Scaling Into Nano Devices

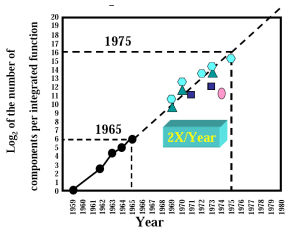
Ken Stevens

Moore's Law

"I see no reason to expect the rate of progress in the use of smaller dimensions in complex circuits to decrease in the near future"

"The new slope might approximate a doubling every two years"

Gordon Moore, December 1975



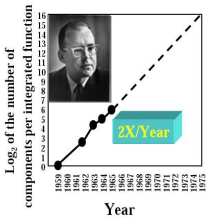
Where Are We Today?

- Does Moore's Law still hold?
  - in my day: $1\mu m$ barrier
- Will it continue to hold?
- What are the current issues?

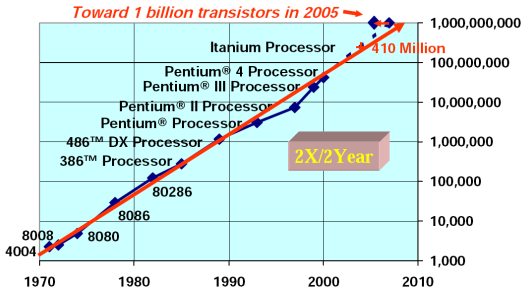
Moore's Law

"Reduced cost is one of the big attractions of integrated electronics, and the cost advantage continues to increase as the technology evolves toward the production of larger and larger circuit functions on a single semiconductor substrate."

Gordon Moore, April 1965



Moore's Law

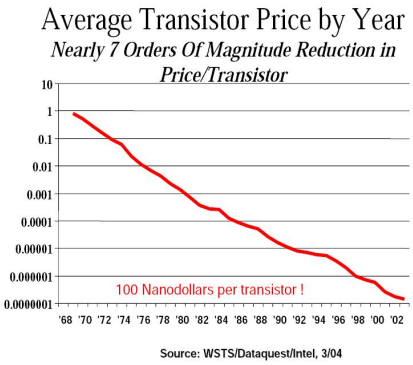


Seems to hold for transistors...

transistors per chip $= 2^{(t-1959)}$ for $1959 \leq t \leq 1975$

$= 2^{16} \times 2^{(t-1959)/1.5}$ for $t \geq 1975$

Moore's Law



Seems to hold for price...

Moore's Law

But *does* it hold for price?

Modern fabrication facility:

- Basketball court size areas packed with litho equipment
- "Stockers" filled with wafers
  - The wafers are idle inventory
  - worth $10,000,000 – $100,000,000
- Cost of Fab plant: ????

Moore's Law

But *does* it hold for price?

Modern fabrication facility:

- Basketball court size areas packed with litho equipment
- "Stockers" filled with wafers
  - The wafers are idle inventory
  - worth $10,000,000 – $100,000,000
- Cost of Fab plant: **5 billion dollars**
- Daily cost for 5-year amortization: ???

Moore's Law

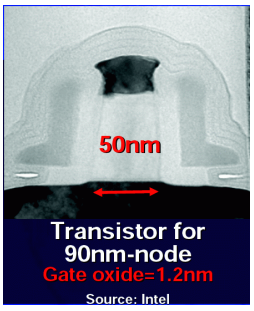
But *does* it hold for price?

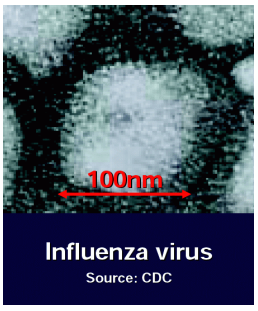
Modern fabrication facility:

- Basketball court size areas packed with litho equipment
- "Stockers" filled with wafers
  - The wafers are idle inventory
  - worth $10,000,000 – $100,000,000
- Cost of Fab plant: 5 billion dollars
- Daily cost for 5-year amortization: **$3,000,000**

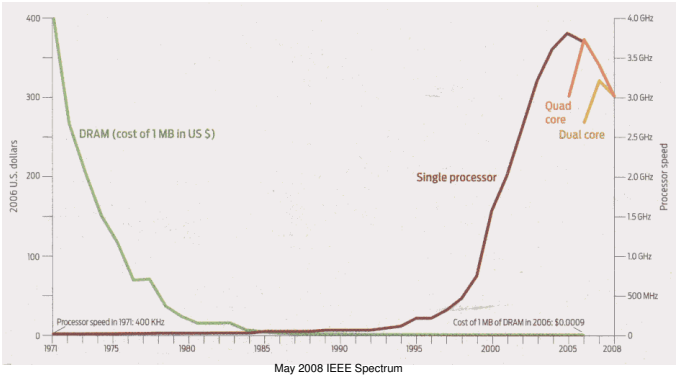
Moore's Law in Action

Our three-generation-old technology has a channel length approximately half the size of an influenza virus!





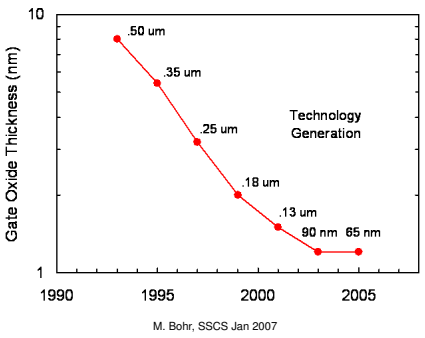
Moore's Law Limits?



May 2008 IEEE Spectrum

Why is there a frequency drop-off?

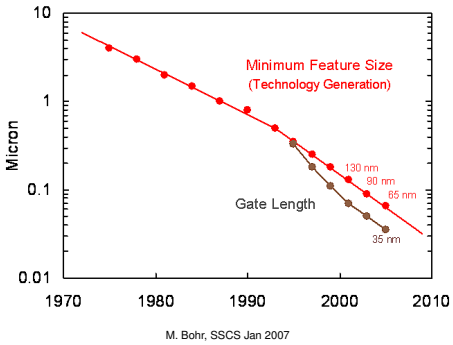
Moore's Law Limits?



M. Bohr, SSCS Jan 2007

Why is there a halt in oxide thickness scaling?

Moore's Law Limits?



M. Bohr, SSCS Jan 2007

Why is there an *increase* in feature size scaling?

Scaling Theory

Applying Scaling Theory To modern process technology will answer all these questions and Moore...

- There are numerous tradeoffs
  - physical limits (many other $1\mu m$ barriers)
  - cost vs almost everything
  - performance vs reliability
  - performance vs power
  - ...
- Will scaling now halt?

Quick overview will follow

Transistor Tradeoffs

$$I_{DSat} \approx \frac{1}{2}\mu C_{ox}\frac{W(V_{dd}-V_{th})^2}{L}$$

- Performance $\approx I_{DSat}$
  - Increase C_{ox}
    $C_{ox} = \frac{\varepsilon_{ox}\varepsilon_{si}}{t_{ox}}$
  - Increase μ
  - Reduce L
- Power $= V_{dd}^2 C_{ox} F_{Max}$
  - Reduce V_{dd}

What about V_{th}??

