Fundamentals of VLSI

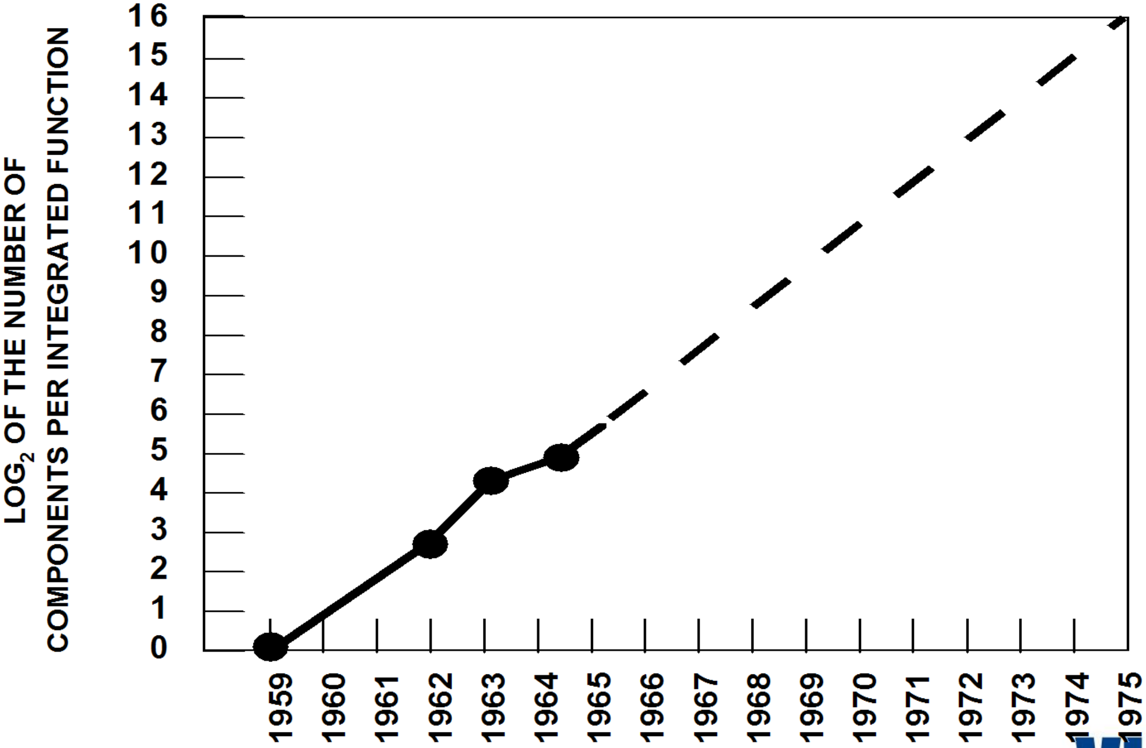
Slides by Adam Teman

Technology Scaling

Moore's Law

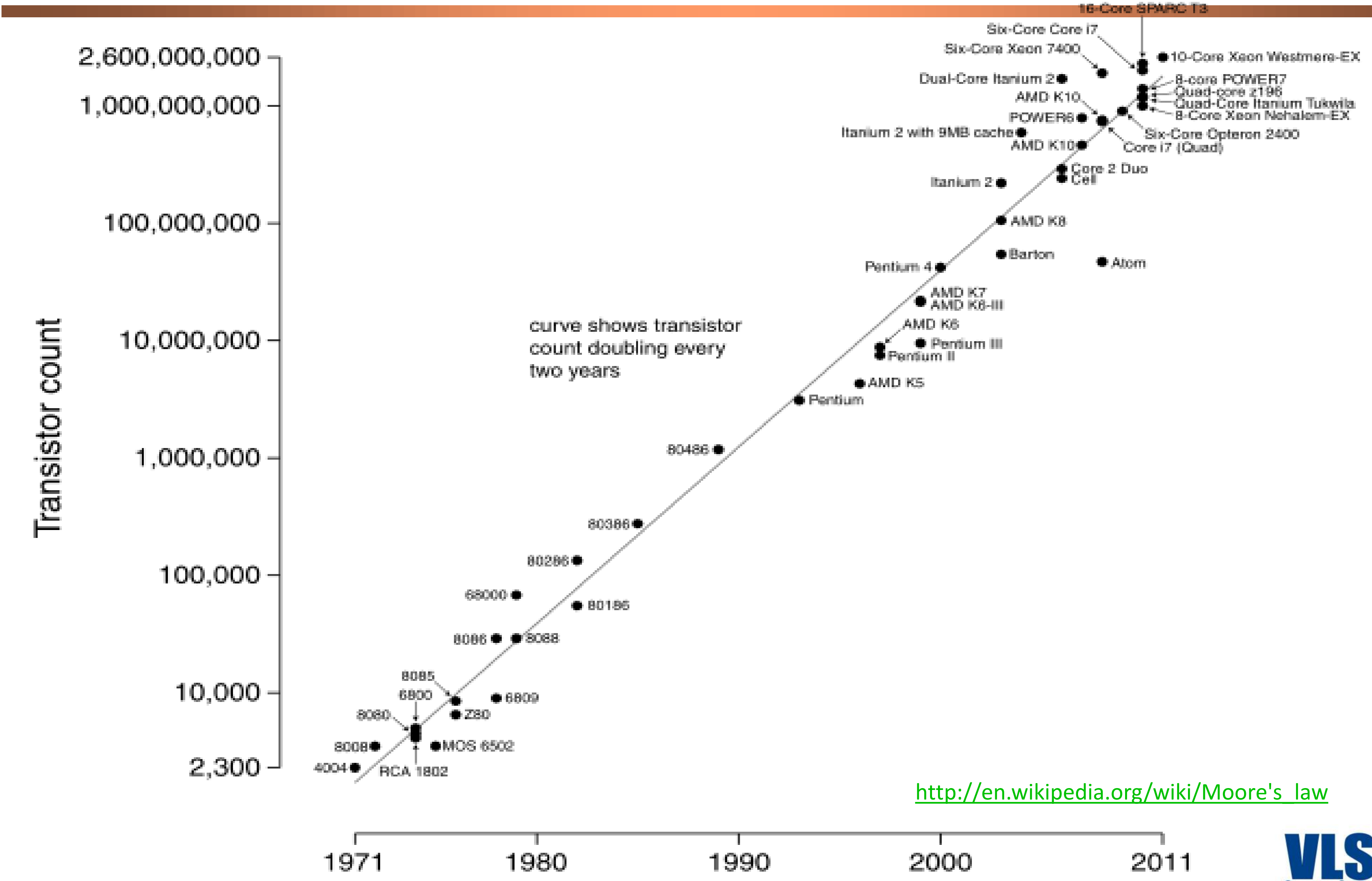
In 1965, Gordon Moore noted that the number of components on a chip doubled every 18 to 24 months.

He made a prediction that semiconductor technology will double its effectiveness every 18 months



Electronics, April 19, 1965.

Moore’s Law 1971-2011

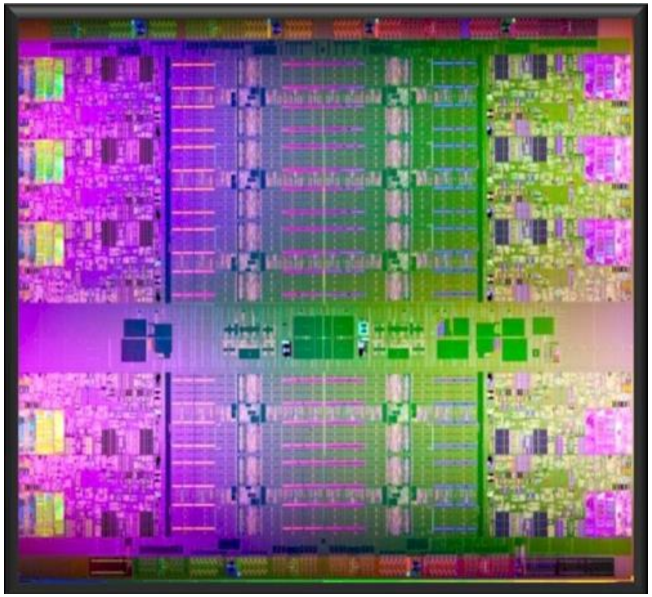


curve shows transistor count doubling every two years

http://en.wikipedia.org/wiki/Moore's_law

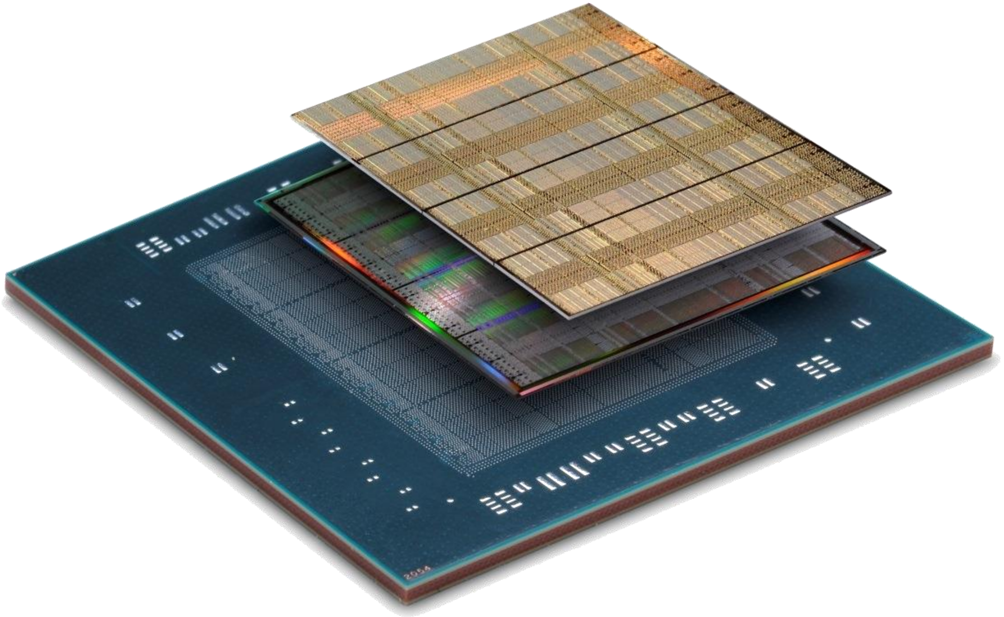
Moore's Law Today (2012)

