IC Technology at New Nodes Made Easy

Alvin Loke

05-Dec-2013

AUTHORIZATION
All copyrights to the material contained in this document are retained by me and my employer.

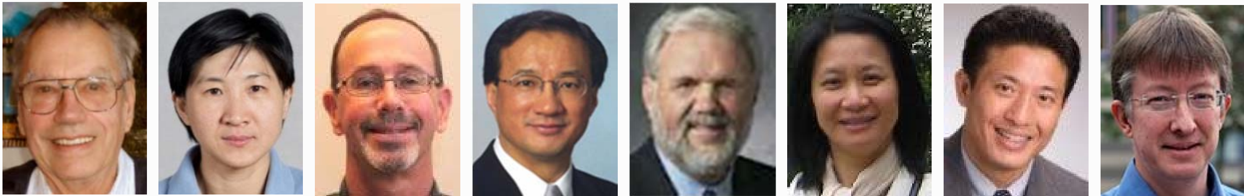
My Teachers

Bob Barnes | Larry Bair | John Bravman | Tom Bryan | Tom Cynkar | Dick Dowell | Bruce Doyle | Emerson Fang | John Faricelli | Dennis Fischette | Phil Fisher | Mike Gilsdorf

Jung-Suk Goo | Bob Havemann | Rick Hernandez | Tim Hollis | Mark Horowitz | Reza Jalilizeinali | Ron Kennedy | Takamaro Kikkawa | Greg Kovacs | Steve Kuehne | Tom Lee | Justin Leung

Ying-Keung Leung | Tom Lii | Shawming Ma | Joe McPherson | Charles Moore | Don Morris | Bich-Yen Nguyen | Michael Nix | Michael Oshima | Chintamani Palsule | Jim Pfiester | Jim Plummer

Dave Pulfrey | Gary Ray | Behzad Razavi | Jeff Rearick | Changsup Ryu | Krishna Saraswat | Shawn Searles | Ray Stephany | Gerry Talbot | Tom Tiedje | Paul Townsend | Ram Venkatraman



Martin Wedepohl | Tin Tin Wee | Jeff Wetzel | Simon Wong | Bruce Wooley | Joanne Wu | Patrick Yue | Carl-Mikael Zetterling

Matt Angyal

Qi-Zhong Hong

Wei-Yung Hsu

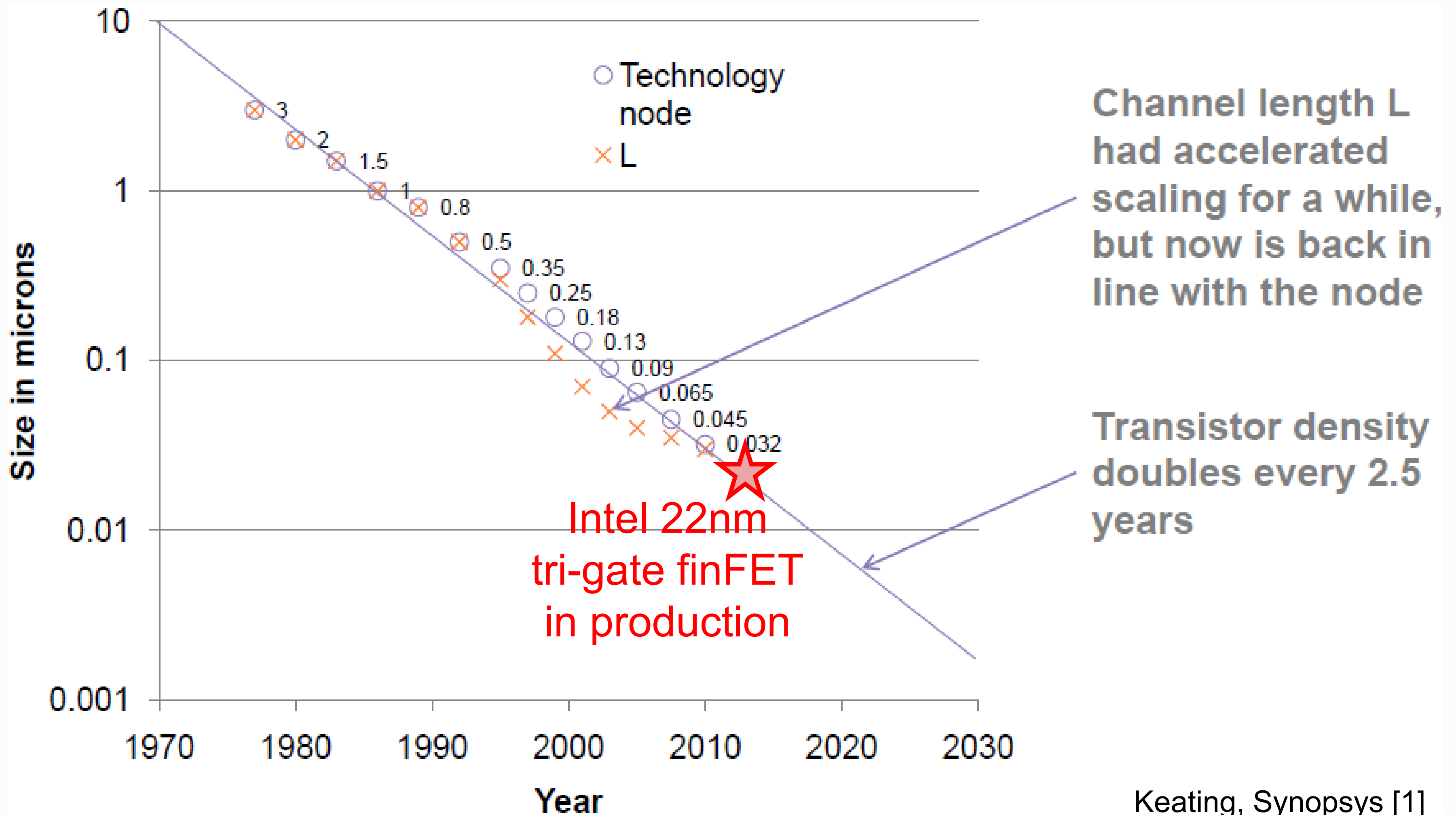
Andy Wei

The 10000-Foot View… A Switch

small, fast, thrifty

Scaling Performance Energy-Efficient

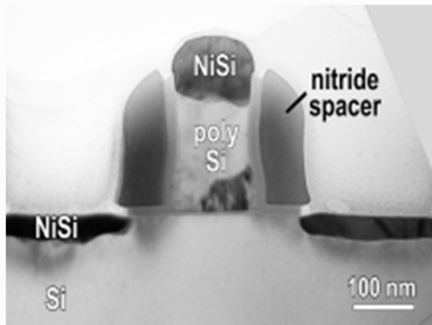
CMOS Scaling Still Alive

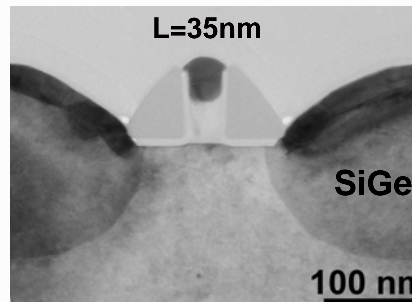


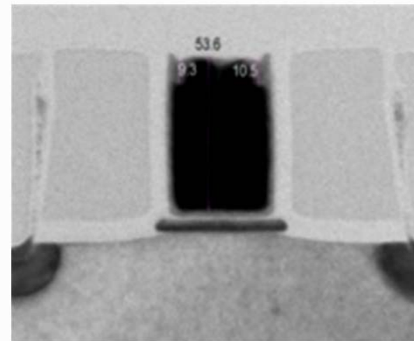
- Leading foundries frantically after manufacturable tri-gate
- Intel already demonstrated 14nm Broadwell

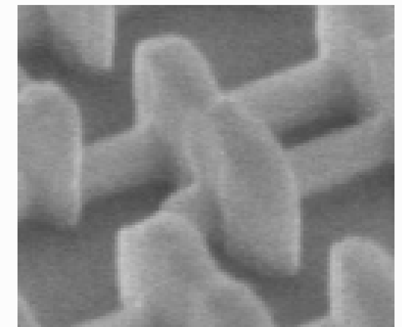
Our Objective

- Understand how MOSFET structure has evolved
- Understand why it has evolved this way









Words of Wisdom

People get lost

because they cannot be found.

Theodorus Loke

Outline

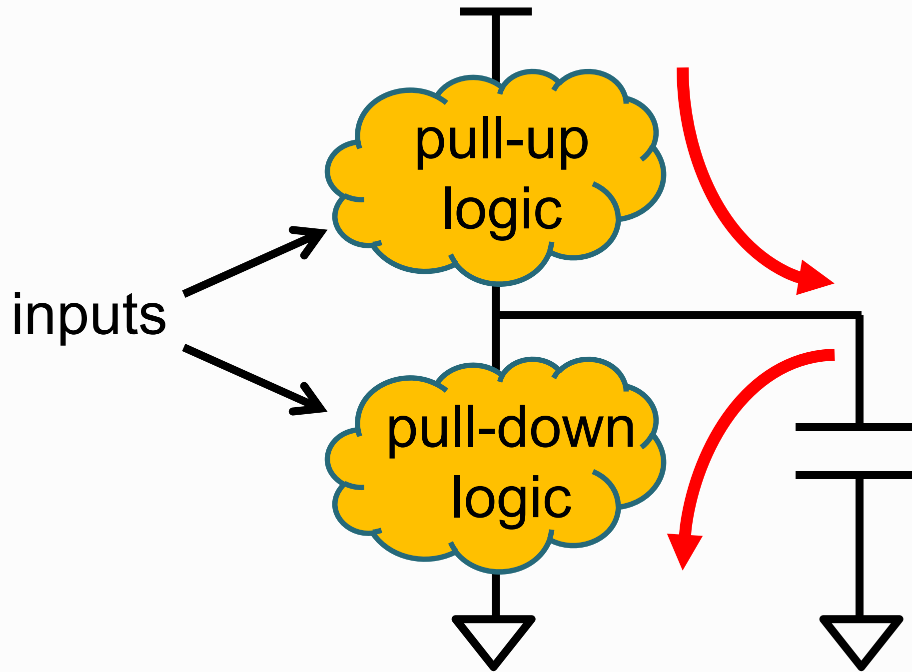
Part 1

- Motivation
- MOSFET & Short-Channel Fundamentals
- 130nm Fabrication
- More MOSFET Fundamentals
- Lithography
- Partially-Depleted SOI

Part 2

- Strain Engineering (90nm & Beyond)
- High-K / Metal-Gate (45nm & Beyond)
- Migrating to Fully-Depleted (22nm & Beyond)
- Tri-Gate FinFETs
- Conclusions

The Basis of All CMOS Digital ICs

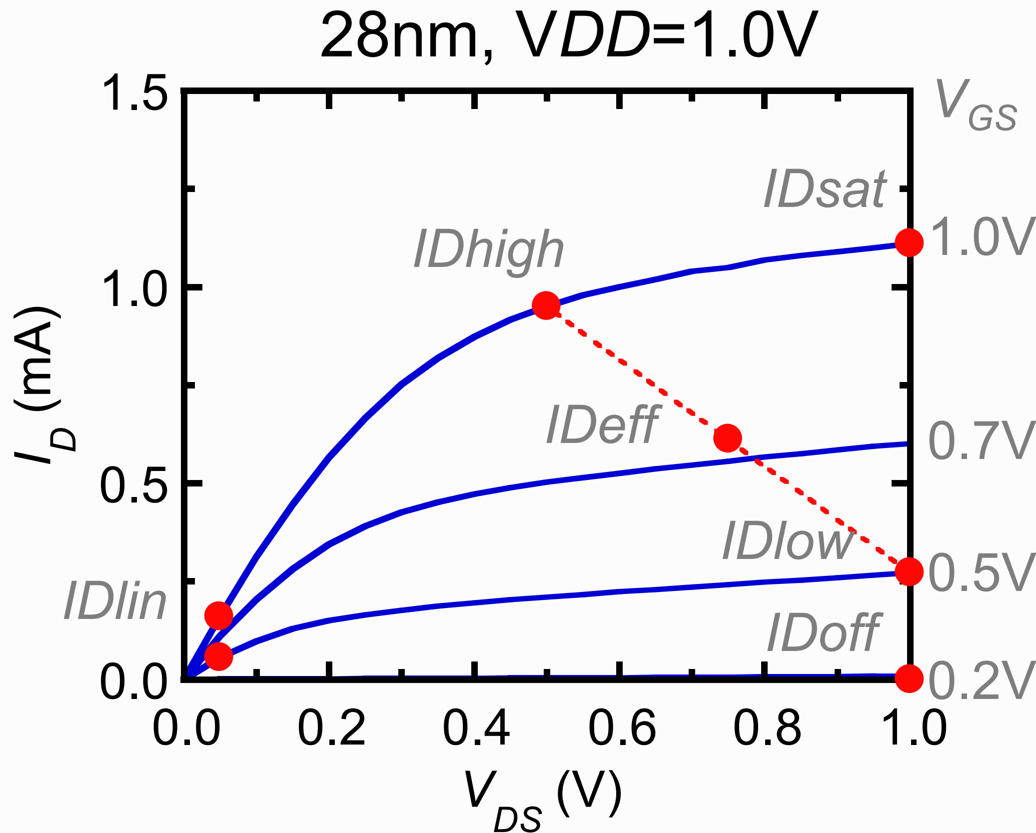


$$t_{delay} \approx \frac{Q_{load}}{I_{eff}} = \frac{C_{load}V_{DD}}{I_{eff}}$$

$$P_{dynamic} \approx \alpha\, C_{load} V_{DD}^2\, f$$

- Charging and discharging a capacitor… very quickly!
- Shorter delay and lower power

Effective Inverter Drive Current



- *IDeff* estimates effective inverter current drawn during switching
- More realistic and way less optimistic than *IDsat*

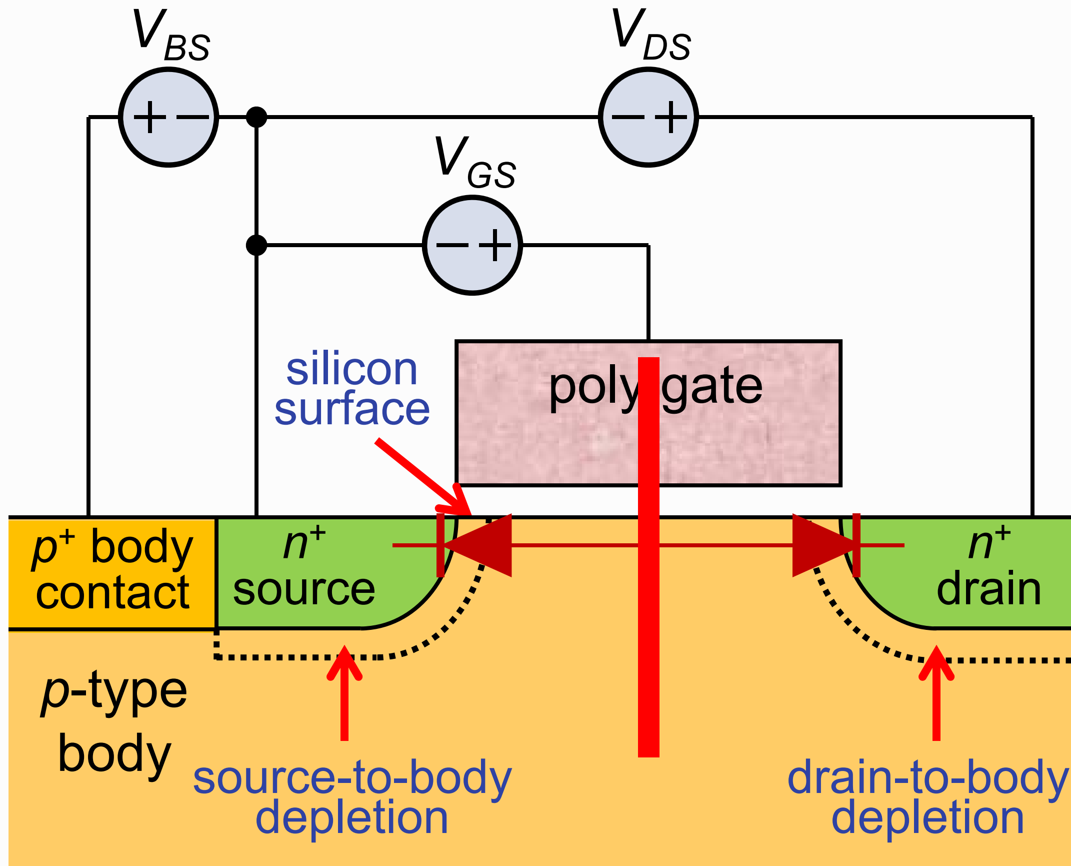
$$IDeff = \frac{IDlow + IDhigh}{2}$$

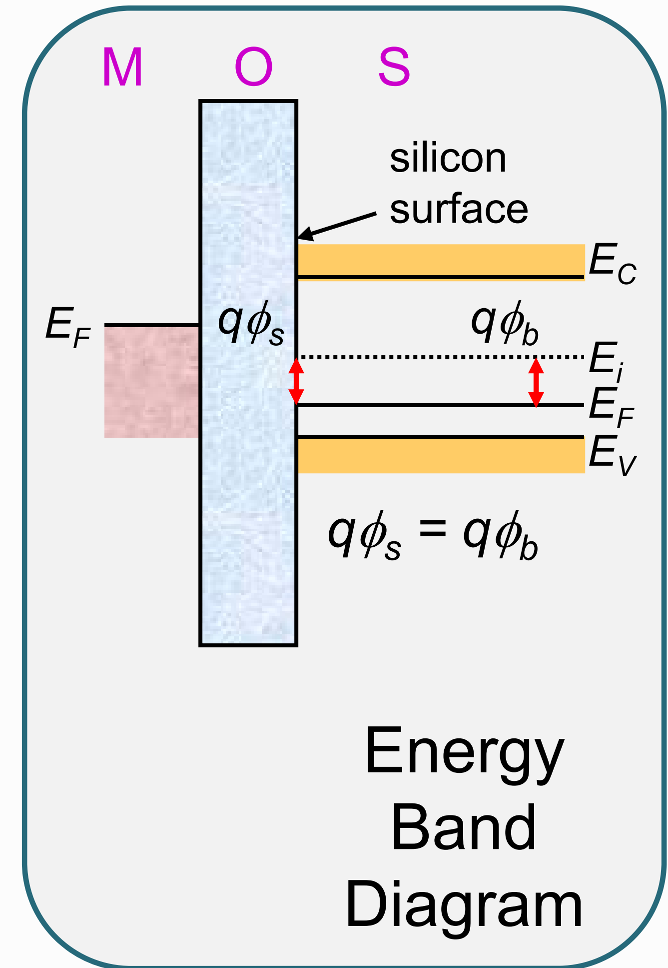
$$IDlow = ID\left(V_{GS} = \frac{V_{DD}}{2}, V_{DS} = V_{DD}\right)$$

$$IDhigh = ID\left(V_{GS} = V_{DD}, V_{DS} = \frac{V_{DD}}{2}\right)$$

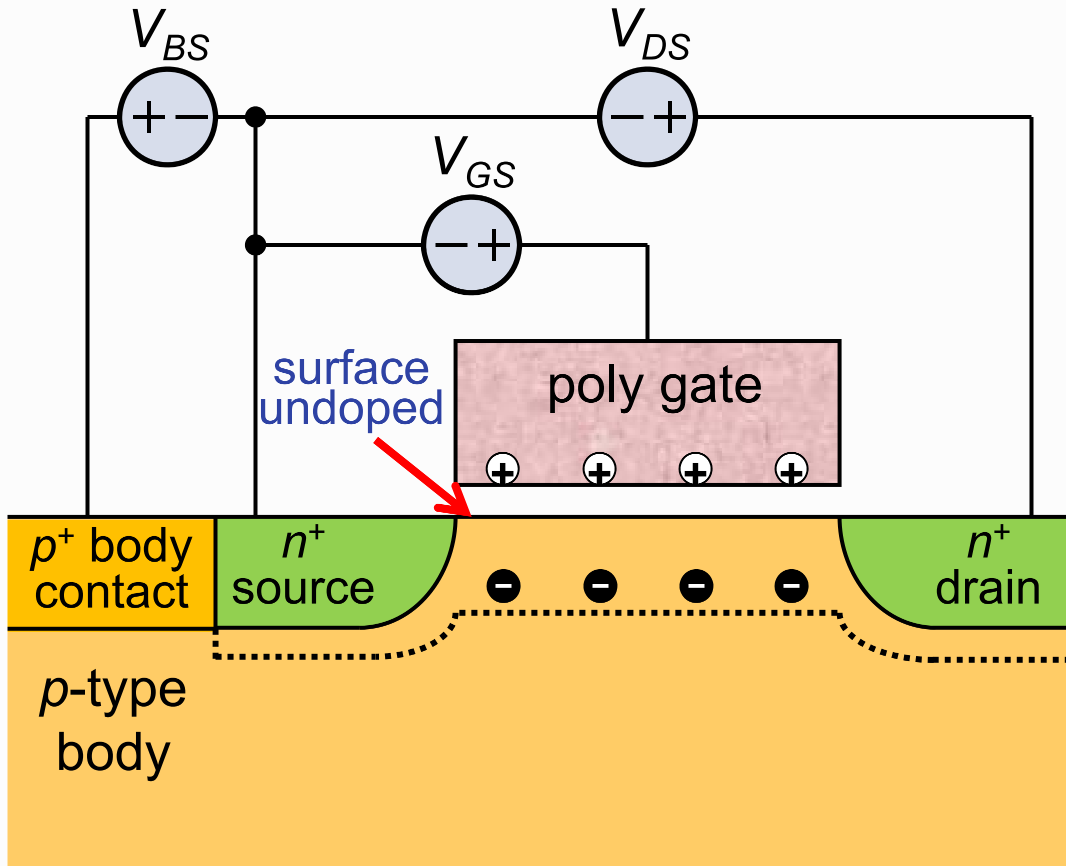
Na *et al.*, IBM [2]

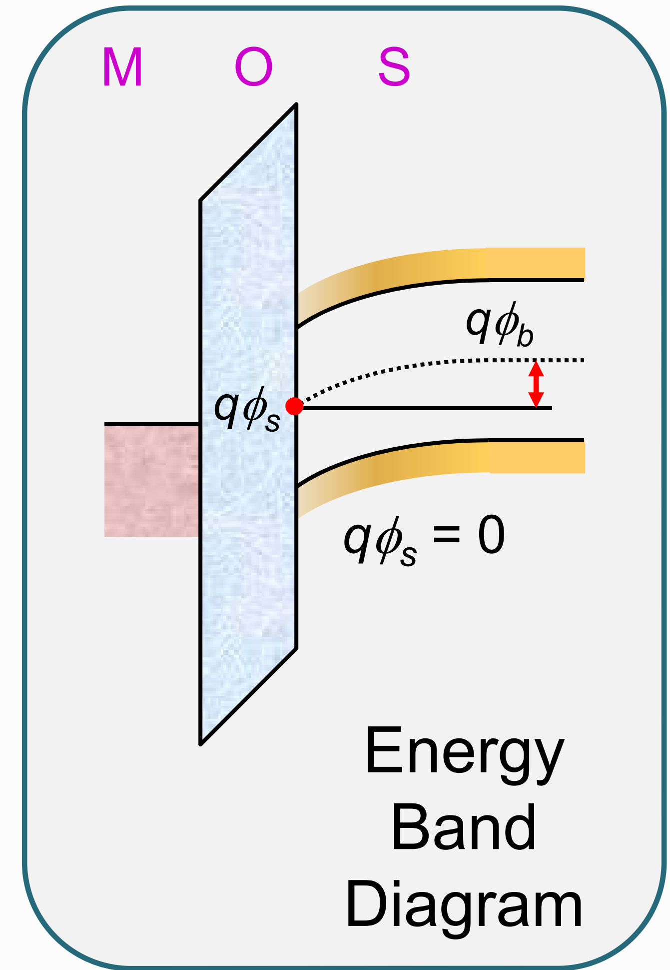
Flatband Condition ($V_{GS}=V_{FB}$)





Onset of Surface Inversion (ϕ_s=0)





+ charge terminating on – charge

Onset of Strong Surface Inversion ($V_{GS}=V_T$)

V_{BS} V_{DS} V_{GS}

poly gate

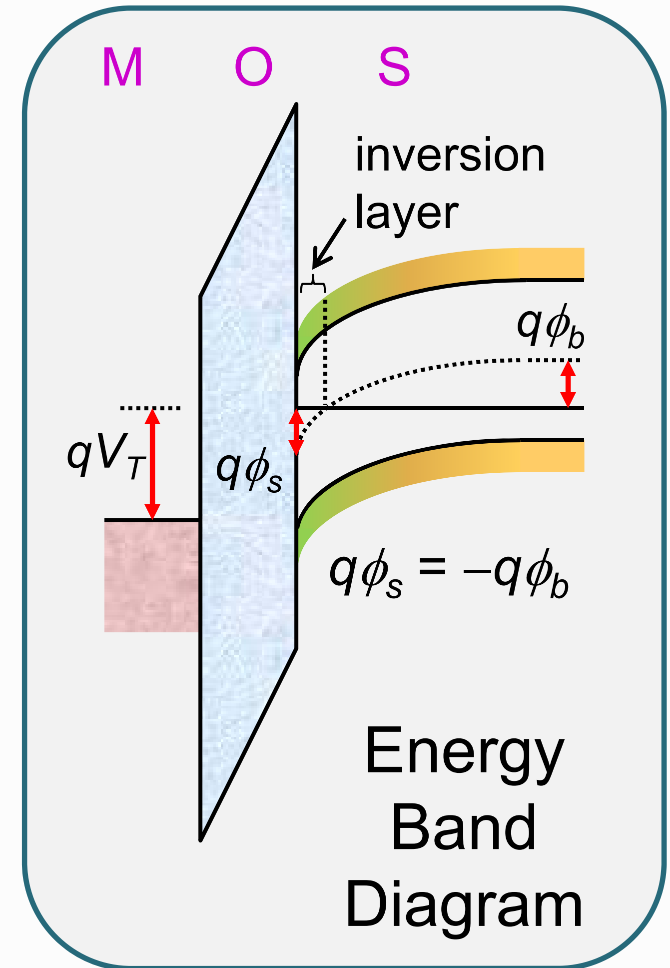
p^+ body contact

n^+ source

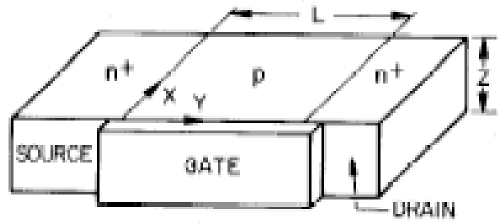
n^+ drain

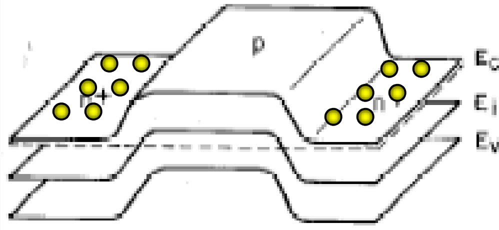
p-type body

$$V_T = V_{FB} + 2\phi_b + \frac{Q_{dep}}{C_{ox}}$$

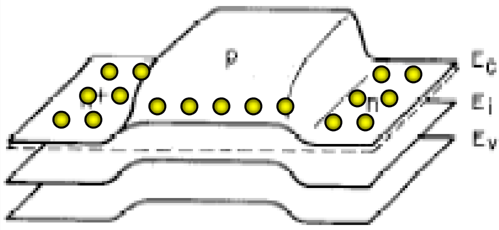


Lower the Surface Barrier

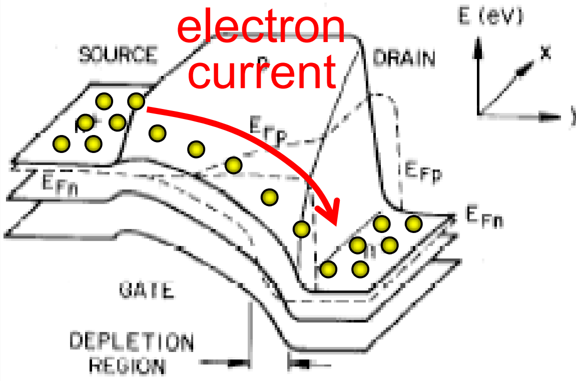




$V_{GS} = 0$
$V_{DS} = 0$ (no current)
Large source barrier
(back-to-back diodes)



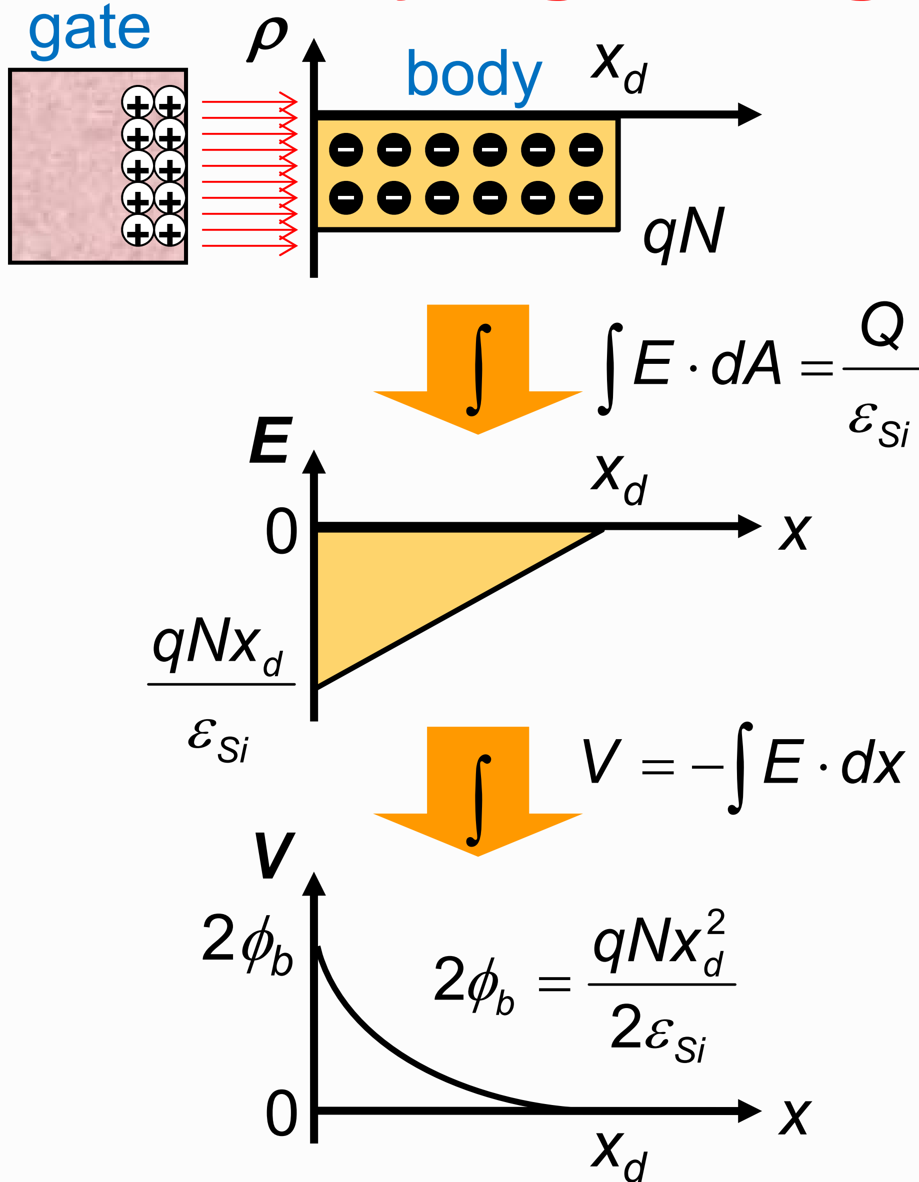
$V_{GS} \approx V_T$
$V_{DS} = 0$ (no net current)
Source barrier lowered
Surface is inverted



$V_{GS} > V_T$
$V_{DS} > 0$ (net source-to-drain current flow)
Carriers easily overcome source barrier
Surface is strongly inverted

Sze [3]

Quantifying Charge to Move ϕ_s by $2\phi_b$

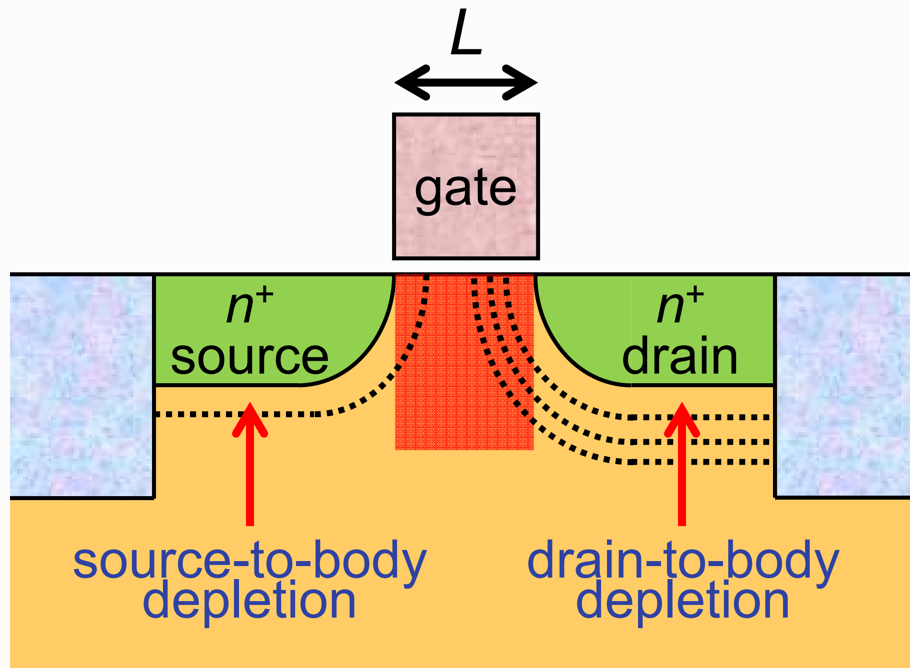


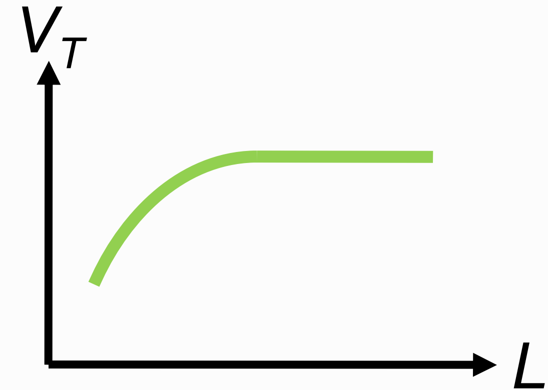
- Assume *uniformly doped* p-type body
- How much body must be depleted to reach strong inversion?