Transistor Tradeoffs

$$I_{DSat} \approx \frac{1}{2}\mu C_{ox}\frac{W(V_{dd}-V_{th})^2}{L}$$

- Performance $\approx I_{DSat}$
  - Increase C_{ox}
    $C_{ox} = \frac{\varepsilon_{ox}\varepsilon_{si}}{t_{ox}}$
  - Increase μ
  - Reduce L
- Power $= V_{dd}^2 C_{ox} F_{Max}$
  - Reduce V_{dd}

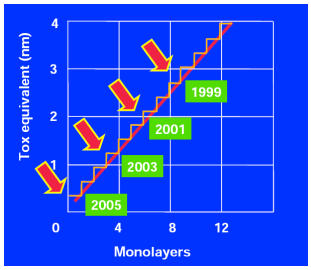
What about V_{th}: You'll learn there are physical limits to its reduction.

Increase Performance

Increase C_{ox}
$\Longrightarrow$ Reduce t_{ox}

$C_{ox} = \frac{\varepsilon_{ox}\varepsilon_{si}}{t_{ox}}$

Can we keep reducing t_{ox} ???

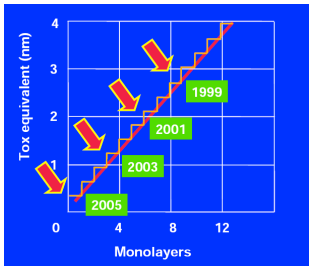


Increase Performance

Increase C_{ox}
$\Longrightarrow$ Reduce t_{ox}

$C_{ox} = \frac{\varepsilon_{ox}\varepsilon_{si}}{t_{ox}}$

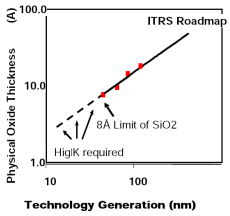
Can we keep reducing t_{ox} ???
Not without new materials!
In 2005 we were 2-3 atomic layers thick.



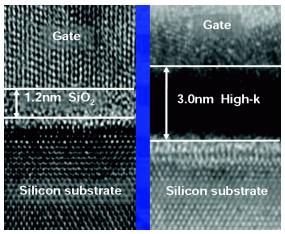
Increase Performance

Increase C_{ox}

$$C_{ox} = \frac{\varepsilon_{ox}\varepsilon_{si}}{t_{ox}}$$



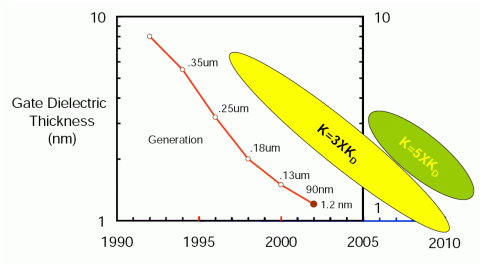
Intel currently uses Hafnium Oxide (starting in the 65nm process)



Increase Performance

Increase C_{ox}

$$C_{ox} = \frac{\varepsilon_{ox}\varepsilon_{si}}{t_{ox}}$$

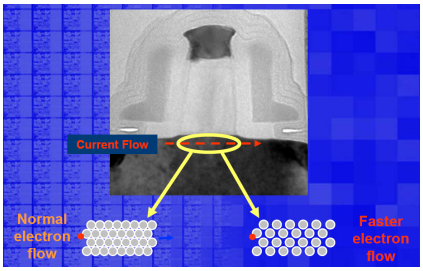


- High-κ dielectrics don't change the slope
- They thus give us a few generations of scaling

Increase Performance

Increase μ

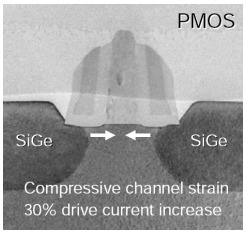
Strained Silicon increases device mobility

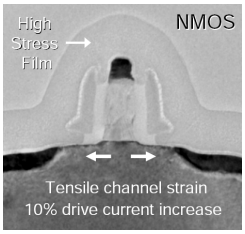


Increase Performance

Increase μ

Strained Silicon increases device mobility

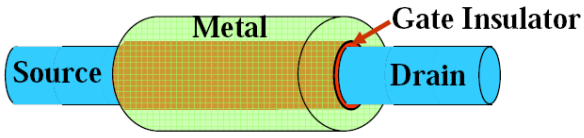


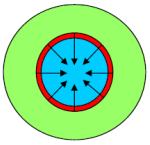


Increase Performance

Increase μ

Optimal field injection is *not* planar!



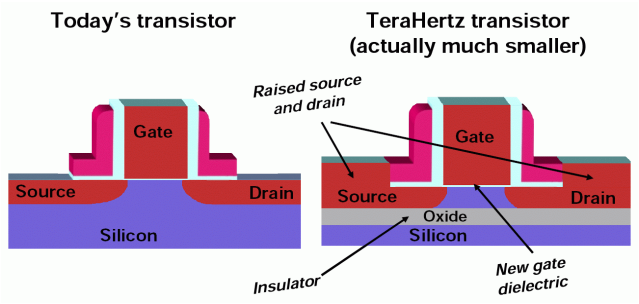


Increase Performance

Increase μ

- Progression
  - TeraHertz Transistor
  - finfet
  - trigate
  - nanowires

TeraHertz Transistor

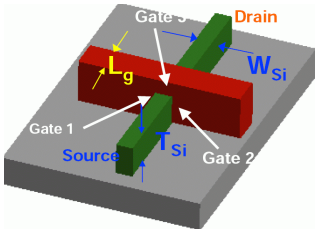


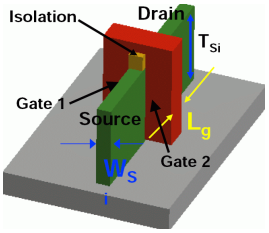
TeraHertz Transistor

- Problem
  1. Tunneling current across dielectric
  2. Leakage current from source to drain
  3. High voltage requirements
- Reason
  1. thin t_{ox}
  2. deep sub$-\mu$ devices
  3. oxide barrier increases source and drain resistance
- Solution
  1. high κ dielectric creates thicker t_{ox}
  2. oxide barrier under channel blocks current path and reduces current by $100\times$
  3. make source and drain thicker, which reduces the resistance by 30%, lowering voltage requirements