Intel Xeon E5-2600



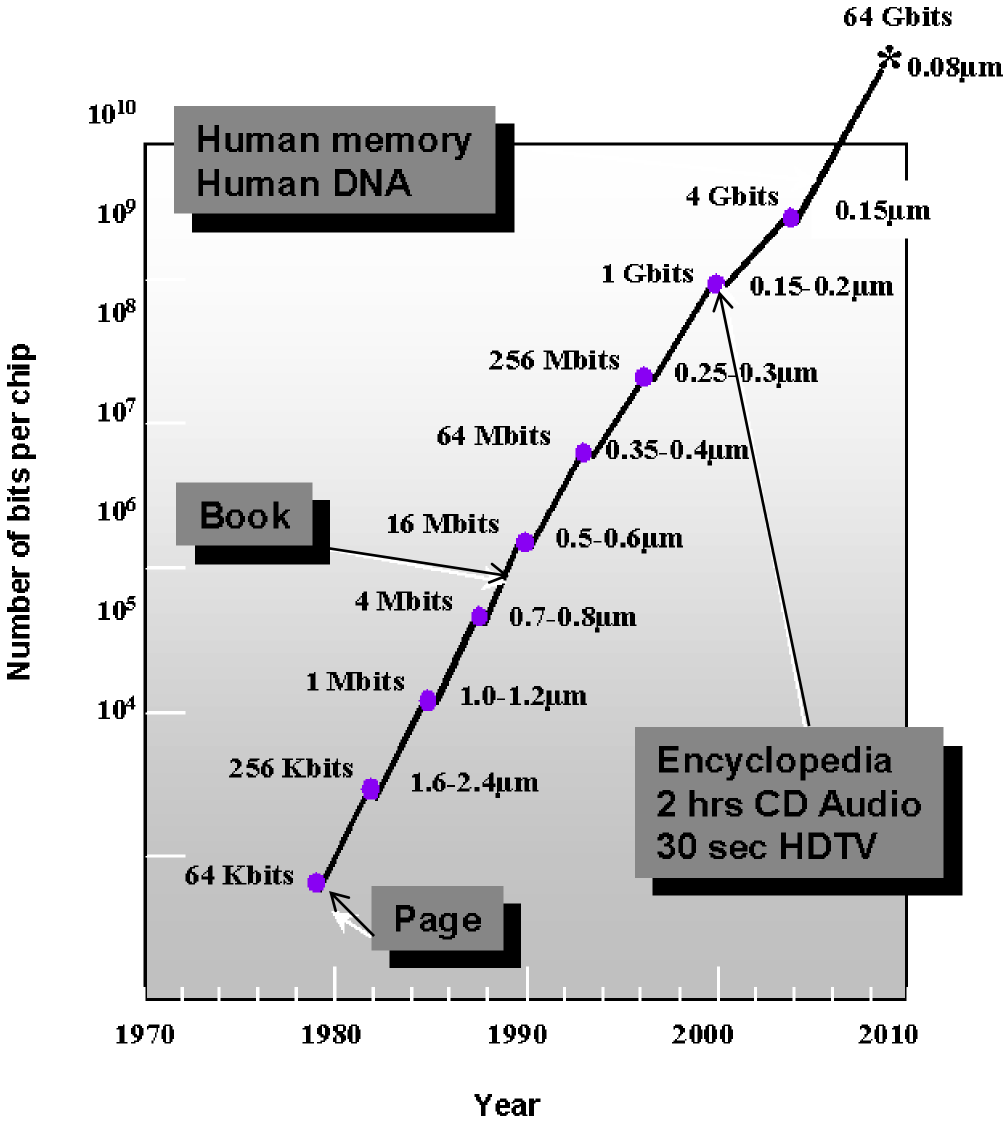
- 32nm "Sandy Bridge"
- 8 Cores
- 32KB L1 Cache
- 256KB L2 Cache
- 416 mm^2
- 2.2 Billion Transistors
- Introduced March 2012

Xilinx Virtex-7 FPGA

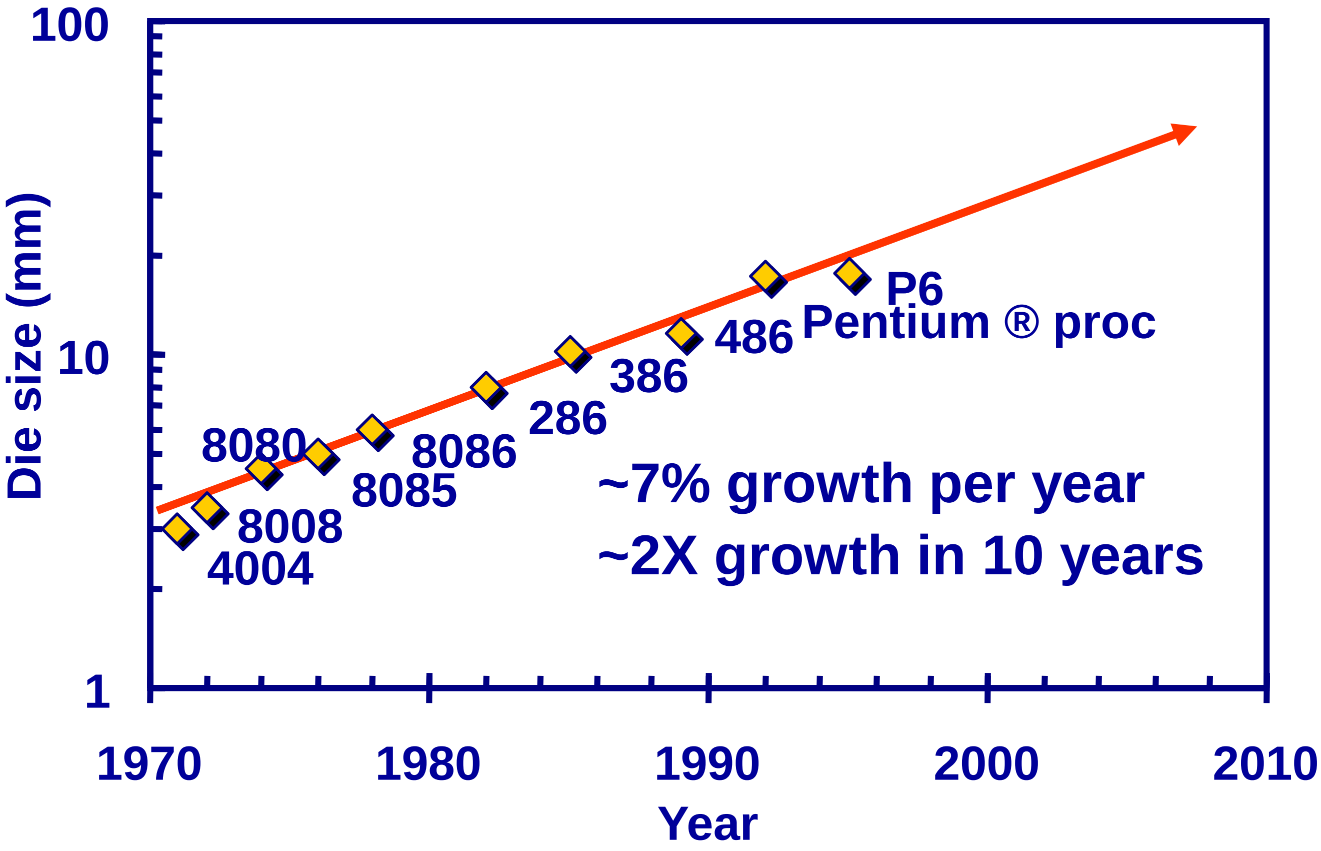


- 28nm – 2.5D IC Stacking
- 416 mm^2
- 6.8 Billion Transistors (World Record!)
- 2 million logic cells
- 12.5 Gb/s serial transceivers
- Introduced October 2011