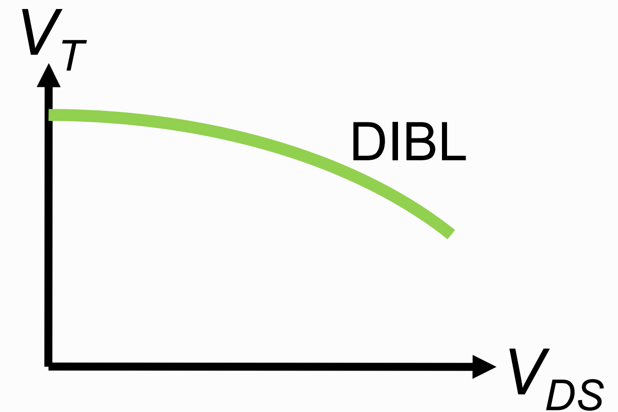
$$x_d = \sqrt{\frac{2\varepsilon_{Si} \cdot 2\phi_b}{qN}} \propto \frac{1}{\sqrt{N}}$$

$$Q_{dep} = qNx_d$$

Short-Channel Effects (SCEs)







V_{DD} not scaling as aggressively as L

→ Higher channel electric fields

- Velocity saturation
- Mobility degradation

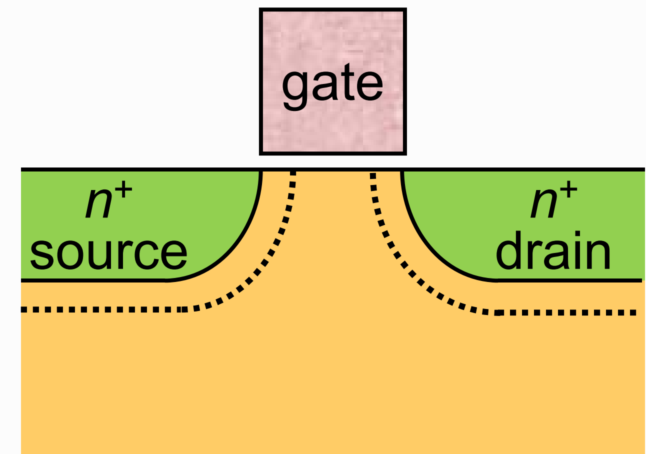
Overcoming Short-Channel Effects

Improve gate electrostatic control of channel charge

- Higher body doping but higher V_T
- Shallower source/drain but higher R_s
- Thinner t_{ox} but higher gate leakage
- High-K dielectric to reduce tunneling
- Metal gate to overcome poly depletion
- Fully-depleted structures (e.g., fins)

Stressors for mobility enhancement

$$x_j \propto \frac{1}{\sqrt{doping}}$$



Outline

Part 1

- Motivation
- MOSFET & Short-Channel Fundamentals
- 130nm Fabrication
- More MOSFET Fundamentals
- Lithography
- Partially-Depleted SOI

Part 2

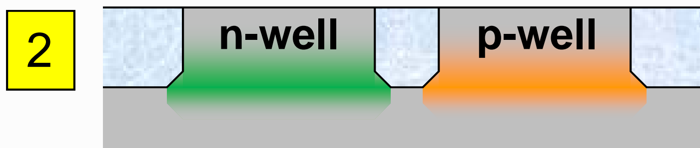
- Strain Engineering (90nm & Beyond)
- High-K / Metal-Gate (45nm & Beyond)
- Migrating to Fully-Depleted (22nm & Beyond)
- Tri-Gate FinFETs
- Conclusions

130nm MOSFET Fabrication

1

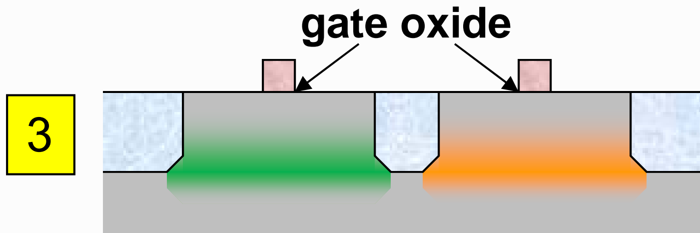
Shallow Trench Isolation

2



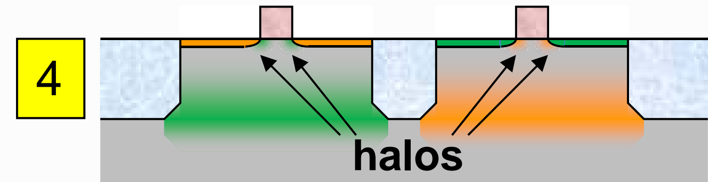
Well Implantation

3



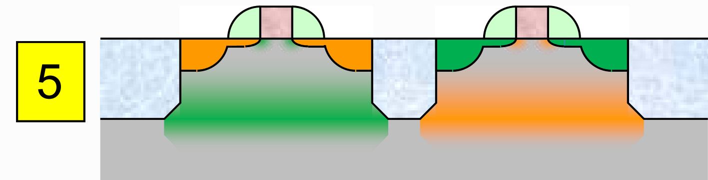
Gate Oxidation & Poly Definition

4



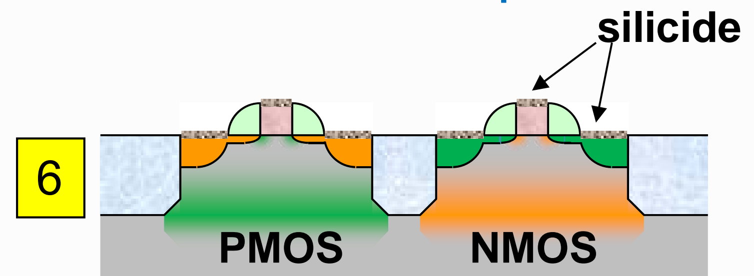
Source/Drain Extension & Halo Implantation

5



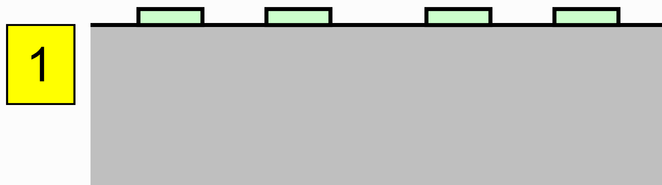
Spacer Formation & Source/Drain Implantation

6

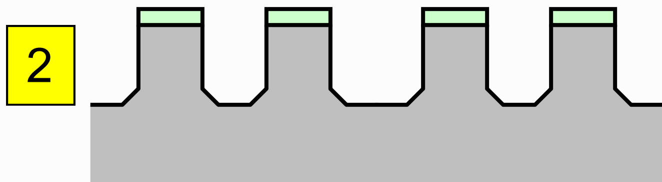


Salicidation

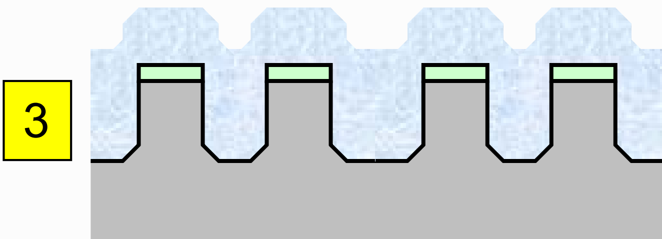
Shallow Trench Isolation



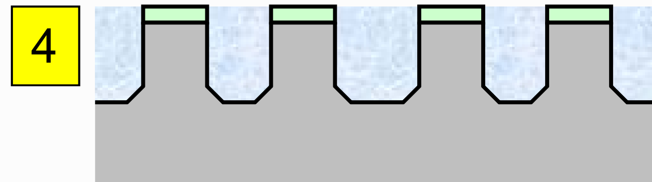
Deposit & pattern thin Si_3N_4 etch mask & polish stop



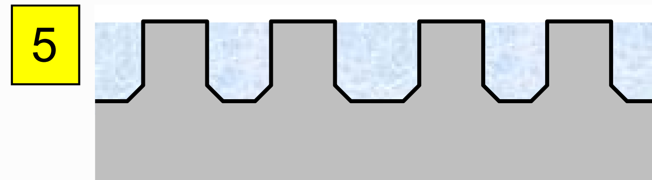
Etch silicon around active area – profile critical to minimize stress



Grow liner SiO_2, then deposit conformal SiO_2 – void-free deposition is critical



CMP excess SiO_2



Recess SiO_2
Strip Si_3N_4 polish stop

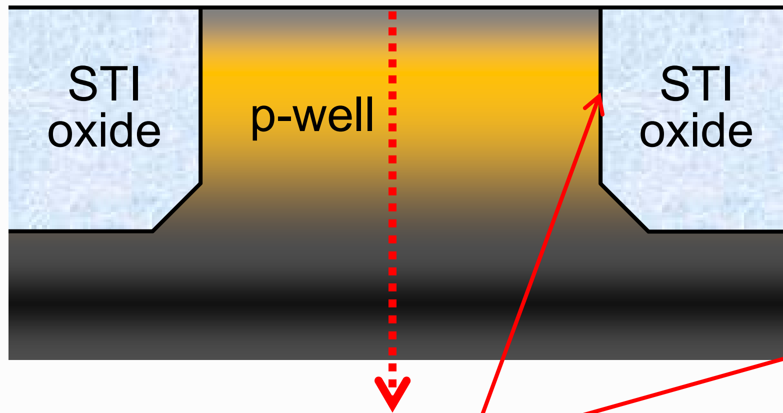
Advantages over LOCOS

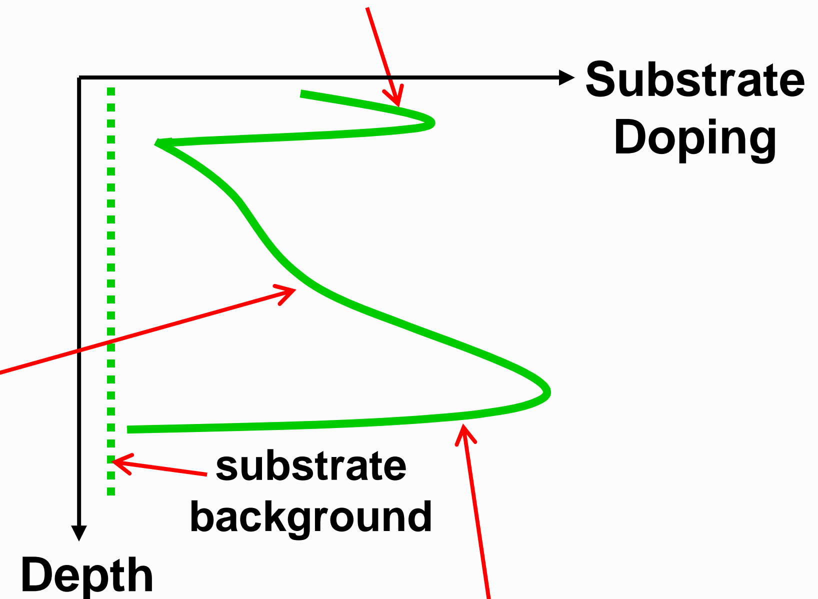
- Reduced active-to-active spacing (no bird's beak)
- Planar surface for gate lithography

Well Implant Engineering

Retrograded well dopant profile (implants before poly deposition)

Shallow/steep surface channel implant
- V_T control
- Slow diffusers critical (In, Sb)





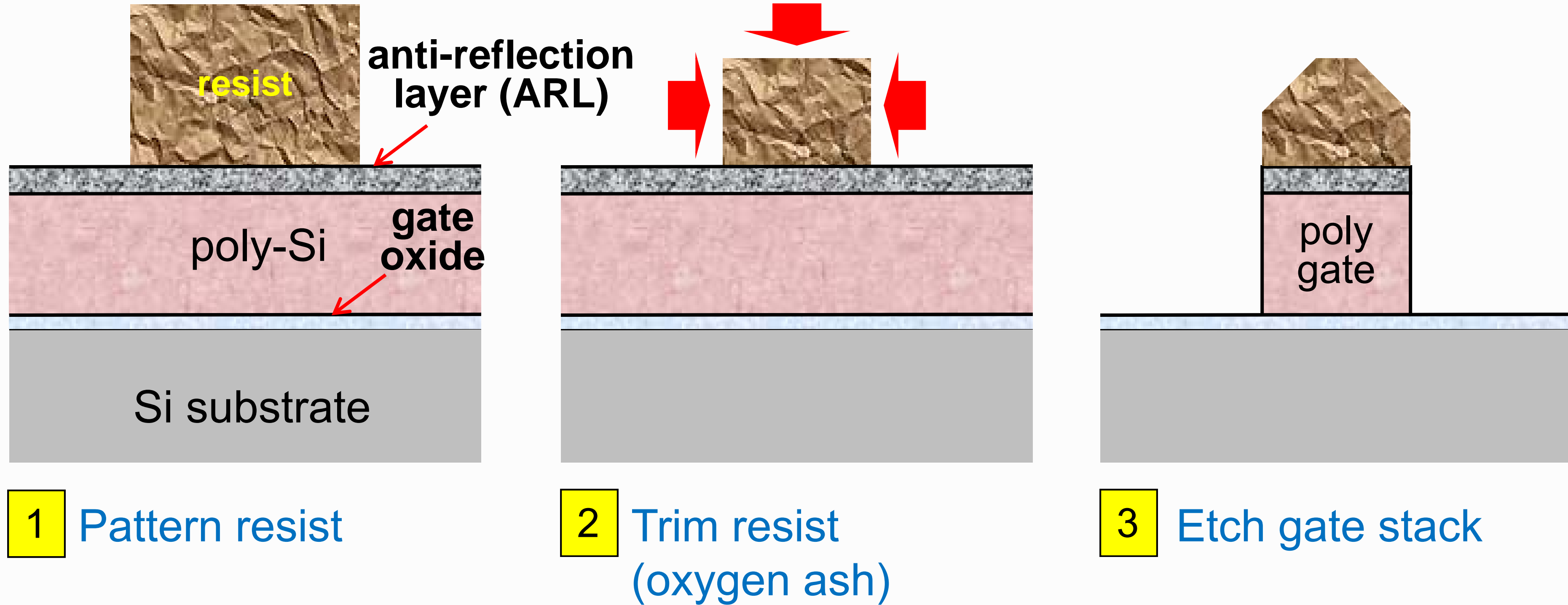
Deeper subsurface implant
- Extra dopants to prevent subsurface punchthrough under halos
- Prevent parasitic channel inversion on STI sidewall beneath source/drain
- Faster diffusers (B, As/P)

Very deep high-dose implant
- Latchup prevention
- Noise immunity
- Faster diffusers (B, As/P)

Sequence implant to reduce ion channeling, especially for shallow implant

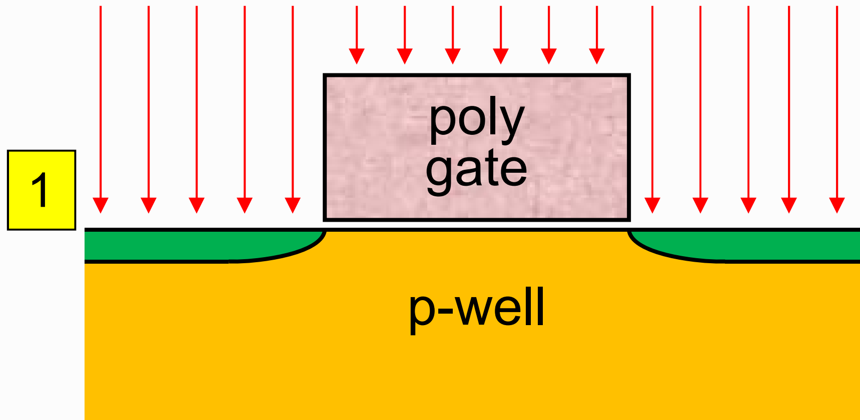
Poly Gate Definition

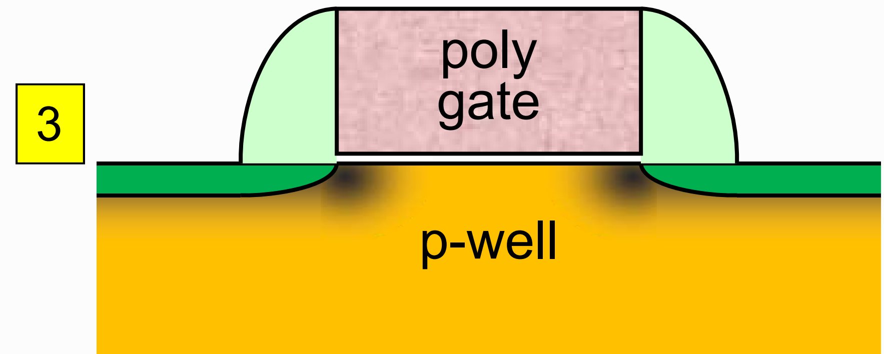
- Gate CD way smaller than lithography capability

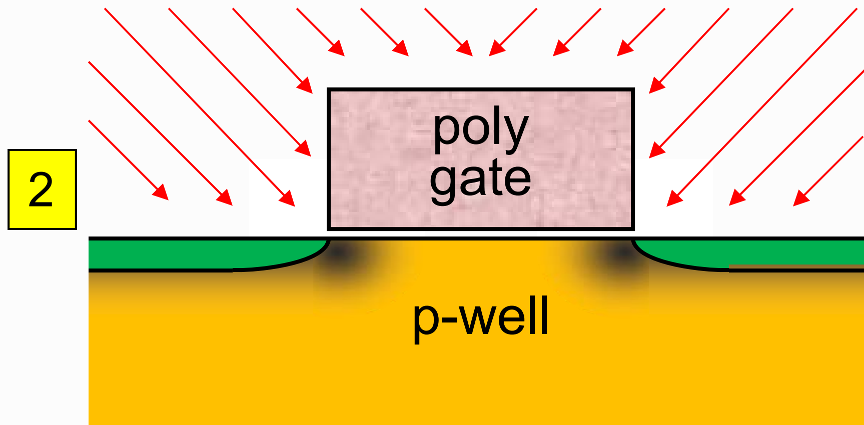


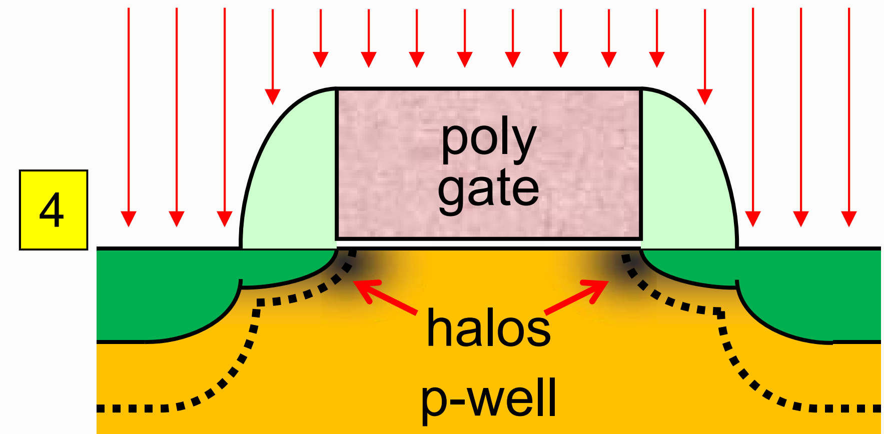
- *Process control is everything* – resist & poly etch chamber conditioning is critical (don't clean residues in tea cups or woks)
- Trim more for smaller CD (requires tighter control)
- Less trimming if narrower lines can be printed

Channel & Source/Drain Engineering





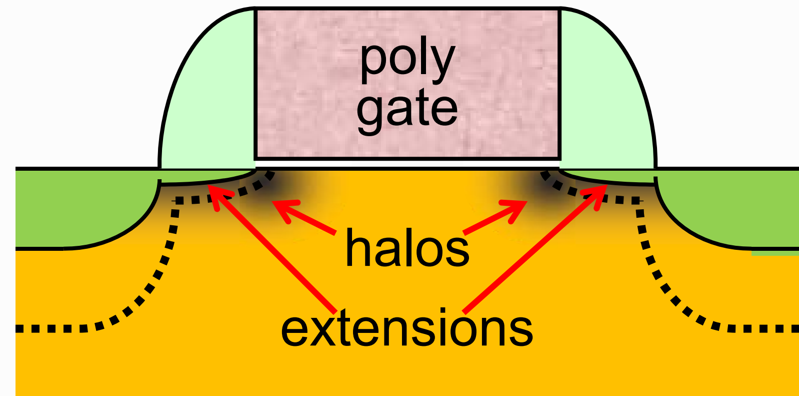




Benefits of Halo and Extension

Resulting structure

- Less short-channel effect
- Shallow junction where needed most



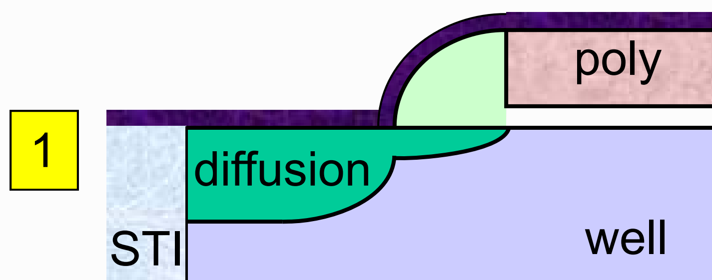
Not to be confused with LDD in I/O FET

- Same process with spacers but lightly doped drain (LDD) is used for minimizing peak electric fields that cause hot carriers & breakdown
- Extensions must be heavily doped for low series resistance

Self-Aligned Silicidation (Salicidation)

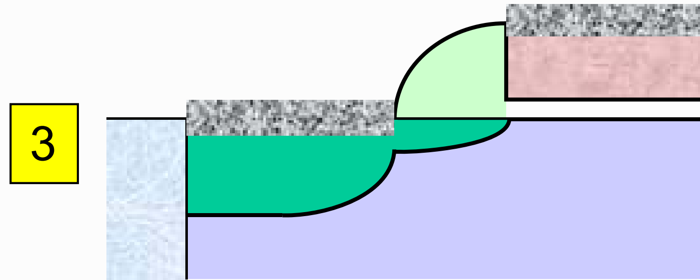
- Need to reduce poly & diffusion R_s, or get severe I_{FET} degradation

1



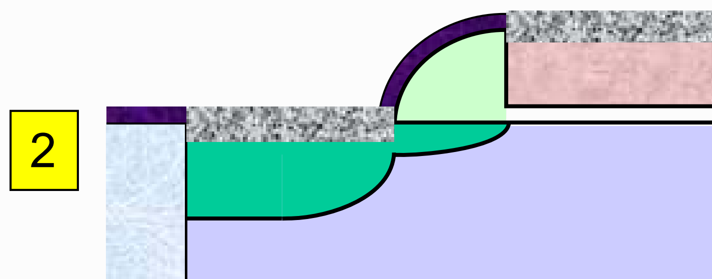
Deposit sicilide metal (Ti, Co, Ni)

3



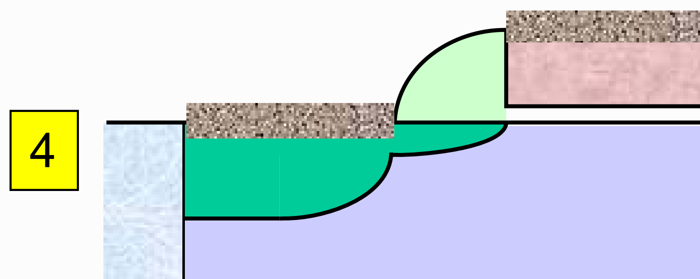
Strip unreacted metal

2



RTA1 (low temperature)
Selective formation of metal silicide from metal reaction with Si

4



RTA2 (high temperature)
Transforms silicide into low-ρ phase by consuming more Si

- $TiSi_x$ → $CoSi_x$ → $Ni/PtSi_x$
  - Scaling requires smaller grain size to minimize R_s variation

Outline

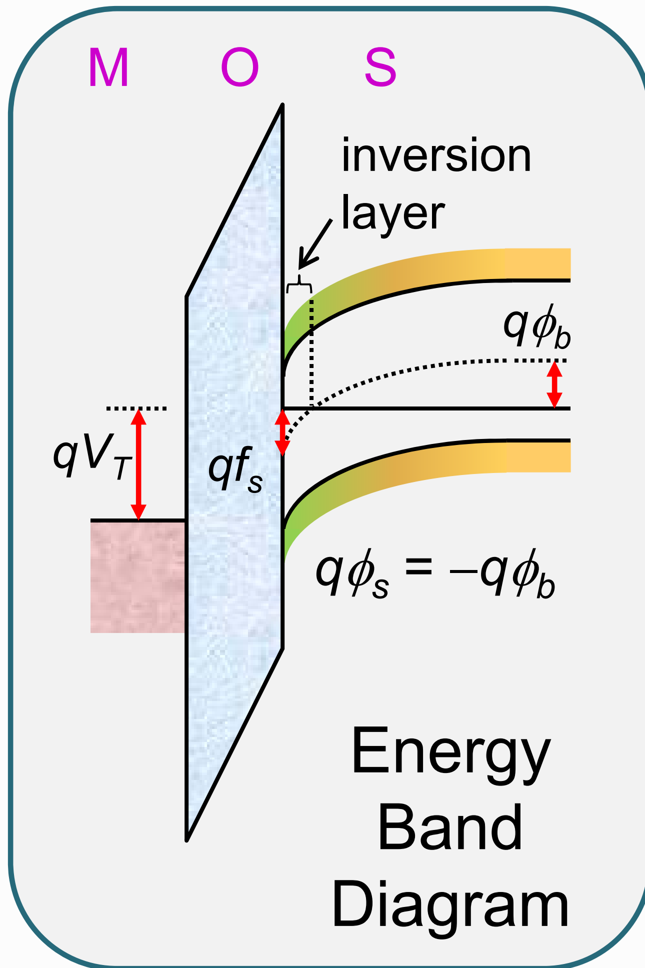
Part 1

- Motivation
- MOSFET & Short-Channel Fundamentals
- 130nm Fabrication
- More MOSFET Fundamentals
- Lithography
- Partially-Depleted SOI

Part 2

- Strain Engineering (90nm & Beyond)
- High-K / Metal-Gate (45nm & Beyond)
- Migrating to Fully-Depleted (22nm & Beyond)
- Tri-Gate FinFETs
- Conclusions

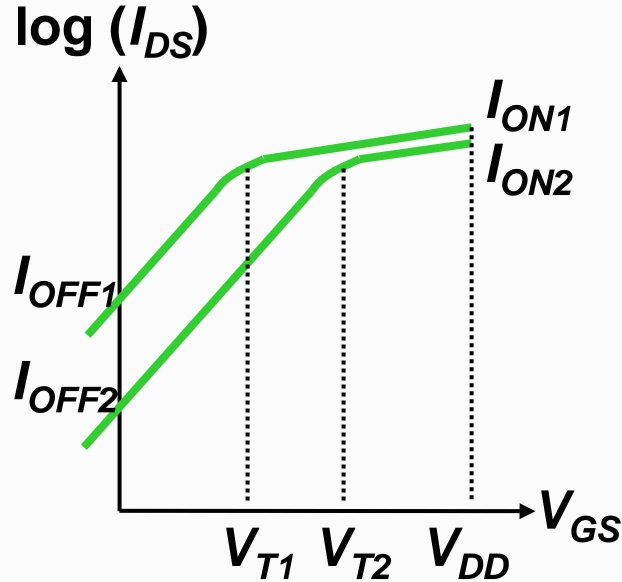
Not So Fundamental After All

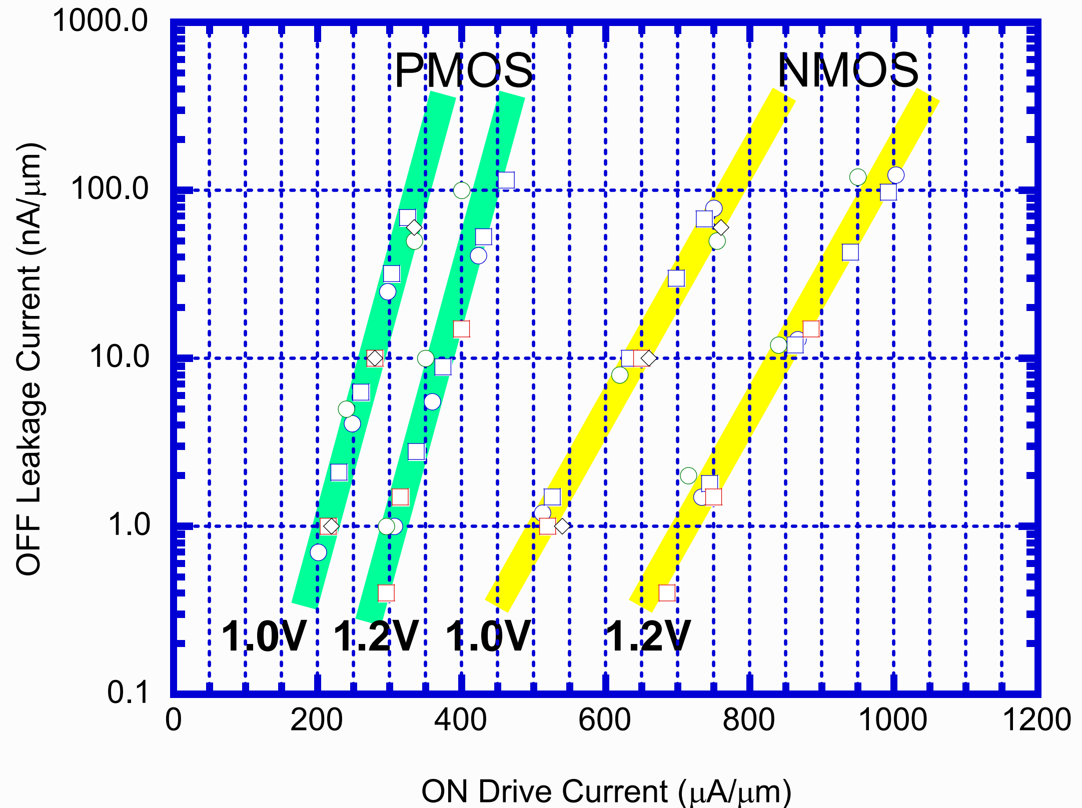


$$V_T = V_{FB} + 2\phi_b + \frac{Q_{dep}}{C_{ox}}$$

- Body doping has increased by 2–3 orders of magnitude over the decades
- Surface way more conductive at strong inversion condition using “fundamental” V_T definition
- What matters is how much OFF leakage you get for a given ON current
- *IDoff* vs. *IDsat* (or *IDeff*) universal plots have become more useful to summarize device performance