Increase Performance

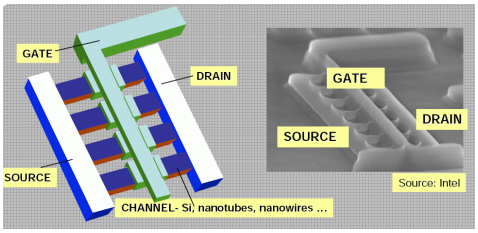
Increase μ





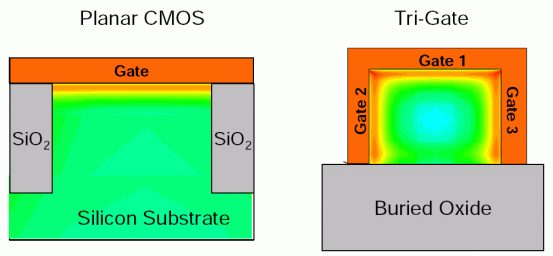
Increase Performance

Increase μ



Increase Performance

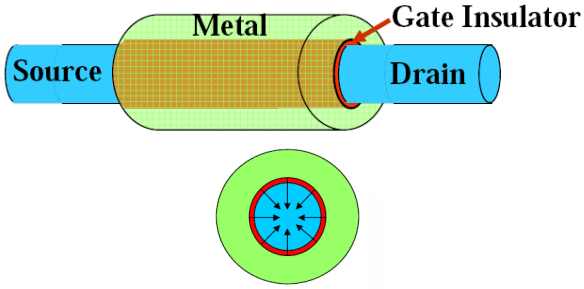
Increase μ



Increase Performance

Increase μ

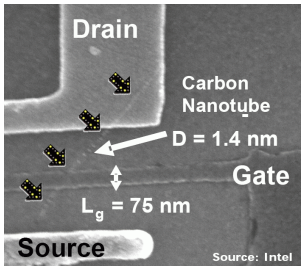
Ideal transistor topology is *round*



Increase Performance

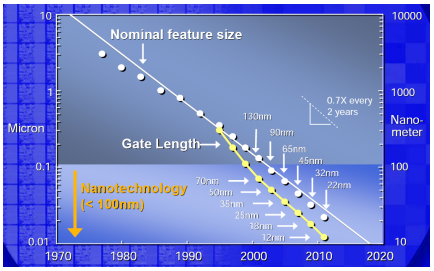
Increase μ

Ideal transistor topology is *round*



Performance

Reduce L



New process naming convention is *not* based on traditional scaling values (0.7$\times$ previous generation)

Transistor Tradeoffs

$$I_{DSat} \approx \frac{1}{2}\mu C_{ox}\frac{W(V_{dd}-V_{th})^2}{L}$$

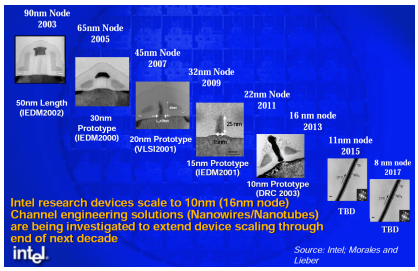
- Performance $\approx I_{DSat}$
  - Increase C_{ox}
    $C_{ox} = \frac{\varepsilon_{ox}\varepsilon_{si}}{t_{ox}}$
  - Increase μ
  - Reduce L
- Power $= V_{dd}^2 C_{ox} F_{Max}$
  - Reduce V_{dd}

Power Scaling

Silicon technology has evolved to increase power efficiency

Mid 1960's	Bipolar, PMOS
Mid 1970's	NMOS
Mid 1980's	CMOS
Mid 1990's	CMOS, Voltage Scaling
Mid 2000's	CMOS, Power efficient Process

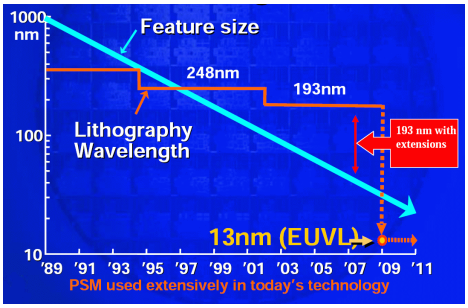
Transistor Scaling



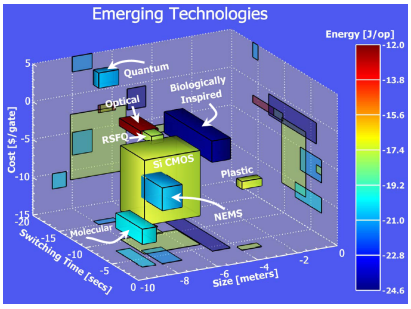
Future Technology Changes

Process Node	250nm	180nm	130nm	90nm	65nm	45nm	32nm	22nm
Introduced	1997	1999	2001	2003	2005	2007	2009	2011
Wafer Size	200mm	200mm	300mm	300mm	300mm	300mm	300mm	450mm
Interconnect	Al	Al	Cu	Cu	Cu	Cu	Cu	?
Channel	Si	Si	Si	strained Si	strained Si	strained Si	trigate?	trigate?
Gate dielec.	SiO_2	SiO_2	SiO_2	SiO_2	SiO_2	high-k	high-k	high-k
Gate electr.	Poly-Si	Poly-Si	Poly-Si	Poly-Si	Poly-Si	metal	metal	metal
ILD	SiO_2	SiOF	SiOF	SiOC	SiOC	SiOC	SiOC	?
Litho	248nm	248nm	248nm	193nm	193nm	193nm	13nm	13nm

Lithography



Silicon Still the Best

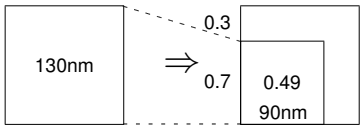


Process Scaling

Ken Stevens

Moore's Law

- Every generation transistors double or area halves



- Therefore, under ideal scaling:
  1. scaling factor $s = 0.7$
  2. area $= s^2 = 0.49$
  3. capacitance $c = 0.7$
  4. cap / area $= 0.7/0.49 = 1.43$

How Does a μ–Processor Scale?

Processor	Data		solve using $P = CV^2f$
Alpha 21164	50W @ 300MHz 3.3V 0.5μ 16.5 × 18.1mm 300mm^2	$C/mm^2 \approx$	$C = P/V^2f$ 51pF/mm^2
Power PC (mobile)	5W @ 250MHz 2.5V 0.3μ 67mm^2	$C/mm^2 \approx$	48pF/mm^2
UltraSparc III	70W @ 600MHz 1.8V 0.25μ 360mm^2	$C/mm^2 \approx$	100pF/mm^2

How many process generations between designs??

How does a design change per process generation??

What C/mm^2 should a similar design have to the Alpha??

How Does a μ–Processor Scale?

Processor	Data		solve using $P = CV^2f$
Alpha 21164	50W @ 300MHz 3.3V 0.5μ 16.5 × 18.1mm 300mm^2	$C/mm^2 \approx$	$C = P/V^2f$ 51pF/mm^2
Power PC (mobile)	5W @ 250MHz 2.5V 0.3μ 67mm^2	$C/mm^2 \approx$	48pF/mm^2
UltraSparc III	70W @ 600MHz 1.8V 0.25μ 360mm^2	$C/mm^2 \approx$	100pF/mm^2

How many process generations between designs?? **2**

How does a design change per process generation??

What C/mm^2 should a similar design have to the Alpha??

How Does a μ–Processor Scale?

Processor	Data		solve using $P = CV^2f$
Alpha 21164	50W @ 300MHz 3.3V 0.5μ 16.5 $\times$ 18.1mm 300mm^2	$C/mm^2 \approx$	$C = P/V^2f$ 51pF/mm^2
Power PC (mobile)	5W @ 250MHz 2.5V 0.3μ 67mm^2	$C/mm^2 \approx$	48pF/mm^2
UltraSparc III	70W @ 600MHz 1.8V 0.25μ 360mm^2	$C/mm^2 \approx$	100pF/mm^2

How many process generations between designs?? **2**

How does a design change per process generation?? scaling theory

What C/mm^2 should a similar design have to the Alpha??

How Does a μ–Processor Scale?

Processor	Data		solve using $P = CV^2f$
Alpha 21164	50W @ 300MHz 3.3V 0.5μ 16.5 $\times$ 18.1mm 300mm^2	$C/mm^2 \approx$	$C = P/V^2f$ 51pF/mm^2
Power PC (mobile)	5W @ 250MHz 2.5V 0.3μ 67mm^2	$C/mm^2 \approx$	48pF/mm^2
UltraSparc III	70W @ 600MHz 1.8V 0.25μ 360mm^2	$C/mm^2 \approx$	100pF/mm^2

How many process generations between designs?? **2**

How does a design change per process generation?? scaling theory

What C/mm^2 should a similar design have to the Alpha?? s^2/s^4

How Does a μ–Processor Scale?