Evolution in Memory Complexity



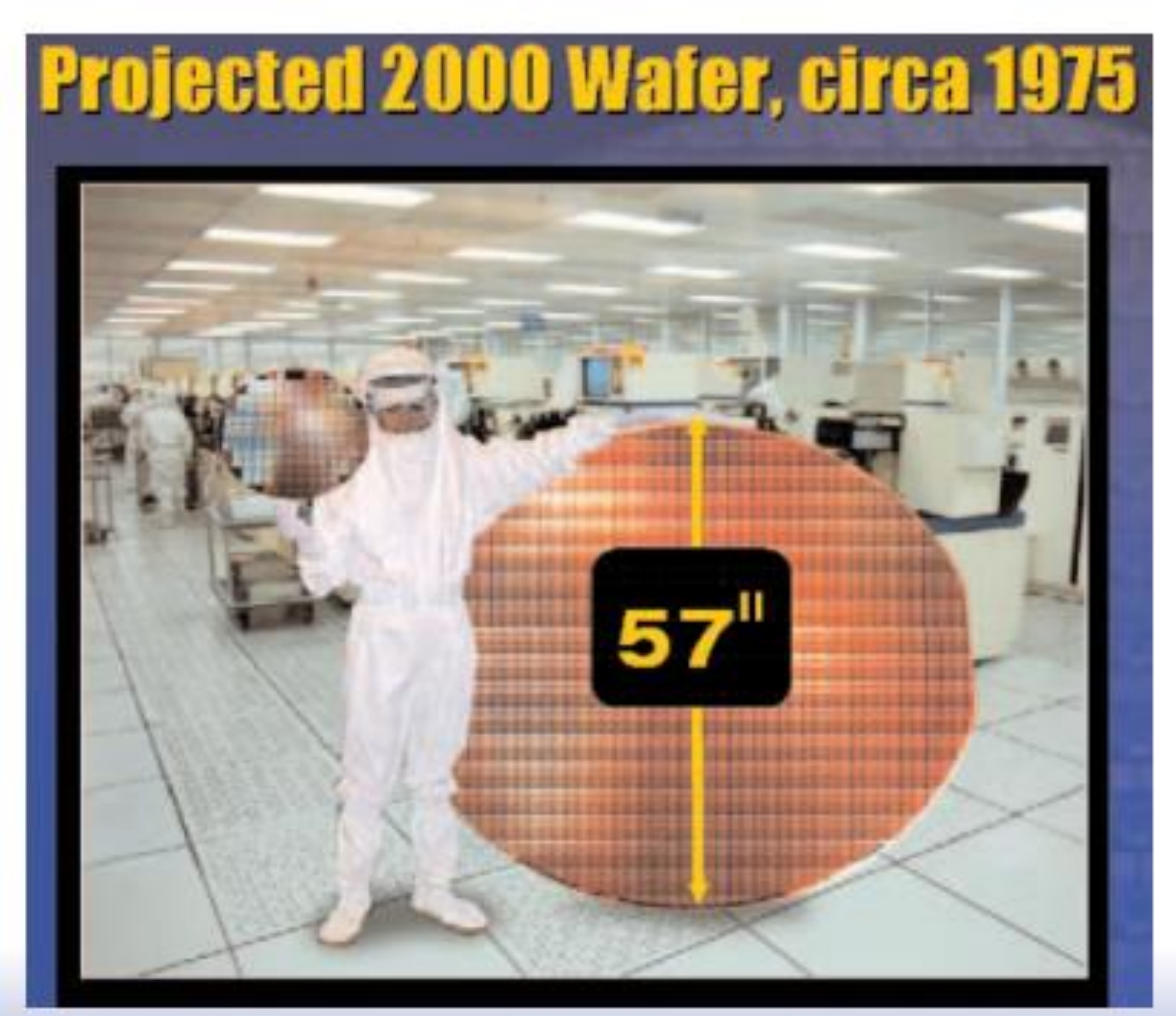
Die Size Growth



Die size grows by 14% to satisfy Moore's Law

Courtesy, Intel

Moore was not always accurate

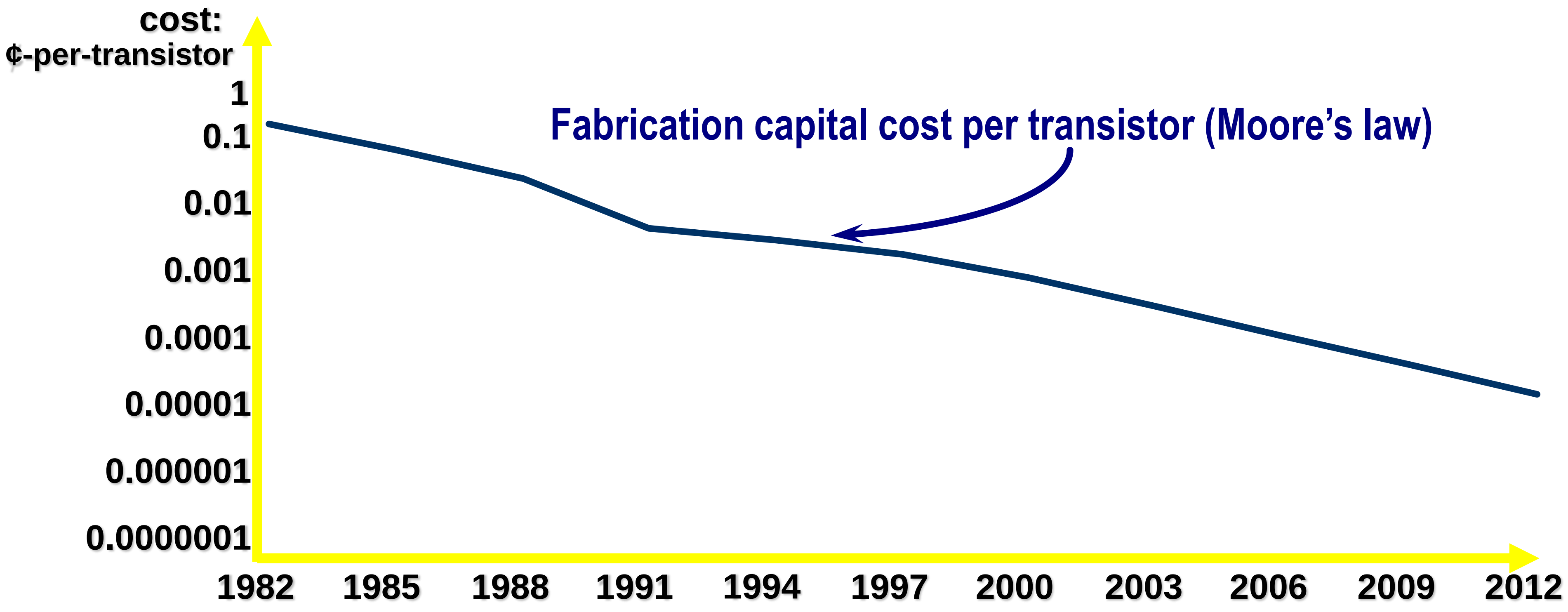


Increasing wafer size:

- More chips per wafer
- Less overhead due to round wafers

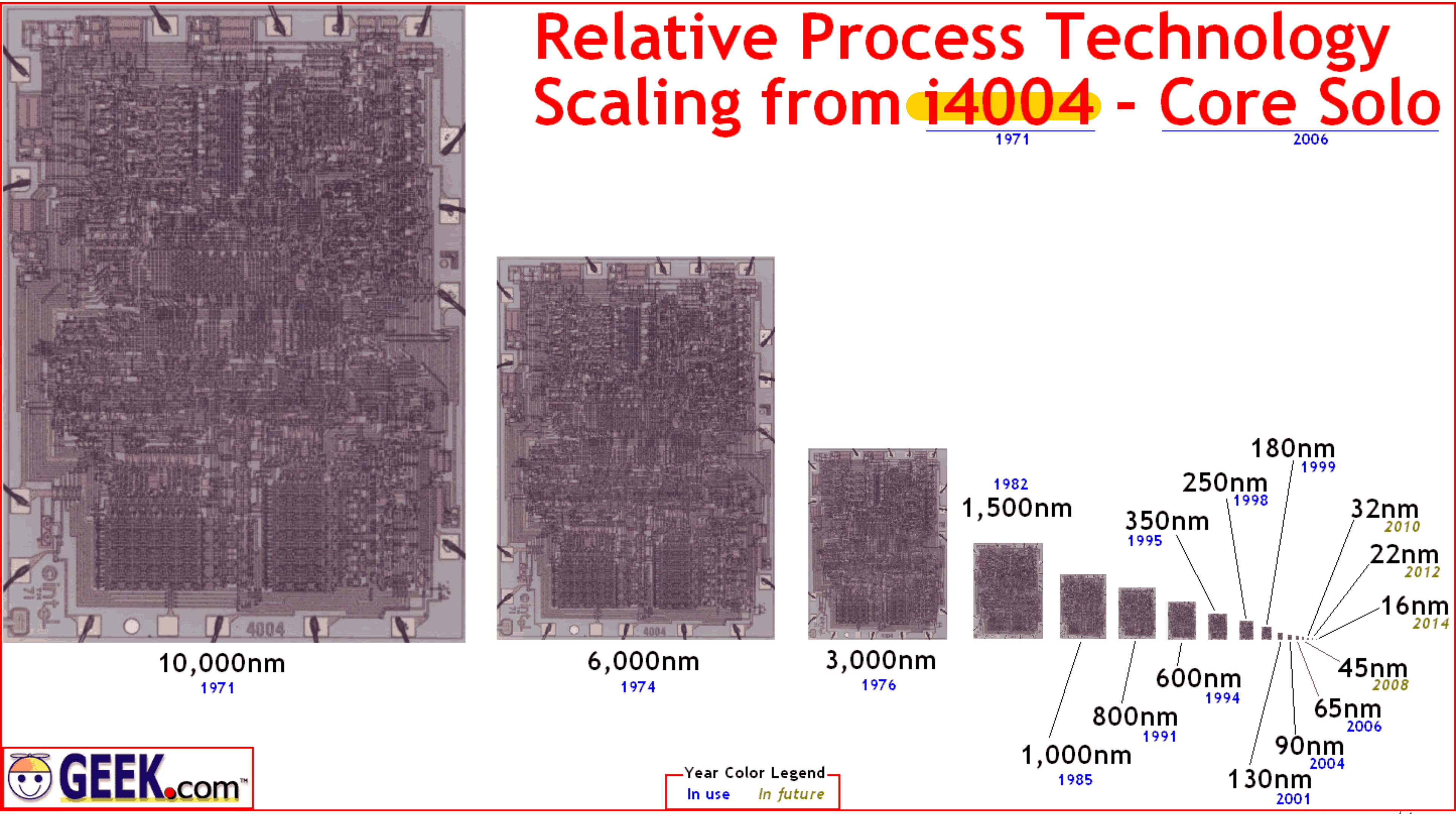
Reality today (2015): 12''
Slowly pushing toward 15''

Cost per Transistor



- Does not include effect of rapidly increasing NREs
- Only valid for very high volume

Scaling…


Relative Process Technology
Scaling from i4004 - Core Solo
1971
2006
10,000nm
1971
6,000nm
1974
3,000nm
1976
1982
1,500nm
1,000nm
1985
800nm
1991
600nm
1994
350nm
1995
250nm
1998
180nm
1999
130nm
2001
90nm
2004
65nm
2006
45nm
2008
32nm
2010
22nm
2012
16nm
2014
Year Color Legend
In use
In future
GEEK.com

Goals of Technology Scaling

- Make things cheaper:
  - Want to sell more functions (transistors) per chip for the same money
  - Build same products cheaper, sell the same part for less money
  - Price of a transistor **has to be reduced**
- But also want to be faster, smaller, and lower power

Technology Scaling – Dennard's Law

- ❑ Technology generation spans 2-3 years

- ❑ Benefits of scaling the dimensions by 30% (Denard):
  - » Double transistor density
  - » Reduce gate delay by 30%