I_{OFF}–I_{ON} Universal Plots

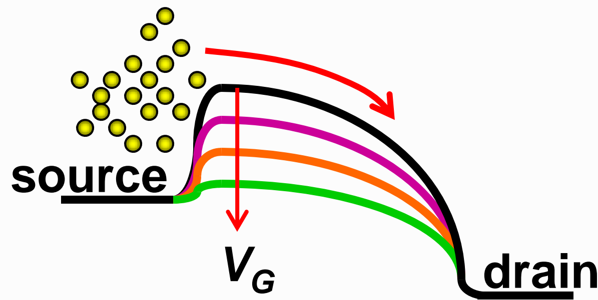




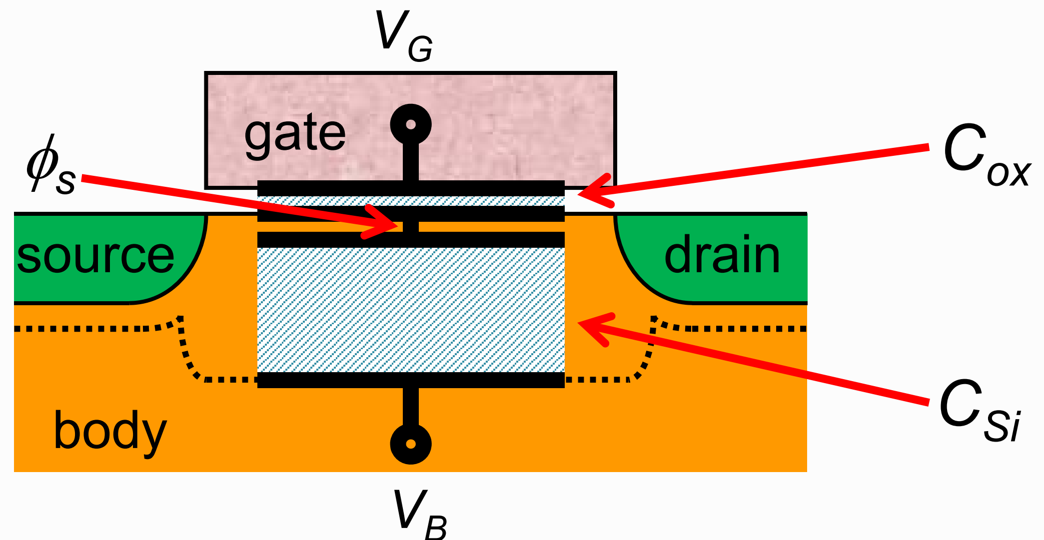
- High I_{ON} → high I_{OFF} & low I_{ON} → low I_{OFF}
- OFF leakage prevents V_T from scaling with gate length
- Several V_T's enable trade-off between high speed vs. low leakage

Subthreshold Leakage

- MOSFET is not perfectly OFF below V_T
- $V_G \uparrow \rightarrow \phi_s \uparrow \rightarrow$ lower source-to-channel barrier
- Gradually more carriers diffuse from source to drain
- Capacitive divider between gate and undepleted body



$$\Delta\phi_s = \Delta V_G \cdot \frac{C_{ox}}{C_{ox} + C_{Si}}$$

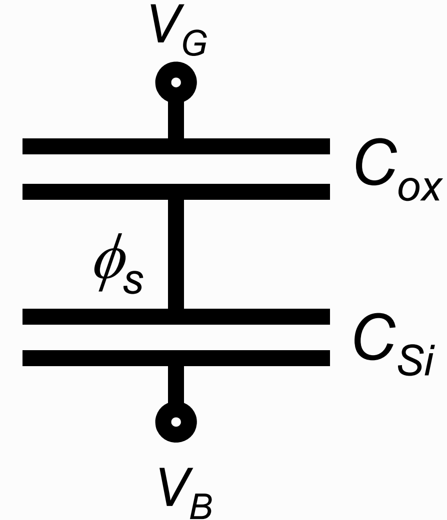


Subthreshold Slope

- V_G needed for 10× change in current

$$S = \frac{k_B T}{q}\ln(10)\cdot\frac{C_{ox}+C_{Si}}{C_{ox}}$$

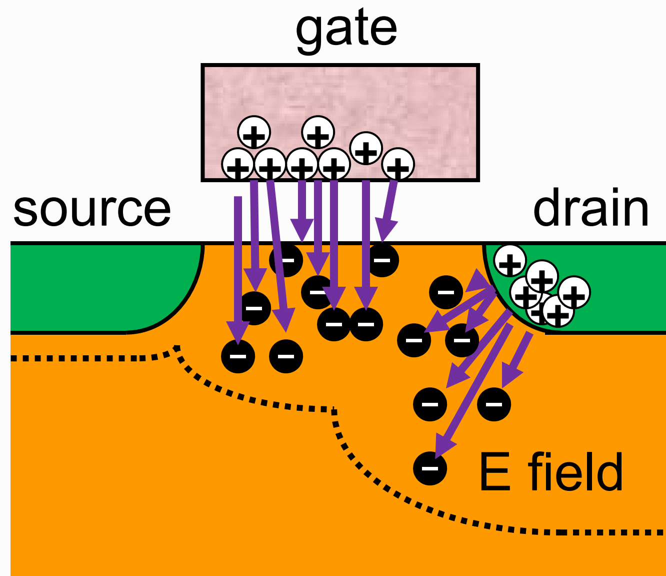
$$S = (60mV/dec)\cdot\frac{C_{ox}+C_{Si}}{C_{ox}} \quad \text{at } 25°\text{C}$$

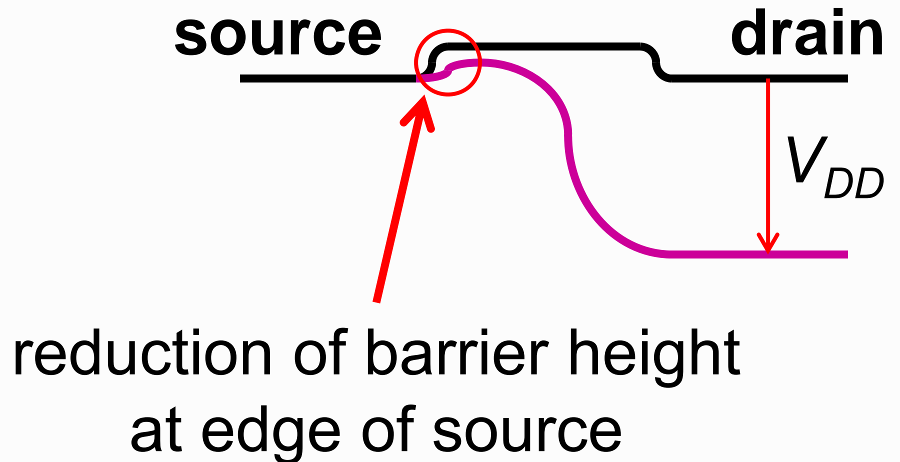


- **Planar 28nm: *S* = 100–110mV/dec at 25°C**
- Want tight coupling of V_G to ϕ_s but have to overcome C_{Si}
  - Large C_{ox} → thinner gate oxide, HKMG
  - Small C_{Si} → lower body doping, FD-SOI, finFET
  - Get diode limit when $C_{ox} \rightarrow \infty$ & $C_{Si} \rightarrow 0$ ($\eta = 1$)
- Reducing S enables lower V_T , V_{DD} & power for same I_{OFF}

Drain-Induced Barrier Lowering (DIBL)

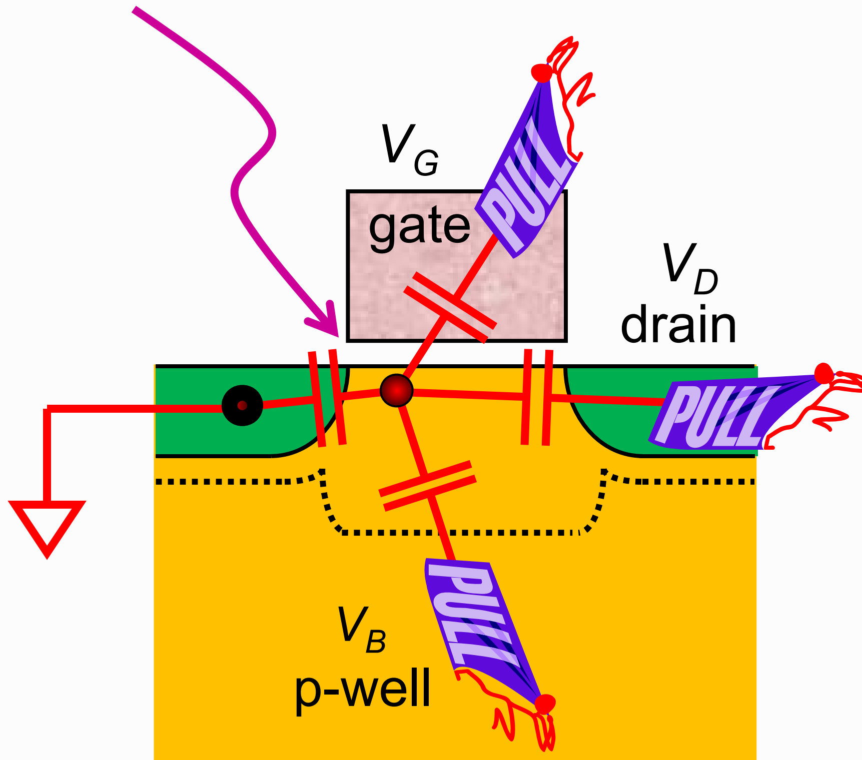
- OFF leakage gets worse at higher V_D
- E field from drain charge terminating in body, reducing gate charge required to reach V_T
- Characterized as V_T reduction for some ΔV_D
- **Planar 28nm: 150–160mV for ΔV_D=1V**
- Reducing DIBL also enables lower V_{DD} & power for same I_{OFF}

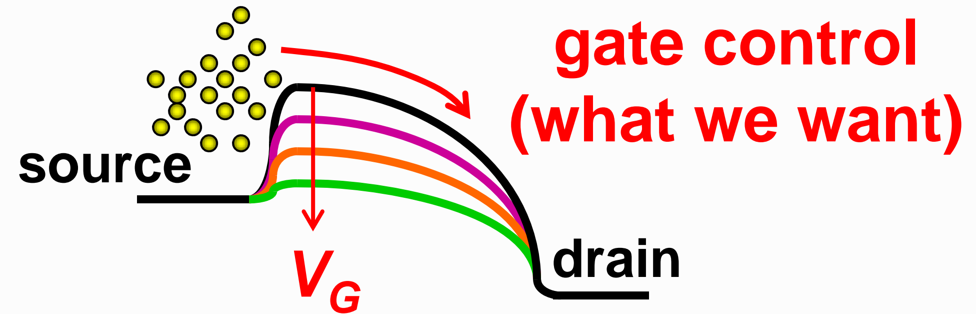


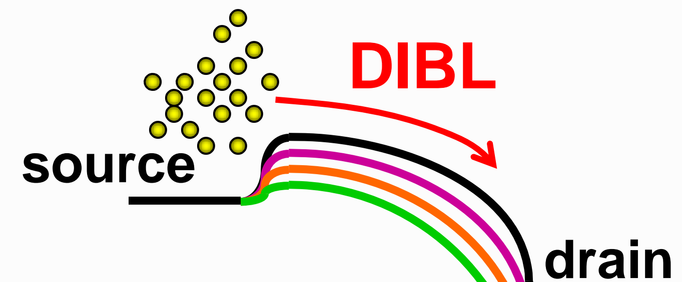


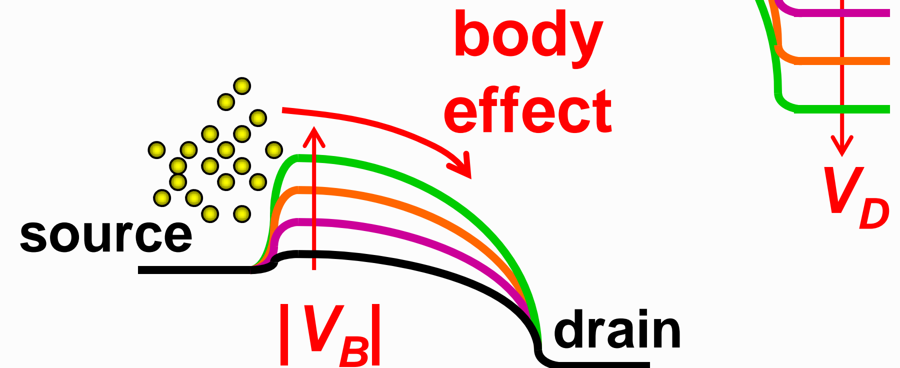
3-Way Competition for Body Charge

What's happening to surface potential?

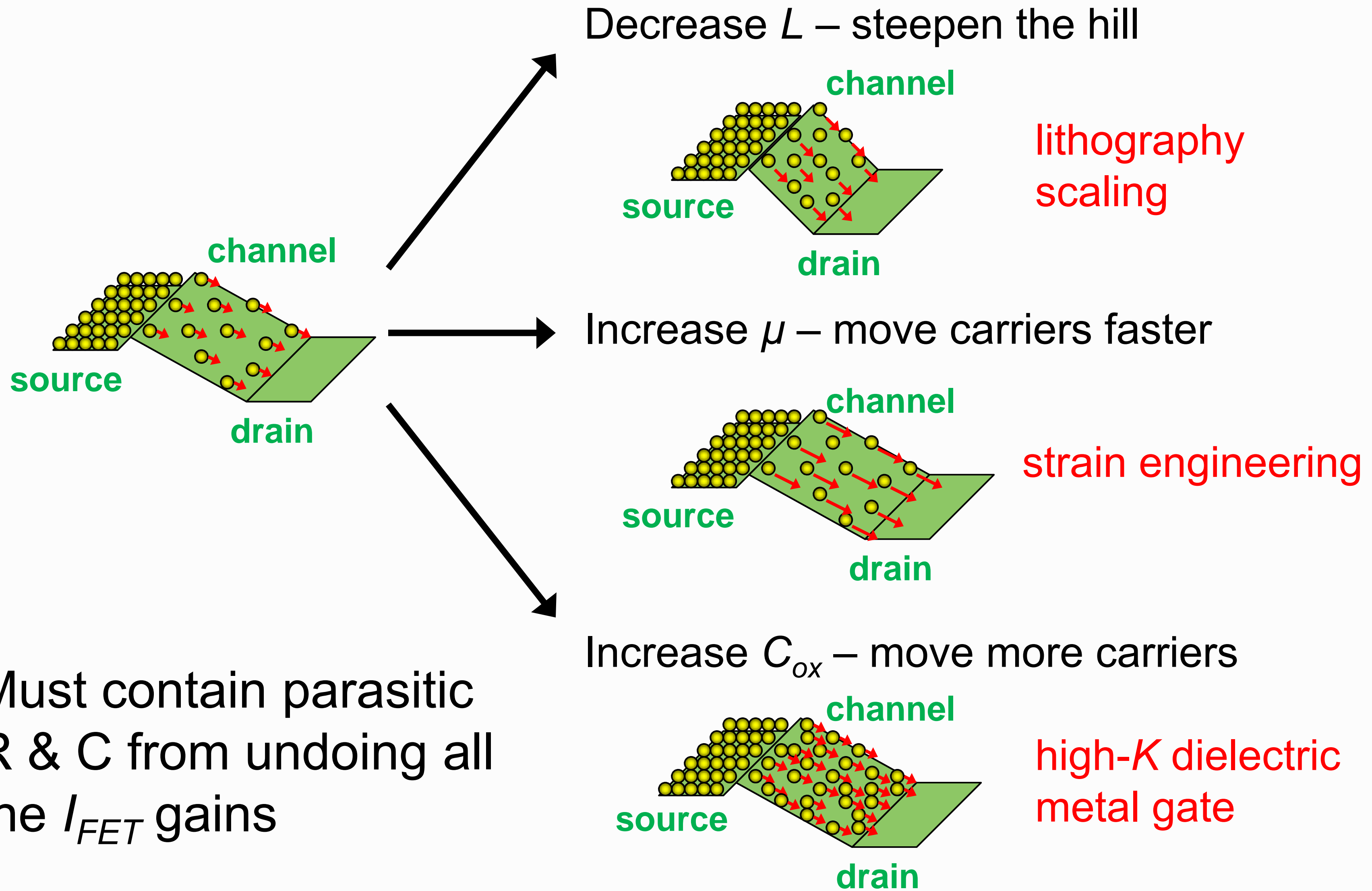








The Roads to Higher Performance



Must contain parasitic R & C from undoing all the I_{FET} gains

Outline

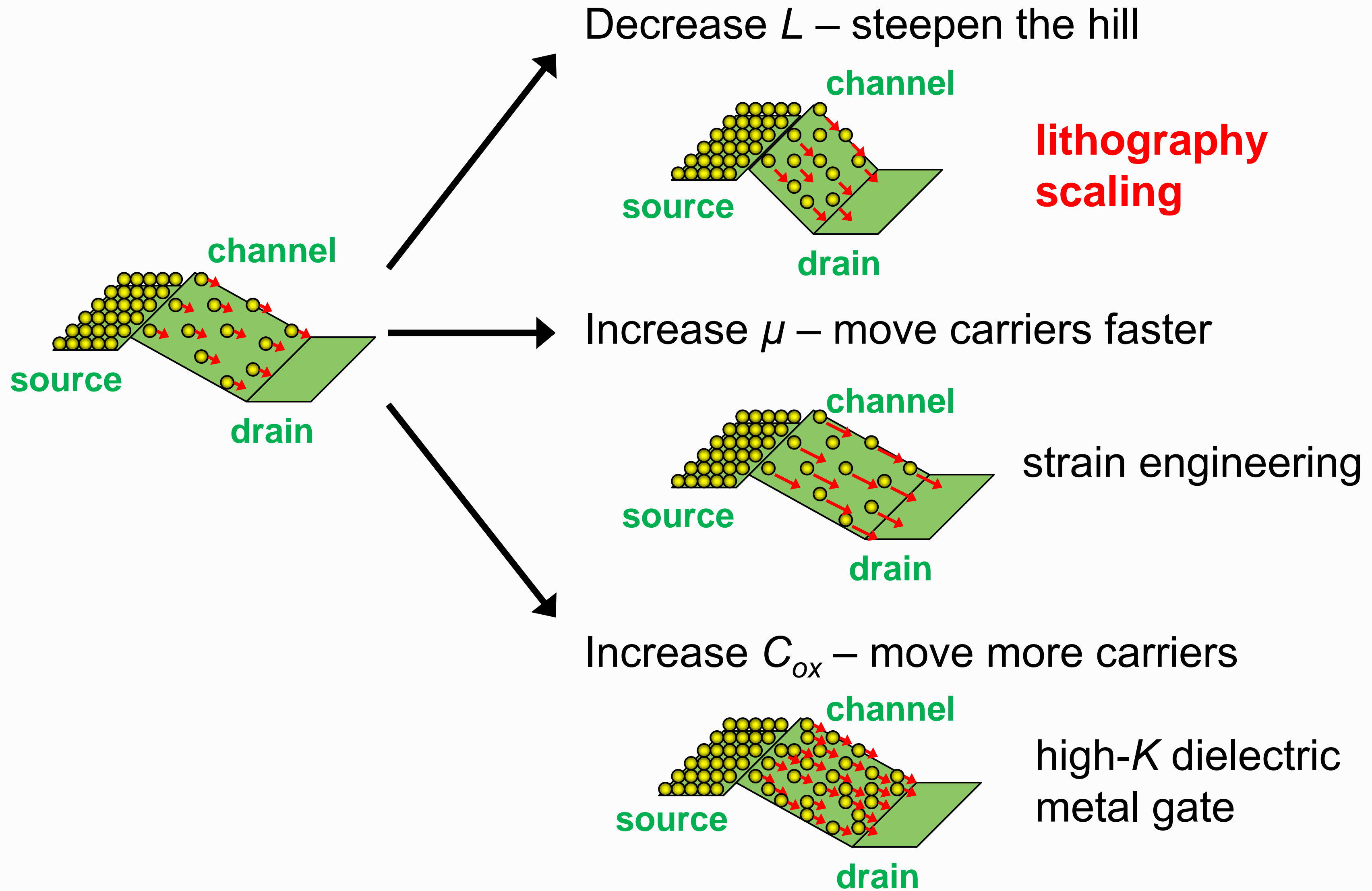
Part 1

- Motivation
- MOSFET & Short-Channel Fundamentals
- 130nm Fabrication
- More MOSFET Fundamentals
- Lithography
- Partially-Depleted SOI

Part 2

- Strain Engineering (90nm & Beyond)
- High-K / Metal-Gate (45nm & Beyond)
- Migrating to Fully-Depleted (22nm & Beyond)
- Tri-Gate FinFETs
- Conclusions

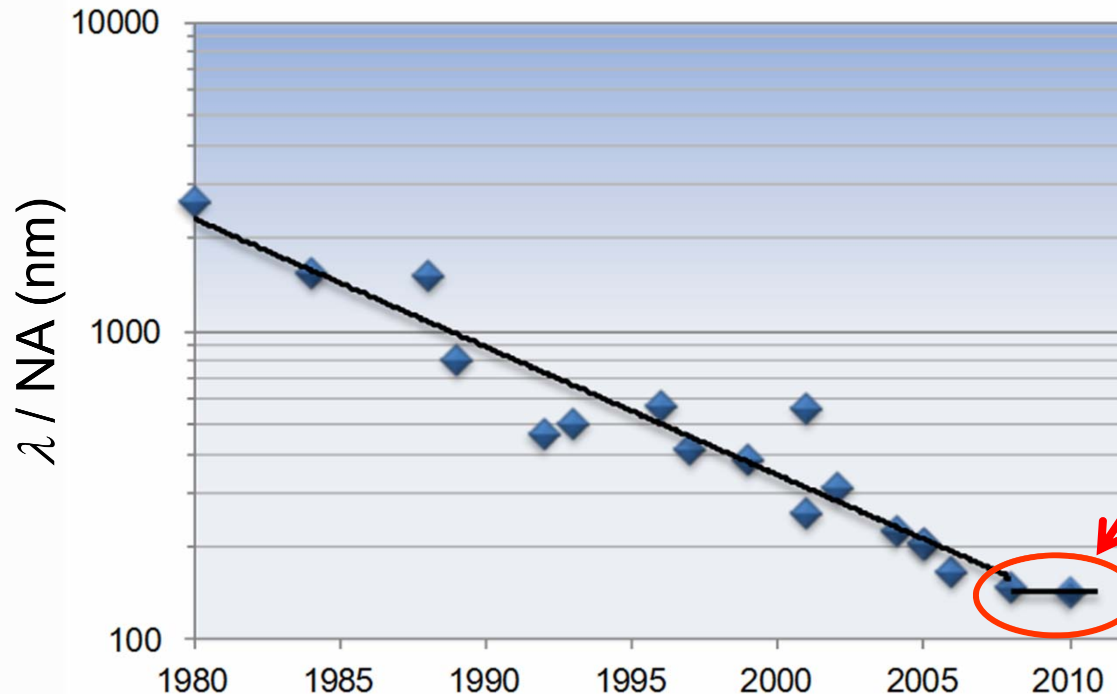
The Roads to Higher Performance



Let There Be Light

Resolution = $\frac{k_1 \lambda}{NA}$

- Tooling has traditionally driven resolution scaling
- Shorter λ : 436nm → 365nm → 248nm → 193nm
- Higher *NA* lenses → capping at 1.35



- Both λ and NA have hit a wall
- No new litho tool for 22/20nm nodes (EUV not primetime yet)
- Single patterning limited to ~80nm pitch

Wei, GlobalFoundries [4]

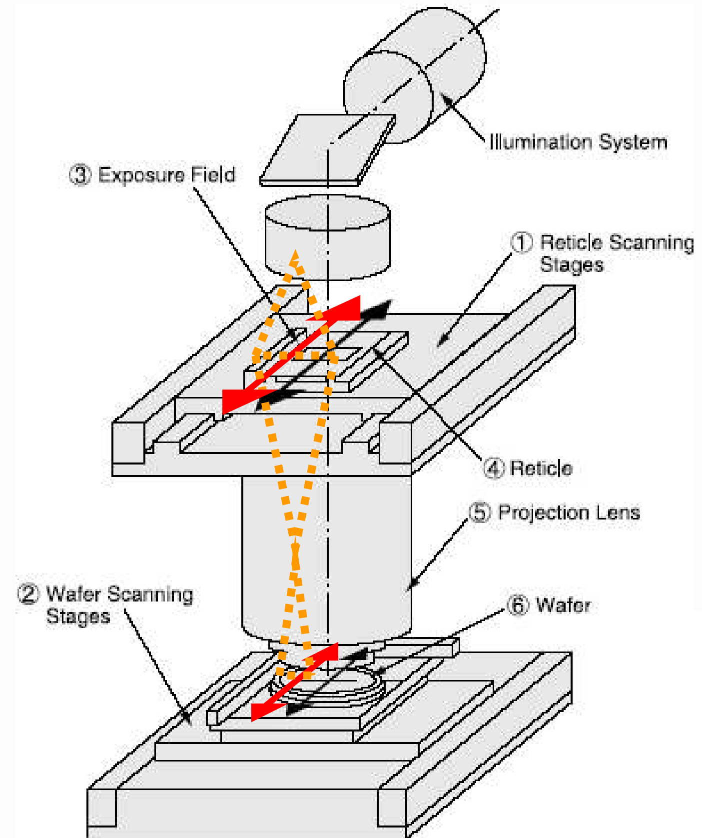
Step-and-Scan Projection Lithography

- Slide both reticle & wafer across narrow slit of light
- Only need high-NA optics orthogonal to scan but now high-precision constant-speed stages to move mask & wafer
- Cheaper than high-*NA* 2-D optics
- 6” x 6” physical reticle size (4× reduction)
- 25 x 33mm or 26 x 32mm field size
- Weak intensity of deep-UV source requires sensitive *chemically-amplified* resists for better throughput
- Enables dose mapping (adjust light dose during scan to compensate for loading)

Slit Source

Excimer Laser

KrF (248nm) or ArF (193nm)

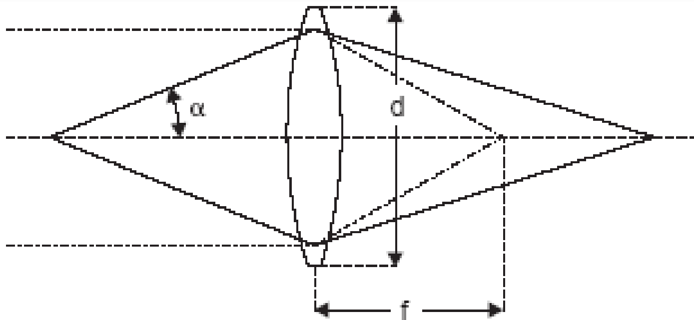


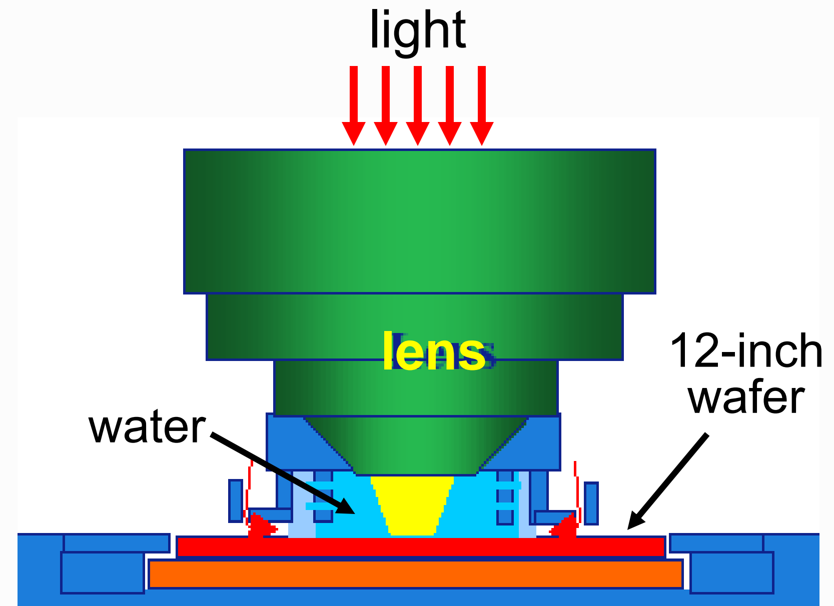
Immersion Lithography

- Remember oil immersion microscopy in biology class?
- Extend resolution of refractive optics by squirting water puddle on wafer surface prior to exposure
  - n_{water} ~1.45 vs. n_{air} ~ 1
  - Tedious but EUV is not primetime yet

$$\text{Resolution} = \frac{k_1 \lambda}{NA}$$

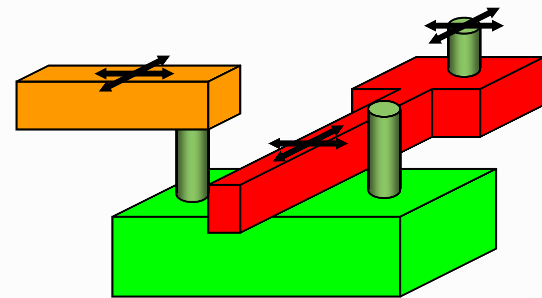
$$NA = \mathbf{n} \sin \alpha = d / 2 f$$





Lithography Misalignment / Overlay

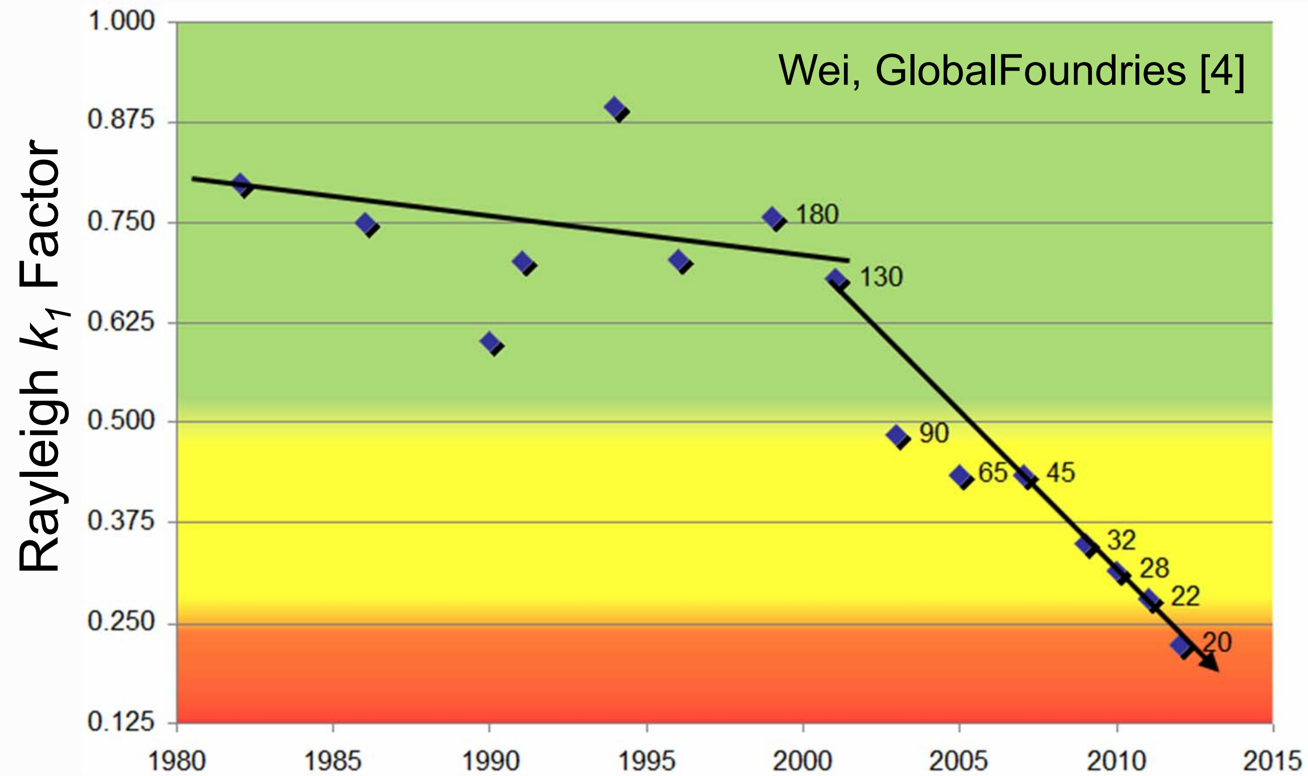
- Mask misalignment tolerance is not keeping pace with gate CD scaling
- ASML has near monopoly on lithography tools largely because of good overlay control (global zero layer patterns)
- Many layout enclosure & spacing rules not scaling with CD