Processor	Data		solve using $P = CV^2f$
Alpha 21164	50W @ 300MHz 3.3V 0.5μ 16.5 $\times$ 18.1mm 300mm^2	$C/mm^2 \approx$	$C = P/V^2f$ 51pF/mm^2
Power PC (mobile)	5W @ 250MHz 2.5V 0.3μ 67mm^2	$C/mm^2 \approx$	48pF/mm^2
UltraSparc III	70W @ 600MHz 1.8V 0.25μ 360mm^2	$C/mm^2 \approx$	100pF/mm^2

How many process generations between designs?? **2**

How does a design change per process generation?? scaling theory

What C/mm^2 should a similar design have to the Alpha?? s^2/s^4

where $s = 0.7$: 2×51pf/mm$^2 \Longrightarrow$ the UltraSparc

Cap / Area of μ–processors

- What are inaccuracies, approximations?
  - process scaling
  - area
  - frequency
  - logic styles
  - architecture (pipe depth, parallelism, speculation, ...)
  - physical design style (CBD, full custom, ...)
- Is this reasonable for direct *process shrink*?

Cap / Area of μ–processors

- What are inaccuracies, approximations?
  - process scaling
  - area
  - frequency
  - logic styles
  - architecture (pipe depth, parallelism, speculation, ...)
  - physical design style (CBD, full custom, ...)
- Is this reasonable for direct *process shrink*?
  Yes, much more so than across designs.

Cap / Area of μ–processors

- What are inaccuracies, approximations?
  - process scaling
  - area
  - frequency
  - logic styles
  - architecture (pipe depth, parallelism, speculation, ...)
  - physical design style (CBD, full custom, ...)
- Is this reasonable for direct *process shrink*?
  Yes, much more so than across designs.
- Why might the Power PC be so different than the other processors?

Cap / Area of μ–processors

- What are inaccuracies, approximations?
  - process scaling
  - area
  - frequency
  - logic styles
  - architecture (pipe depth, parallelism, speculation, ...)
  - physical design style (CBD, full custom, ...)
- Is this reasonable for direct *process shrink*?
  Yes, much more so than across designs.
- Why might the Power PC be so different than the other processors? Mobile design using low power techniques

Power Trends

- Classic scaling keeps power per unit area constant
- Not true in reality
- Three main problems
  - cooling
  - current
  - δI/δt
- Other significant issues
  - glitching (static and dynamic hazards)
  - coupling and noise

Power Trends

- First Order:
  $P = CV_{dd}^2 f$
- ignores:
  1. short circuit current
  2. leakage
  3. DC current (sense amps, etc.)

Power Scaling

In ideal process scaling applied to power: $(P = CV_{dd}^2 f)$

- Capacitance $= s$
- Frequency $= \frac{1}{s}$
- Voltage $= s$

so

- $P = s \times s^2 \times \frac{1}{s} = s^2$

but

- $P = s \times V_{dd}^2 \times \frac{1}{s}$
- $P = V_{dd}^2$

Therefore area and voltage determine power with ideal scaling!

Relation between Power and Performance

Two main equations used (ε = energy, τ = delay):

- $\varepsilon \cdot \tau$
- $\varepsilon \cdot \tau^2$

What are they both best at representing??

- $\varepsilon \cdot \tau$ *think C and f*
  - performance under ideal scaling
- $\varepsilon \cdot \tau^2$ *what happens to τ as V drops?*
  - power at same performance with voltage scaling
  - good metric when using aggressive voltage scaling
    - ("Speed-step" technology)

Active Power Control

- Power = heat
- CPUs now have temperature sensors
- Throttle when too hot
- What happens if heat sink removed??

Active Power Control

- Power = heat
- CPUs now have temperature sensors
- Throttle when too hot
- What happens if heat sink removed??
  - P4 ran slower
  - Athlon caught on fire

Basic Transistor Scaling Theory

- If $V_{in} = V_{cc}$ then CONDUCT else INSULATE
- If $V_{in} > V_{th}$ then CONDUCT else INSULATE
- ...

Basic concept:

- transconductance for fixed W (width) increases with reduced L (length)
- intrinsic capacitance lowers with L

So to scale, we want to reduce L!

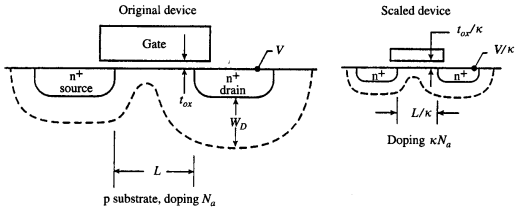
Basic Transistor Scaling Theory

- Cannot arbitrarily reduce L due to:
  - short channel effects
  - body effect
  - hot carrier injection
  - variation

Basic Scaling Theory

Ideal scaling is based on the concept of:
Constant Field Scaling

- scale all dimensions by s
- scale voltage by s
- scale doping by $1/s$



Scaling Table - Constant Field Scaling

I prefer $s = 0.7$ (Moore's Law), book uses $\kappa = 1/s = 1.43$

Constant Field Scaling is broken into three sections

1. scaling assumptions
2. derived scaling behaviors of device parameters based on scaling assumptions.
3. derived scaling behaviors of circuit parameters

Device Scaling Assumptions

assumptions	device dimensions (t_{ox}, W, L)	s
	voltage (V)	s
	doping concentration (N_a, N_d)	$\frac{1}{s}$
device params	electric field (ε)	1
	carrier velocity (v, $\mu\varepsilon$)	1
	depletion layer width (W_d)	s
	capacitance (C, $\frac{\varepsilon A}{t_{ox}}$)	s
	area (A)	s^2
	inversion layer density (Q_i)	1
	current, drift (I)	s
	channel resistance (R)	1
	threshold (V_{th})	s
	parasitics	1

Wire Scaling and System Parameters

assumptions	wire dimensions (T_w, L_w, W_w, t_{ins}, W_{sp})	s
	conductor resistivity (ρ_w)	1
	insulator permitivity (ε_{ins})	1
derived wires	cap per unit length (C_w)	1
	resistance per unit length (R_w)	$\frac{1}{s^2}$
	wire RC delay (τ_w)	1
	wire current density ($\frac{I}{W_w t_w}$)	$\frac{1}{s}$
system params	circuit delay (τ, $\frac{CV}{I}$, $R_c C$)	s
	power dissipation (VI)	s^2
	charge per transistor (CV)	s^2
	power-delay product ($\tau P, e$)	s^3
	circuit density ($\frac{1}{A}$)	$\frac{1}{s^2}$
	power density ($\frac{P}{A}$)	1

Implications

Process (nm)	250	180	130	90	65	45	32
Length L (nm)	250	175	117	77	49	30	20
voltage	2.2	1.8	1.8	1.2	1.2	1.0	1.0
delay (ps/nm)	0.4	.41	.44	.46	.52	.57	.61
delay (ps)	100	72	51	40	28	20	14
clock (GHz)	0.7	1.4	2.8	3.6	5	7	10
stages/clock	25	17	12	12	12	12	12
repeaters/clk	14	10	7	7	7	8	8

Scaling Trend Observations

- length scaling more aggressive
- delay per nm worsens
- frequency and pipelining
- voltage scaling (or lack) should produce higher performance

some effects cause substantial deviations from ideal scaling which we will discuss.

