) = ***S^2***!!!!!
- Once we max out chip power
  - We cannot allow clocks to increase
  - We use all remaining tricks to keep power constant

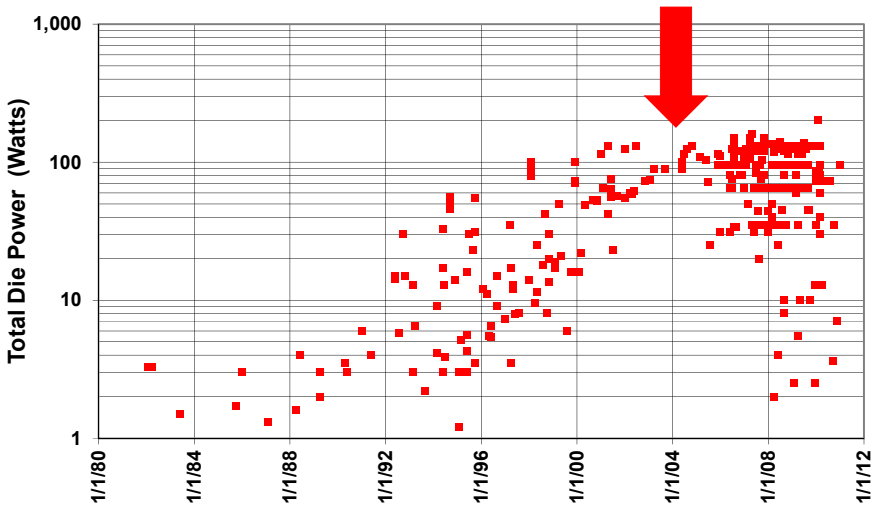
V_{dd} Has Flattened



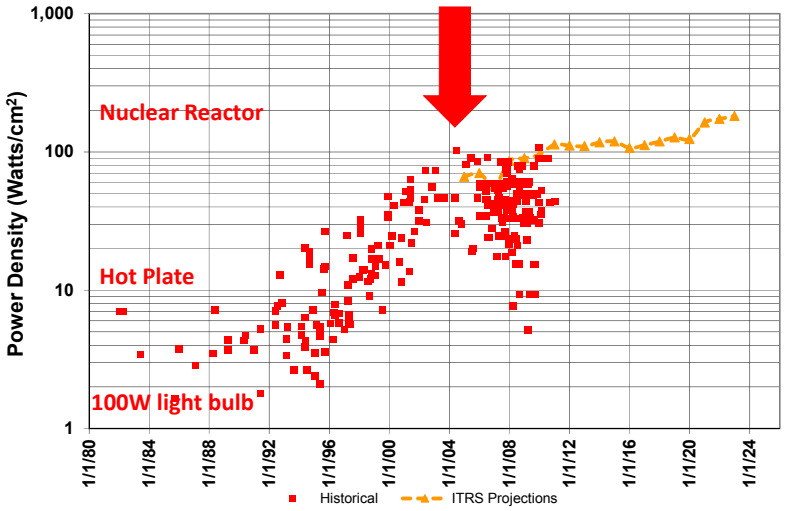
Real Scaling

- t_{ox} scaling has slowed since 65 nm
  - Limited by gate tunneling current
  - Gates are only about 4 atomic layers thick!
  - High-k dielectrics have helped continued scaling of effective oxide thickness
- V_{DD} scaling has slowed since 65 nm
  - SRAM cell stability at low voltage is challenging
- Dennard scaling predicts cost, speed, power all improve
  - Below 65 nm, some designers find they must choose just two of the three