- Examples:
  - Poly overhang beyond active
  - Contact spacing to poly
  - Active enclosure around contact
  - Metal enclosure around vias
- Layout for matching must be robust against overlay errors

Resolution Enhancement Technology

$$\text{Resolution} = \frac{k_1 \lambda}{NA}$$



- Reducing k_1 is the remaining ticket to better resolution
- Attack problem from all fronts: mask, source & wafer
- Imposes significant restrictions on layout design rules

Outline

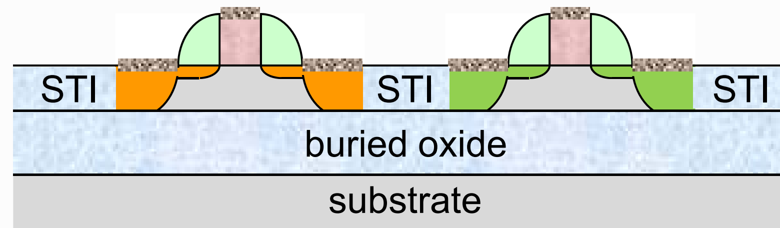
Part 1

- Motivation
- MOSFET & Short-Channel Fundamentals
- 130nm Fabrication
- More MOSFET Fundamentals
- Lithography
- Partially-Depleted SOI

Part 2

- Strain Engineering (90nm & Beyond)
- High-K / Metal-Gate (45nm & Beyond)
- Migrating to Fully-Depleted (22nm & Beyond)
- Tri-Gate FinFETs
- Conclusions

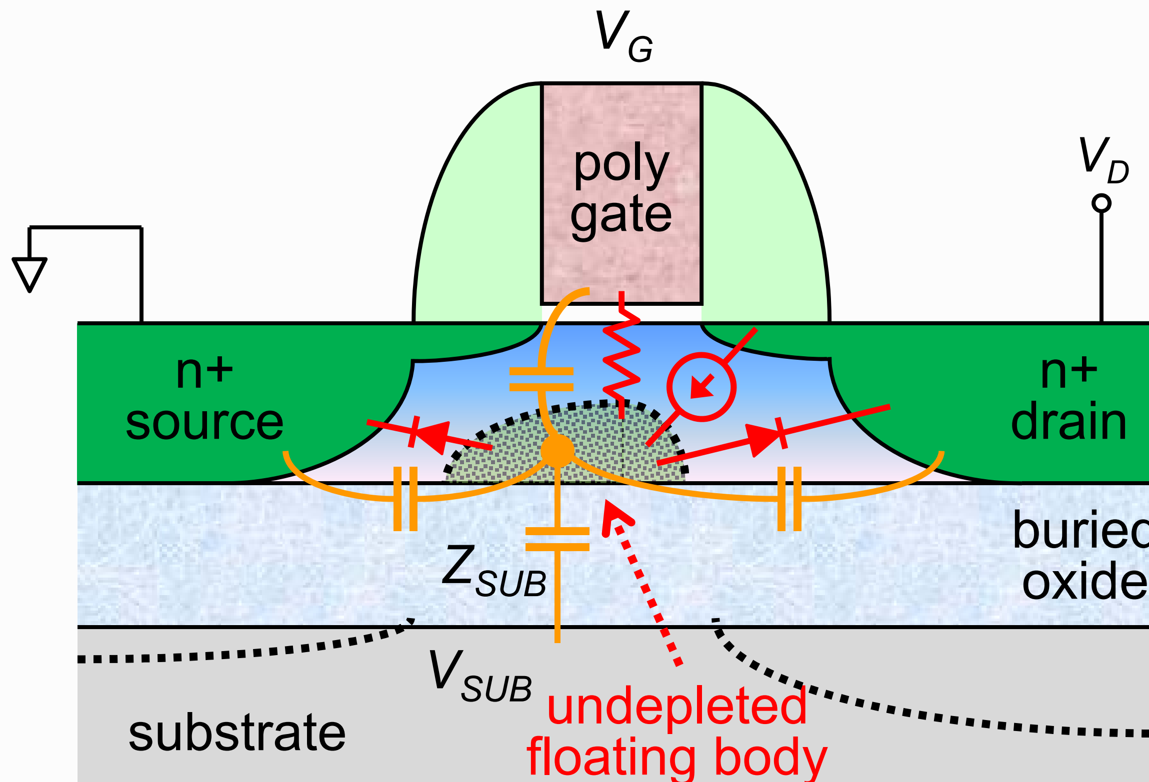
Partially-Depleted Silicon-On-Insulator



- Pros
  - Dynamic threshold effect (ϕ_s coupling to V_G edge)
  - Low junction area capacitance
  - No V_T increase for stacked FETs
  - 4× lower SRAM soft-error rate
  - Body isolation from substrate noise
  - Simpler isolation process, reduced well proximity effect
- Cons
  - Body hysteresis effect – floating body gets kicked around
    - Requires conservative margining for digital timing
    - Major pain for analog/mixed-signal design
  - Substrate heating – buried oxide is good insulator
  - More expensive substrate, and from a single supplier [5]

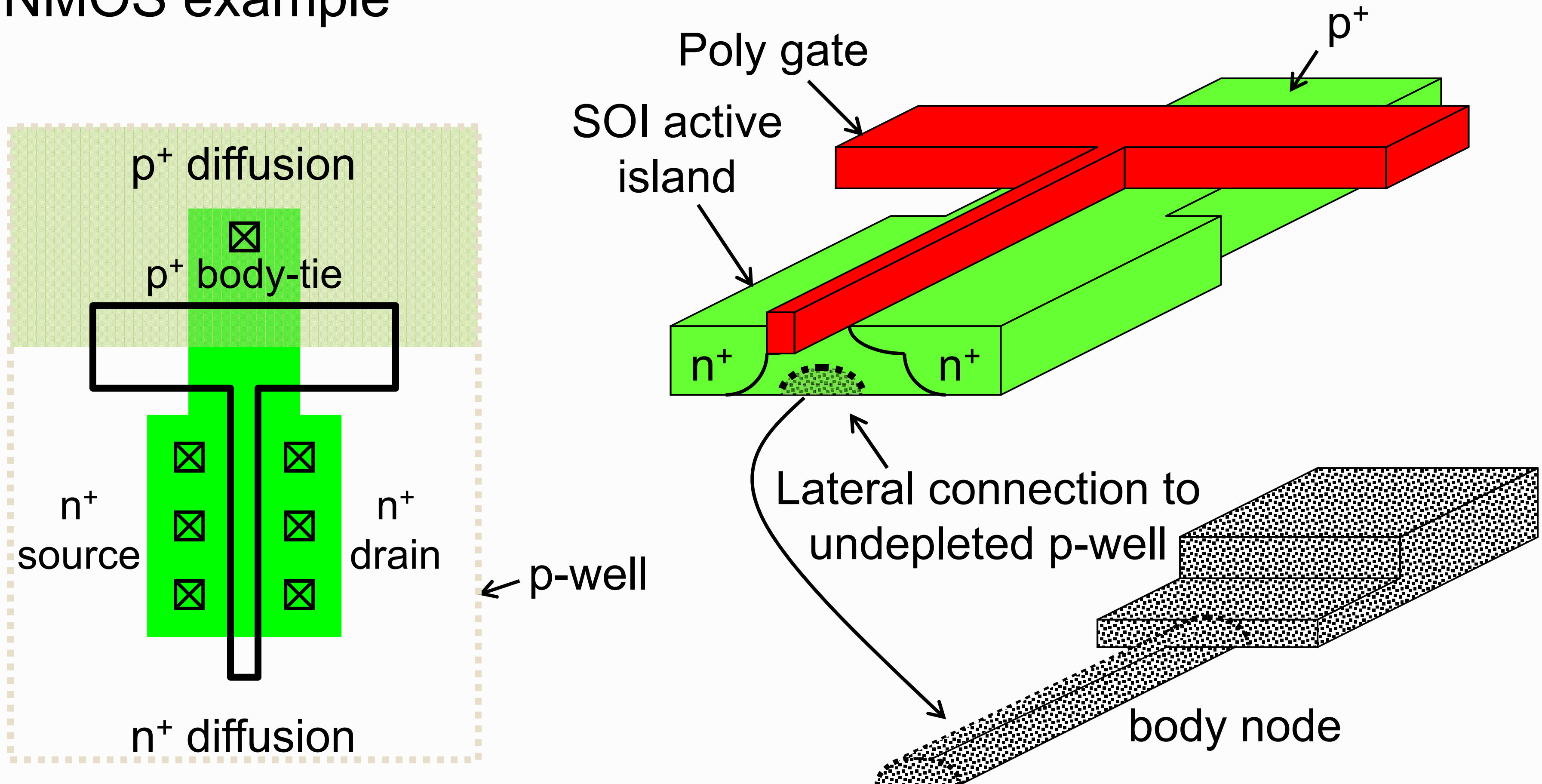
The Dreaded Hysteresis Effect

- Floating body is coupled to source, gate, drain & body
- Body voltage has memory or history of other terminals, analogous to intersymbol interference in wireline I/O
- Floating body voltage noise → V_T noise → I_D noise
- Can get hysteresis in bulk if Z_{SUB} is too high



Body-Tied PD-SOI MOSFET (T-Gate)

- Enables body connection to undepleted FET well
- High R_{body} and extra C_{gate} limits bandwidth of body connection
- NMOS example



Outline

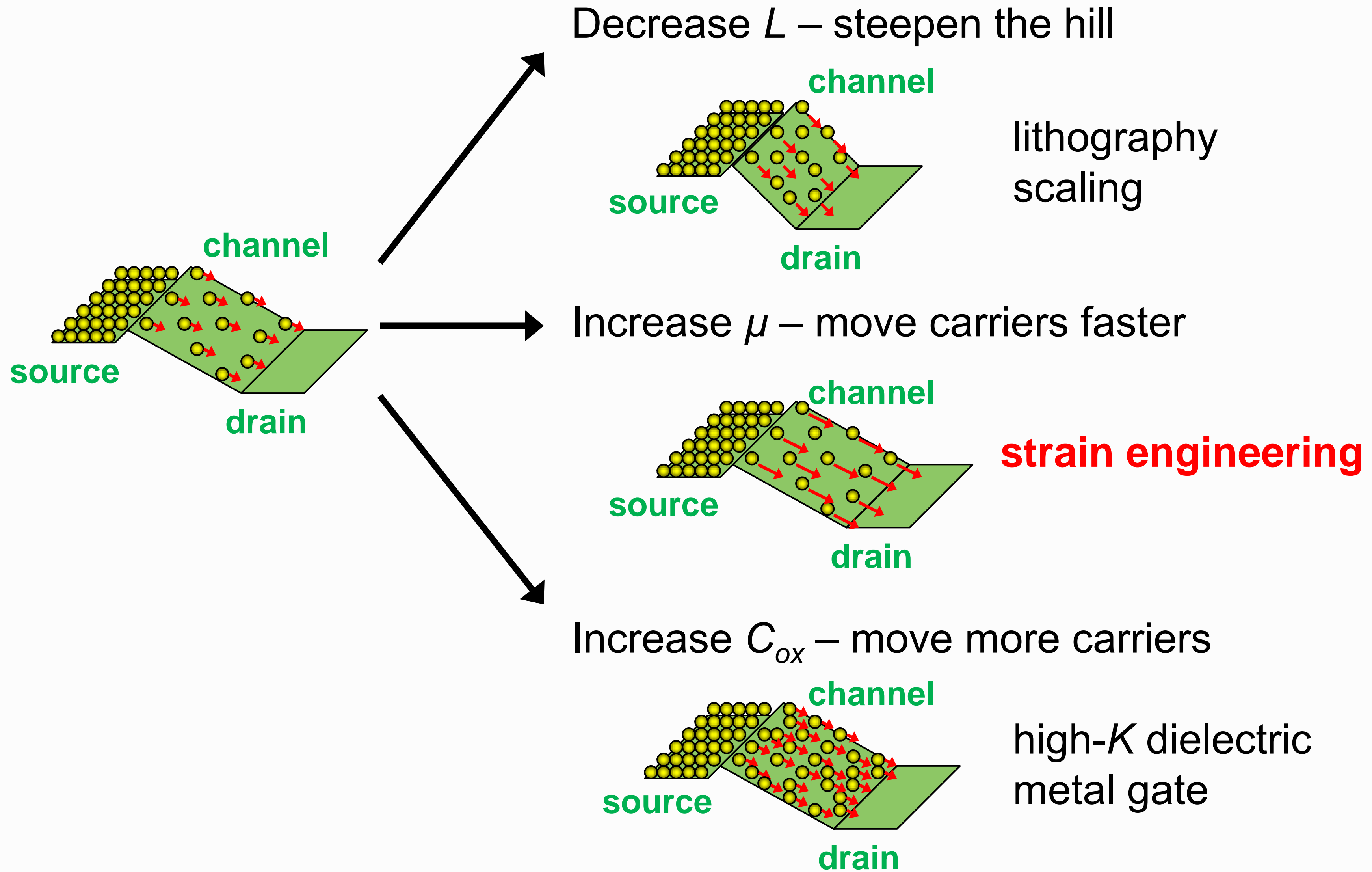
Part 1

- Motivation
- MOSFET & Short-Channel Fundamentals
- 130nm Fabrication
- More MOSFET Fundamentals
- Lithography
- Partially-Depleted SOI

Part 2

- Strain Engineering (90nm & Beyond)
- High-K / Metal-Gate (45nm & Beyond)
- Migrating to Fully-Depleted (22nm & Beyond)
- Tri-Gate FinFETs
- Conclusions

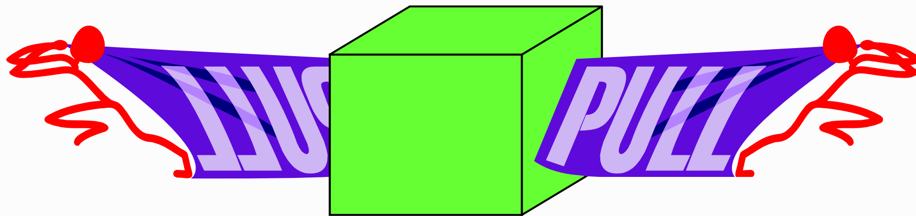
The Roads to Higher Performance

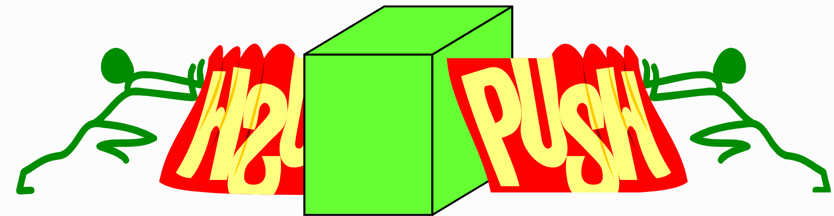


Mechanical Stresses & Strains

$$Stress(\sigma) = \frac{Force}{Area} \qquad Strain(\varepsilon) = \frac{\Delta \ell}{\ell_0}$$

Tension (positive stress) vs. Compression (negative stress)





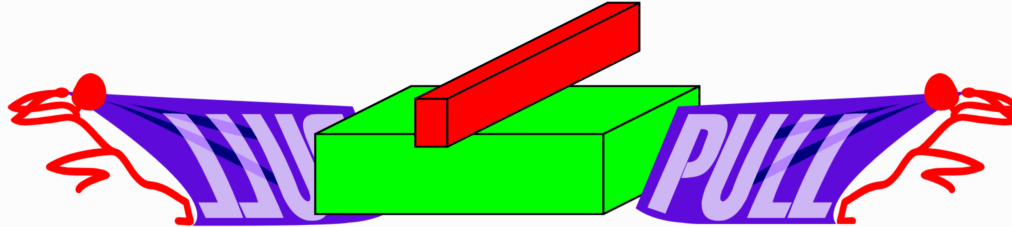
atomic spacing > equilibrium spacing

atomic spacing < equilibrium spacing

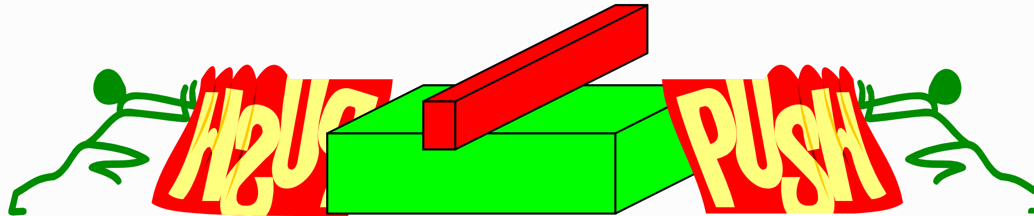
- Stretching / compressing FET channel atoms by *as little as 1%* can improve electron / hole mobilities by *several times*
- Strain perturbs crystal structure (energy bands, density of states, etc.) → changes effective mass of electrons & holes
- Increase I_{ON} for the same I_{OFF} without increasing C_{OX}

Longitudinal Uni-Axial Strain

tension (stretch atoms apart) → faster NMOS

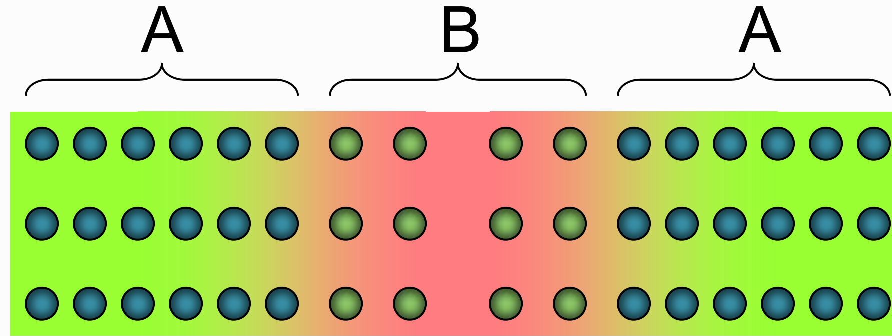


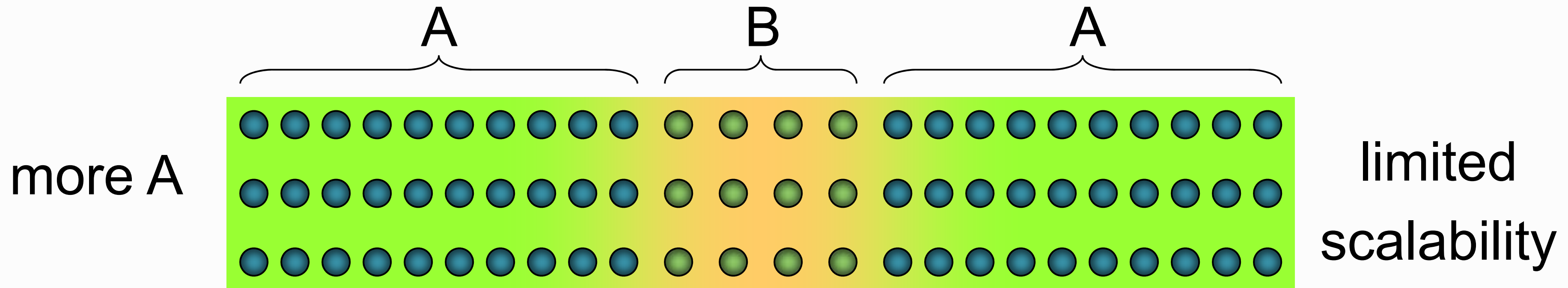
compression (squeeze atoms together) → faster PMOS

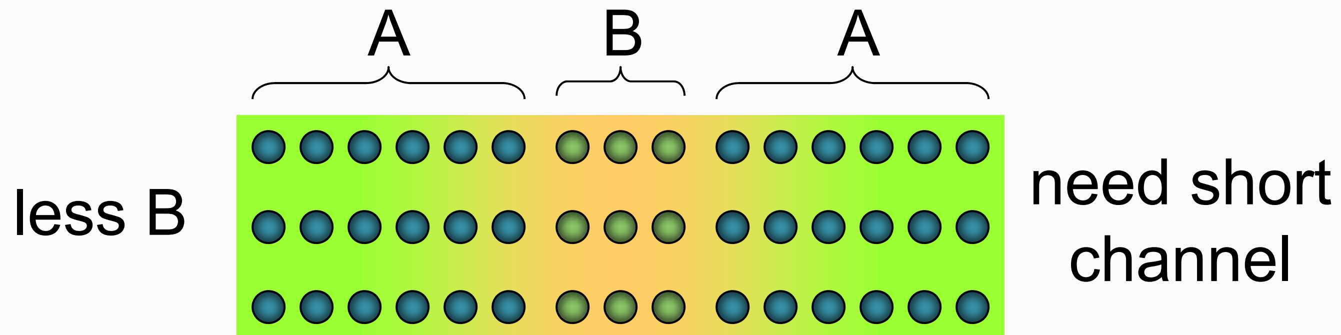


- Most practical means of incorporating strain for mobility boost
- Want 1-3GPa (high-strength steel breaks at 0.8GPa)
- How? Deposit strained materials around channel
  - Material in tension wants to relax by pulling in
  - Material in compression wants to relax by pushing out

Transferring Strain from Material A to B





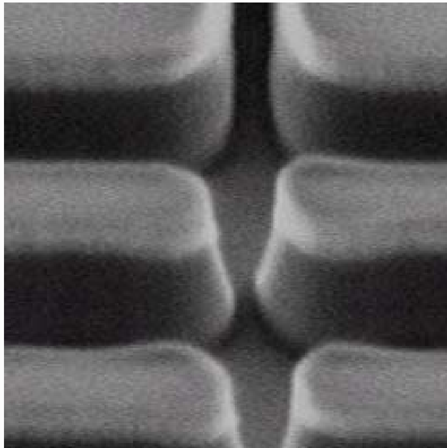


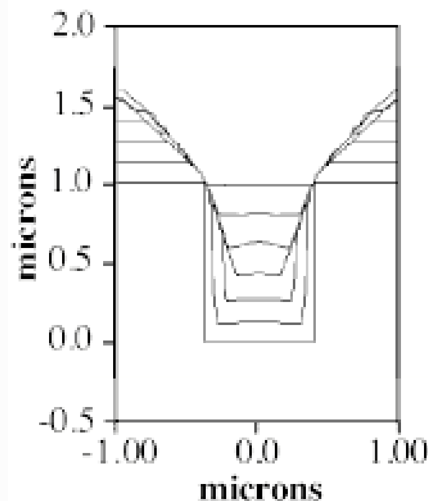
Ways to Incorporate Uni-Axial Strain

- NMOS wants tension, PMOS wants compression
- Un-Intentional (comes for *free*)
  - Shallow Trench Isolation – **NMOS ☹** / **PMOS ☺**
- Intentional (requires extra processing)
  - Stress Memorization Technique – **NMOS ☺**
  - Embedded-SiGe Source/Drain – **PMOS ☺**
  - Embedded-SiC Source/Drain – **NMOS ☺**
  - Dual-Stress Liners – **NMOS ☺ & PMOS ☺**
  - Compressive Gate Fill – **NMOS ☺** / **PMOS ☹**
- Strain methods are additive

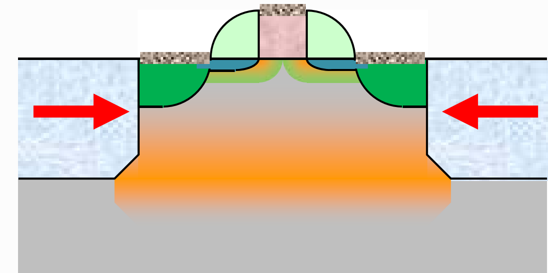
Shallow Trench Isolation (STI)
NMOS ☹ & PMOS ☺

- STI oxide under compression
  - High-Density Plasma CVD SiO_2 process (alternating deposition/etch) deposits intrinsically compressive oxide for good trench fill
  - 10× CTE mismatch between Si & SiO_2 increases compression when cooled from deposition temperature
- Migrated to High Aspect Ratio Process (HARP) fill in recent nodes → less compressive oxide



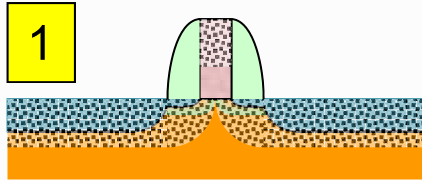


Plummer *et al*., Stanford [6]

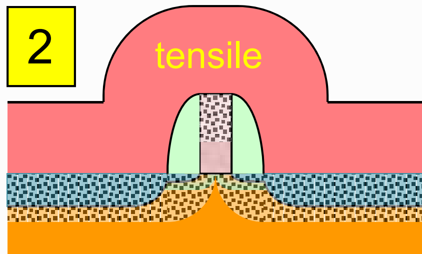


Stress Memorization Technique (SMT)

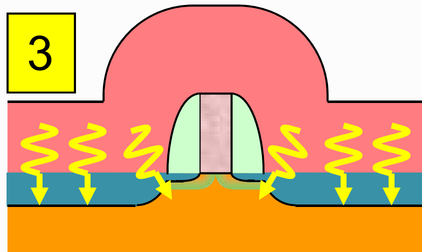
NMOS ☺



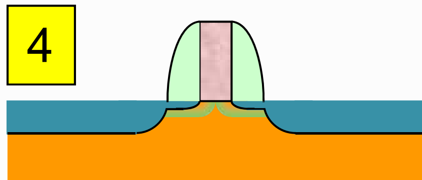
Amorphize poly & diffusion with silicon implant



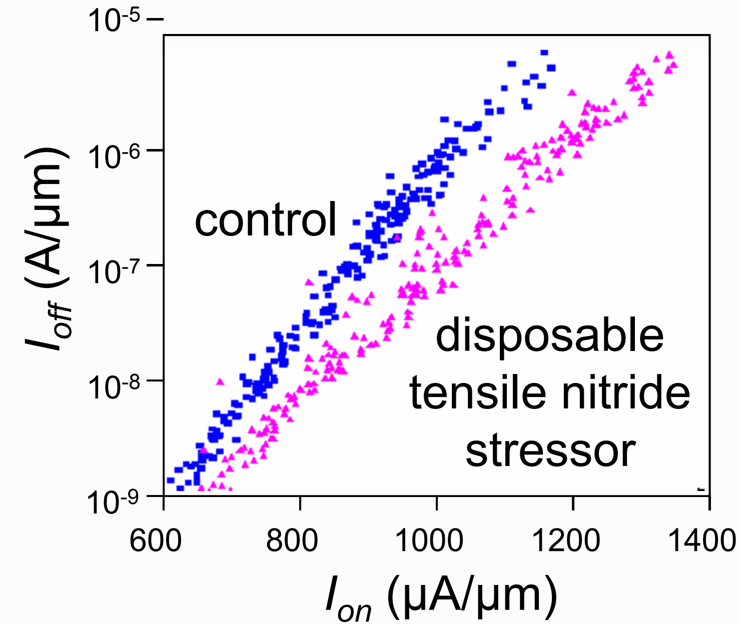
Deposit tensile nitride



Anneal to *make nitride more tensile* and transfer nitride tension to crystallizing amorphous channel



Remove nitride stressor (tension now frozen in diffusion)



Chan *et al.*, IBM [7]

Periodic Table Trends

Periodic Table of the Elements

	IA	IIA	IIIB	IVB	VB	VIB	VIIB		VII		IB	IIB	IIIA	IVA	VA	VIA	VIIA	O
1	1 H																	2 He
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	57 *La	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	89 +Ac	104 Rf	105 Ha	106 Sg	107 Ns	108 Hs	109 Mt	110 110	111 111	112 112	113 113					

* Lanthanide Series	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
+ Actinide Series	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

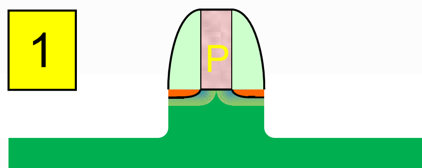
lattice spacing ↑
bandgap ↓

- Compound semiconductor like Si_xGe_{1-x} has lattice spacing & bandgap between Si & Ge
- Same idea with Si_xC_{1-x}

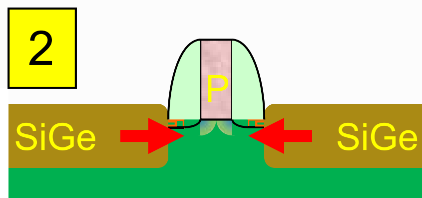
Embedded-SiGe Source/Drain (e-SiGe)

PMOS ☺

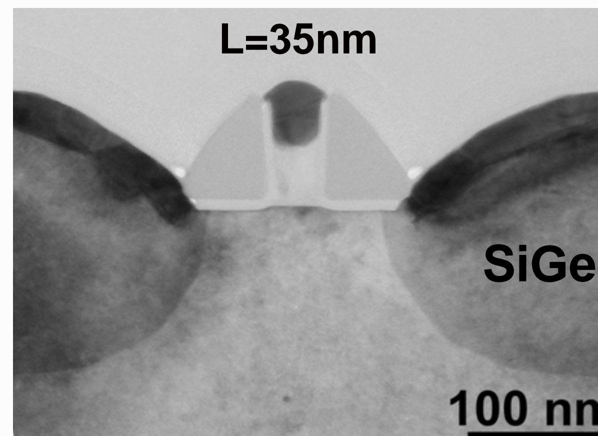
- SiGe constrained to Si lattice will be in compression
- Compressive SiGe source/drain transfers compression to Si channel



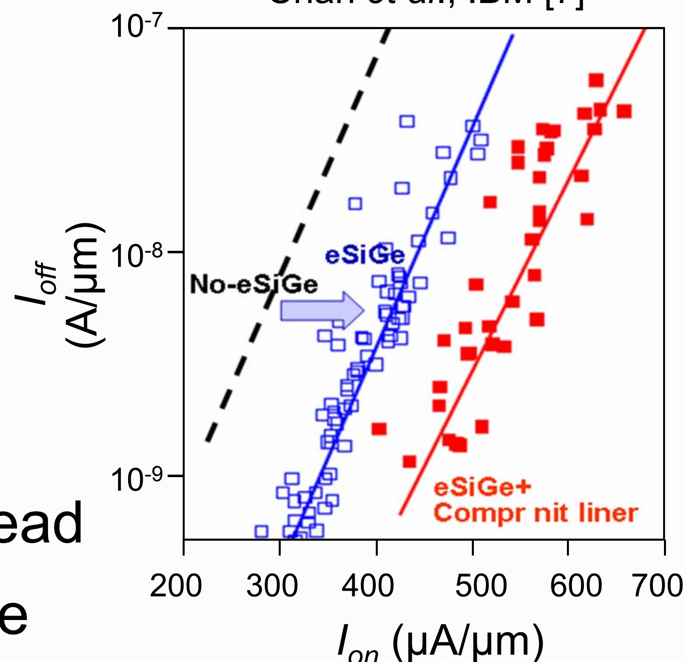
Etch source/drain recess



Grow SiGe epitaxially in recessed regions



Chan *et al*., IBM [7]



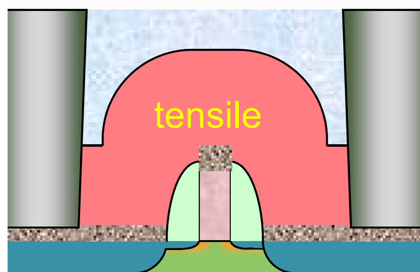
- e-SiC is similar but introduces tension instead
- Epitaxial SiC much tougher to do than SiGe

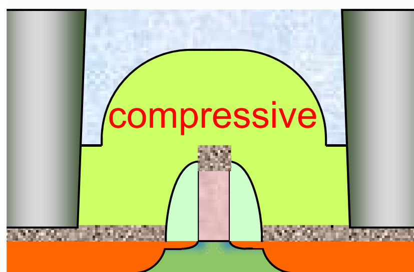
Dual-Stress Liners

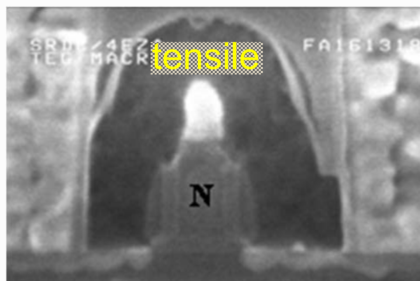
NMOS ☺ & PMOS ☺

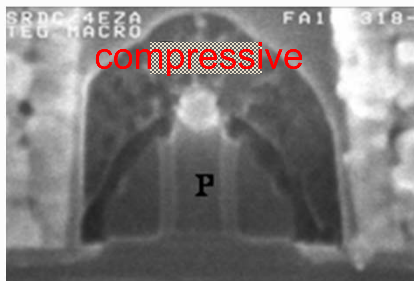
- Deposit tensile/compressive PECVD SiN (PEN) liners on N/PMOS
- Liner stress is dialed in by liner deposition conditions (gas flow, pressure, temperature, etc.)