    (increase operating frequency by 43%)
  - » Reduce energy per transition by 65%

    (50% power savings @ 43% increase in frequency

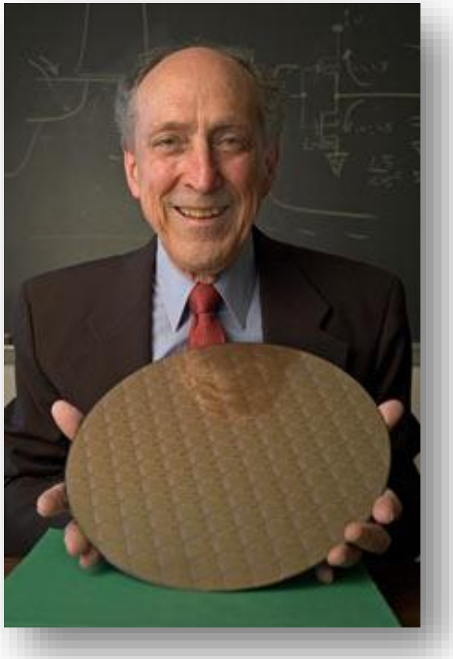
- ❑ Die size used to increase by 14% per generation
  - » Flattens out at 1-4cm^2 (mostly limited by yield issues)

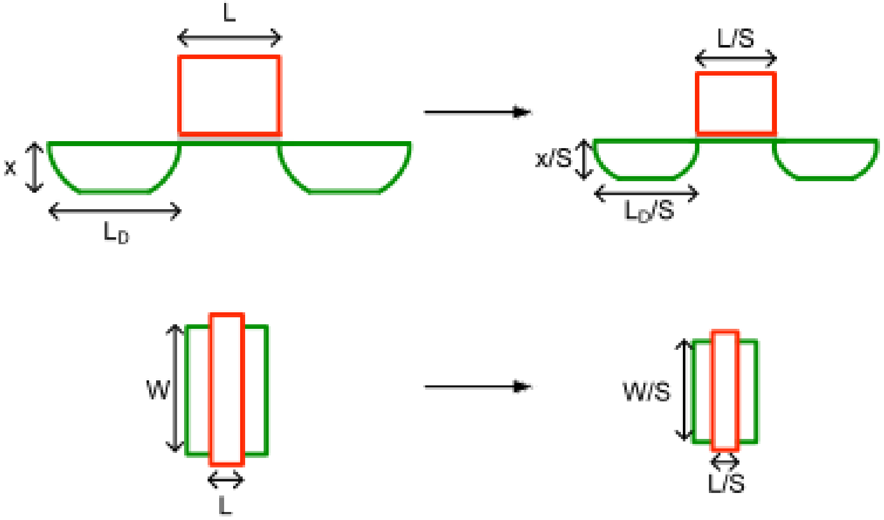
Technology Scaling Models

- Predicting future developments and potentials
- Comparing circuits across different technologies

Dennard Scaling (Constant Field Scaling)

- In 1974, Robert Dennard of IBM described the MOS scaling principles that have accompanied us for forty years.
- As long as we scale all dimensions of a MOSFET by the same amount (*S*), we will arrive at better devices and lower cost, while **maintaining a constant electric field**
  - » L – 1/S
  - » W – 1/S
  - » t_{ox} – 1/S
  - » Na – S
  - » Vdd – 1/S
  - » V_T – 1/S





Dennard (Full) Scaling for Long Transistors

Property	Sym	Equation	Calculation	Scaling	Good?
Oxide Capacitance	C_{ox}	ε_{ox}/t_{ox}	$1/S^{-1}$	S	
Device Area	A	$W \cdot L$	$S^{-1} \cdot S^{-1}$	$1/S^{2}$	☺
Gate Capacitance	C_g	$C_{ox} \cdot W \cdot L$	$S \cdot S^{-1} \cdot S^{-1}$	$1/S$	☺
Transconductance	K_n	$\mu_n C_{ox} W/L$	$S \cdot S^{-1}/S^{-1}$	S	☺
Saturation Current	I_{on}	$K_n V_{GT}^{2}$	$S \cdot S^{-2}$	$1/S$	
On Resistance	R_{on}	V_{DD}/I_{on}	S^{-1}/S^{-1}	1	
Intrinsic Delay	t_{pd}	$R_{on} C_g$	$1 \cdot S^{-1}$	$1/S$	☺
Power	P_{av}	$f \cdot C \cdot V_{DD}^{2}$	$S \cdot S^{-1} \cdot S^{-2}$	$1/S^{2}$	☺☺
Power Density	PD	P_{av}/A	S^{-2}/S^{-2}	1	☺

$$L \propto S^{-1},\ W \propto S^{-1},\ t_{ox} \propto S^{-1},\ V_{\mathrm{DD}} \propto S^{-1},\ V_{\mathrm{T}} \propto S^{-1},\ N_A \propto S$$

Dennard (Full) Scaling for Short Transistors

$V_{DSat} = \xi_{crit} L$

Property	Sym	Equation	Calculation	Scaling	Good?
Oxide Capacitance	C_{ox}	ε_{ox}/t_{ox}	$1/S^{-1}$	S	
Device Area	A	$W \cdot L$	$S^{-1} \cdot S^{-1}$	$1/S^2$	☺
Gate Capacitance	C_g	$C_{ox} \cdot W \cdot L$	$S \cdot S^{-1} \cdot S^{-1}$	$1/S$	☺
Transconductance	K_n	$\mu_n C_{ox} W/L$	$S \cdot S^{-1}/S^{-1}$	S	☺
Saturation Current With velocity saturation	I_{on}	$K_n V_{DSat}(V_{GT} - V_{DSat})$ $K_n V_{GT}^2$	$S \cdot S^{-1} \cdot S^{-1}$ $S \cdot S^{-2}$	$1/S$	
On Resistance	R_{on}	V_{DD}/I_{on}	S^{-1}/S^{-1}	1	
Intrinsic Delay	t_{pd}	$R_{on} C_g$	$1 \cdot S^{-1}$	$1/S$	☺
Power	P_{av}	$f \cdot C \cdot V_{DD}^2$	$S \cdot S^{-1} \cdot S^{-2}$	$1/S^2$	☺☺
Power Density	PD	P_{av}/A	S^{-2}/S^{-2}	1	☺

$V_{GT} = (V_{gs} - V_T)$ ali

$L \propto S^{-1},\ W \propto S^{-1},\ t_{ox} \propto S^{-1},\ V_{DD} \propto S^{-1},\ V_T \propto S^{-1},\ N_A \propto S$

But what if we want more speed?

- We saw that $t_{pd} \propto C_g \cdot V_{DD} / I_{on}$

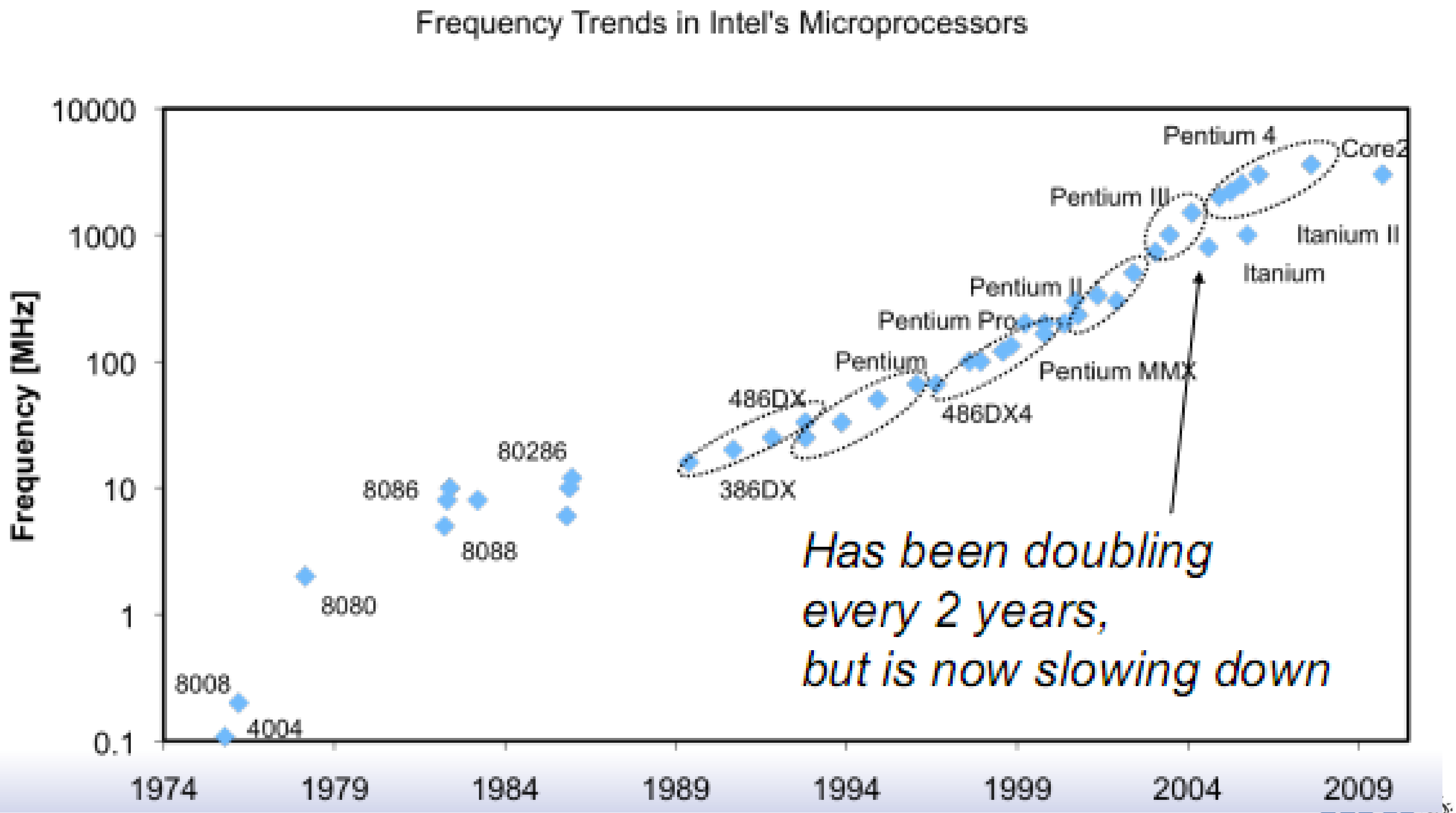
- We can aggressively increase the speed by keeping the voltage constant.
  - » Long channel devices:

$$I_{on} \propto K_n V_{GT}^2 \propto S \;\Rightarrow t_{pd} \propto S^{-1} \cdot 1/S = 1/S^2$$