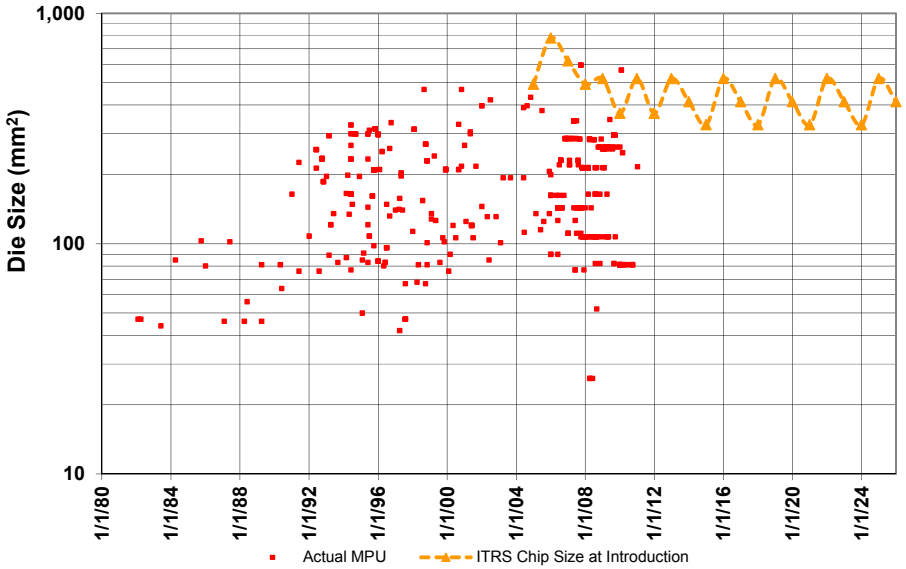
Real World Chip Power



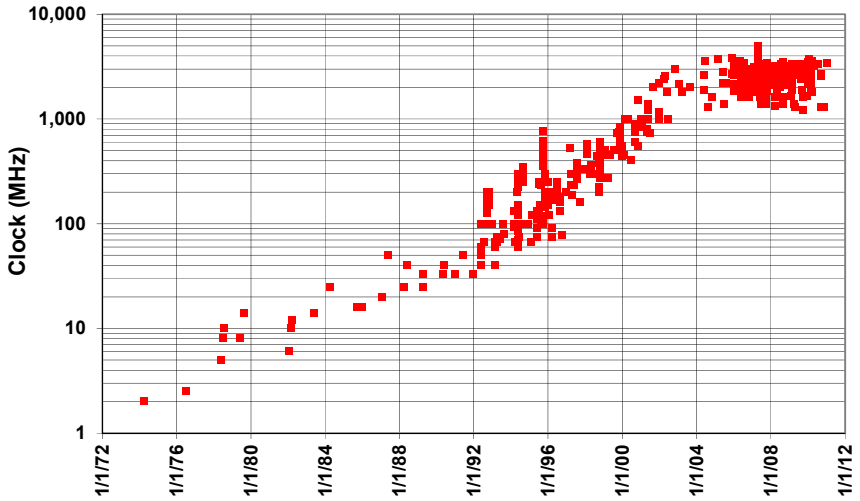
Real World Rise in Power Density



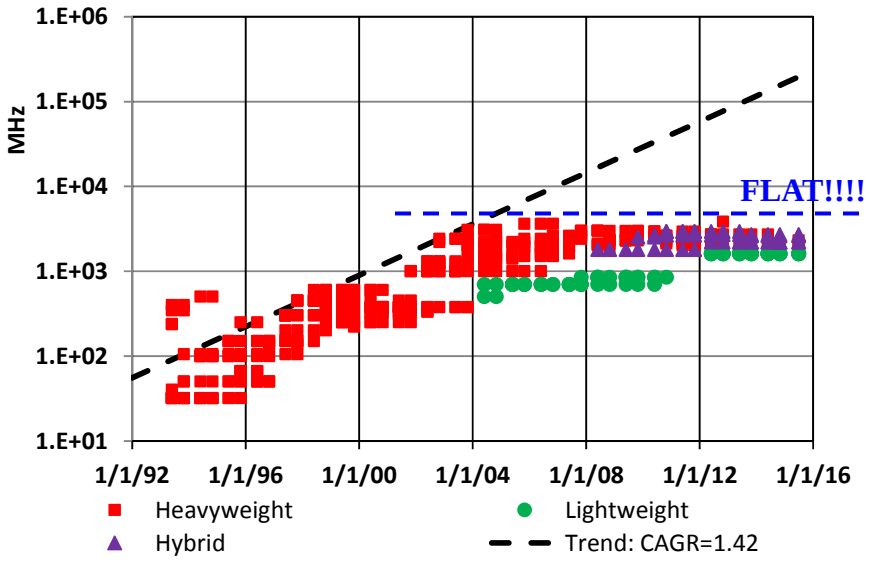
Die Size Became Constant



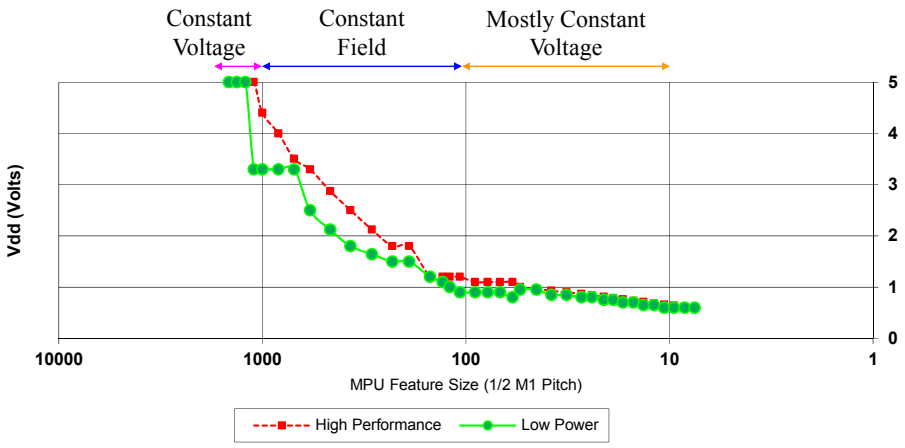
Microprocessor System Clocks



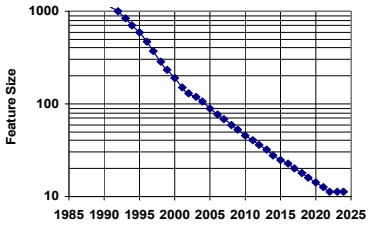
Supercomputer System Clocks