TPEN for NMOS *CPEN for PMOS*

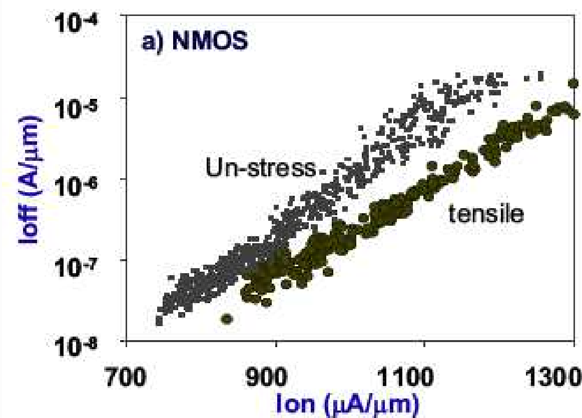


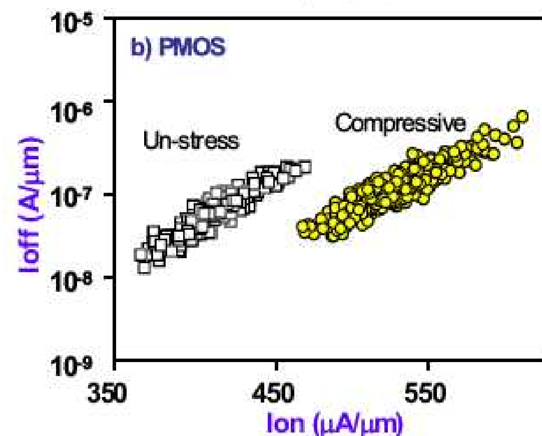






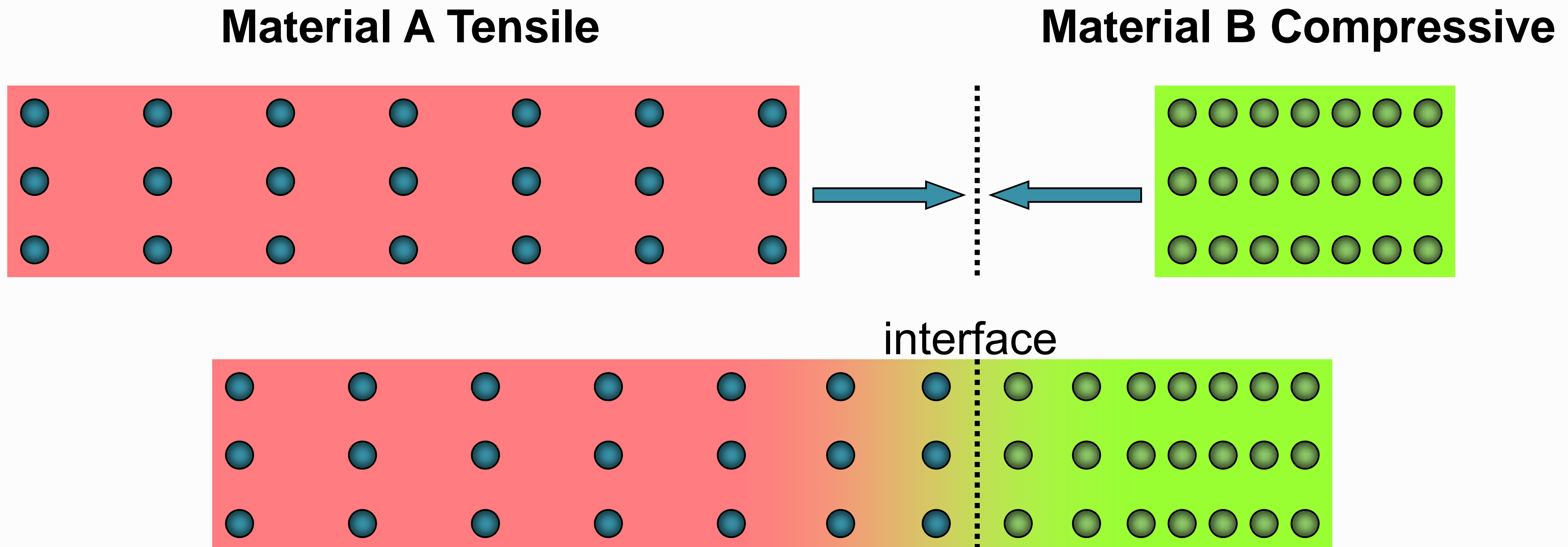
Chan *et al.*, IBM [7]





Strain Relaxation

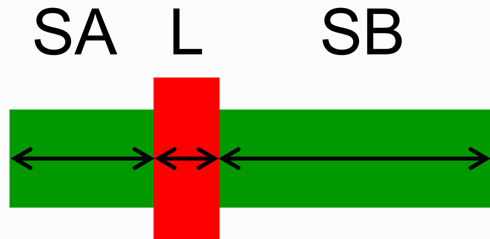
When materials of different strain come together…



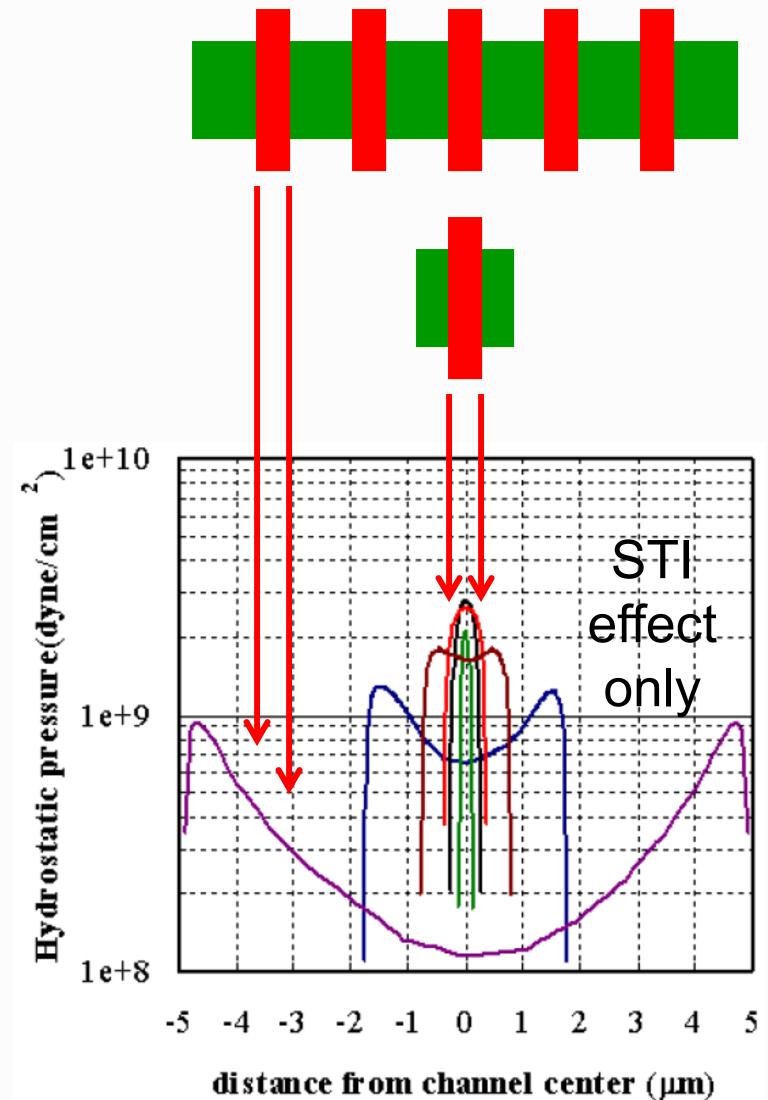
- Both materials will relax at the interface
- Extent of relaxation is gradual, depends on distance from interface
- No relaxation far away from interface

Strain Depends on Channel Location

- SA, L & SB specify where channel is located along active area



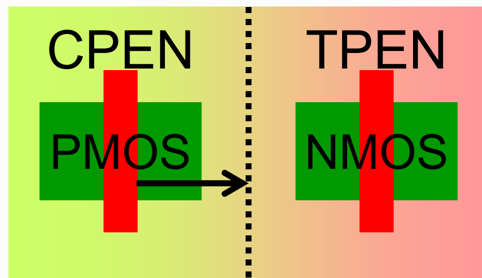
- Critical for modeling device mobility change due to STI, SMT, e-SiGe, etc.
- Strain at source & drain ends of channel may be different
- Important consideration for matching, e.g., current mirrors
- Concavity & stress polarity will vary with stressors in given technology but concept still applies

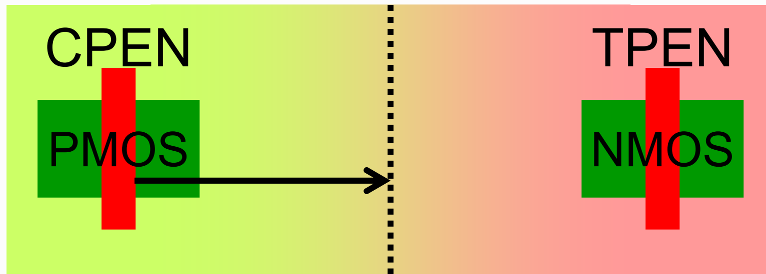


Xi *et al.*, UC Berkeley [8]

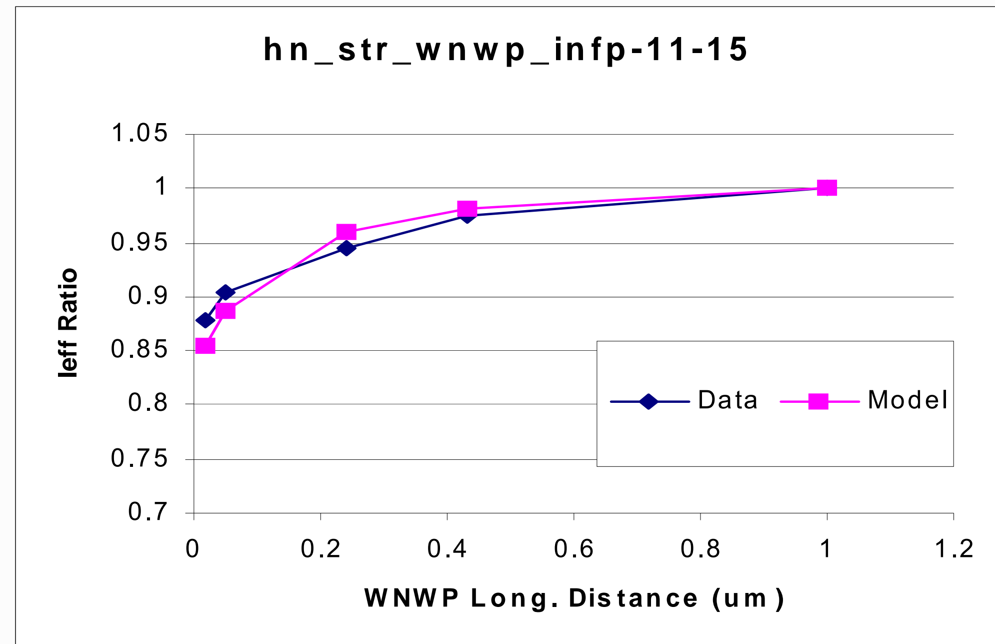
Longitudinal DSL Proximity

- Opposite device type nearby in longitudinal direction reduces impact of stress liner → *mutually slow each other down*
- Opposite PEN liner absorbs/relieves stress introduced by PEN





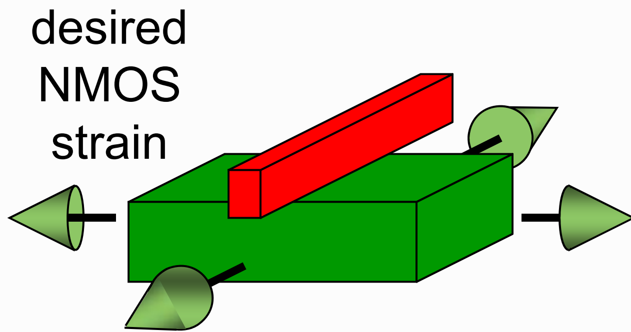
PMOS Longitudinal Proximity

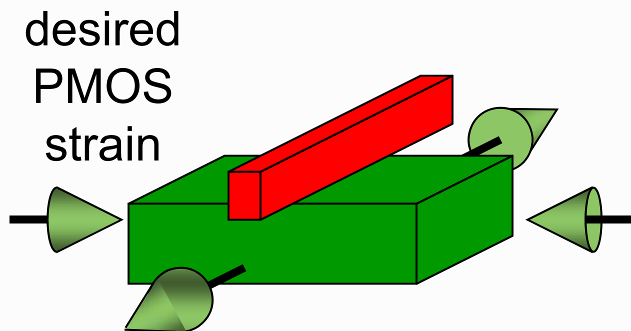


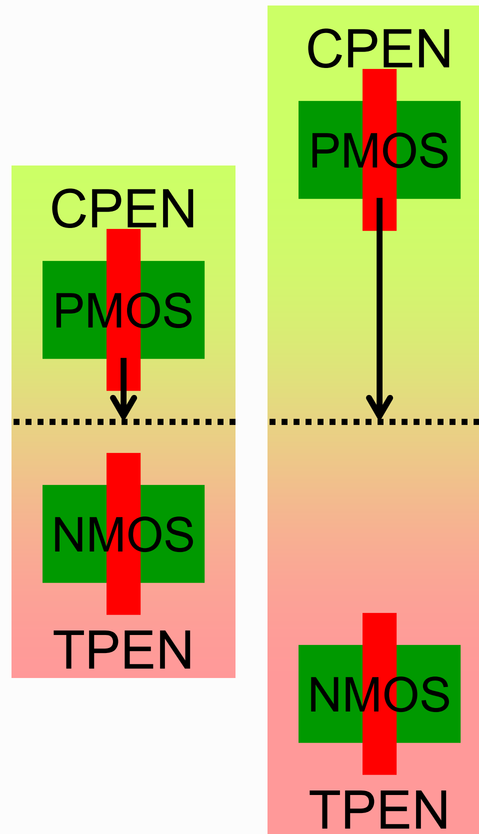
Faricelli, AMD [9]

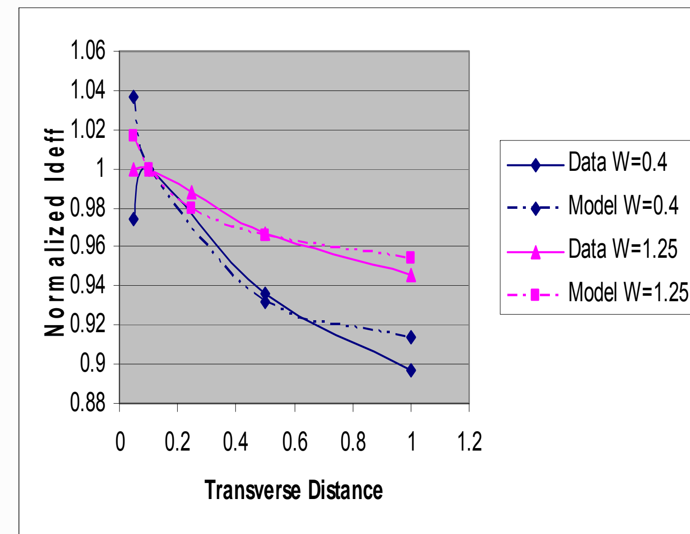
Transverse DSL Proximity

- Both NMOS & PMOS like tension in transverse direction, unlike longitudinal direction
- NMOS near PMOS in width direction → helps PMOS, hurts NMOS









Faricelli, AMD [9]

Outline

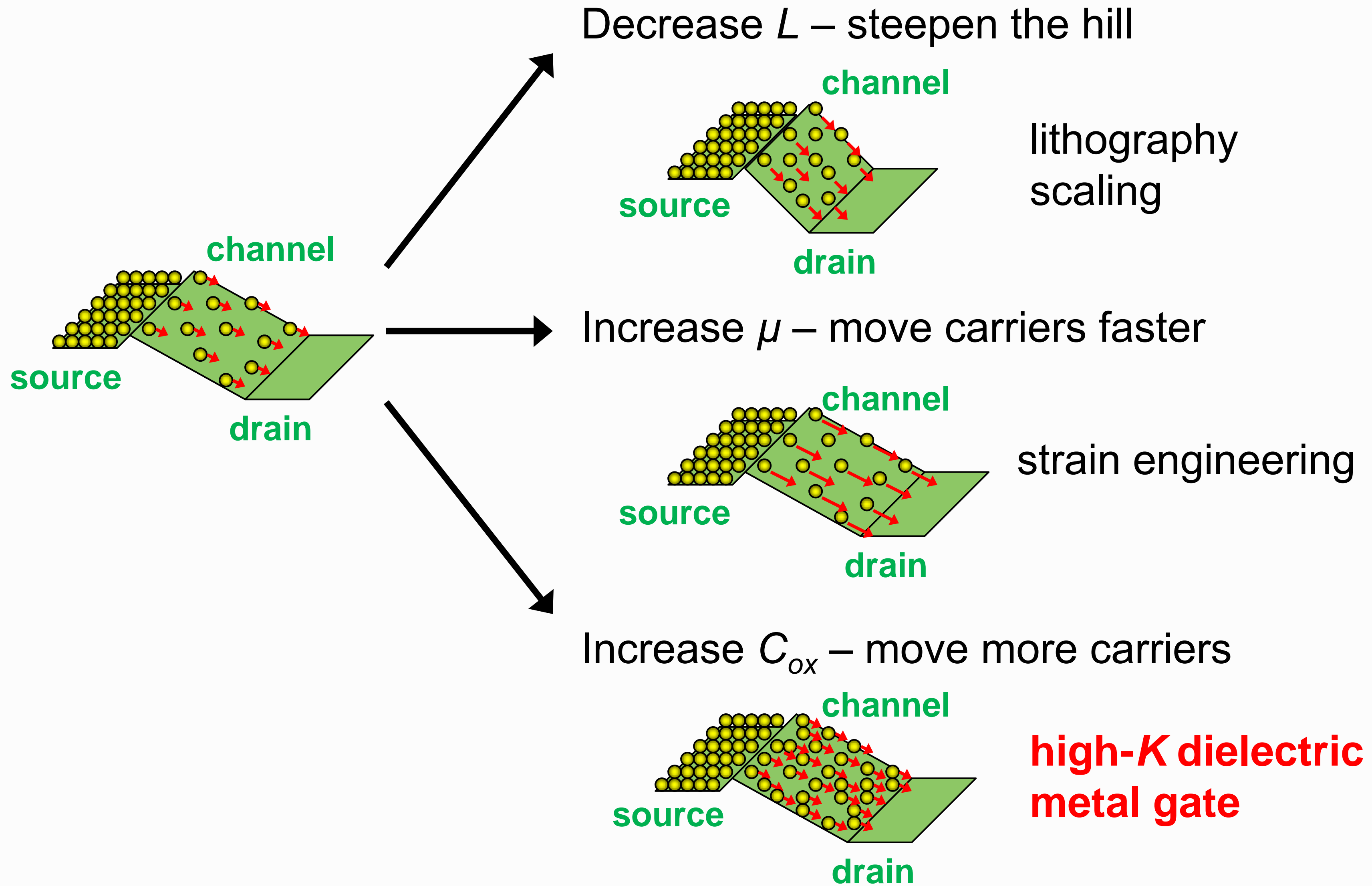
Part 1

- Motivation
- MOSFET & Short-Channel Fundamentals
- 130nm Fabrication
- More MOSFET Fundamentals
- Lithography
- Partially-Depleted SOI

Part 2

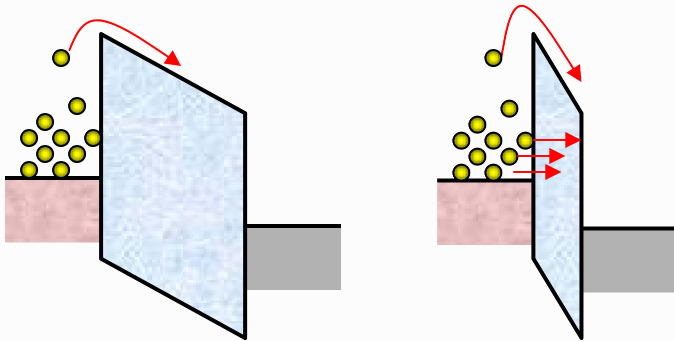
- Strain Engineering (90nm & Beyond)
- High-K / Metal-Gate (45nm & Beyond)
- Migrating to Fully-Depleted (22nm & Beyond)
- Tri-Gate FinFETs
- Conclusions

The Roads to Higher Performance



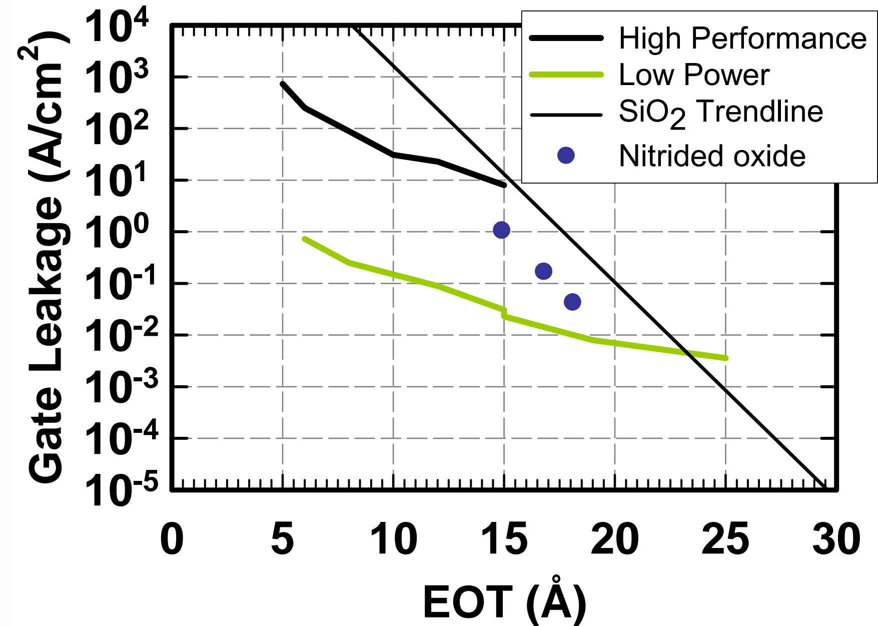
Direct Tunneling Gate Leakage

- t_{ox} had to scale with channel length to maintain gate control
  - Less SCE
  - Better FET performance
- Significant direct tunneling for t_{ox} < 2nm



- High-*K* gate dielectric achieves same C_{ox} with much thicker t_{ox}

McPherson, Texas Instruments [10]



EOT = Equivalent Oxide Thickness

$$C_{ox} = \frac{\varepsilon_{gate}}{t_{gate}} = \frac{\varepsilon_{ox}}{EOT}$$

Poly Depletion & Charge Centroid

Dielectric Only Half the Story

poly depletion (band bending)

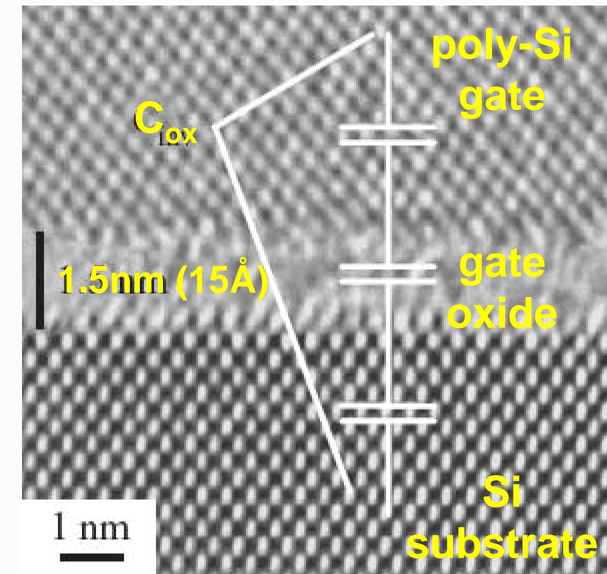
surface charge centroid few Å's away from oxide interface

gate charge centroid few Å's away from oxide interface

p-well

n^+ poly gate

gate oxide



Wong, IBM [11]

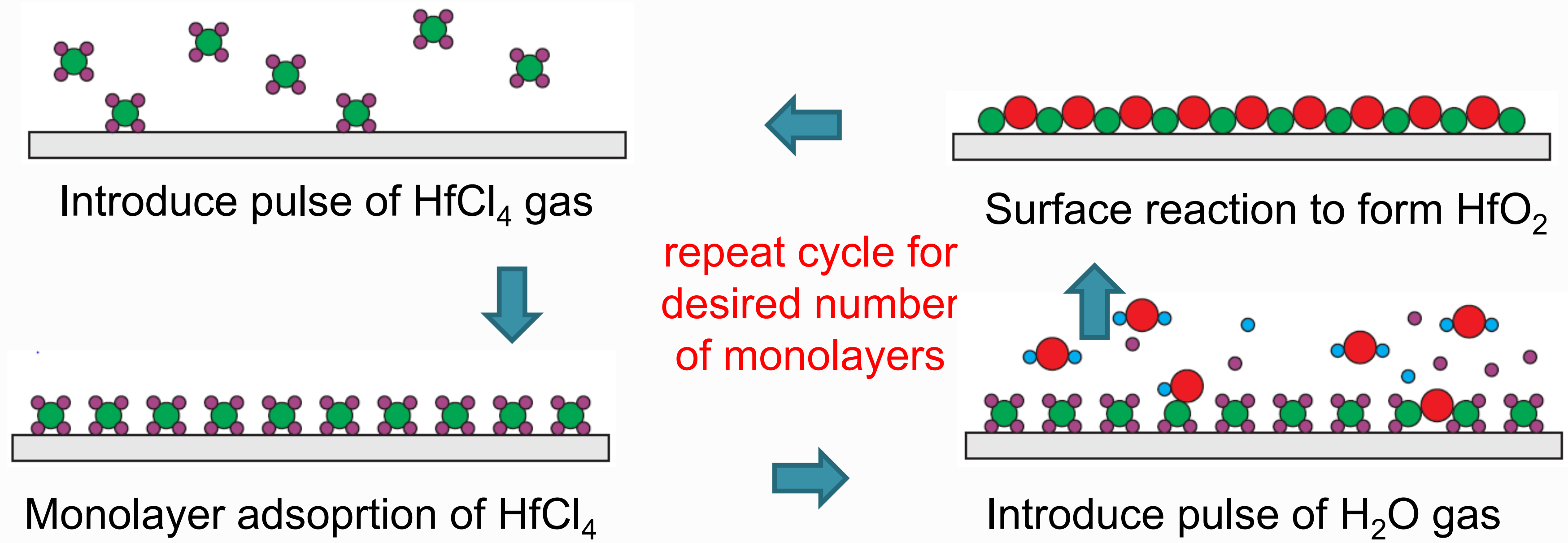
- Even heavily-doped poly is a limited conductor
- Discrepancy between electrical & physical thicknesses since charge is not intimately in contact with oxide interface

Enter High-*K* Dielectric + Metal-Gate

- High-*K* Dielectric (HK)
  - Hf-based material with K~20–30 (Zr-based also considered)
  - Need to overcome hysteretic polarization
  - High deposition temperature for good film quality
- Metal-Gate (MG)
  - Thin conductive film intimately in contact with high-*K* dielectric to set gate work function Φ_M → V_{FB} → V_T
  - Want band-edge Φ_M, i.e., NMOS @ E_C & PMOS @ E_V (just like n^+ poly & p^+ poly) → different MG for NMOS & PMOS
  - Typically complex stack of different metal layers →secret sauce
  - Conductive *fill metal* on top of Φ_M-setting metal-gate
- Key challenges
  - INTEGRATION, INTEGRATION, INTEGRATION
  - Φ_M shifts when exposed to dopant activation anneals
  - Getting the right V_T for both NMOS & PMOS

Atomic Layer Deposition

- Deposit monolayer at a time using sequential pulses of gases
- Introduce one reactant at a time & purge before introducing next reactant
- Key to precise film thickness control of HKMG stack
- e.g., SiO_2 ($SiCl_4$+H_2O) → HfO_2 ($HfCl_4$+H_2O) → TiN ($TiCl_4$+NH_3)



ICKnowledge.com [12]

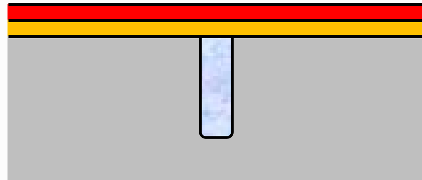
HK-First / MG-First Integration

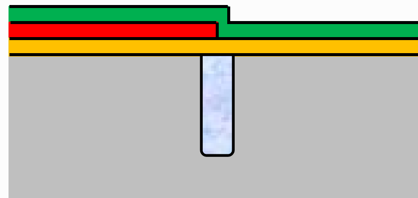
1
Deposit HK
Deposit MG1

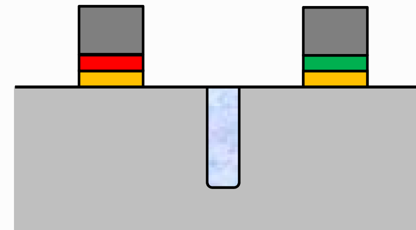
2
Pattern MG1
Deposit MG2

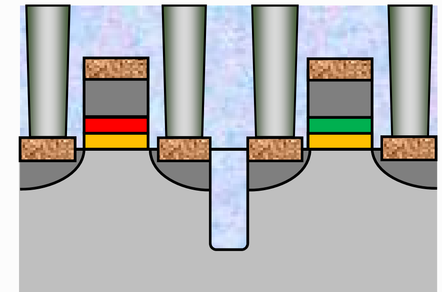
3
Pattern MG2
Deposit gate
Pattern gates / MGs / HK