Modeling and Scaling

depletion width $W_D = \sqrt{\frac{2\varepsilon_{si}(\psi_{bi}+V_{dd})}{qN_a}}$

where

- ε_{si} – the permitivity of silicon
- ψ_{bi} – built-in potential (band gap)
- V_{dd} – scales by s
- N_a – substrate doping, scales by $1/s$

$W_D \approx \sqrt{ks^2}$ (if band gap $\ll V_{dd}$)

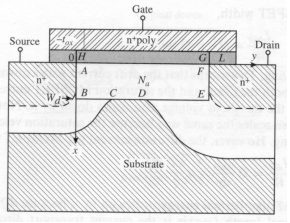
$W'_D \approx sW_D$

Issue: $\psi_{bi} \approx IV$, so scaling starts breaking down, as boundary conditions of built-in potential of P-N junctions don't scale right.

Short Channel Effects

Often referred to as **DIBL** – Drain-induced barrier lowering

This affect allows voltage of drain or source to partially control the field in the channel.



When "box" is twice as wide as tall, behaves like long channel transistor.

Otherwise 2D effect applies.

Short Channel Effects

- Field normally controlled by gate
- Aspect ratio determines:
  - to determine height ($W_{dm}+t_{ox}$) at boundary conditions:
    $\varepsilon_{si}\xi_{x,si} = \varepsilon_{ox}\xi_{x,ox}$
    vertical field ξ based on permitivity of oxide or silicon
  - currently, $\varepsilon_{si}/\varepsilon_{ox} \approx 3$
    (higher κ dielectrics will increase this!)
    $W_{dm}+3t_{ox}$ is box height
  - UNLESS $L > 2(W_{dm}+3t_{ox})$ we get short channel effects!

W_{dm} is maximum gate depletion width

Gate Insulator

Assume vertical field dominates interface to channel

- Insulator thickness t_{ox} dependent on dielectric constant κ
- High κ materials allow thicker insulator
  - good for manufacturing
    - now only a few molecules thick!
    - hot carrier effects
    - tunneling (gate leakage)
  - can effect vertical field
- ratio of ε_i/t_i important

Gate Insulator Thickness

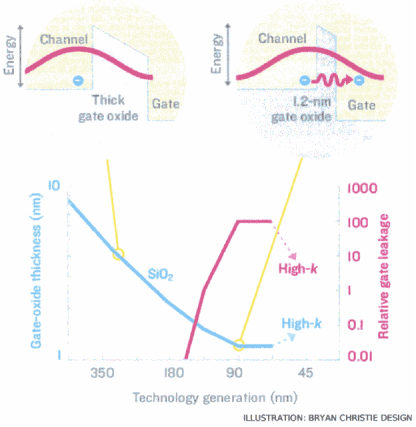
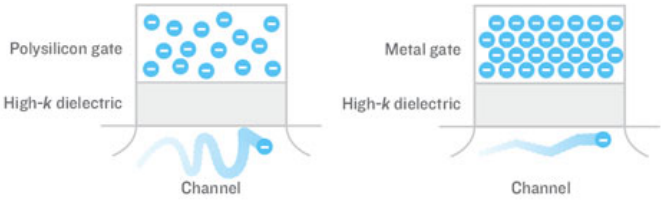


ILLUSTRATION: BRYAN CHRISTIE DESIGN

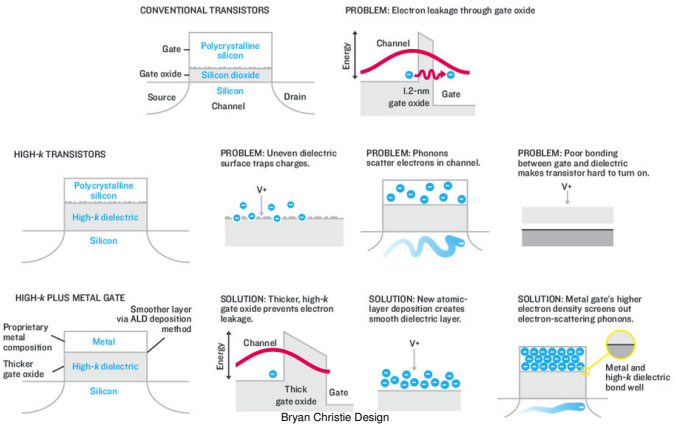
Aside: Metal vs Polysilicon Gate



Bryan Christie Design

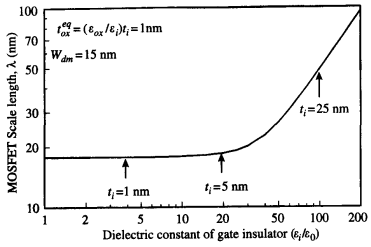
Metal gate electron density reduces channel electron scatter.

Gate Insulator and Gate Composition



Bryan Christie Design

Gate Insulator Thickness Limit



Plots smallest useable L vs. dielectric constant of gate insulator

- DIBL (lateral field) becomes important as it limits the ability to scale down a transistor $(\varepsilon_i/\varepsilon_0) > \approx 15$
- Insulator thickness (even high–κ) must be less than half of L.

Voltage Scaling

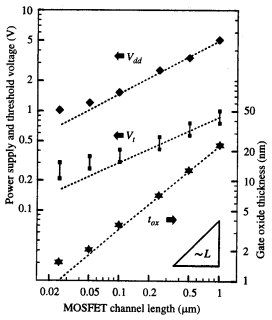
Constant Field Scaling too restrictive

- W_{dm} and DIBL - short channel effects
- degraded performance
- standard voltages for parts

But higher voltage has problems

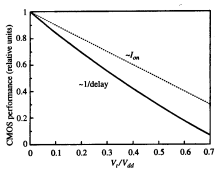
- hot carrier injection
- electromigration of metals
- increased current density
- temperature and cooling
- reliability

Voltage Scaling



Note scaling difference between V_{dd}, V_{th}, and t_{ox}
Ratio of V_{th}/V_{dd} is increasing.

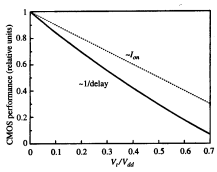
Threshold Voltage Scaling



- Threshold has direct effect on device performance: gate overdrive $= V_{dd} - V_{th}$
- stronger dependence than current, fitting curves: performance $\propto 0.7 - \frac{V_{th}}{V_{dd}}$ vs $\propto 1.0 - \frac{V_{th}}{V_{dd}}$

Lowering threshold substantially increases transistor performance.
Is this desirable?

Threshold Voltage Scaling

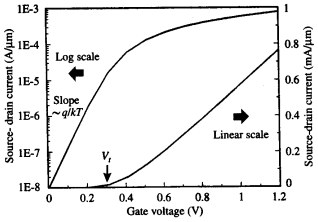


- Threshold has direct effect on device performance: gate overdrive $= V_{dd} - V_{th}$
- stronger dependence than current, fitting curves: performance $\propto 0.7 - \frac{V_{th}}{V_{dd}}$ vs $\propto 1.0 - \frac{V_{th}}{V_{dd}}$

Lowering threshold substantially increases transistor performance.
Is this desirable? Only when power is not an issue!