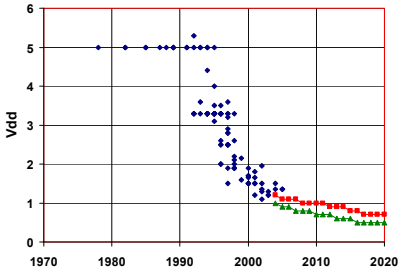
Scaling Time Periods

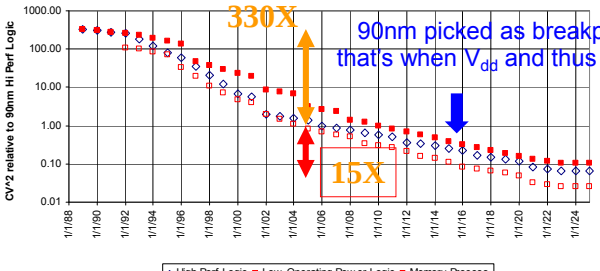


How Did CV^2 Improve With Time?



Assume capacitance of a circuit scales as feature size





90nm picked as breakpoint because that's when V_{dd} and thus clocks flattened

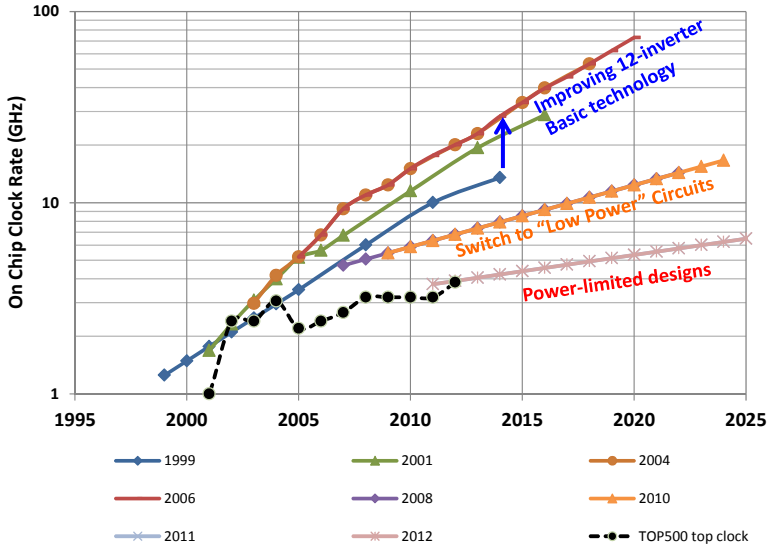
Today

Looking Forward: ITRS Projections (from 2013)