4
Implant/anneal S/ D
Form silicide
Deposit/CMP ILD0
Form contacts



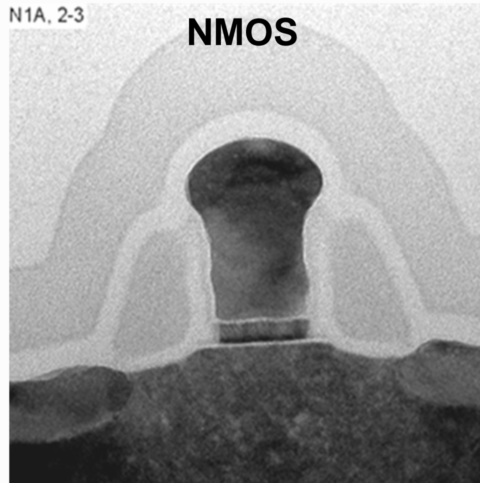


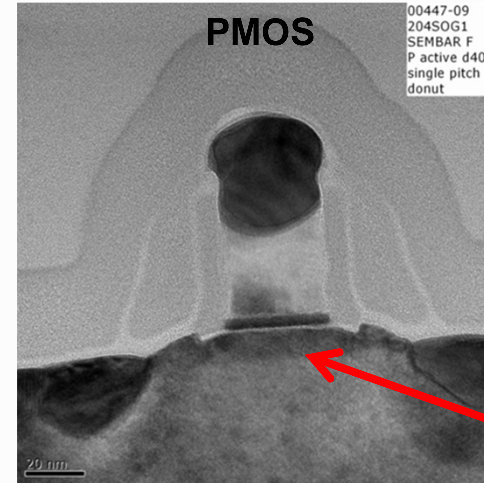




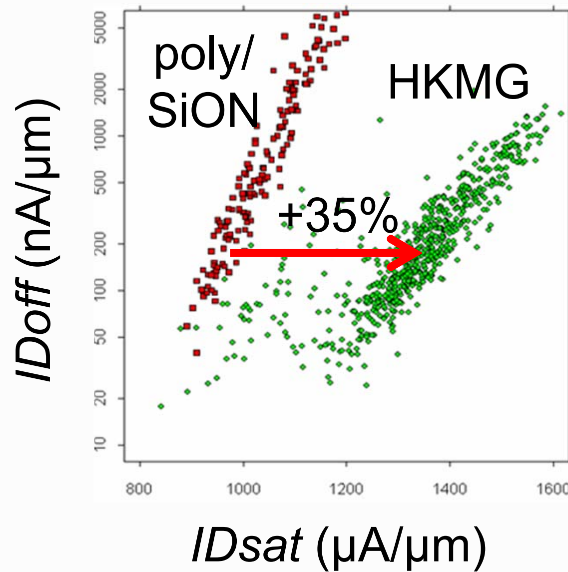
- Obvious extension of poly-Si gate integration
- Seems obvious & “easy” at first but plagued with unstable work function when HKMG is exposed to activation anneals
- Especially problematic with PMOS V_T coming out too high

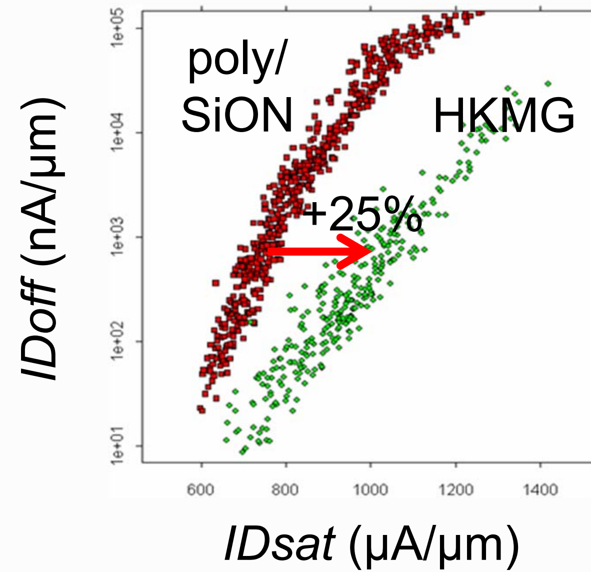
GlobalFoundries 32nm-SOI





epi-cSiGe to set channel Φ_M





Horstmann *et al.*, GlobalFoundries [13]

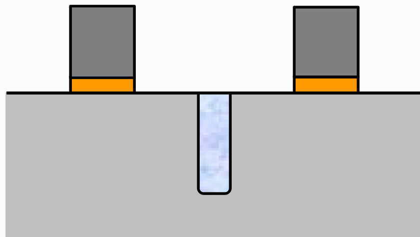
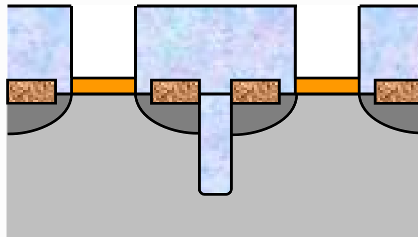
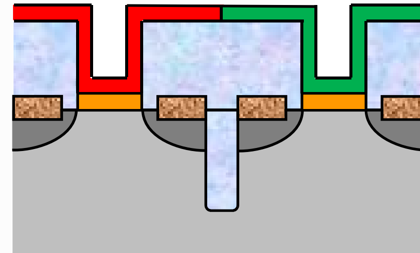
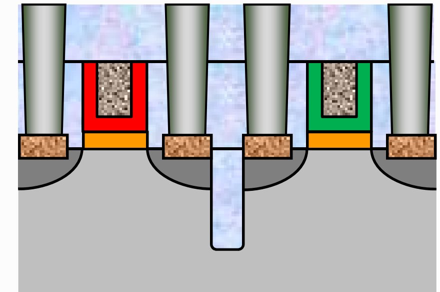
HK-First / MG-Last Integration

1
Deposit HK / gate
Pattern gate / HK

2
Implant/anneal S / D
Form silicide
Deposit ILD0
CMP ILD0 to expose top of gate
Remove gate

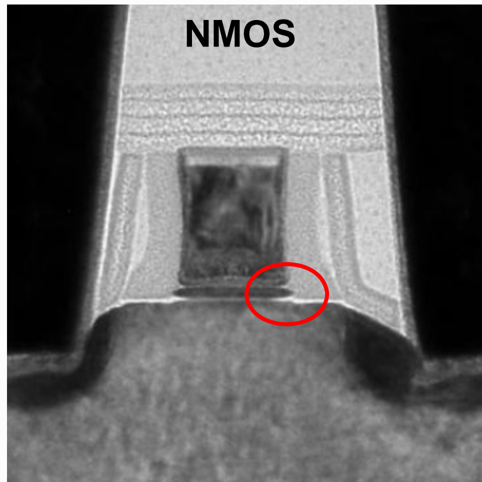
3
Deposit MG1
Pattern MG1
Deposit MG2
Pattern MG2

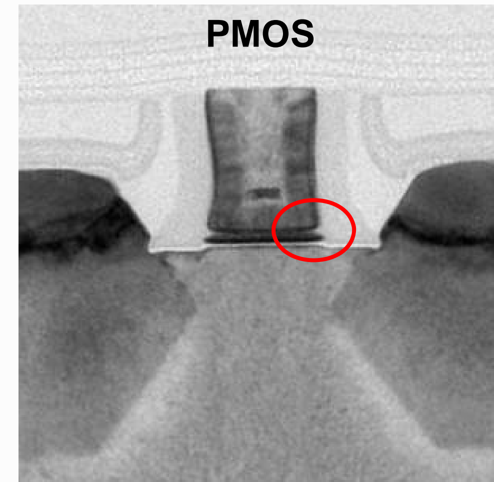
4
Deposit gate-fill
CMP gate-fill / MGs
Deposit more ILD0
Form contacts

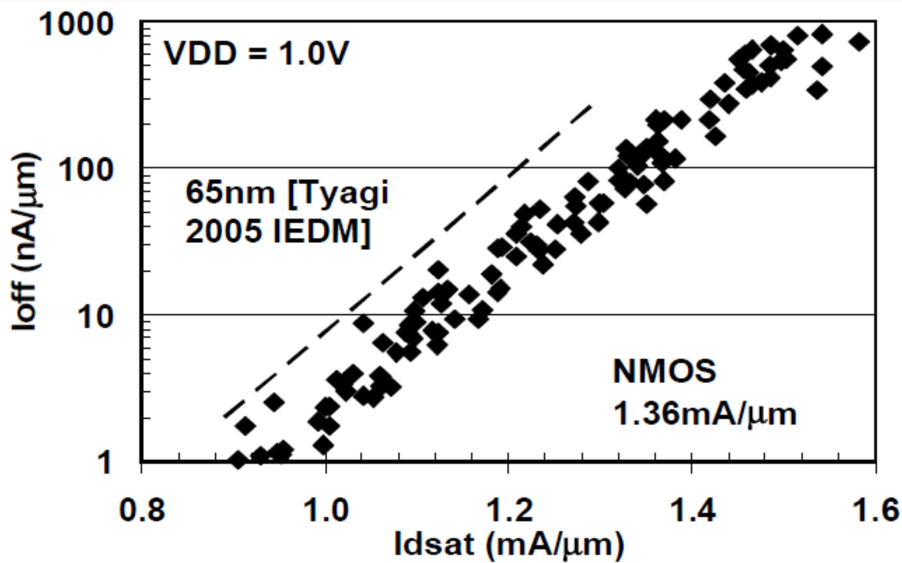





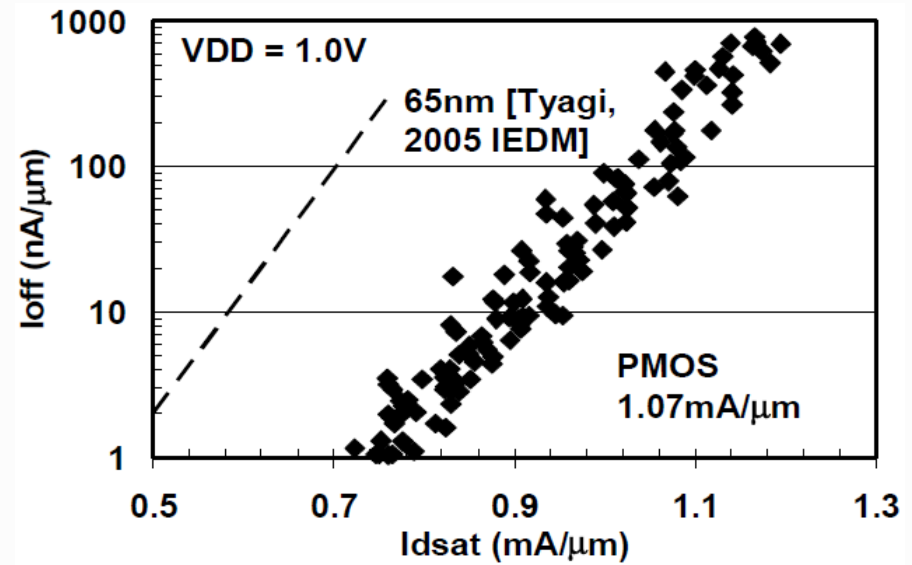
- High thermal budget available for middle-of-line
- Low thermal budget for metal gate → more gate metal choices
- Enhanced strain when sacrificial poly is removed & resulting trench is filled with gate fill metal

Intel 45nm









Auth *et al.*, Intel [14]

HK-Last / MG-Last Integration

1

Deposit oxide / gate
Pattern gate / oxide

2

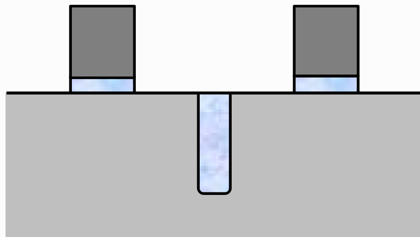
Implant/anneal S / D
Deposit ILD0
CMP ILD0 to expose top of gate
Remove gate/oxide

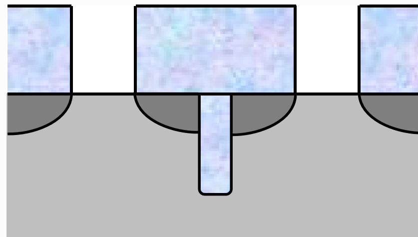
3

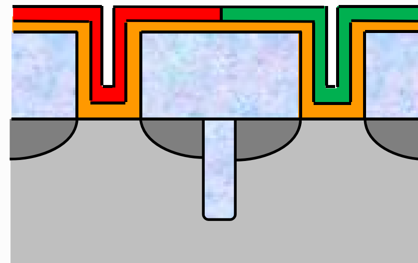
Deposit HK
Deposit MG1
Pattern MG1
Deposit MG2
Pattern MG2

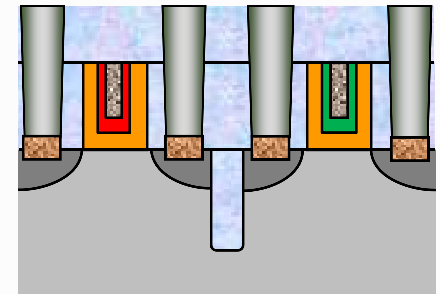
4

Deposit gate-fill
CMP gate-fill / MGs
Cut to expose active
Form silicide
Deposit / CMP ILD0
Pattern/form contacts



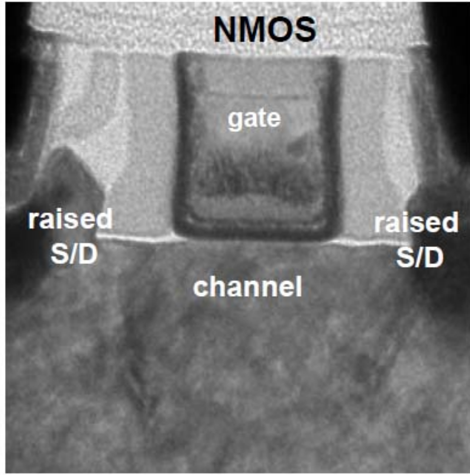


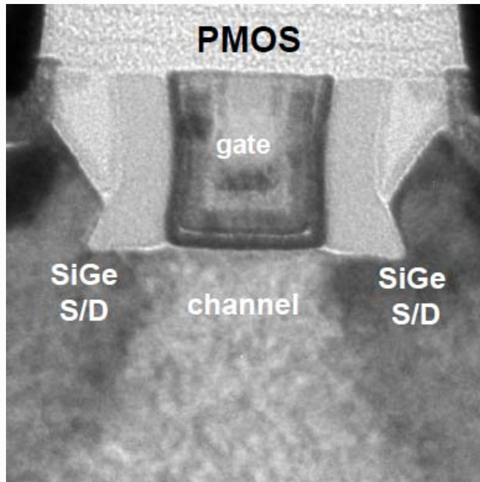


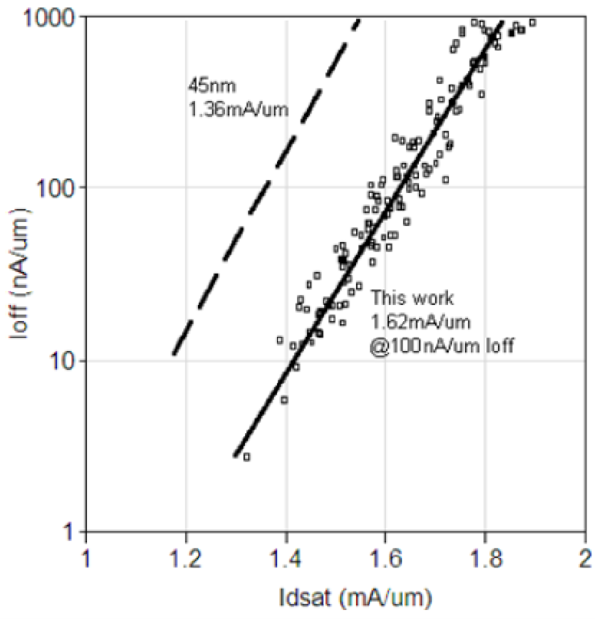


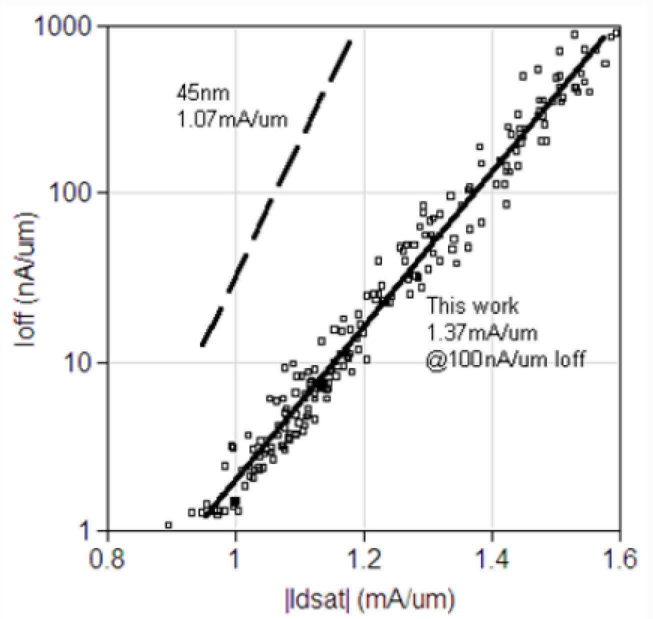
- Same advantages as HK-first / MG-last integration
- Overcomes EOT scaling limitations in HK-first / MG-last
- Need to postpone silicidation to after opening source/drain etch
- DSL relax & no longer useful since contacts cut through FET width

Intel 32nm









Packan *et al.*, Intel [15]

Outline

Part 1

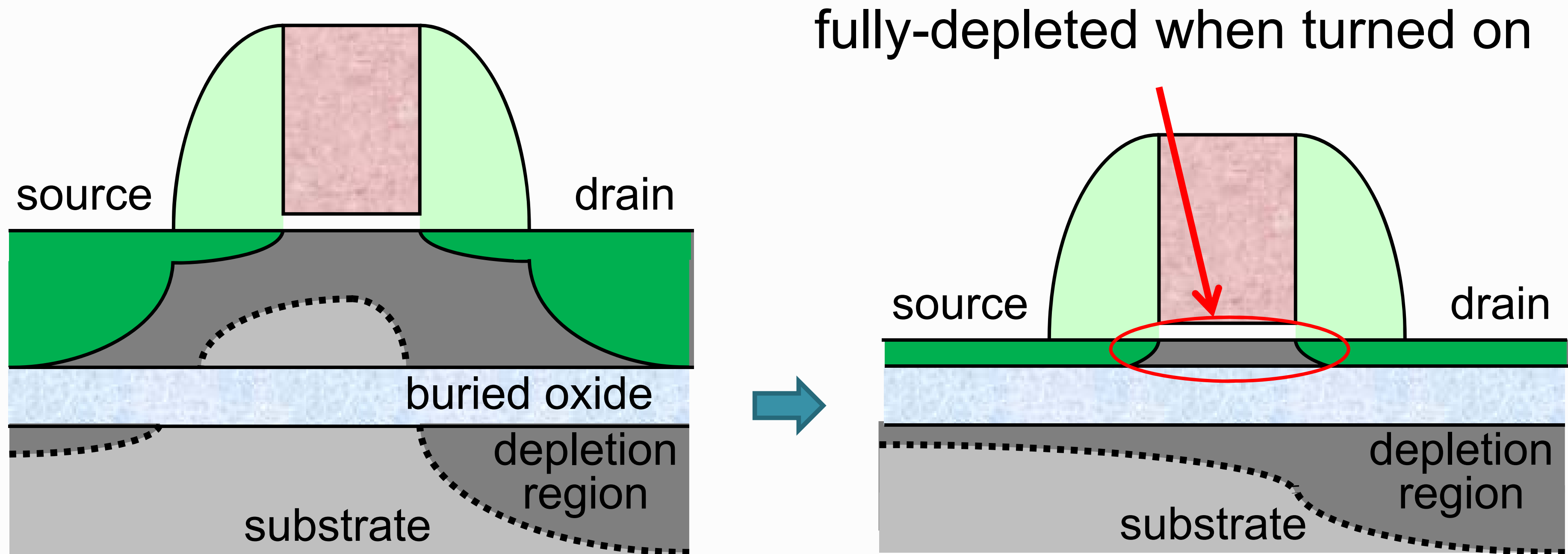
- Motivation
- MOSFET & Short-Channel Fundamentals
- 130nm Fabrication
- More MOSFET Fundamentals
- Lithography
- Partially-Depleted SOI

Part 2

- Strain Engineering (90nm & Beyond)
- High-K / Metal-Gate (45nm & Beyond)
- Migrating to Fully-Depleted (22nm & Beyond)
- Tri-Gate FinFETs
- Conclusions

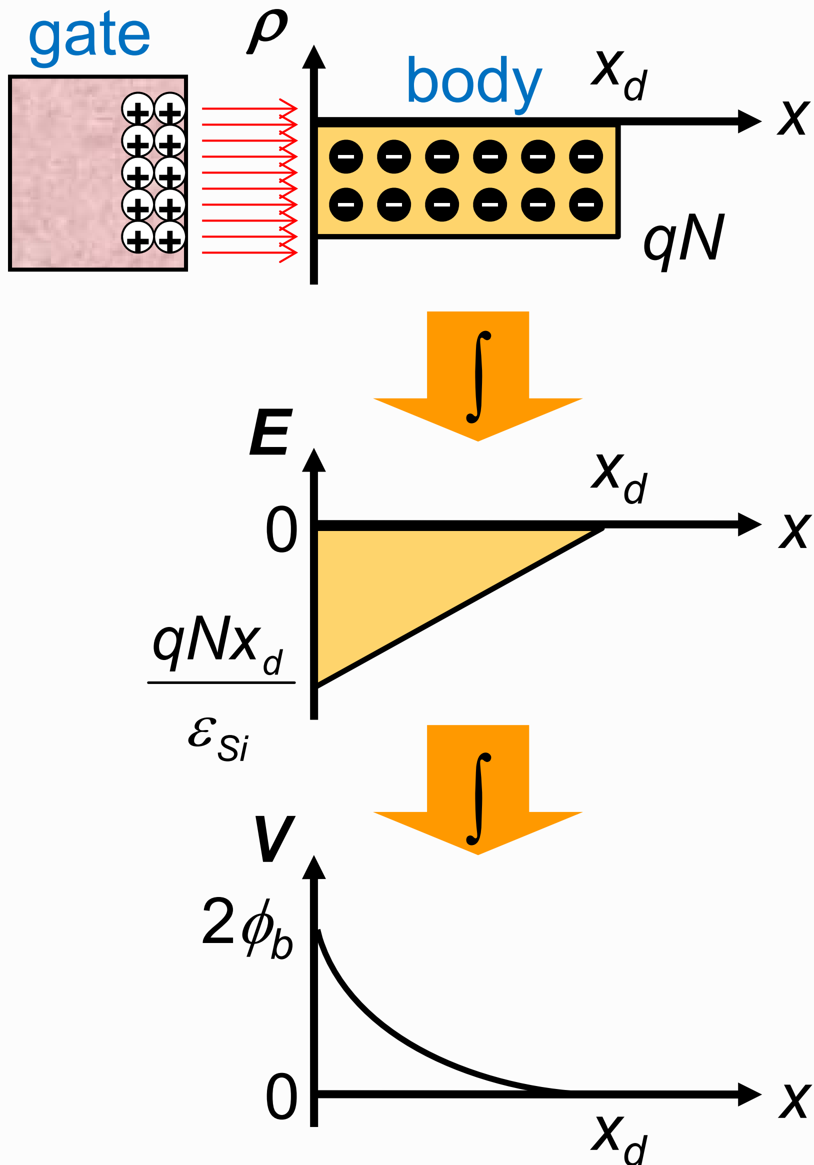
What Does Fully-Depleted Really Mean?

- Consider what happens when SOI layer thins down



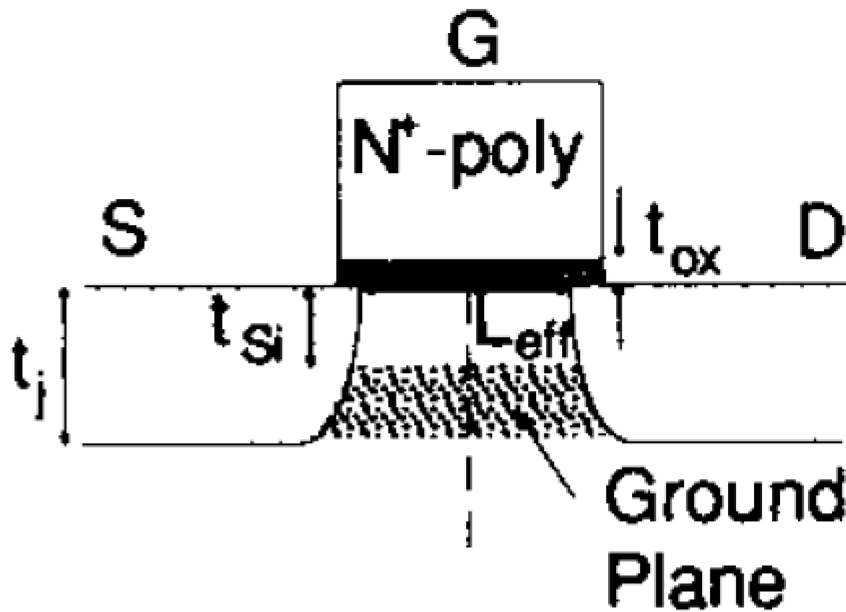
- Conservation of charge cannot be violated
- So once body is fully depleted, extra gate charge must be balanced by charge elsewhere, e.g., beneath buried oxide
- If substrate is insulator, then charge must come from source/drain
- No floating body in fully-depleted → no hysteresis

Requirement for Field-Effect Action

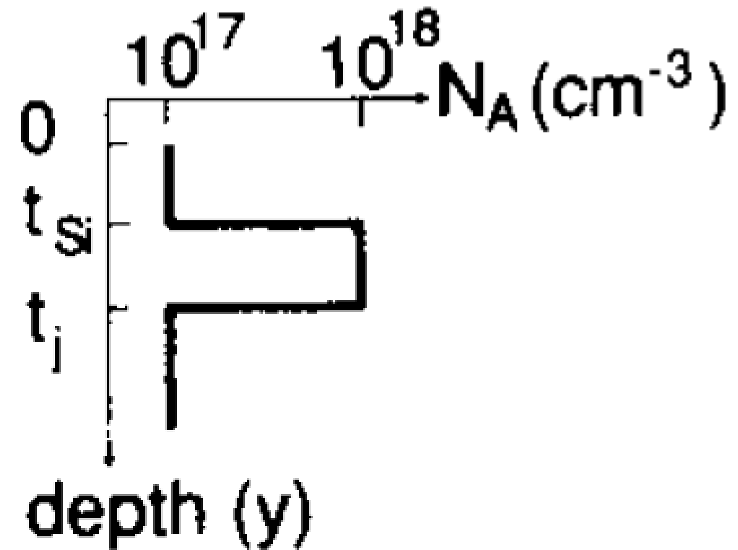


- V_{GS} modulates surface charge density under gate dielectric
  - → Modulate I_{DS} when $V_{DS} \neq 0$
  - → Need band bending at surface
  - → Need electric field for band bending
  - → Need + & – charge separation between gate & body beneath surface
- Do we really need dopants in the body to create field effect?

Ground-Plane MOSFET

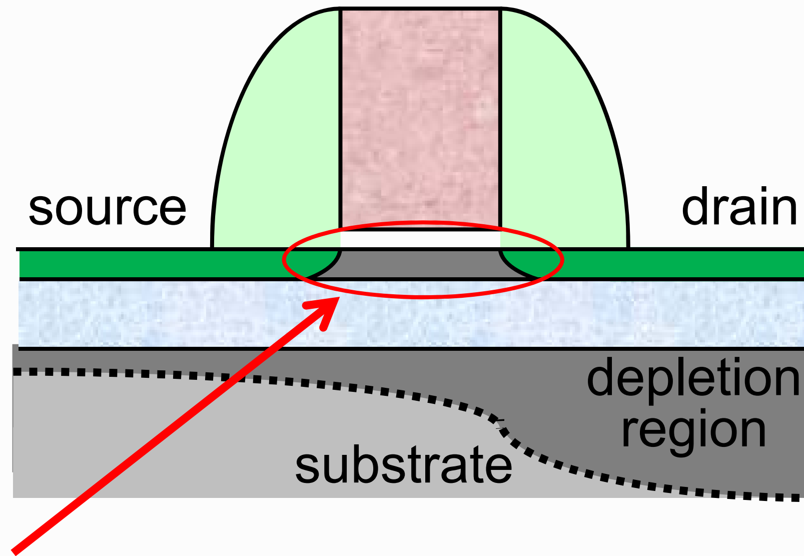


Yan *et al*., Bell Labs [16]



- Extremely retrograded well profile with no surface dopants
- Depletion region cannot extend beyond buried pulse of dopants
- All you fundamentally need for field-effect action is a parallel-plate capacitor with gate dielectric & *undoped* semiconductor in between plates → dopants are not required in the body

Why Fully-Depleted Suppresses SCE



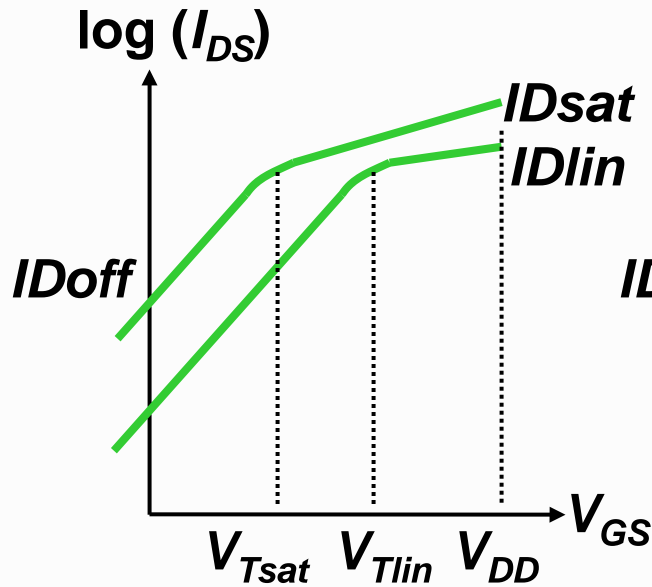
- Basic idea: effectively no charge in body
  - → Body cannot terminate field lines from source & drain
  - → Field lines from source & drain forced to move down to substrate
  - → Source to body surface barrier not impacted by shorter gate length
- Substrate must be close to source & drain to prevent field lines from drain to terminate to source
- Side benefit: no dopants → less scattering → higher μ

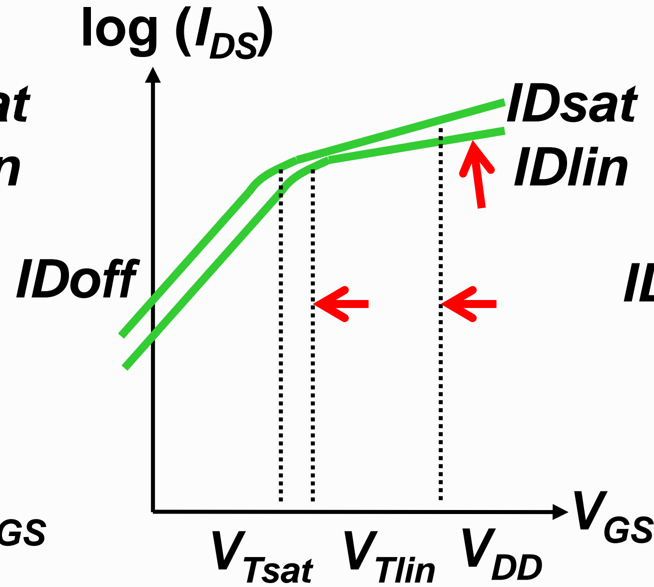
Benefits of Lower *DIBL* & *S*

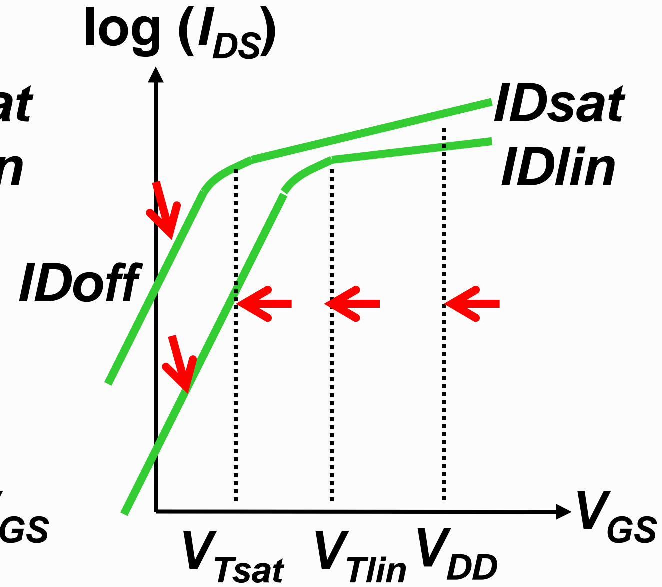
Maintain
IDsat & *IDoff*

Same *S*
Lower *DIBL*

Lower *S*
Same DIBL







- Fully-depleted options
  - Planar: FD-SOI, Bulk with retrograded well
  - 3-D: FinFET or Tri-Gate – SOI or Bulk