Threshold Voltage Scaling

Threshold voltage scales even worse than voltage



Two primary regions for current

1. above threshold
2. subthreshold

Threshold Voltage Scaling

Current when transistor is off $\propto$ thermal energy kT

$I_{off} = I_0 e(-\frac{qV_{th}}{mkT})$

- $m \approx 1.2$, the body effect that modifies V_{th}
- $I_0 \approx 1 - 10\mu A/\mu m$, current at threshold

I_0 is proportional to inversion charge density at threshold:

$Q_i \sim (1-2)\frac{kT}{q}C_{ox}$

This makes I_{off} proportional to $\frac{1}{t_{ox}}$ and $\frac{W_{tot}}{L}$

Leakage

Leakage current equation:

$P_{off} = W_{tot} V_{dd} I_{off}$

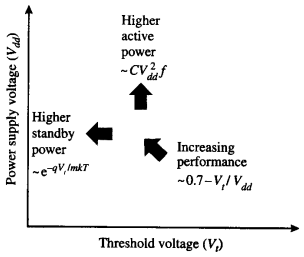
- threshold directly controls leakage current
- leakage directly proportional to number and size of transistors

Leakage increases by about $10\times$ for every 0.1V reduction in V_{th}

Becoming *serious* limiter to our designs!

Fin FETs made a significant improvement in this area.

Threshold Voltage Limitations

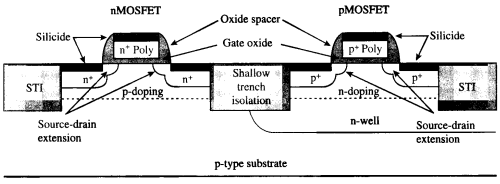


- V_{th} cannot physically be reduced below 0.2V
- performance is a function of V_{th}/V_{dd}
  - implies V_{dd} cannot be reduced much below 1.0V

Leakage

- Gate oxide key for short channel effects
- Gate oxide $\approx \frac{1}{25} - \frac{1}{50}$ of L
- Gate oxide tunneling
  - More prevalent in N-type due to lack of electrons in P^+ poly gate. Lessened with metal gates.
- With high κ dielectrics, most leakage is between source and drain, and to the body

Channel Profile



- Low voltage threshold:
  - requires n^+ and p^+ polysilicon gates
  - midgap work function metal gate gives high threshold

Doping

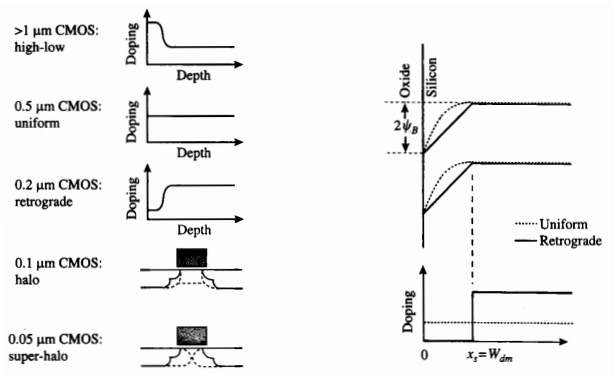
- To control short channel effects, need proportional scaling (s) of:
  - oxide thickness t_{ox}
  - gate controlled depletion width W_{dm}
  - channel length L
- This in turn, for uniformly doped N_a channel:
  - requires increased channel doping concentration
  - increases the potential across oxide
  - $\Longrightarrow$ causes the threshold to increase
    $V_t = V_{fb} + 2\psi_B + \frac{\sqrt{4\varepsilon_{si} q N_a \psi_B}}{C_{ox}}$
    where band gap $\psi_B = \frac{kT}{q}\ln(\frac{N_a}{ni})$
    (C_{ox}, N_a scale as $\frac{1}{s}$)

Doping Profiles to the Rescue

To reduce gate controlled depletion width and support V_t reduction:

- simultaneously optimize gate deletion depth W_{dm} and threshold V_t
- $\geq$ 500 nm
  - high–low and low–high doping profiles
- $\geq$ 250 nm
  - uniform
- $\geq$ 100 nm
  - retrograde
- $\geq$ 50 nm
  - halo
- $\leq$ 50 nm
  - super-halo

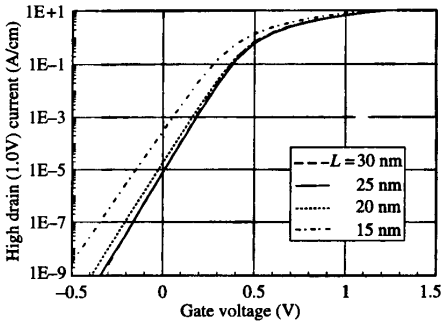
Doping Diagrams



Doping Effect

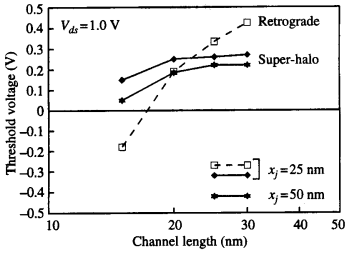
- Retrograde
  - decouple threshold voltage V_t and gate controlled depletion width W_{dm}
  - body effect and inverse subthreshold slope keep dependence on gate depletion width W_{dm}
- halo
  - both horizontally and vertically (super-halo) nonuniform
  - increases doping concentration N_a in shorter devices
  - decreases leakage sensitivity to channel length variations
  - permits significant lowering of threshold V_t

Super-Halo Subthreshold Currents



Super-halo doping robustness to L_e variation in 25nm process

Super-Halo Threshold Roll-off



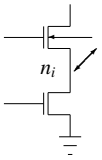
Short-channel threshold roll-off for super halo and retrograde doping profiles. The Threshold voltage is defined as gate voltage where $I_{ds} = 1\frac{\mu A}{\mu m}$

Body Effect

- Body effect:
  - increases with doping
  - decreases with diode depletion
  - coupled to gate controlled depletion width:
    $m - 1 = \frac{\varepsilon_{si} t_{ox}}{\varepsilon_{ox} W_{dm}} \approx \frac{3t_{ox}}{W_{dm}}$
  - typically $m - 1 \leq 0.5$

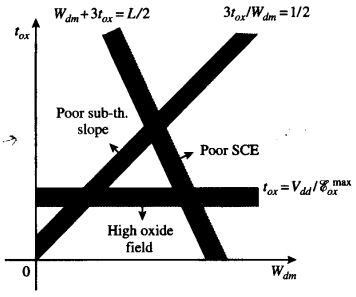
Body Effect

- Structural Dependency
  - source-to-substrate bias substantially increases threshold V_t



Body effect in series devices

Body Effect vs Short Channel Effect



t_{ox} limited by maximum oxide field, body effect limits $\frac{3t_{ox}}{W_{dm}} \leq 0.5$, and short channel effects constrain $W_{dm} + 3t_{ox} \leq \frac{L}{2}$

Hot Carriers

- Charge carriers accelerated from source to drain
- Scattering causes some to have higher energy (*hot*)
- Can inject and stick into transistor oxide
- Significantly effects threshold
  - n-type: on-current degraded
  - p-type: off-current increases
- Related to radiation tolerance of a circuit

Hot Carriers

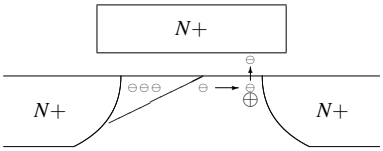


Figure 1: Electrons/holes accelerate towards drain, resulting in impact ionization and carrier injection into gate oxide

Scaling Review

- Active power: $P = CV_{dd}^2 f$
- Leakage power: $P_{\text{off}} = W_{\text{tot}} V_{\text{dd}} I_{\text{off}}$
- Reasons for power not scaling in ideal fashion
- Energy-delay products and applications: $e\tau$ and $e\tau^2$
- Power performance tradeoff, and relation to V_t
- Subthreshold voltage and leakage
- Short channel effect
- Body effect
- Hot carriers
- Channel doping and thresholds

Ideal Scaling

- Constant field scaling: s, ξ
- Scaling assumptions (constant field scaling)
  - device dimensions, doping, voltage
- Derived parameters
  - electric field, depletion layer width, capacitance,
  - channel resistance carrier velocity
- Derived circuit parameters
  - charge per transistor, threshold voltage, circuit delay
  - circuit density, power density, power-delay product
  - power dissipation per circuit

Historical Scaling, limits

- Scaling related to s and limits of:
  - threshold voltage, oxide thickness, voltage

Scaling Into Nano Devices

Ken Stevens

Dielectrics and Atomic Layers

high κ dielectrics allow same capacitance for thicker insulator.