	2000	2005	2010	2015	2020	2022	2024
Feature Size (nm)	82.9	32.0	27.0	16.8	10.7	8.9	7.4
Microprocessor MTransistors/sq. cm	26	97	564	2548	8080	12840	20400
DRAM Gbits per chip	0.25	1	2	8	16	32	32
Production MPU Chip Size (sq. mm)	170	111	99	88	111	140	88
Production DRAM Chip Size (sq. mm)	129	88	47	29	37	23	15
Max ASIC Signal pins/chip	900	2000	2400	2800	3100	3420	3420
Max On-Chip Clock Rate (GHz)	1.0	5.0	5.9	4.4	5.3	5.8	6.2
Max Logic Wiring Levels	7	10	12	13	14	15	15
High Perf MPU Supply Voltage (V)	1.8	1.1	1.0	0.8	0.8	0.7	0.7
Max Watts/sq. mm of chip	0.50	0.66	0.96	1.19	1.24	1.73	#N/A

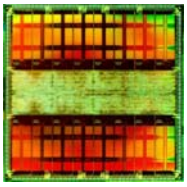
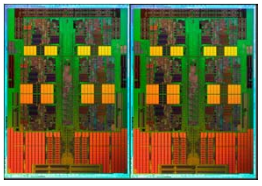
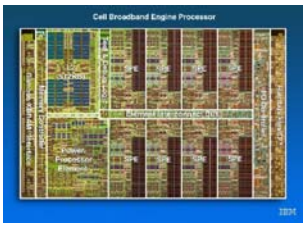
http://www.itrs.net/

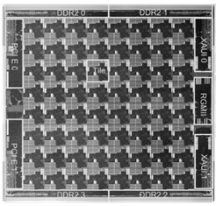
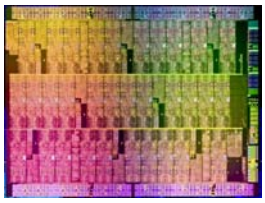
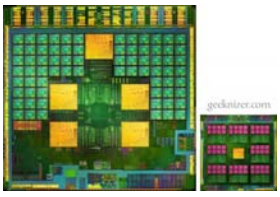
ITRS Clock Predictions