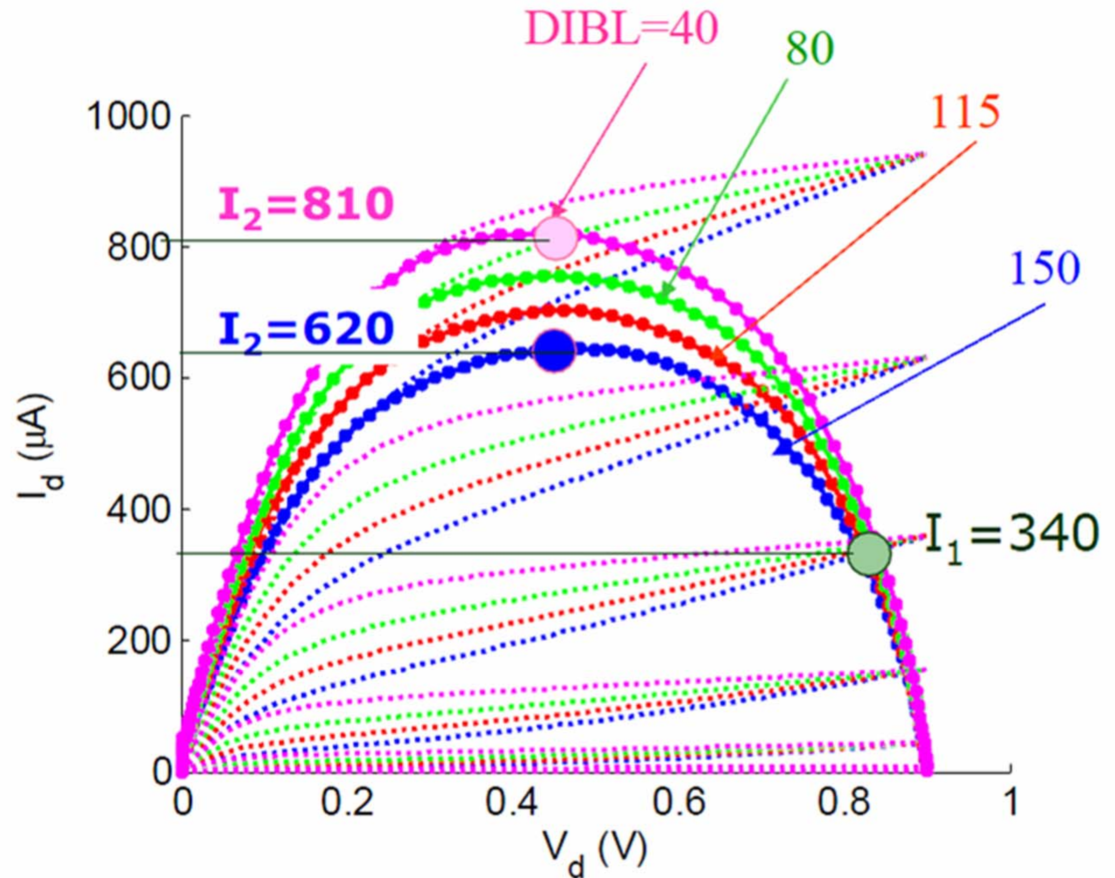
The Big Deal with Lower DIBL

$I_{eff} \approx (340+810)/2=575$

$I_{eff} \approx (340+620)/2=480$

$\Delta f/f=\Delta Ieff/Ieff=95/480 =20\%$

Lower DIBL
=
Higher Performance

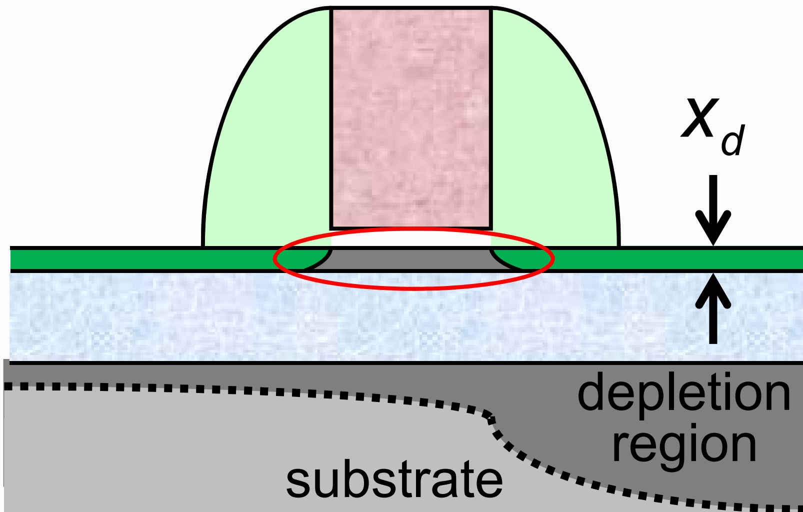


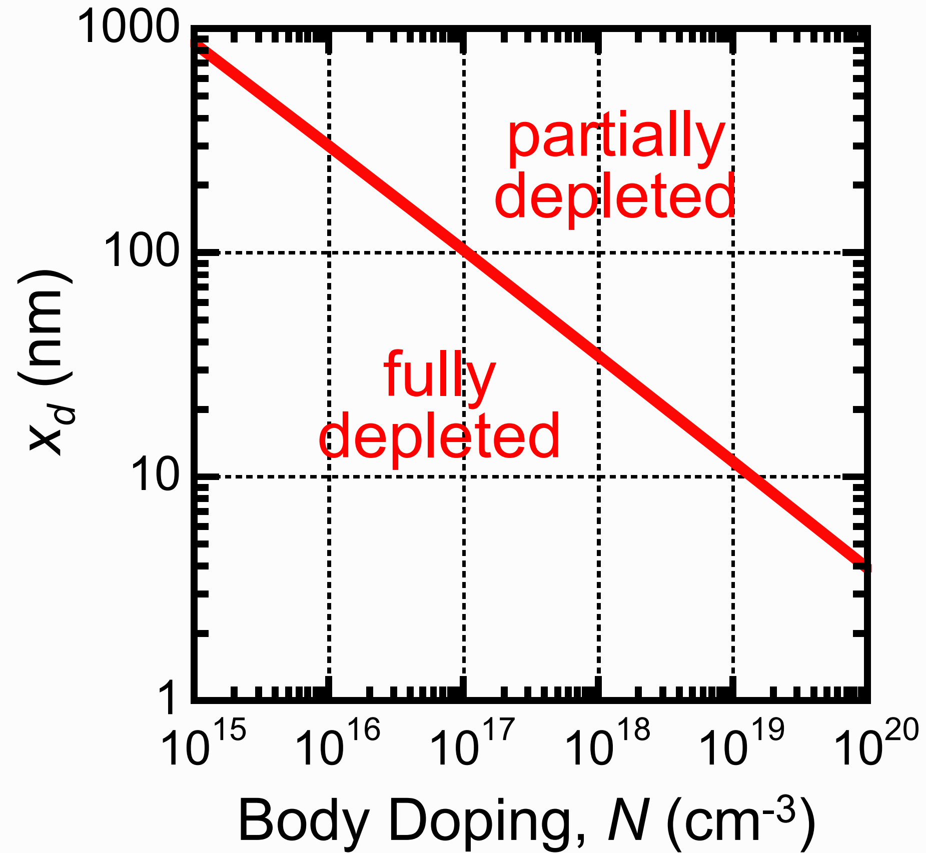
Higher performance for the same *IDsat* & *IDoff*

L. Wei *et al.*, Stanford [17]

Body Thickness for Fully-Depleted

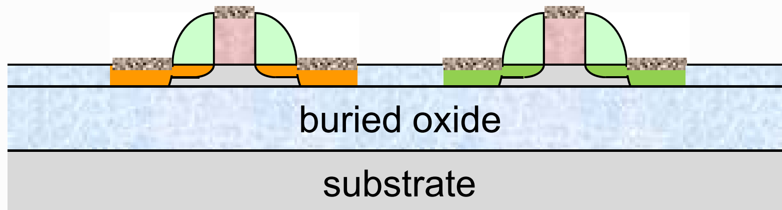
$$x_d < \sqrt{\frac{2\varepsilon_{Si} \cdot 2\phi_b}{qN}}$$

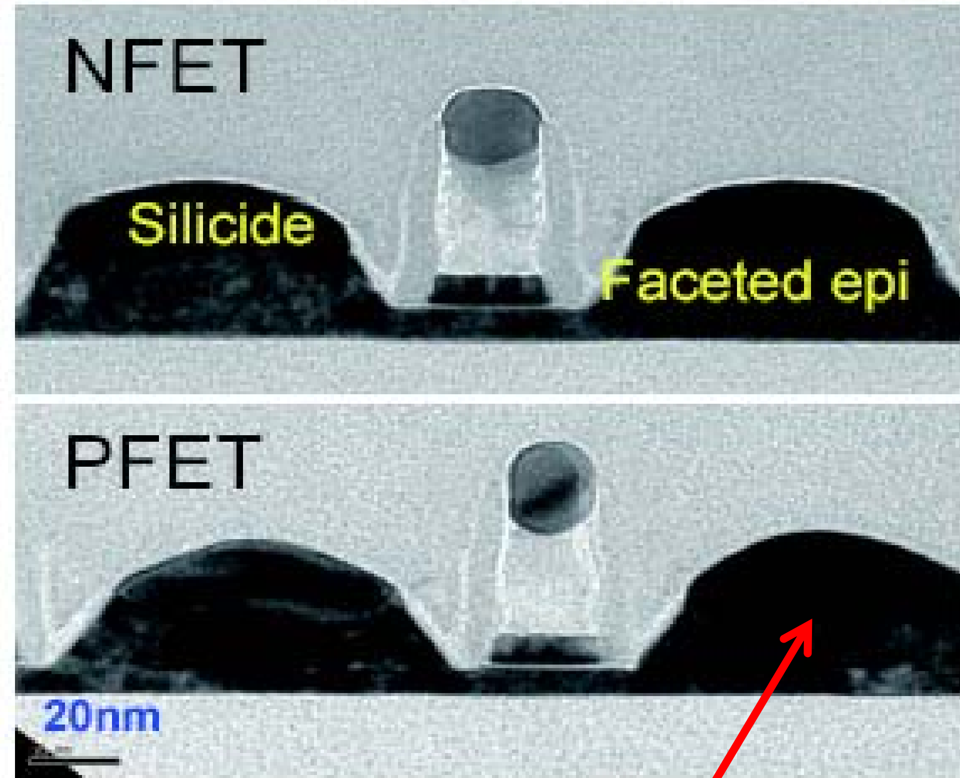




Fully-Depleted Planar on SOI

- a.k.a. ET (Extremely Thin) or UTBB (Ultra-Thin Body & BOX) SOI to refer to very thin SOI and Buried Oxide (BOX) layers
- SOI Si layer is so thin that charge mirroring gate charge comes from beneath BOX

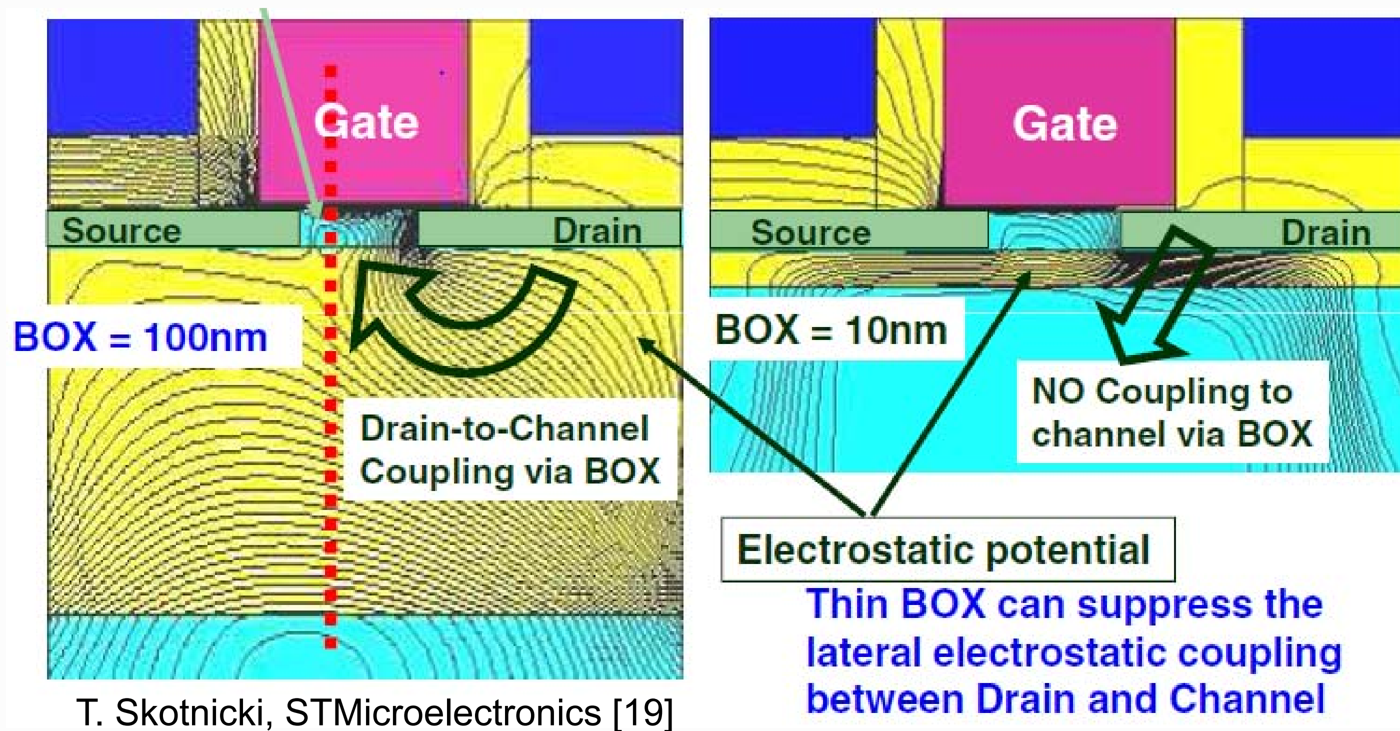




K. Cheng *et al.*, IBM [18]

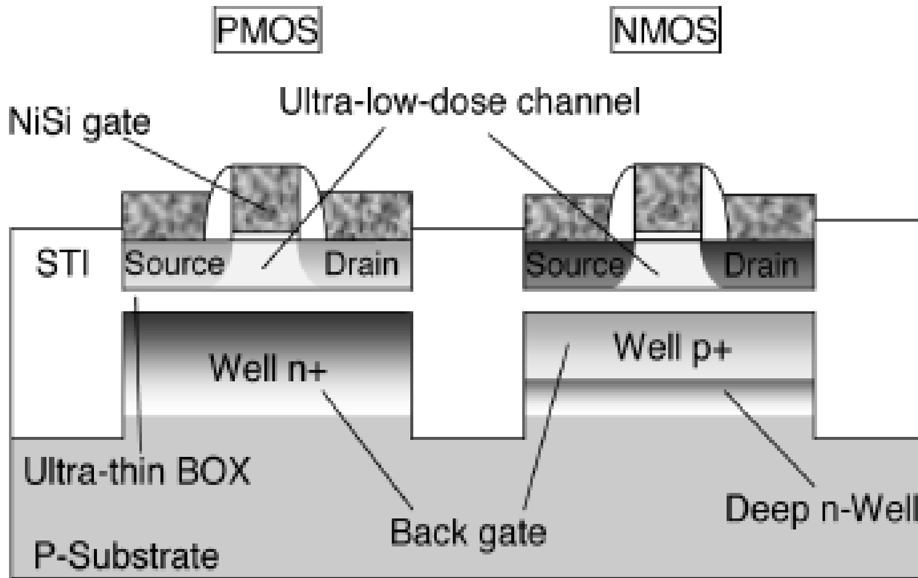
thick to reduce series resistance & apply stress

Thin BOX to Suppress SCE

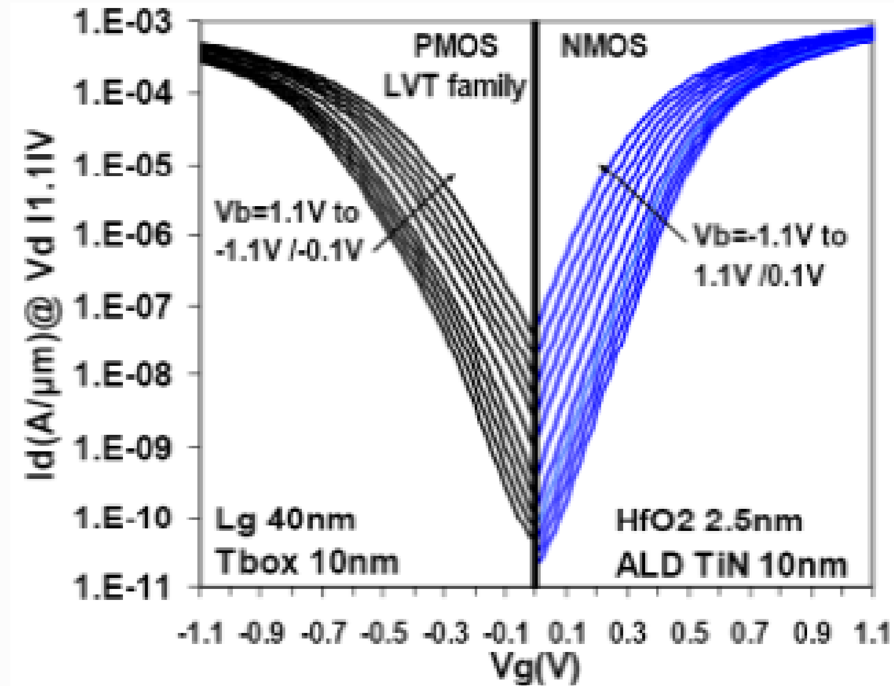


- If body is fully depleted, field lines from drain cannot terminate in the body since there's no charge to terminate to → no DIBL
- Charge elsewhere must be nearby or field lines from drain will terminate on source charge ☹
- However, lateral field always present when $V_{DS} \neq 0$

Performance Tuning with Backgate Bias



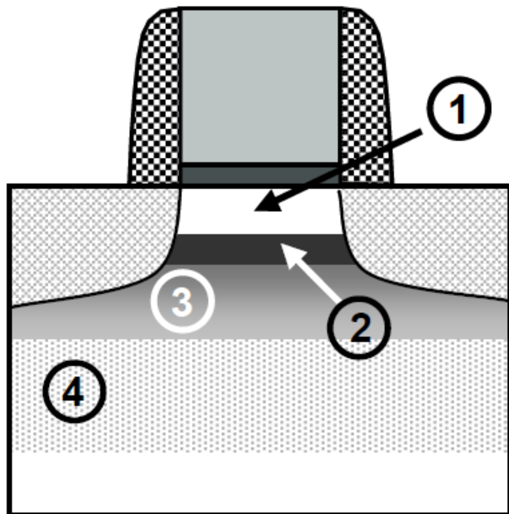
Yamaoka *et al*., Hitachi [20]

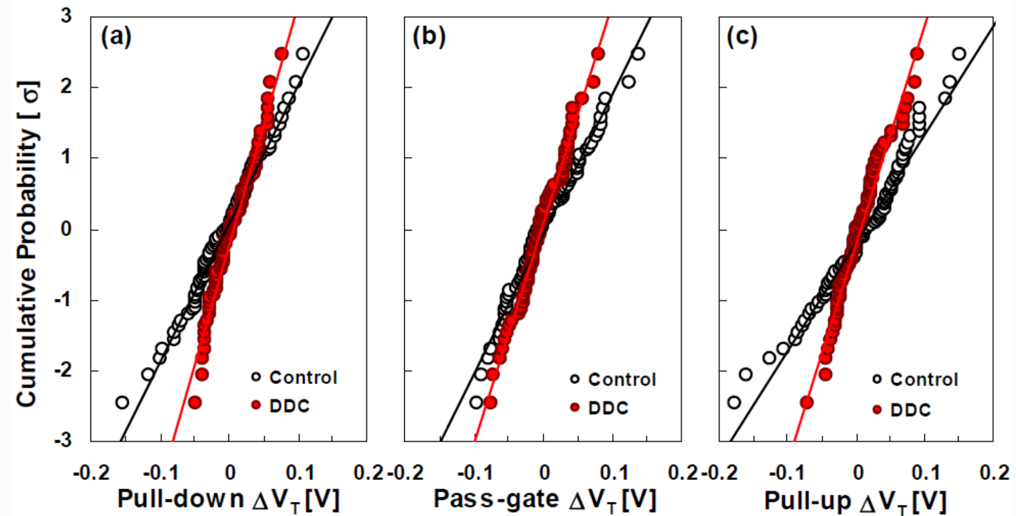


T. Skotnicki, STMicroelectronics [21]

- Like a “body effect” in planar bulk with C_{Si} spanning SOI & BOX
- Backgate bias can modulate both NMOS and PMOS V_T at 80mV/V
- Not option in finFETs but finFET subthreshold slope is better

Fully-Depleted Planar on Bulk





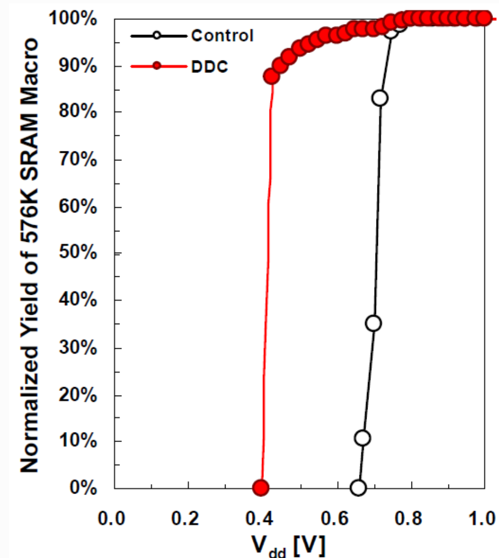
1 Low-doped layer for RDF reduction (fully depleted)

2 V_T setting layer for multiple V_T devices

3 Highly-doped screening layer to terminate depletion

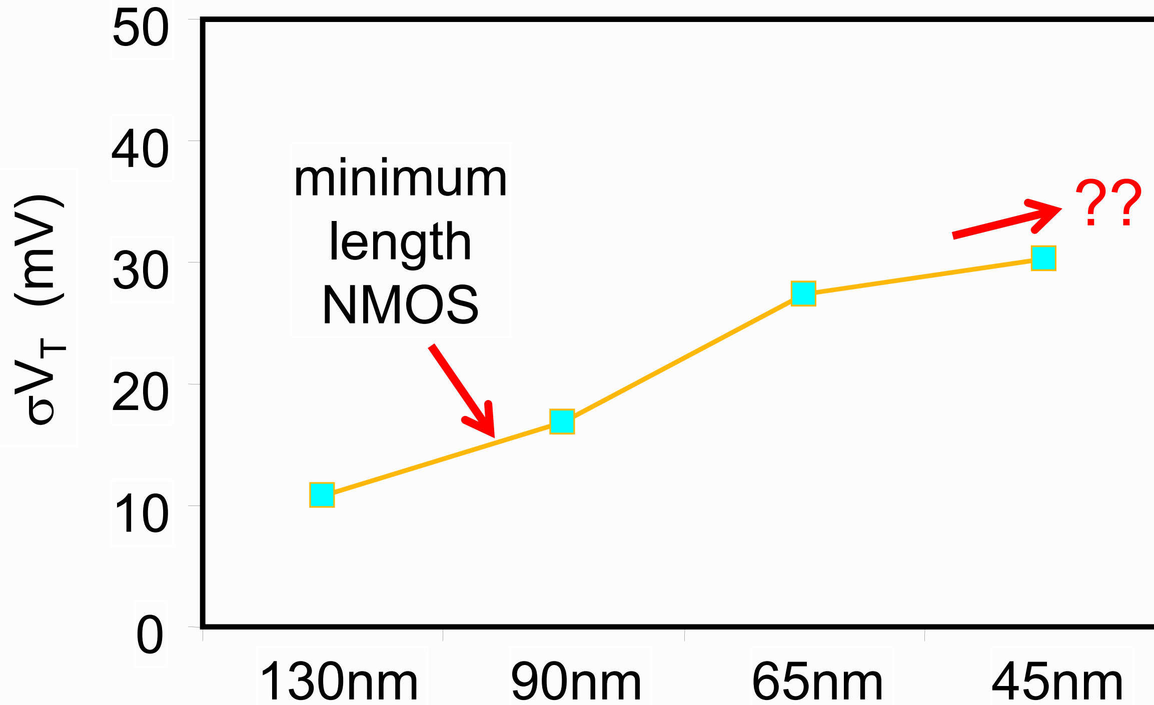
4 Sub-surface punchthrough prevention

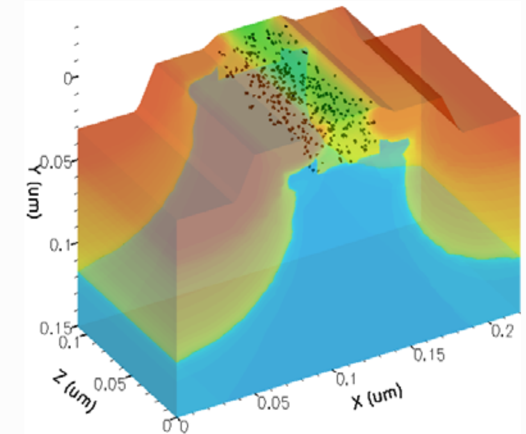
Reduced RDF for tighter V_T control & lower SRAM V_{DDmin}



Fujita *et al*., Fujitsu & SuVolta [22]

Random Dopant Fluctuation (RDF)





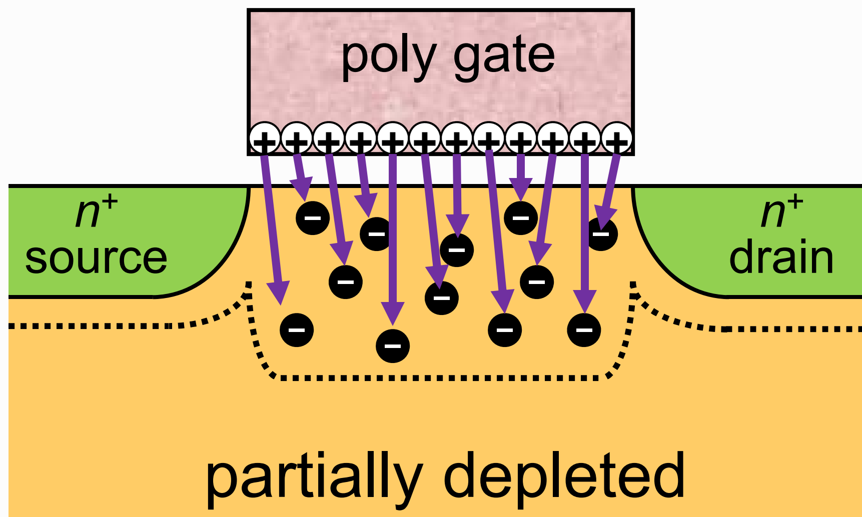
- RDF more prevalent with scaling since number of dopants is decreasing with each MOS generation
- Why does RDF impact magically disappear in fully-depleted?

Auth, Intel [14]

RDF in Conventional MOS

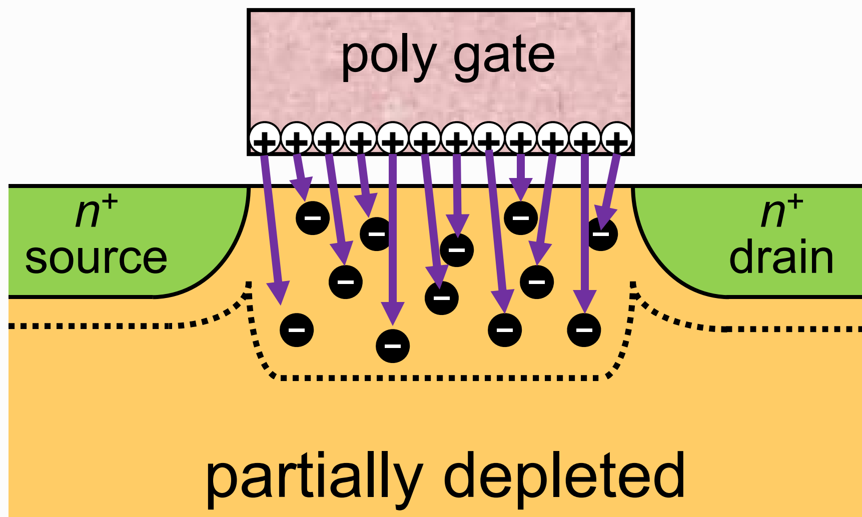
$$V = -\int E \cdot dx$$

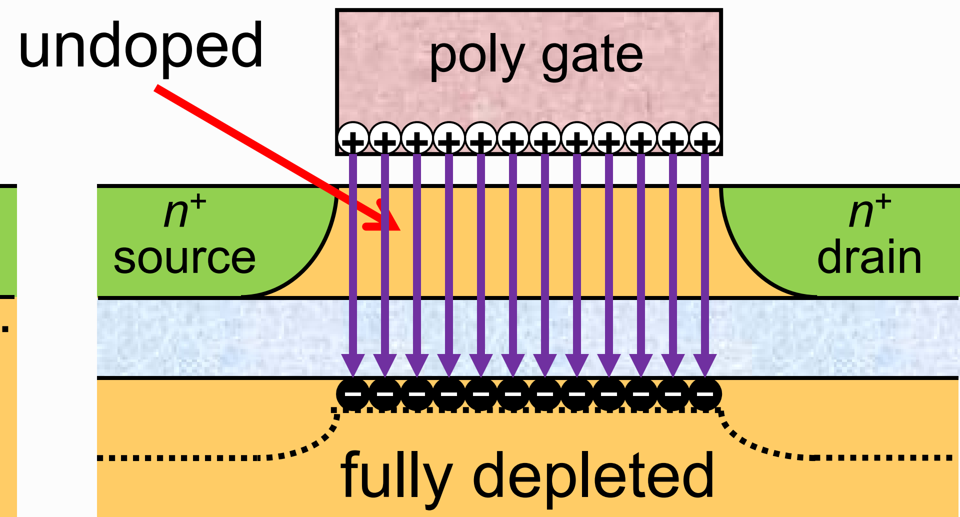
- Back to basics
  - Conservation of charge
  - Electric field lines start at +ve charge & end at –ve charge
- Number of dopant atoms vary from FET to FET
- BUT dopant atoms also vary in location
  - *Lengths* of field lines exhibit variation
  - Integrated field (voltage or band bending) has V_T variation



Why Fully-Depleted Eliminates RDF

- In fully-depleted SOI, field lines from gate cannot terminate in the undoped body (no charge there)
  - Mirror charges are localized beneath BOX
  - Lengths of field lines have tight distribution → small V_T variation
  - However, V_T now very sensitive to dimensional variation, e.g., SOI and BOX thickness
  - Other sources of variation also present, e.g., MG grains





Outline

Part 1

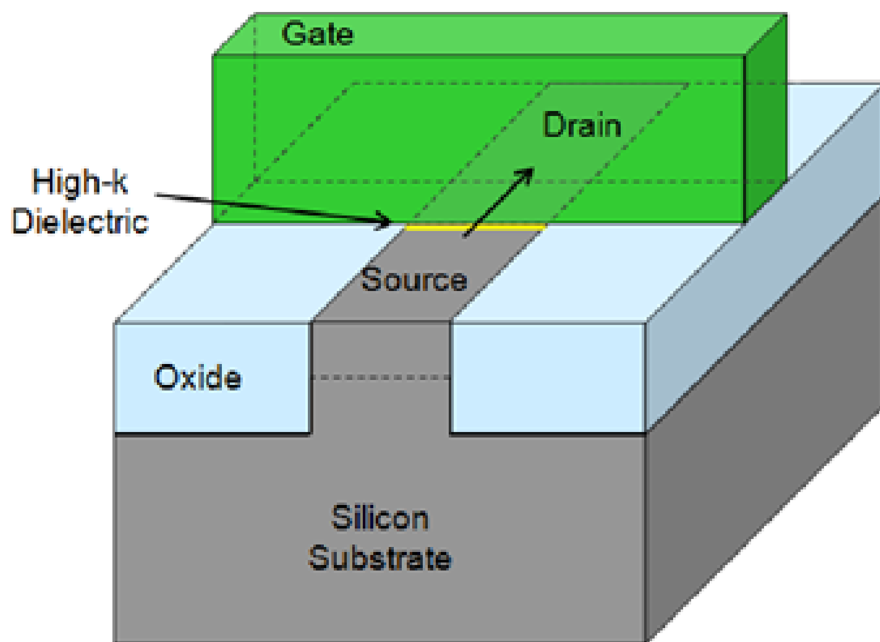
- Motivation
- MOSFET & Short-Channel Fundamentals
- 130nm Fabrication
- More MOSFET Fundamentals
- Lithography
- Partially-Depleted SOI

Part 2

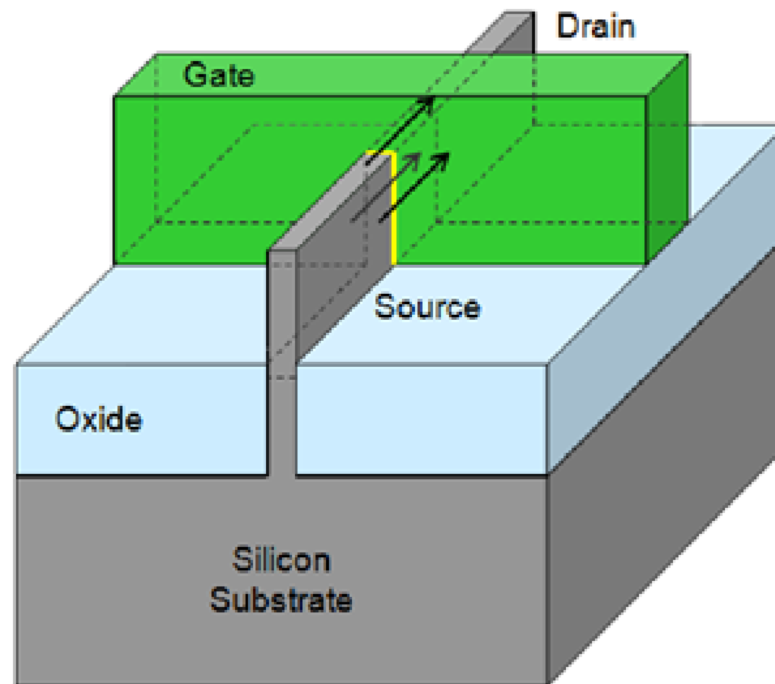
- Strain Engineering (90nm & Beyond)
- High-K / Metal-Gate (45nm & Beyond)
- Migrating to Fully-Depleted (22nm & Beyond)
- Tri-Gate FinFETs
- Conclusions

What is Fully-Depleted Tri-Gate?