- SiO_2 = 4
- Air or vacuum = 1

The dielectric property measures the ability of the charges in a material's atoms to reorient themselves in the direction of the field. The internal charges are more responsive in a high κ than a low κ material.

Dielectric Materials over Time

- Each decade dielectric material shrinks to approximately one sixth the starting thickness.
- A silicon atom is 0.26 nm in diameter.
- Gate oxides (SiO_2) were approximately five atoms in a 90 nm process (about 1.2 nm)
- Quantum effects create gate leakage:
  - Probability of electron location is an area larger than the thickness of the insulator

Dielectric Materials over Time

- New dielectric materials:
  - Aluminum oxide (Al_2O_3), titanium dioxide (TiO_2), tantalum Pentoxide (Ta_2O_5), hafnium dioxide (HfO_2), hafnium silicate ($HfSiO_4$), zirconium oxide (ZrO_2), zirconium silicate ($ZrSiO_4$), lanthanum oxide (La_2O_3)

Dielectric Materials over Time

- Most work poorly due to charge getting trapped in the interface between the gate electrode and dielectric
  - Results in "memory" based on previous transistor operations
  - Solution: make ultra smooth layers where no variations exist to leave gaps and pockets that trap charges
    - used "reactive sputtering" and "metal organic chemical vapor deposition"
    - New technique called "atomic layer deposition" allows building materials one atom at a time. This works by having a gas react with surface of the silicon wafer leaving the whole substrate coated with a single layer of atoms. Second gas added and reacts with the layer of atoms just layed down.
  - Hafnium and zirconium materials seem to work best.

Dopant and Gate Materials

- Polysilicon gates were used from 1969 to 2007
  - The 'M' in CMOS is metal for the metal gate used before 1969
- Doping materials: arsenic, phosphorus, or boron
  - Boron adds positive carriers, making it p-type
  - Arsenic and phosphorus make n-type devices
  - n-type gates use n-type polysilicon

Metal Gates

Once good thicker dielectrics existed, transistors didn't switch well. The charge-carrier mobility was reduced.

- Needed higher voltages – or Fermi-level pinning
- Problem was interaction between poly gate and high-κ dielectrics.
  - High κ dielectrics have dipoles that vibrate
  - Vibrations are conducted through silicon crystal lattice, called PHNONs
  - The phonons deflect electrons reducing mobility
- Dipole vibrations are filtered by electron density in gate electrodes.
  - Metal packs hundreds of times more electrons than SiO_2
  - This filters out phonons and lets current flow smoothly through the transistor channel.
- Bonding between gate and dielectrics also improved with metal gates, reducing Fermi-level pinning.

Metal Gates

Work function:

- Different metals change the “work function” – the energy of an electron in the gate electrode relative to that of an electron in the channel.
  - This effectively changes the gate threshold
- Thus different metals have been engineered to match doped polysilicon work functions for the p and n-type gate electrodes

Metal Gates

Manufacturing:

- Self-aligned transistors use “gate-first” manufacturing
- Problematic with pure metal gate
  - Doping implantation and heat stress due to annealing reduced stability of gate oxide and metal materials
  - p-type transistors more problematic than n-type
- Create silicided gates with “gate-first” manufacturing
- Turn polysilicon gate into metal-silicide gate
  - Essentially replace every other silicon atom with metal (usually nickel)
- Then you can dope nickel silicide to alter the work function for n or p-type devices

Channel Leakage

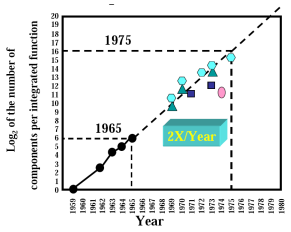
- Short lengths leak more but give higher drive current
- High threshold pinches off leakage, but throttles drive current
- Thick dielectric reduces ability to conduct, increasing threshold
- New dielectrics give 26% increase in current, for 5× reduction in channel leakage.
- Also reduce gate leakage, giving 10× leakage reduction at same drive current

Moore’s Law

“I see no reason to expect the rate of progress in the use of smaller dimensions in complex circuits to decrease in the near future”

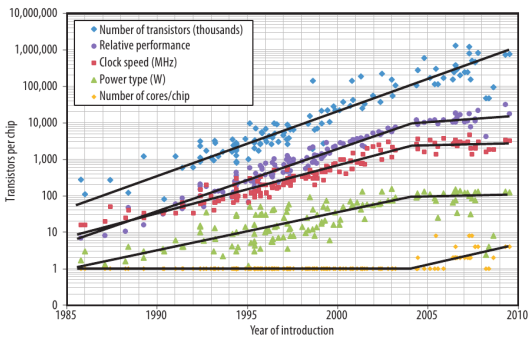
“The new slope might approximate a doubling every two years”

Gordon Moore, December 1975



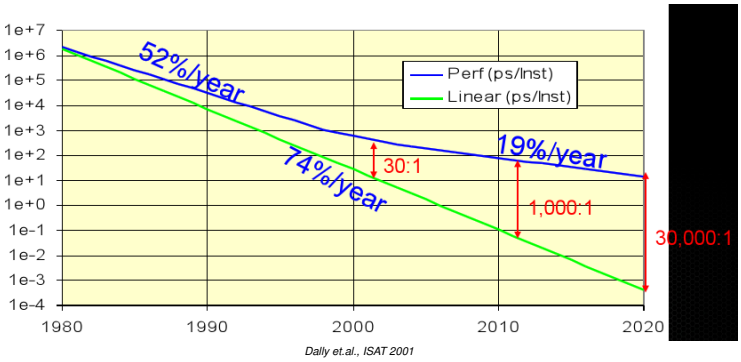
What Happened?