- This led to the *Fixed Voltage Scaling Model* which was used until the 1990s (V_{DD}=5V)

$$V_{GT} = (V_{gs} - V_T)$$

Moore's Law in Frequency



Has been doubling every 2 years, but is now slowing down

Fixed Voltage Scaling

Property	Sym	Equation	Calculation	Scaling	Good?
Oxide Capacitance	C_{ox}	ε_{ox}/t_{ox}	$1/S^{-1}$	S	
Device Area	A	$W \cdot L$	$S^{-1} \cdot S^{-1}$	$1/S^{2}$	🙂
Gate Capacitance	C_g	$C_{ox} \cdot W \cdot L$	$S \cdot S^{-1} \cdot S^{-1}$	$1/S$	🙂
Transconductance	K_n	$\mu_n C_{ox} W/L$	$S \cdot S^{-1}/S^{-1}$	S	🙂
Saturation Current	I_{on}	$K_n V_{GT}^{2}$	$S \cdot 1$	S	🙂
On Resistance	R_{on}	V_{DD}/I_{on}	$1/S$	$1/S$	
Intrinsic Delay	t_{pd}	$R_{on} C_g$	$S^{-1} \cdot S^{-1}$	$1/S^{2}$	🙂 🙂
Power	P_{av}	$f \cdot C \cdot V_{DD}^{2}$	$S^{2} \cdot S^{-1} \cdot 1$	S	❌
Power Density	PD	P_{av}/A	S/S^{-2}	S^{3}	❌ ❌

$$V_{DD} \propto 1,\ L \propto S^{-1},\ W \propto S^{-1},\ t_{ox} \propto S^{-1},\ V_T \propto S^{-1},\ N_A \propto S$$

Fixed Voltage Scaling – Short Channel

- ❑ What happens under velocity saturated devices?

$$I_{on} \propto K_n V_{DSat} \underbrace{\left(V_{GT} - V_{DSat}\right)} \quad \propto S \cdot S^{-1} \cdot 1 = 1$$

Dominated by Vdd

- ❑ So the on current doesn't increase leading to less effective speed increase.

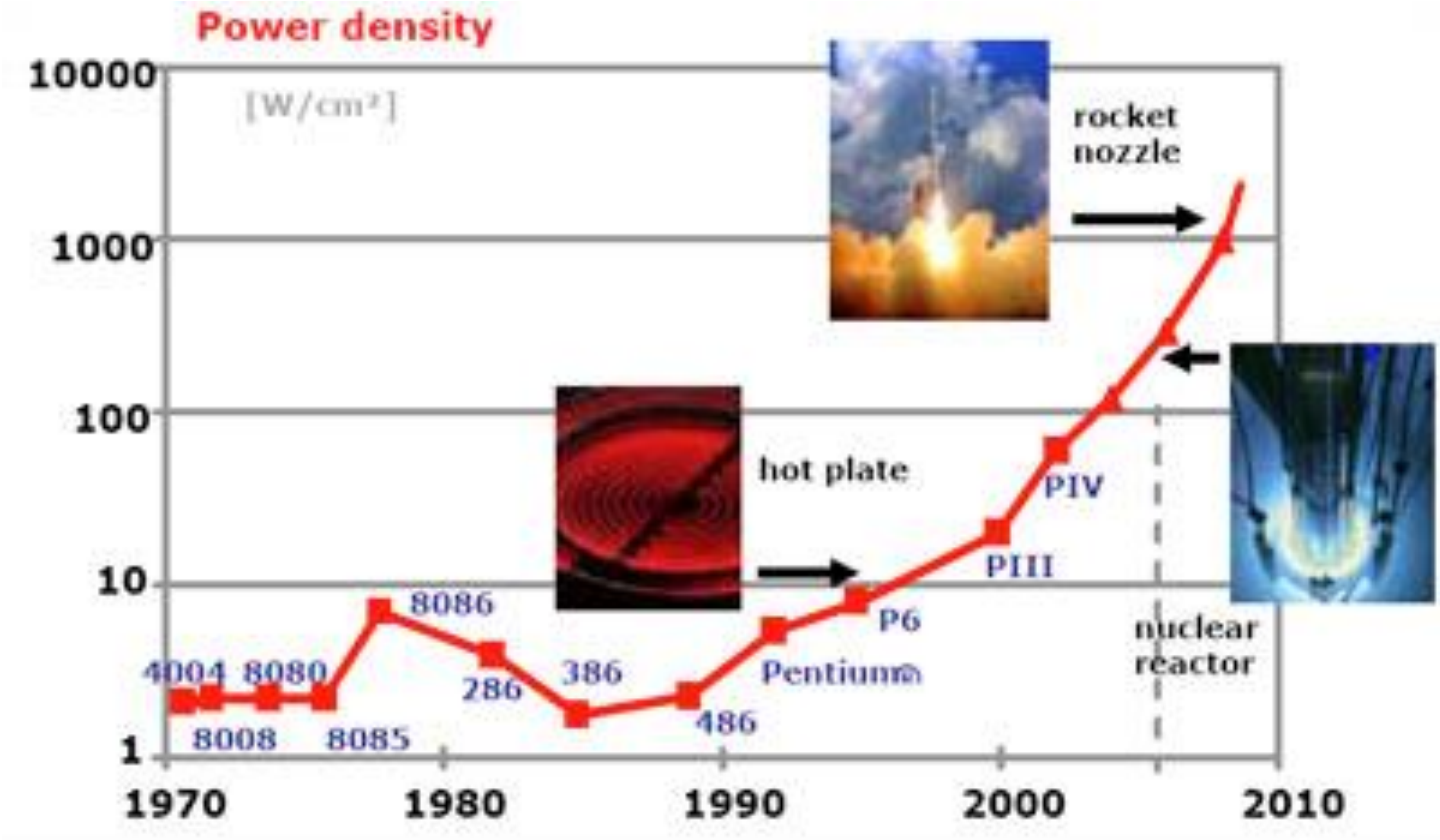
$$t_{pd} \propto R_{on} C_g \propto 1 \cdot S^{-1} = 1/S$$

- ❑ The power density still increases quadratically!

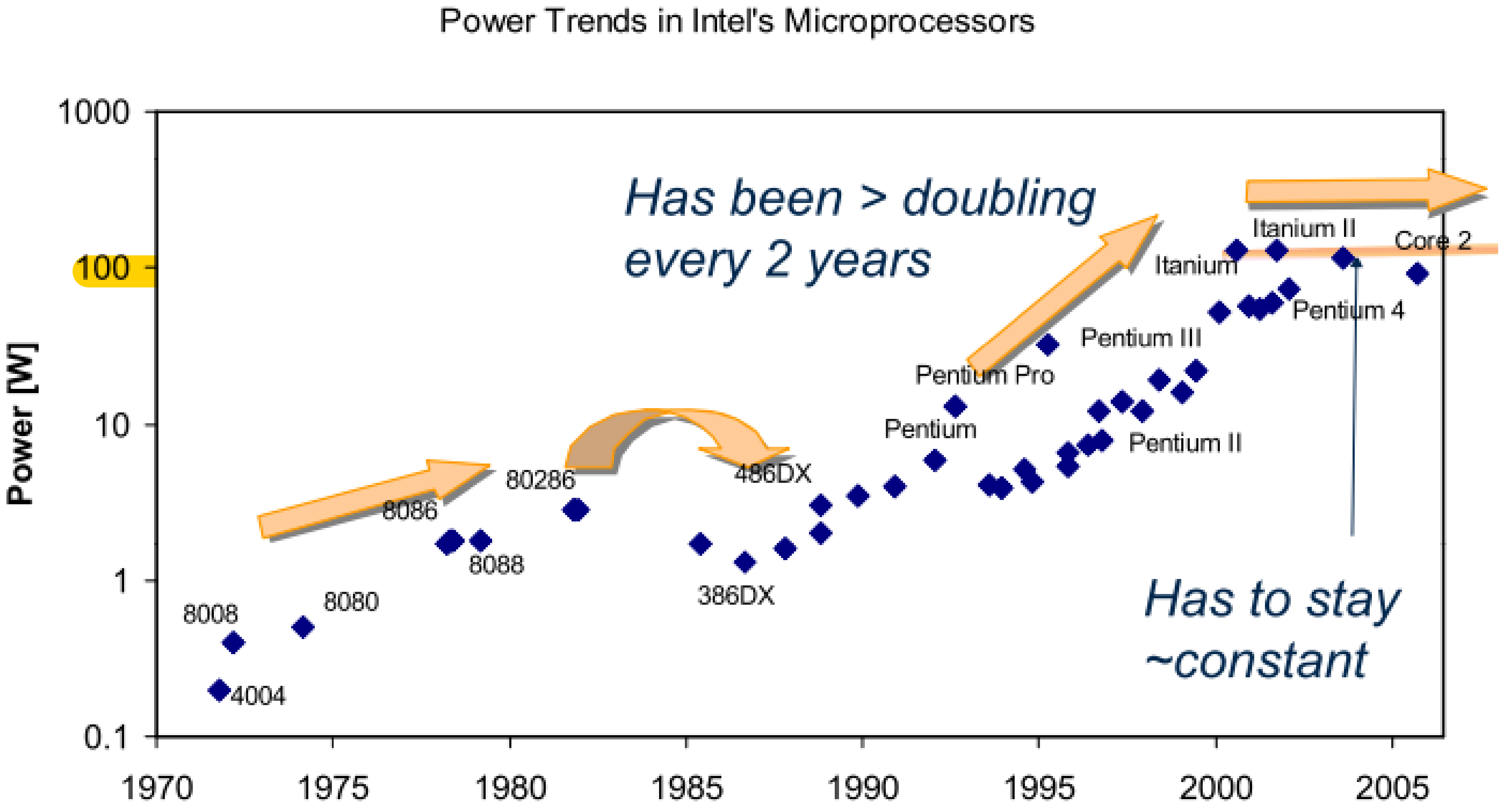
$$PD \propto fCV_{DD}^2 / A \propto S \cdot S^{-1} \cdot 1/S^{-2} = S^2$$

$V_{GT} = (V_{gs} - V_T)$

Power density (2004 expectation)



What actually happened?



Has been > doubling every 2 years

Has to stay ~constant

Technology Scaling Models

- Fixed Voltage Scaling
  - Supply voltages have to be similar for all devices (one battery)
  - Only device dimensions are scaled.
  - 1970s-1990s
- Full “Denard” Scaling (Constant Electrical Field)
  - Scale both device dimensions and voltage by the same factor, *S*.
  - Electrical fields stay constant, eliminates breakdown and many secondary effects.
  - 1990s-2005
- General Scaling –
  - Scale device dimensions by *S* and voltage by *U*.
  - Now!