Scaling in the World of Multi-Core Chips

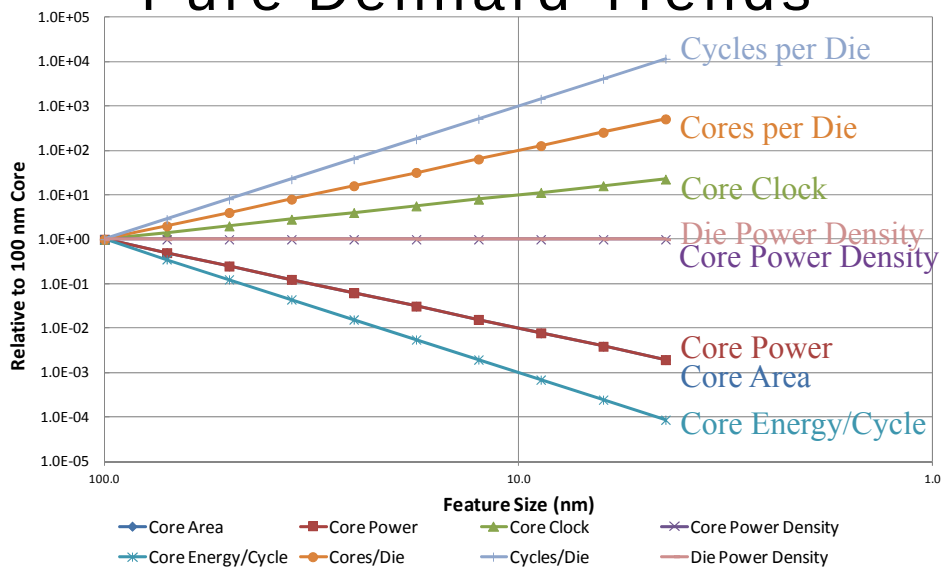
Multi-Core Chips

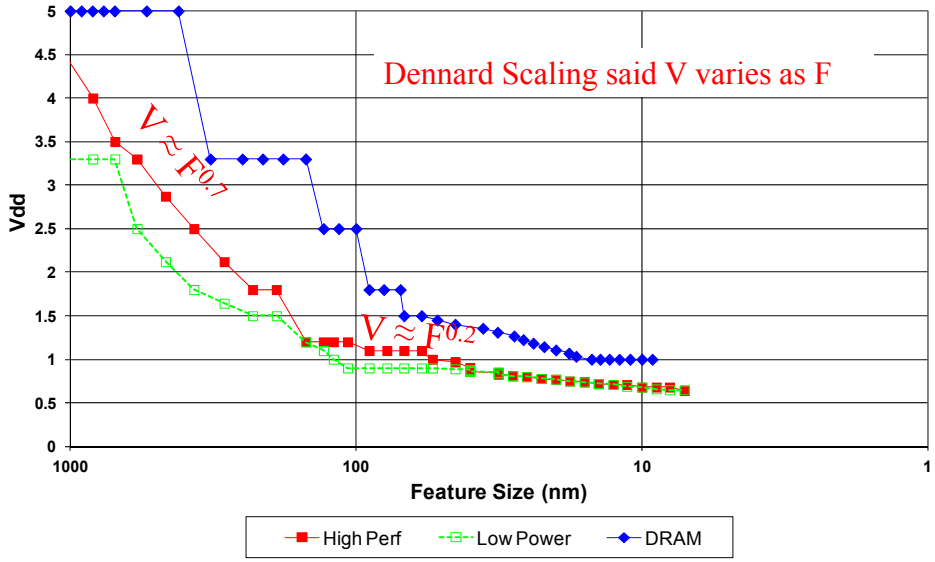




Place many copies of same core on chip, and run parallel programs

Pure Dennard Trends



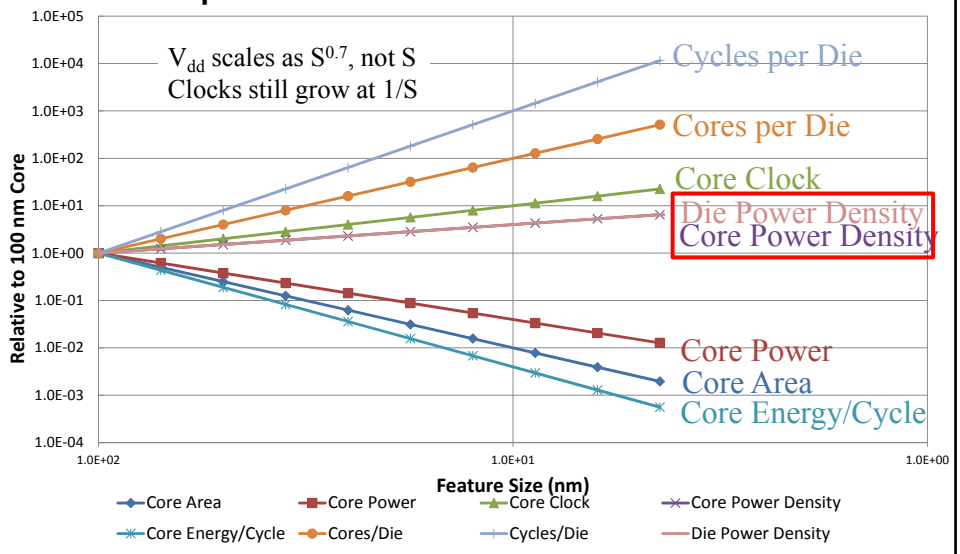
We See Vdd Flattening Clearer vs F



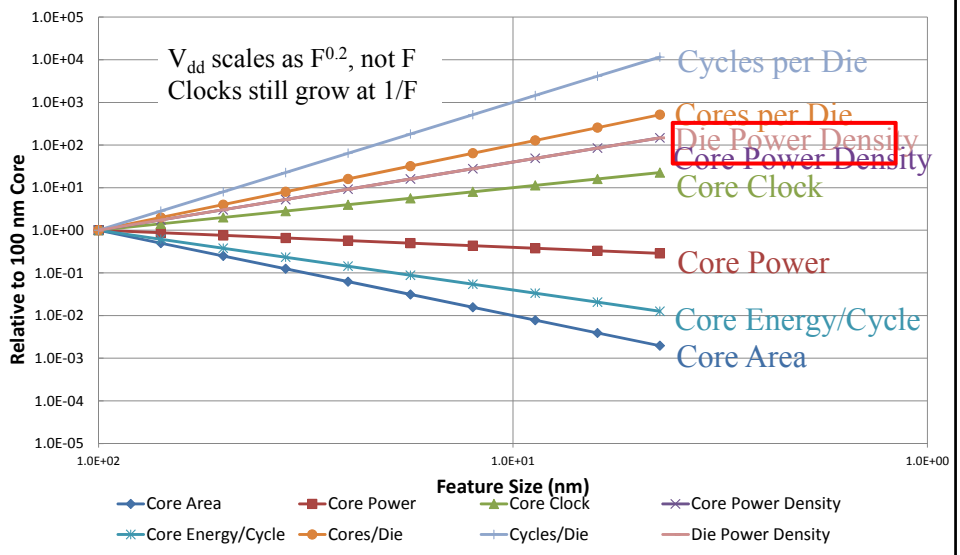
Relating the Regimes: No Change to Clocks