Technology and economic turmoil produce changes and uncertainty



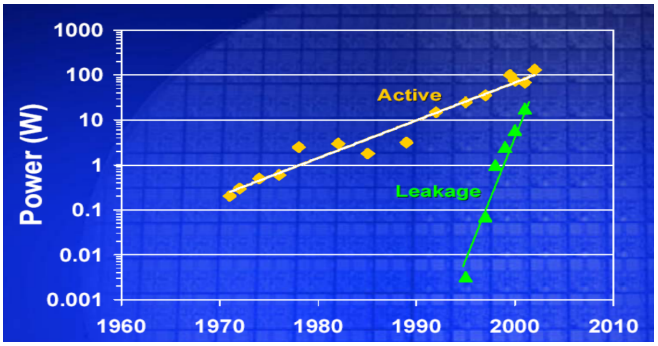
Computing Performance: Game Over or Next Level? by S. H. Fuller and L. I. Millett

CPU Instruction Level Concurrency is Mined Out



Dally et.al., ISAT 2001

Leakage Power Becomes Design Driver

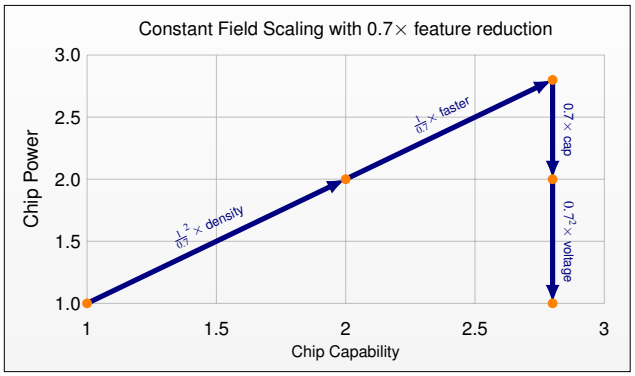


Moore, ISSCC Keynote 2003

(Leakage power was not considered initially in Dennard scaling. . .)

Classic Dennard Scaling

2.8× chip capability in same power



Modern Scaling Values

Density is scaling for planar devices, albeit at a slower pace, but performance and power really tailing off.

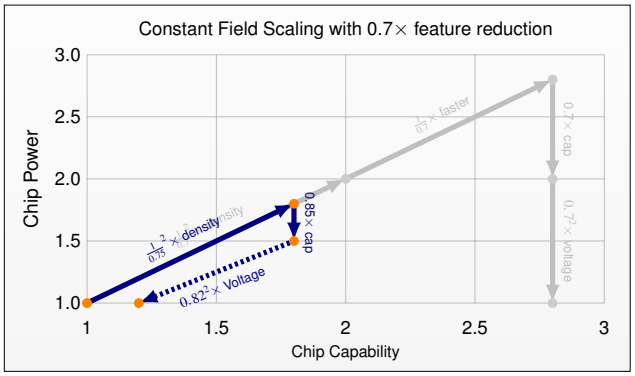
Node	Dim s	Area s^2	Cap C	Freq f	V_{dd} V	Power CV^2f	Power Density
45→32	0.755×	0.57×	0.665×	1.10×	0.925×	0.626×	1.096×
32→22	0.755×	0.57×	0.665×	1.08×	0.950×	0.648×	1.135×
22→14	0.755×	0.57×	0.665×	1.05×	0.975×	0.664×	1.162×
14→10	0.755×	0.57×	0.665×	1.04×	0.985×	0.671×	1.175×

Huang et. al, "Scaling with Design Constraints: Predicting the Future of Big Chips

Designers must deal with leakage to control power, and activity to control thermal.

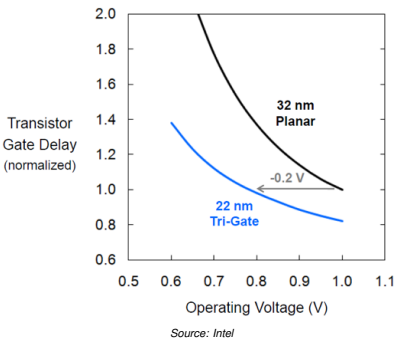
Dennard Scaling Has Ended

1.8× chip capability at 1.5× power
or 1.2× chip capability at same power



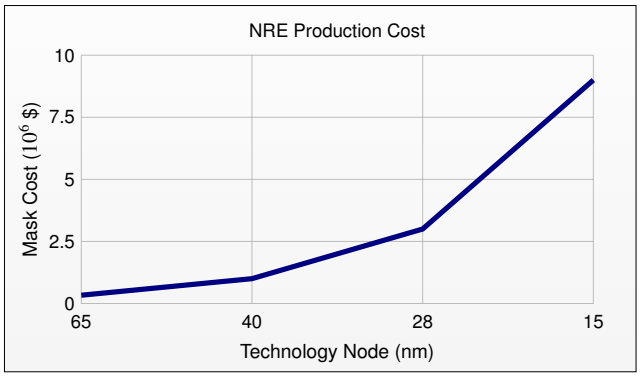
The FinFET Effect

FinFET make a huge one-time impact, keeping scaling going for another generation.



Source: Intel

Foundry Mask Costs



Mask costs increasing approximately 3× per generation

Complete SoC Design Costs

- Average cost for 28 nm SoC: \$30 MM
- Average cost for 14 nm SoC: \$80 MM

It costs \$270 MM to design a 7 nm SoC
Gartner

This is an 9 fold increase.

Foundry Consolidation

Let's ignore research and development costs for now.

- A foundry is a warehouse the size of two basketball courts filled with equipment for projecting a lithographic image onto wafers.
- Stockers are filled with wafers that feed this expensive equipment
  - The wafers are typically worth \$10–100 MM of idle inventory
- Modern fabs cost over \$5 B.
  - Amortized over five years this equals \$3 MM per day
  - Mask costs require modern semiconductor products run volumes of 5-10,000 wafers per month

Foundry Consolidation

Intel's 10 nm FinFET

- Intel will ship 10 nm node products in 2017
- Compared to 14 nm
  - Gate pitch reduces from 70 nm to 54 nm ($s = 0.77$)
  - Production wafer cost increases, but per transistor cost still decreases
    - (This ignores mask cost)
  - Performance and power: Not Yet Reported
- Node introduction will slow
  - smaller node incremental improvements, or "deminodes" 10 nm+, 10 nm++, will be introduced before the 7 nm node

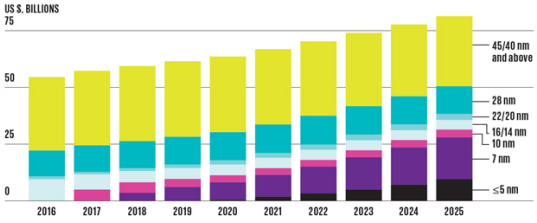
7 nm FinFET

- Intel 14 nm reference: 20 nm gate length, 70 nm pitch
- Different foundry 7 nm specs will vary substantially
  - Projected 12 – 18 nm gate length, 45 – 55 nm gate pitch
  - Fin width of 5 – 6 nm
    - This is near physical limit for fin structure
- 80 photolithograpic passes (193 nm), 66 mask steps
  - 28 nm has 50 passes
- 1 – 1.5 days to process a mask layer
  - 5 months to manufacture a 7nm wafer!
- \$160 MM fab equipment for every 1k wafer per month
  - 60% increase over 28 nm

Future FinFETs

- FinFETs will be re-engineered
  - Possibly taller
    - Gives greater drive current
    - Increases device capacitance
    - Traps more heat
  - Scale the fin
    - Reduces device capacitance
- Other materials III–V such as SiGe

Future Technology Nodes



Source: IBS, 2017

"It's really power reduction or energy efficiency that's the primary goal on these new generations, besides or in addition to transistor cost reduction."

Mark Bohr, Intel

End of Dennard Scaling

- Process improvements greatly reduced
  - One generation performance at equi-power is 1.2×, not 2.8×
- Only high volume products will scale
  - at $30 MM mask costs, you need 300 M sales to reduce mask cost to 10 cents per chip
- We have lots of concurrency
  - Processors are not faster, just wider
- All future systems are energy limited
  - Efficiency *IS* Performance

It is now designer's responsibility to create efficient IC's