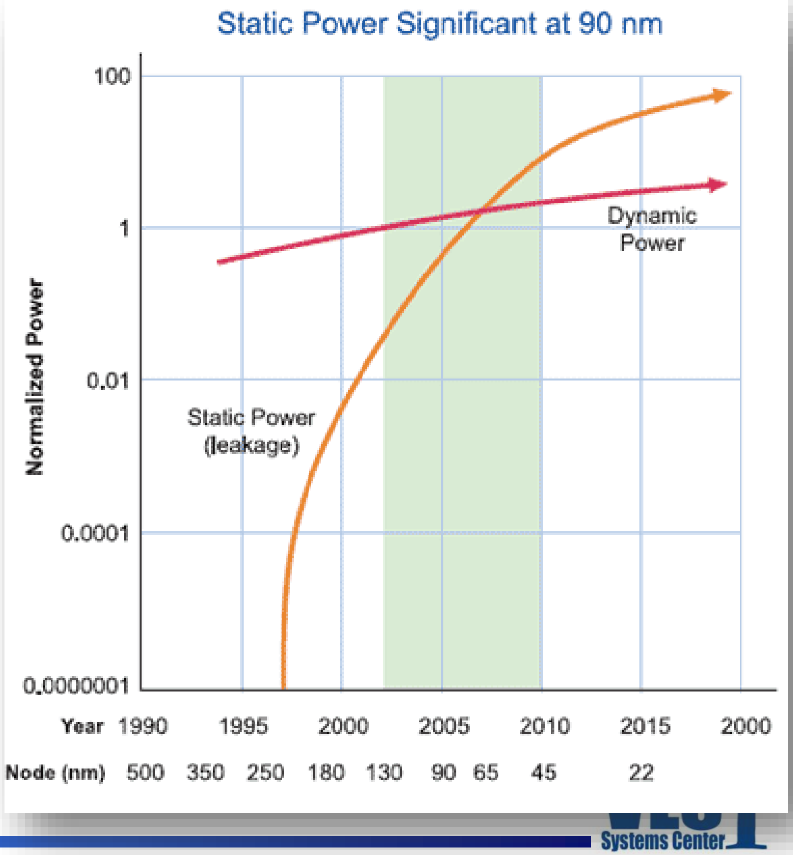
How about Leakage Power?

- The off current is exponentially dependent on the threshold voltage.

$$I_{off} \propto e^{-V_T / n\phi_T}$$

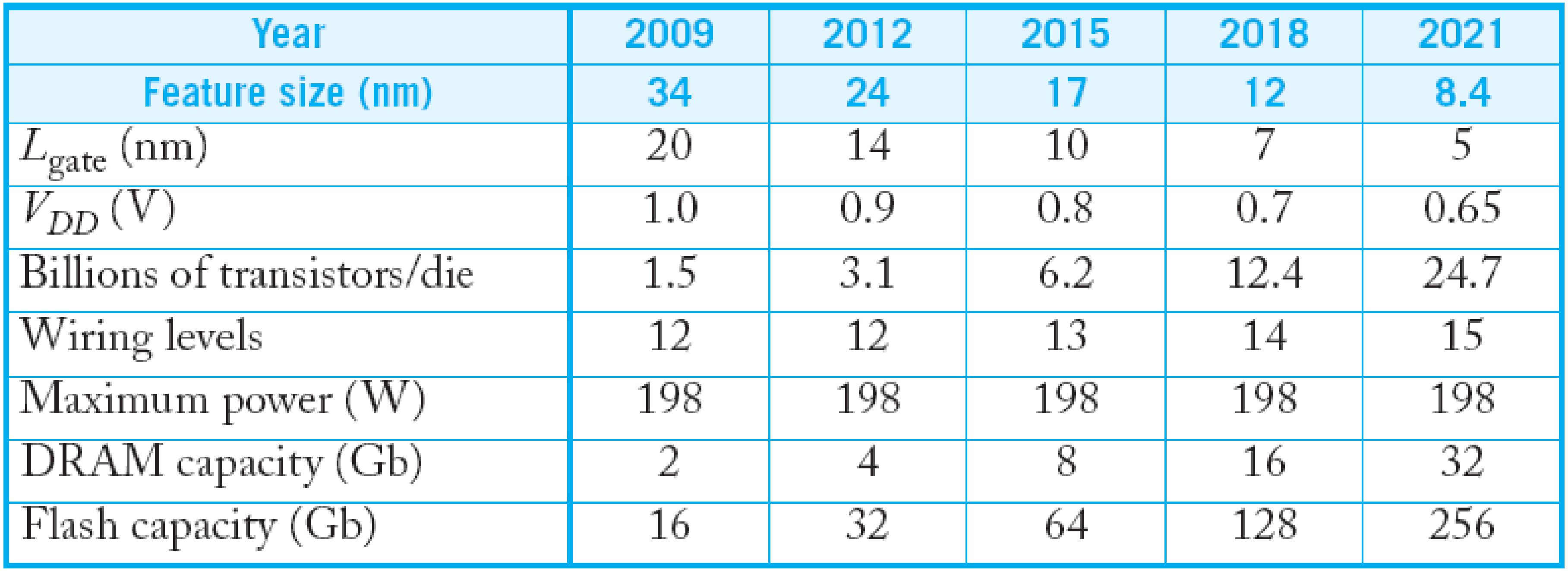
- In the case of *Full Scaling*, the leakage current *increases exponentially* as V_T is decreased!

- Since the 90nm node, static power is one of the major problems in ICs.


Static Power Significant at 90 nm

ITRS

- International Technology Roadmap for Semiconductors



Year	2009	2012	2015	2018	2021
Feature size (nm)	34	24	17	12	8.4
L_{gate} (nm)	20	14	10	7	5
V_{DD} (V)	1.0	0.9	0.8	0.7	0.65
Billions of transistors/die	1.5	3.1	6.2	12.4	24.7
Wiring levels	12	12	13	14	15
Maximum power (W)	198	198	198	198	198
DRAM capacity (Gb)	2	4	8	16	32
Flash capacity (Gb)	16	32	64	128	256

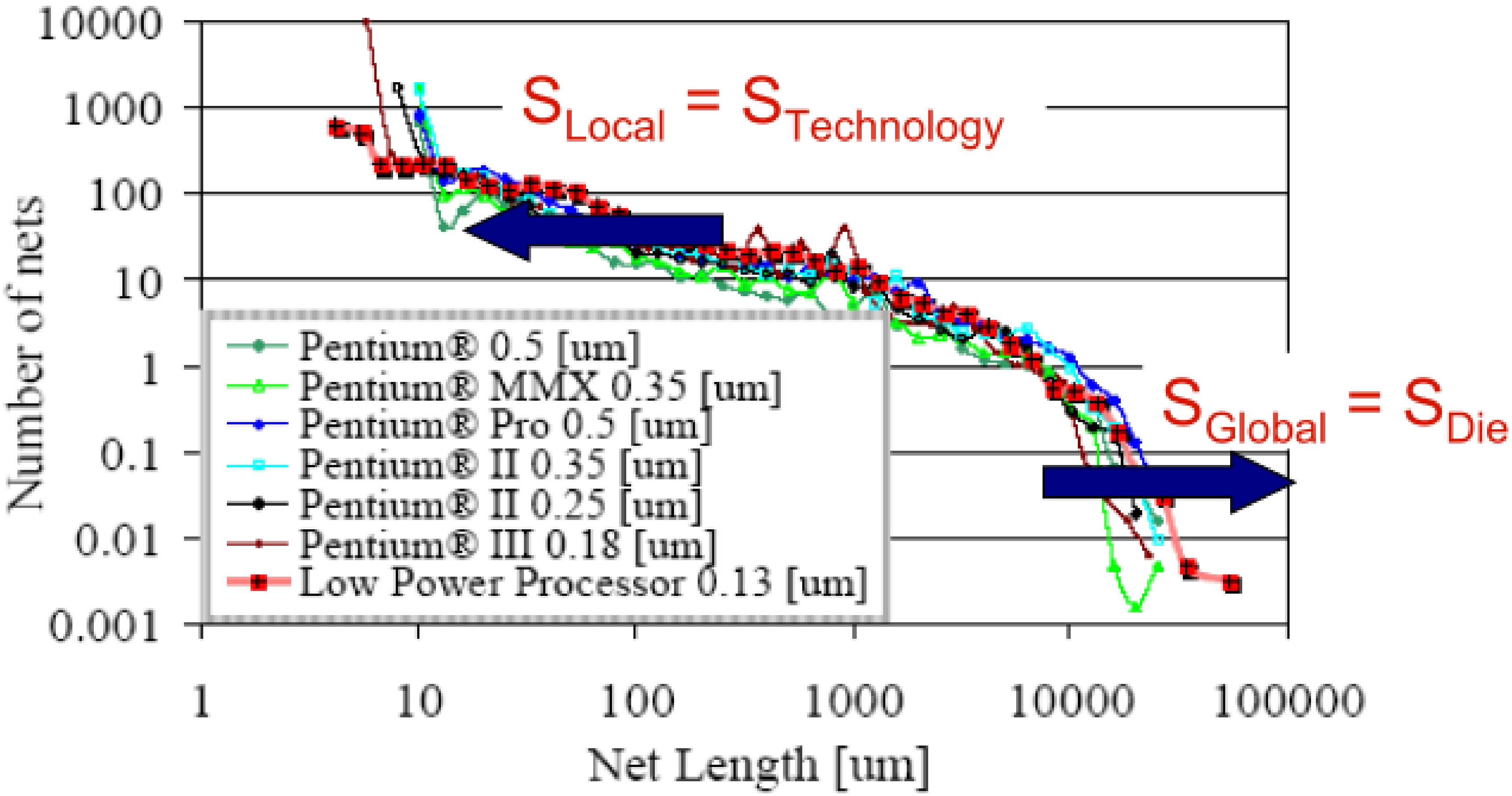
How about Interconnect?

Wire Scaling

- We could try to scale interconnect at the same rate (S) as device dimensions.
  - » This makes sense for *local interconnect* that connects smaller devices/gates.
  - » But *global interconnections*, such as clock signals, buses, etc. won’t scale in *length*.