	Voltage Scales As:		
	1/S	**1/S^0.7**	**1/S^0.2**
	Dennard	**Pre 2004**	**Post 2004**
Area	1/S^2		
Capacitance	1/S		
Clock	S		
Core Power	1/S^2	1/S^1.4	1/S^0.4
Power Density	1	S^0.6	S^1.6
Energy/Cycle	1/S^3	1/S^2.4	1/S^1.4

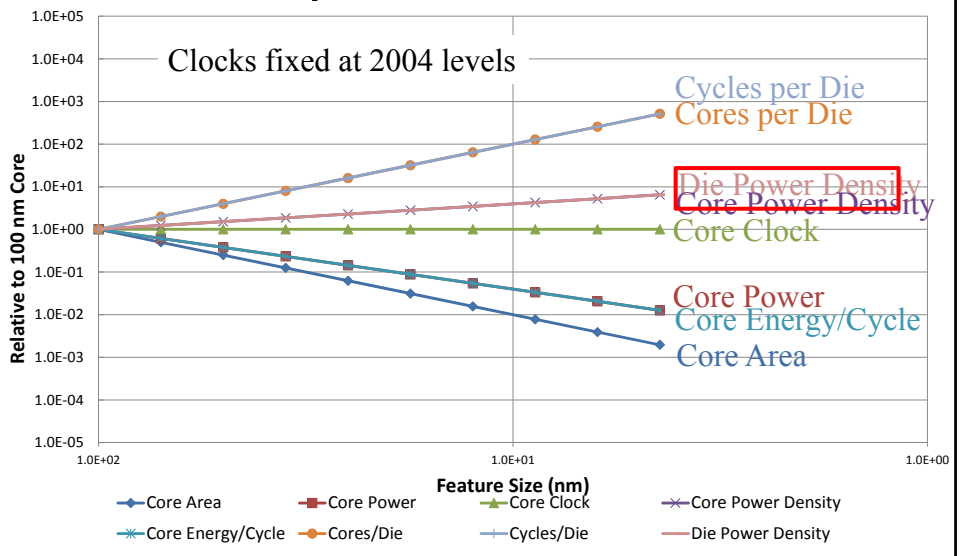
Comparative Trends: Pre 2004



Comparative Trends: Post 2004



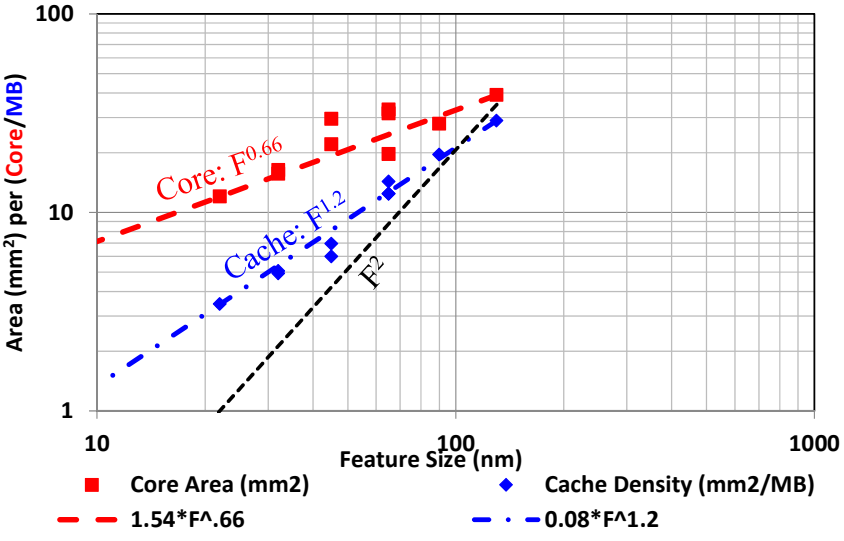
Comparative Trends



Should We Expect Dennard Scaling of Cores?