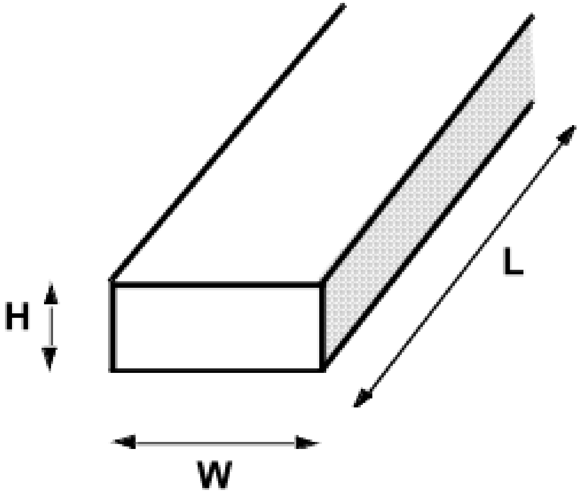
- Length of global interconnect is proportional to *die size* or *system complexity*.
  - » Die Size has increased by 6% per year (X2 @10 years)
  - » Devices have scaled, but complexity has grown!

Nature of Interconnect



From Magen et al., “Interconnect Power Dissipation in a Microprocessor”

Local Wire Scaling



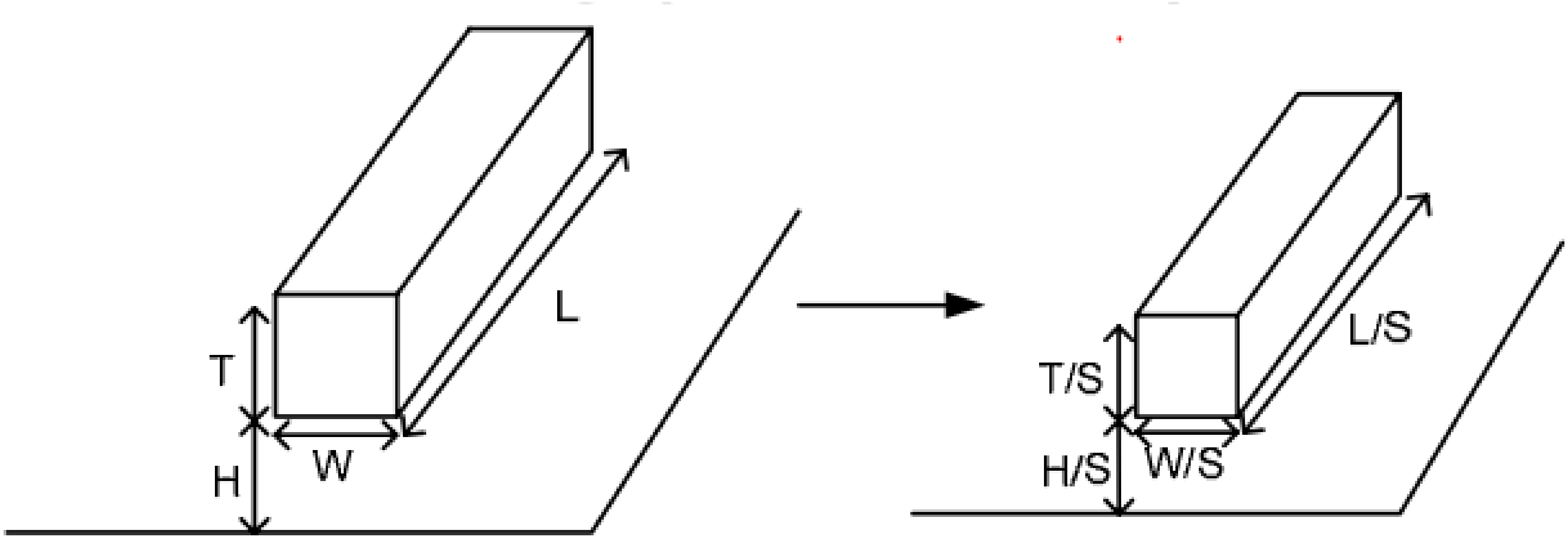
- ❑ Looking at local interconnect:
  - » W, H, t, L all scale at 1/S
  - » C=LW/t→1/S
  - » R=L/WH →S
  - » RC=1

So the delay of local interconnect stays constant.

- ❑ Reminder – Full (Dennard) Scaling of transistors:
  - » Ron=VDD/Ion α 1
  - » tpd=RonCg α 1/S
- ❑ So the delay of local interconnect still increases relative to transistors!

Local Wire Scaling – Full Scaling

- What about fringe cap?

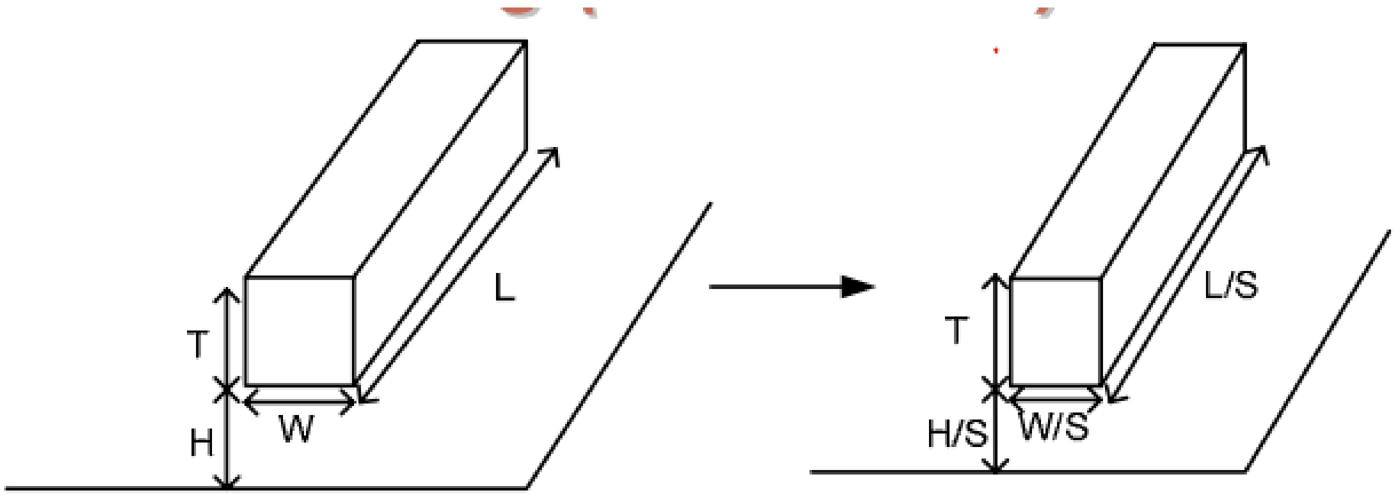


$C_{pp} \alpha WL/H$
$C_{fringe} \alpha \sim L$
$R_w \alpha L/(WT)$
$t_{pwire} \alpha R_w C_w$

$C_{pp}' \alpha 1/S$
$C_{fringe}' \alpha 1/S$
$R_w' \alpha S$
t_{pwire}' const.

Local Wire Scaling - Constant Thickness

- Thickness wasn't scaled!



$C_{pp} \; \alpha \; WL/H$

$C_{fringe} \; \alpha \; \sim L$

$R_w \; \alpha \; L/(WT)$

$t_{pwire} \; \alpha \; R_w C_w$

$C_{pp}' \; \alpha \; 1/S$

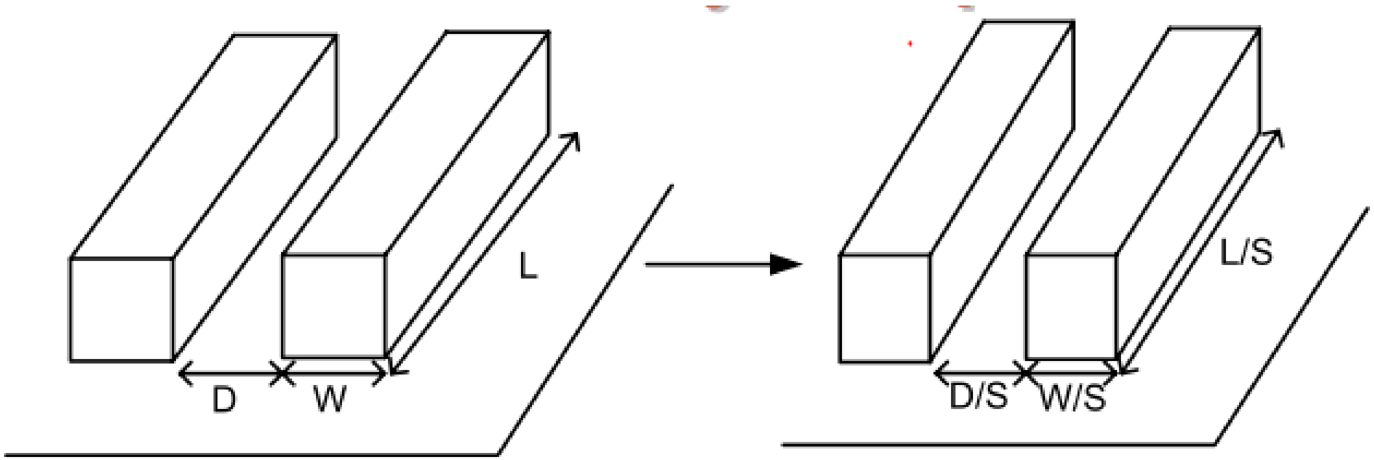
$C_{fringe}' \; \alpha \; 1/S$

R_w' const.

$t_{pwire}' \; \alpha \; 1/S$

Local Wire Scaling – Interwire Capacitance

- Without scaling height, coupling gets much worse.