- **Effective Core Area** = Die area /# cores
- How does # cores affect effective core area?
- How many cores may fit on constant area die?
- Core itself
  - Growth in aggregate cache
  - Modifications to microarchitecture
  - Short SIMD additions
- Multi-core chip
  - Routing usually grows faster than linear
  - Off chip interfaces may be same

Core Area Not Following "Moore's Law"

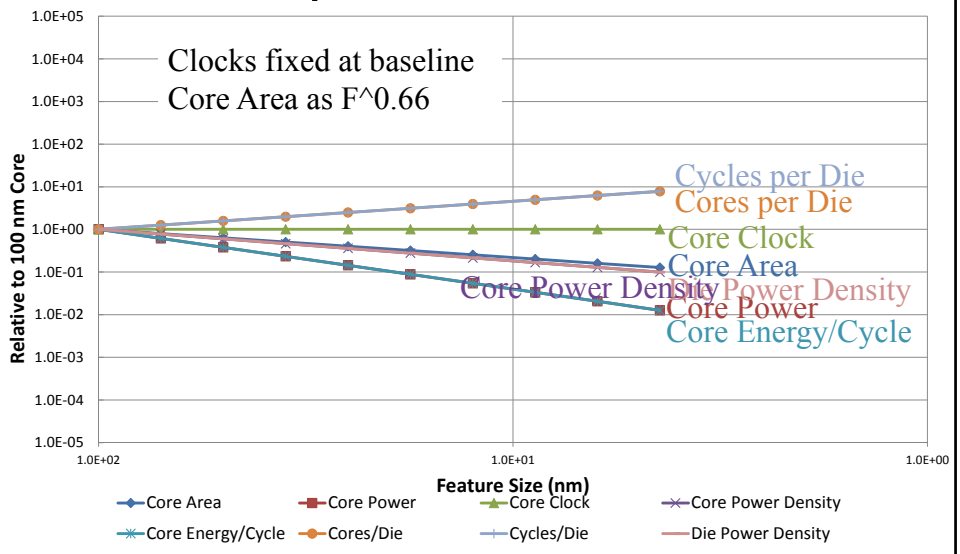


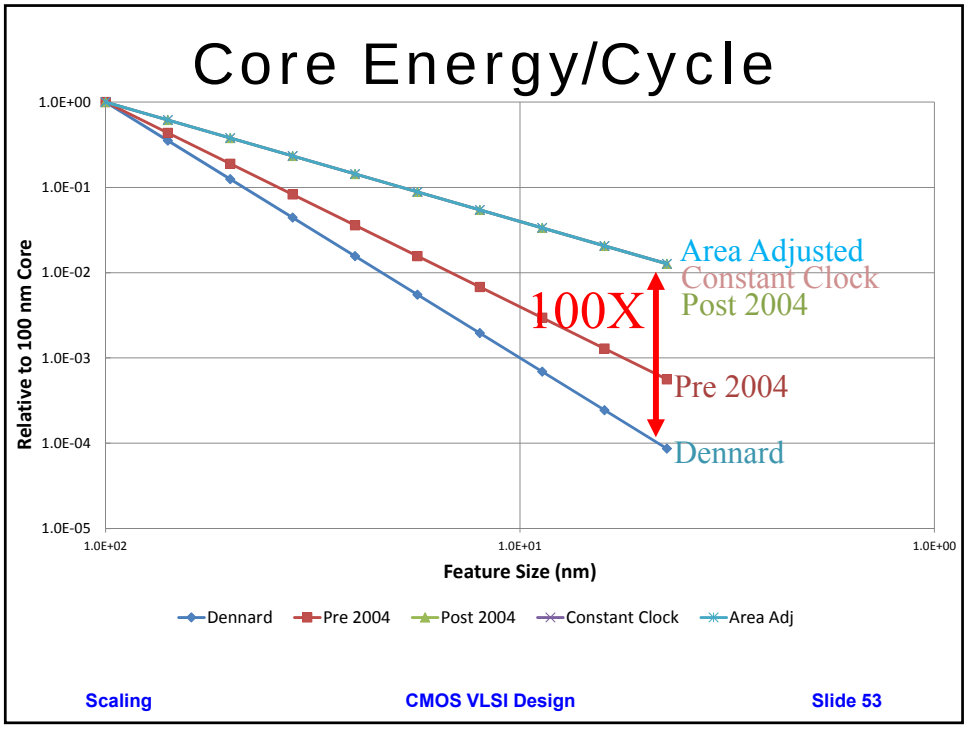
Looking Forward

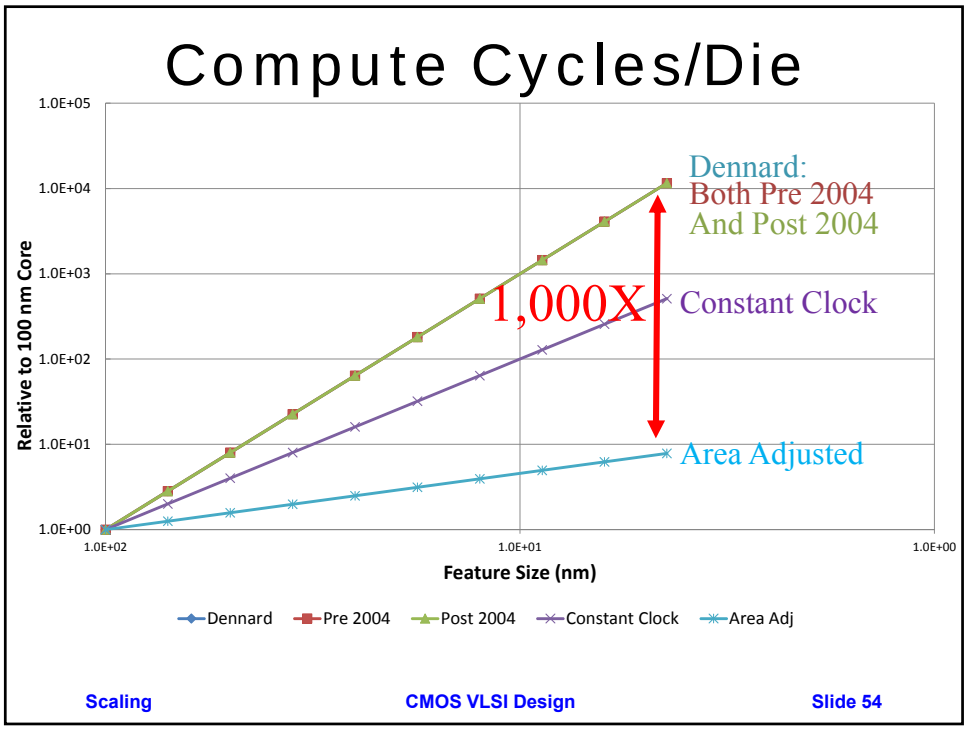
- V_{dd} varies as $1/S^{0.2}$ (not as 1/S)
- Clock remains constant
- Core area varies as $1/S^{0.66}$ (not as $1/S^2$)

Voltage	1/S	$1/S^{0.2}$
Area	$1/S^2$	$1/S^{0.66}$
Capacitance	1/S	
Clock	S	1
Core Power	$1/S^2$	$1/S^{1.4}$
Power Density	1	$1/S^{0.74}$
Energy/Cycle	$1/S^3$	$1/S^{0.74}$

Comparative Trends




Core Energy/Cycle
1.0E+00
1.0E-01
1.0E-02
1.0E-03
1.0E-04
1.0E-05
Relative to 100 nm Core
1.0E+02
1.0E+01
1.0E+00
Feature Size (nm)
Area Adjusted
Constant Clock
Post 2004
100X
Pre 2004
Dennard
Dennard
Pre 2004
Post 2004
Constant Clock
Area Adj
Scaling
CMOS VLSI Design
Slide 53


Compute Cycles/Die
1.0E+05
1.0E+04
1.0E+03
1.0E+02
1.0E+01
1.0E+00
Relative to 100 nm Core
1.0E+02
1.0E+01
1.0E+00
Feature Size (nm)
Dennard:
Both Pre 2004
And Post 2004
1,000X
Constant Clock
Area Adjusted
Dennard
Pre 2004
Post 2004
Constant Clock
Area Adj
Scaling
CMOS VLSI Design
Slide 54

Conclusions

- We have seen the end of Dennard Scaling
  - No more “faster and faster chips”
- Multi-core has taken over processor chips
  - Forcing parallel programming for more performance
- Power has become #1 design issue
- With power of interconnect becoming dominant