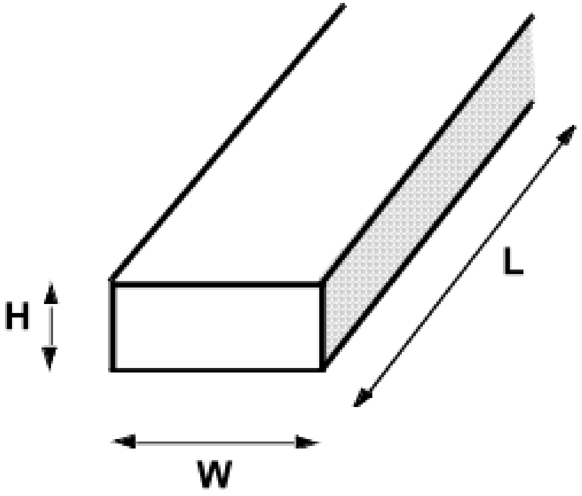
$C_{pp,side} \; \alpha \; LT/D$

$C_{pp,side}'$ const.

- $C_{pp,side}$/Length increases
  → Crosstalk, coupling issues get worse

- Aspect ratio limited – eventually have to scale T
  - Different metal layers have different T

Global Wire Scaling



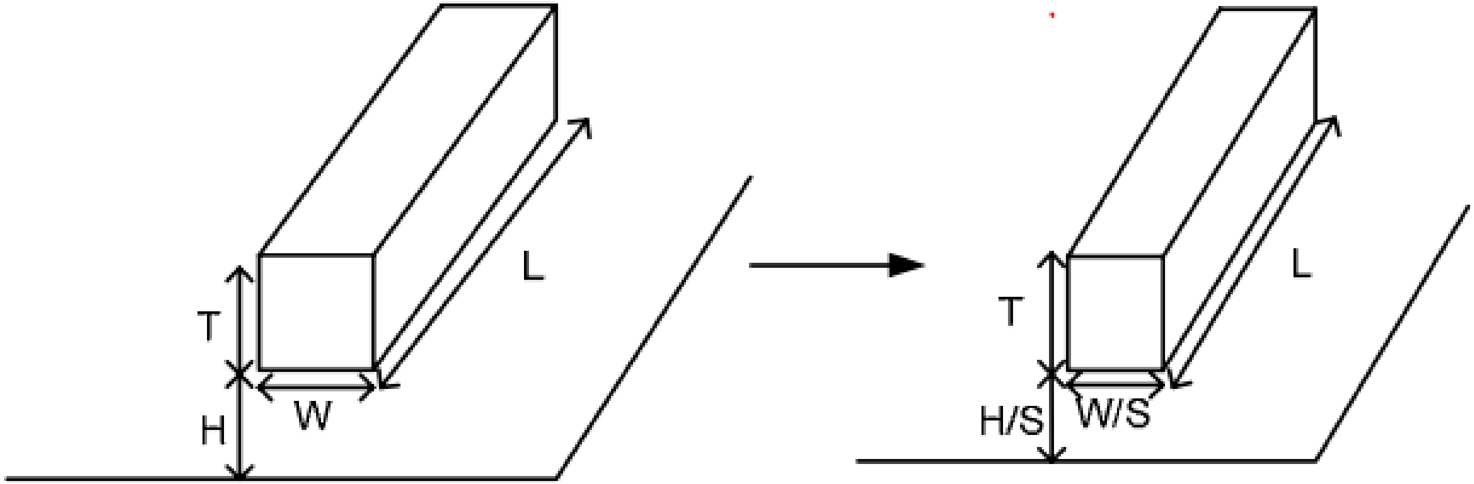
- ❑ Looking at global interconnect:
  - » W, H, t scale at 1/S
  - » L doesn't scale!
  - » C=LW/t→1
  - » R=L/WH →S^2
  - » RC=S^2 !!!

Long wire delay increases quadratically!!!

- ❑ And if chip size grows, *L* actually increases!

Global Wire Scaling – Constant Thickness

- Leave thickness constant for global wires



$C_{pp}\ \alpha\ WL/H$
$C_{fringe}\ \alpha\ {\sim}L$
$R_w\ \alpha\ L/(WT)$
$t_{pwire}\ \alpha\ R_w C_w$

C_{pp}' const
C_{fringe}' ~const
$R_w'\ \alpha\ S$
$t_{pwire}'\ \alpha\ S$

Very bad: wire delay S^2 worse than gates

Wire Scaling

- So whereas device speed increases with scaling:
  - Local interconnect speed stays constant.
  - Global interconnect delays increase quadratically.
- Therefore:
  - Interconnect delay is often the limiting factor for speed.
- What can we do?
  - Keep the wire thickness (*H*) fixed.
  - This would provide *1/S* for local wire delays and *S* for constant length global wires.
  - But fringing capacitance increases, so this is optimistic.

Wire Scaling

- What is done today?
  - Low resistance metals.
  - Low-K insulation.
  - Low metals (M1, M2) are used for local interconnect, so they are thin and dense.
  - Higher metals are used for global routing, so they are thicker, wider and spaced farther apart.

Technology Strategy Roadmap

2000
2005
2010
2015
2020
2025
2030

Plan A: Extending Si CMOS

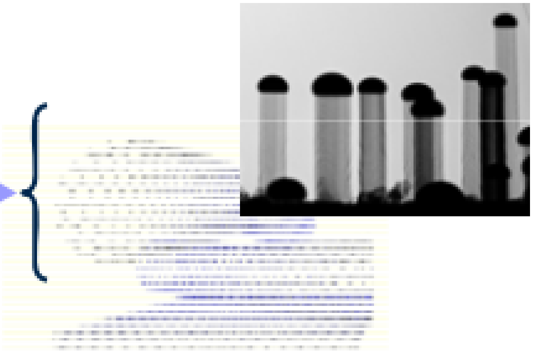
R D

Plan B: Subsytem Integration

R D

Plan C: Post Si CMOS Options

R R&D



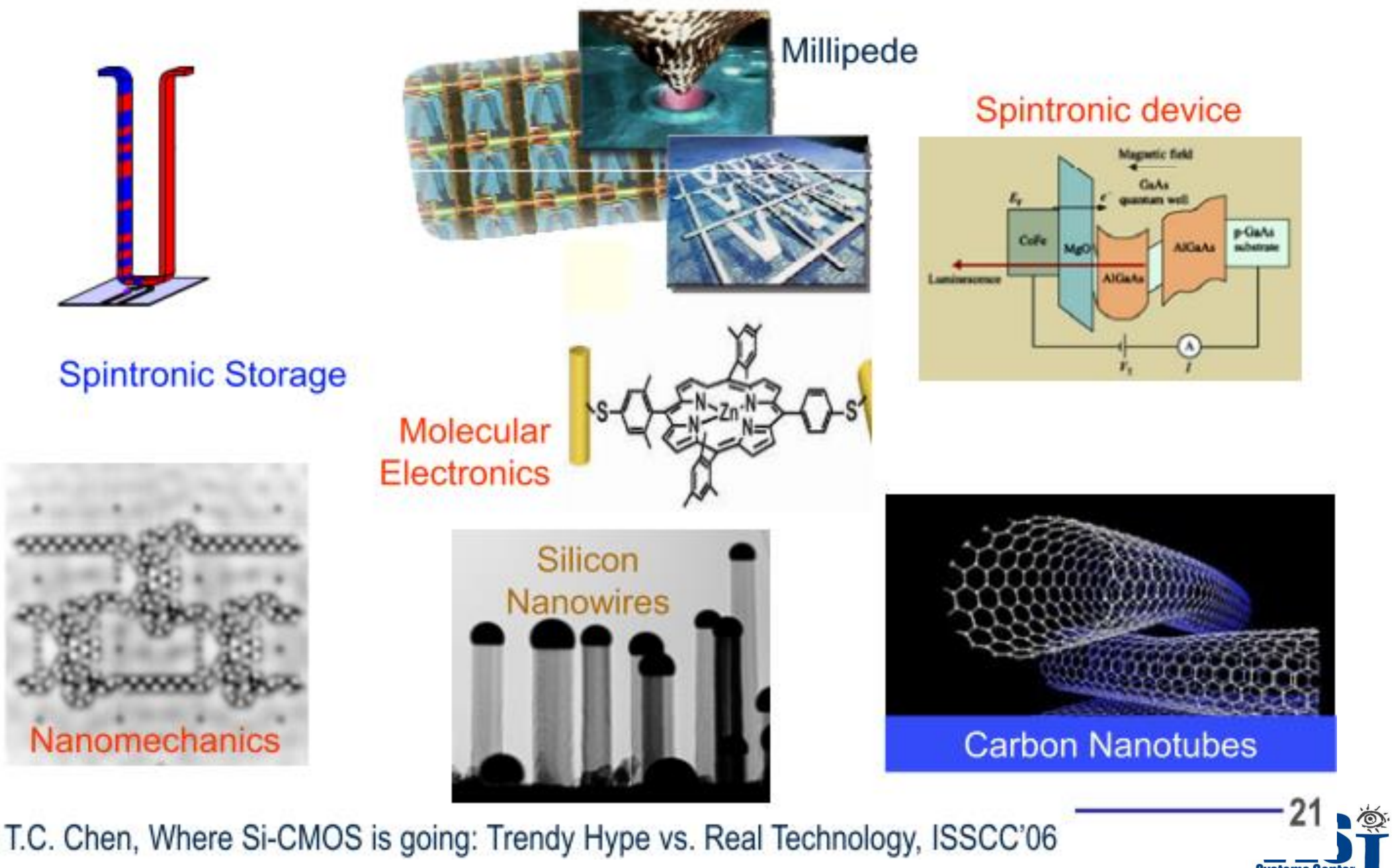
Plan Q: Quantum Computing

R D

22

T.C. Chen, Where Si-CMOS is going: Trendy Hype vs. Real Technology, ISSCC'06

When will Moore’s Law End?



T.C. Chen, Where Si-CMOS is going: Trendy Hype vs. Real Technology, ISSCC’06

Further Reading

- “The International Technology Roadmap for Semiconductors” www.itrs.net
- “The Impact of Dennard’s Scaling Theory”, SSCS Magazine, Winter 2007 http://www.ieee.org/portal/cms_docs_societies/sscs/PrintEditions/200701.pdf
- J. Rabaey, “*Digital Integrated Circuits*” 2003, Chapters 1.1, 3.5, 4, 5.6,
- E. Alon, Berkeley *EE-141*, Lectures 1, 17 (Fall 2009) http://bwrc.eecs.berkeley.edu/classes/icdesign/ee141_f09/
- B. Nicolic, Berkeley EE-241, Lectures 1-5 (Spring 2011) http://bwrc.eecs.berkeley.edu/classes/icdesign/ee241_s11