32nm planar



22nm tri-gate

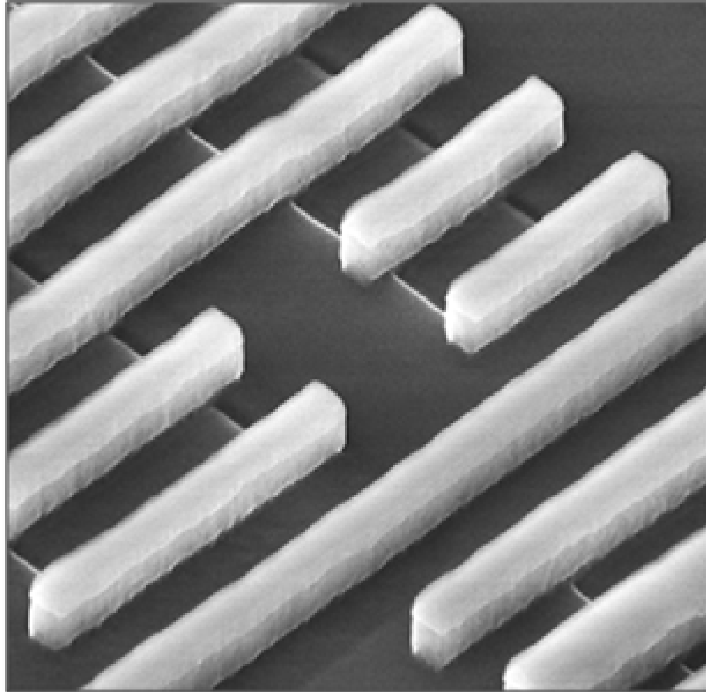


- Channel on 3 sides
- Fin width is *quantized* (SRAM & logic implications)
- Fin so narrow that gate mirror charge must come from fin base

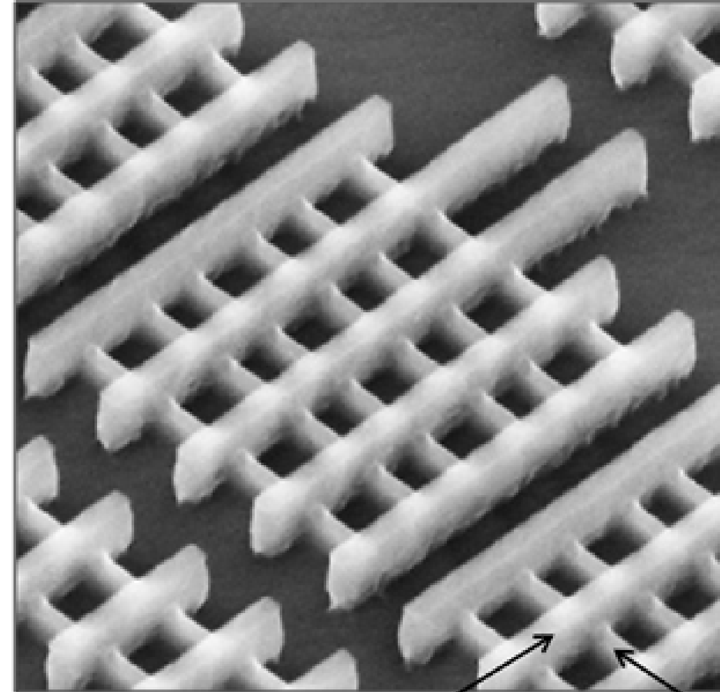
Hu, UC Berkeley [23]
M. Bohr, Intel [24]

Tri-Gate FinFETs in Production

32nm planar



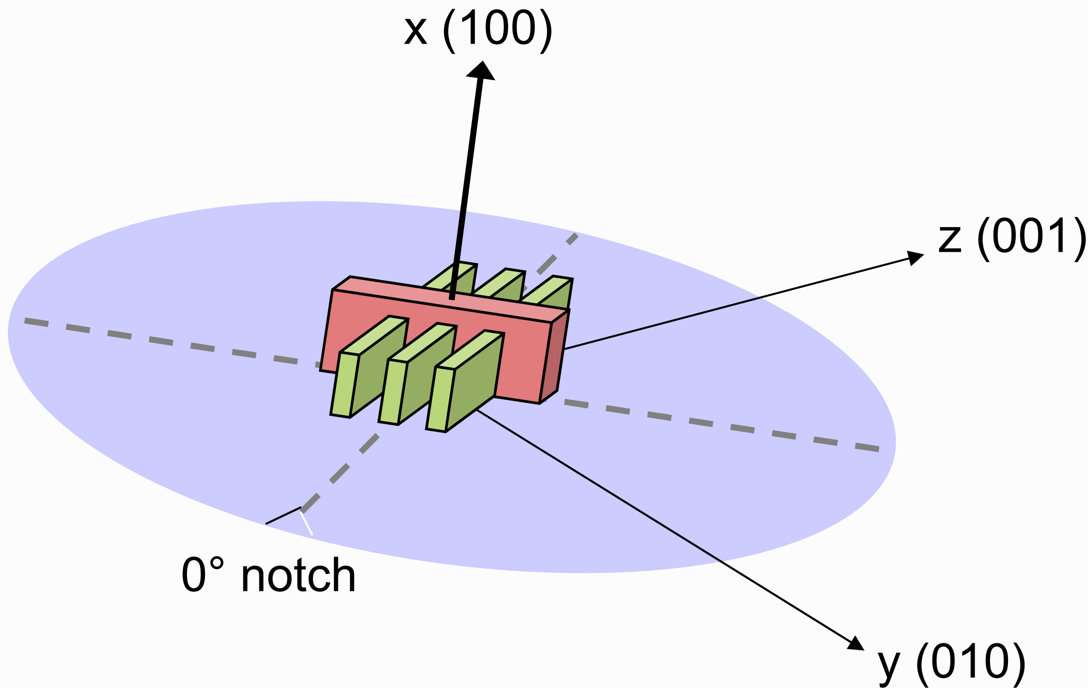
22nm tri-gate

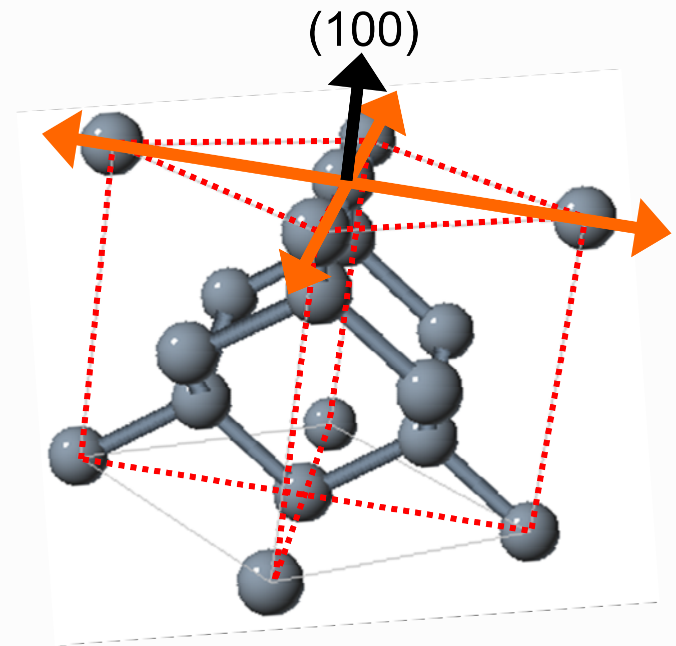


Truly impressive!!!

M. Bohr, Intel [24]

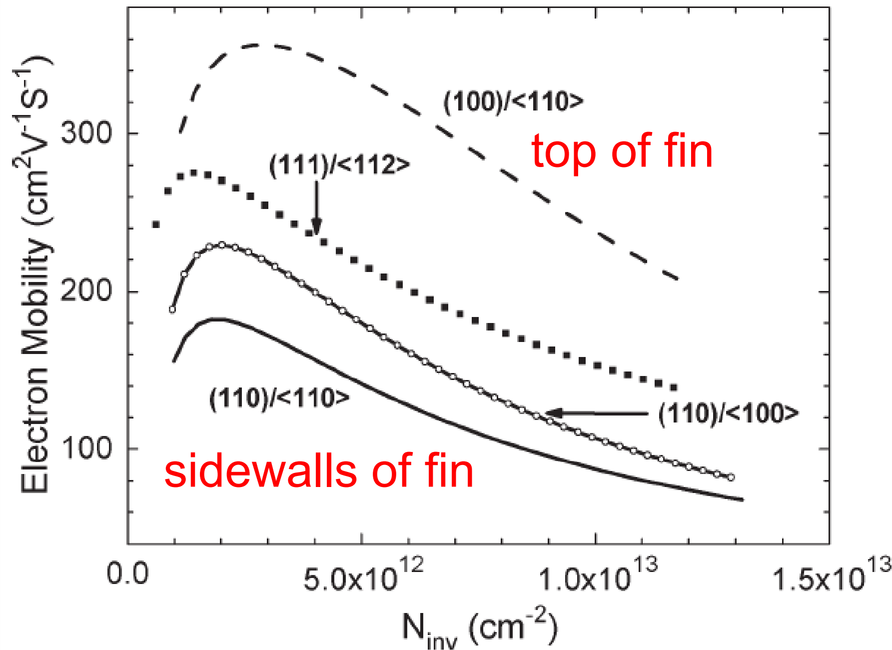
Conventional Wafer Surface Orientation & Channel Direction

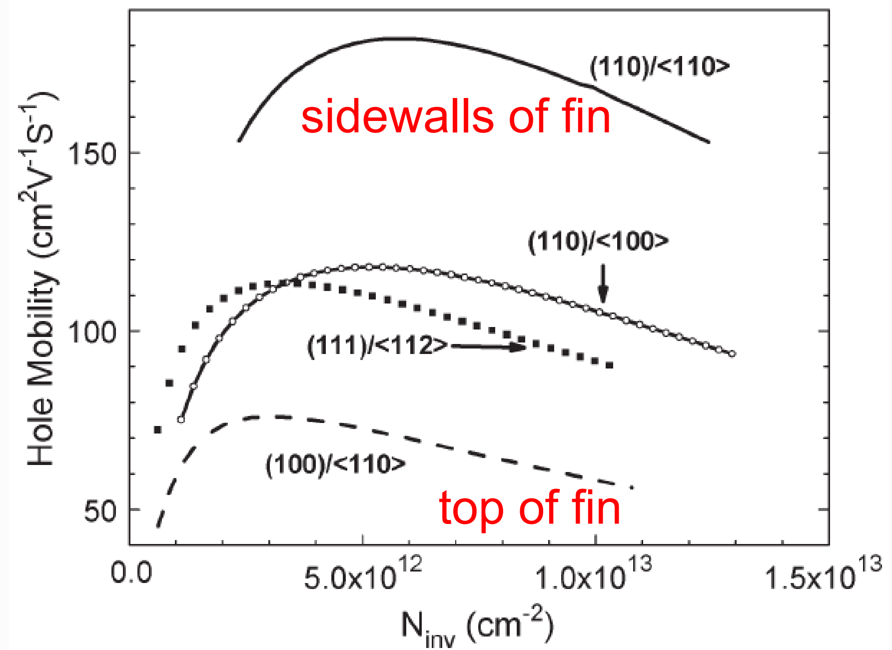




- Wafer normal is (100), current flows in <110> direction
- Tri-Gate FinFET: top surface (100), sidewall surfaces (110)

Mobility Dependence on Surface Orientation & Direction of Current



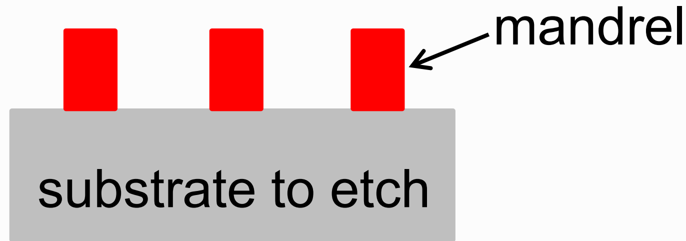


Yang *et al.*, IBM [25]

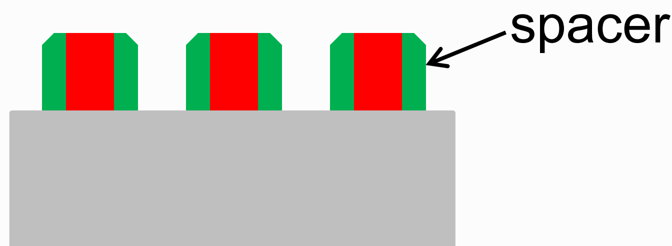
- Strain-induced mobility boost also depends on surface orientation & channel direction – not as strong for current along sidewalls vs. top of fin

Fin Patterning – Sidewall Image Transfer

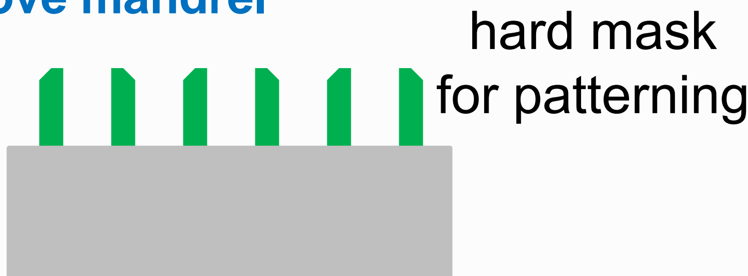
1 **Deposit & pattern sacrificial mandrel**



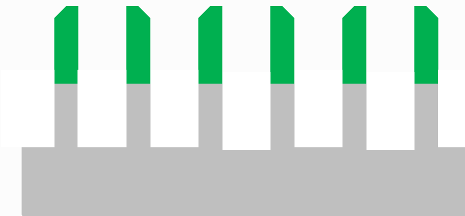
2 **Deposit & etch spacer**



3 **Remove mandrel**



4 **Etch target material using spacer as hard mask**



5 **Remove spacer mask**

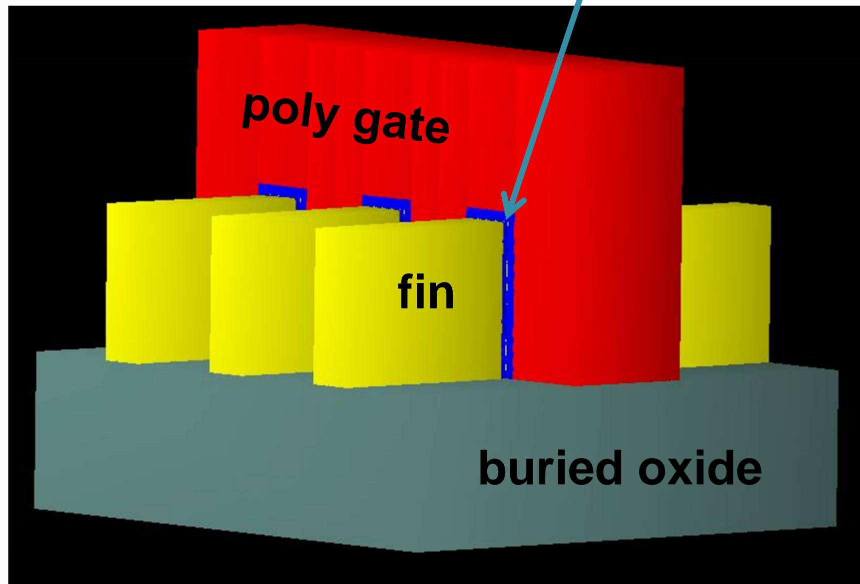
- Standard approach for patterning fins down to 60nm pitch (Intel 22nm)
- In principle, pitch can go down to ~40nm without double patterning

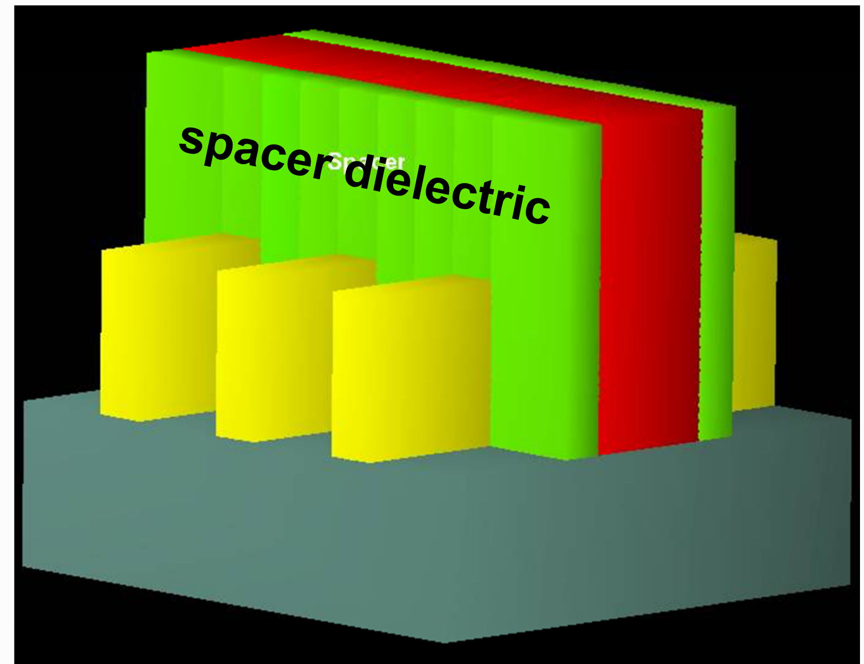
Process Flow Summary I

- Example shows tri-gate on SOI but bulk flow is similar
- Pattern fins using SIT
- Deposit/CMP STI oxide
- Recess STI oxide by fin height
- Deposit, CMP & pattern poly

- Deposit spacer dielectric & etch, leaving spacer on gate sidewalls
- Spacer must be removed on fin sidewall

gate oxide on top & both sidewalls of fin

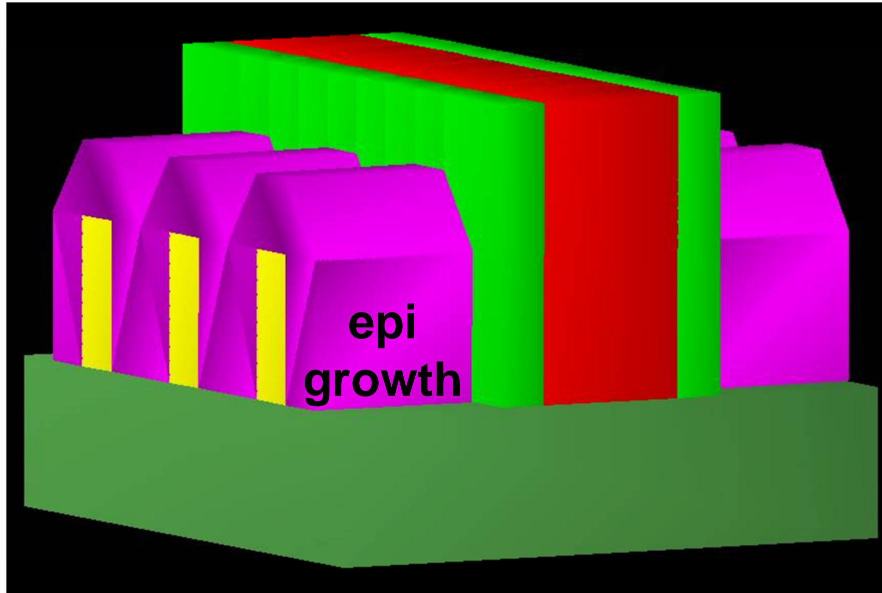


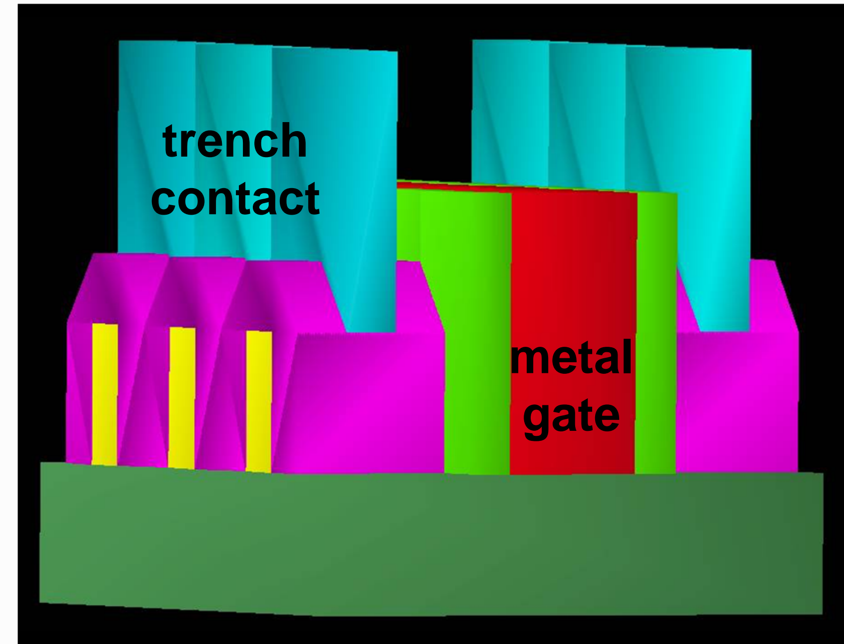


Paul, AMD [26]

Process Flow Summary II

- Recess fins
- Grow Si epitaxially to merge fins together for reduced source/drain resistance
- Induce uni-axial channel strain by growing e-SiGe or e-SiC
- Dope source/drain dopants with *in situ* doping during epi
- Deposit ILD0 & CMP to top of poly
- Do replacement-gate HKMG module
- Deposit & pattern contact dielectric
- Form trench contacts (note overlap capacitance to gate)



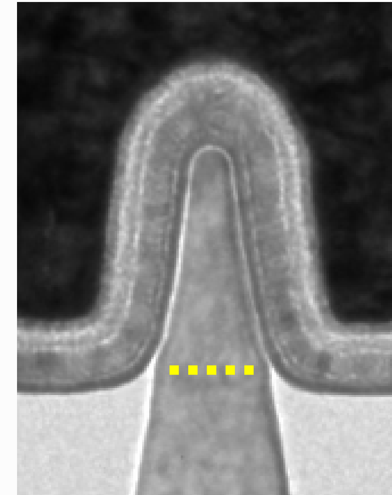


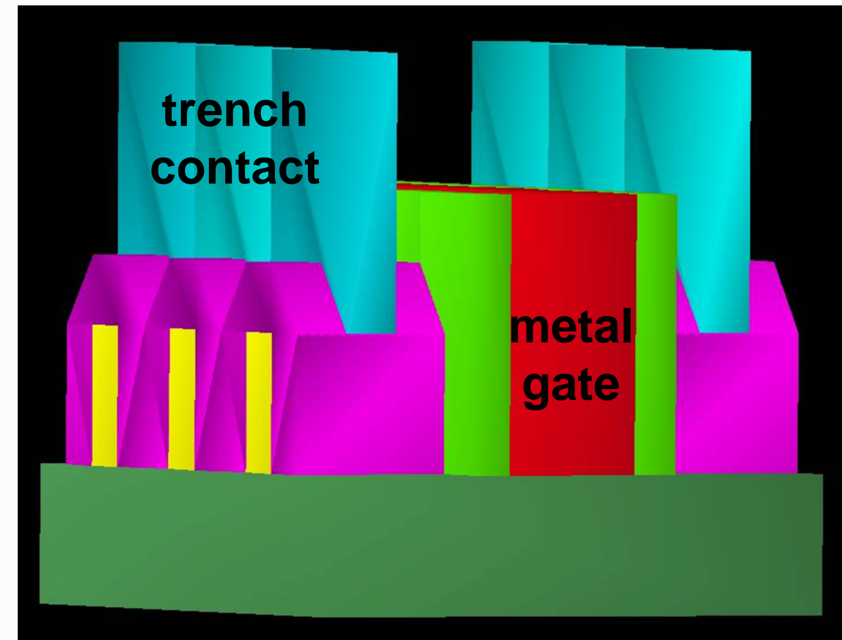
Paul, AMD [26]

Some Tri-Gate Considerations

- Field lines of from gate terminates at base of fins
- Fin base must be heavily doped for fin-to-fin isolation
- Dimensional variation of fins → device variation
- Current density is not uniform along *width* of device – V_T & S varies along sidewall

- Series resistance vs. overlap capacitance
- “Dead” space between fins



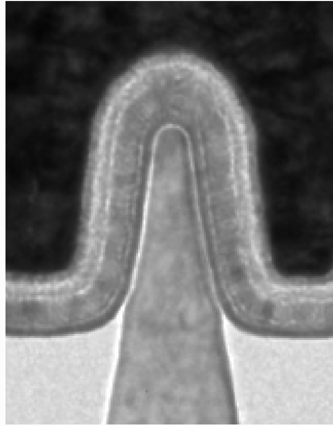


Pacifying The Multi-V_T Addiction

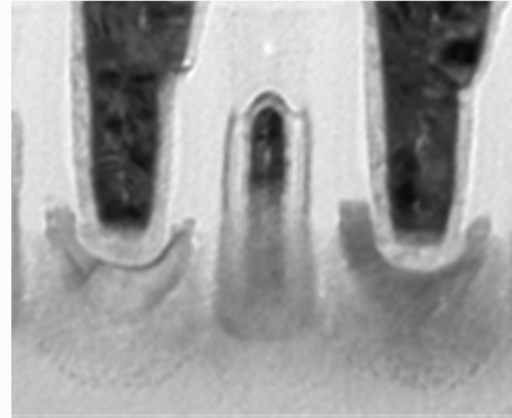
- 8 V_T's typical in 28nm (NMOS vs. PMOS, thick vs. thin oxide)
- Methods of achieving multiple V_T
  1. Bias channel length
     - Exploit SCE (V_T rolloff with shorter L)
     - Increase L for lower I_{ON} & I_{OFF}
  2. Implant fin body with different dose
     - Field lines from gate must "work through" available body dopants before terminating at base of fin
     - Prone to RDF
  3. Integrate more metal gate Φ_M
     - Already 2 Φ_Ms in standard HKMG flow
     - More complex integration

Intel 22nm TEM Cross-Sections

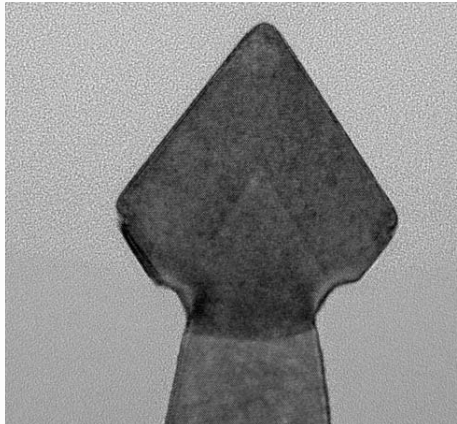
Single fin (along W)



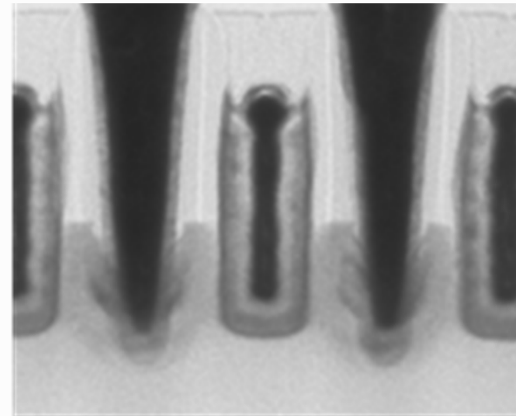
NMOS (along L)



Epi merge (along W)



PMOS (along L)

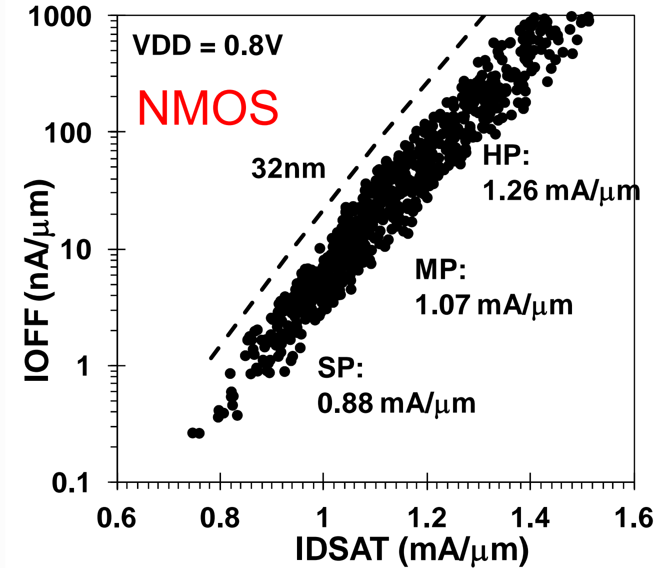


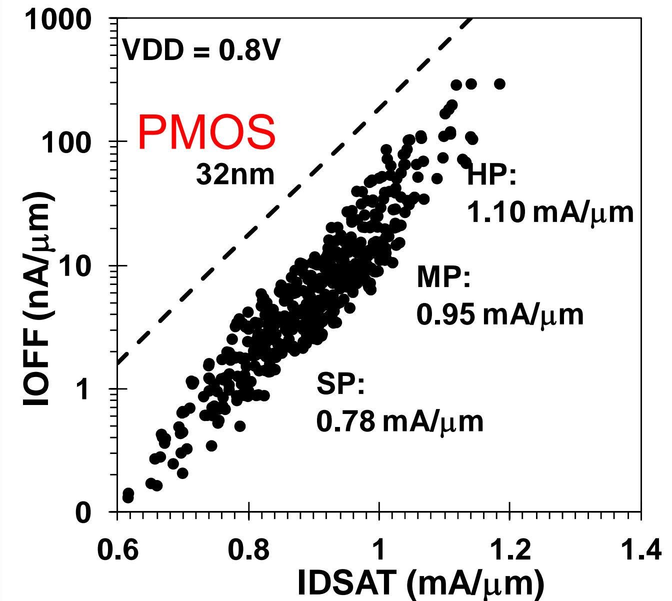
Auth, Intel [27]

Intel 22nm Performance at 0.8V

1.E-02
1.E-03
1.E-04
1.E-05
1.E-06
1.E-07
1.E-08
1.E-09
IDSAT (A/μm)
0.80V
0.05V
PMOS
SS ~72mV/dec
DIBL ~50 mV/V
0.80V
0.05V
NMOS
SS ~69mV/dec
DIBL ~46 mV/V
-1.0 -0.6 -0.2 0.2 0.6 1.0
VGS (V)

Auth, Intel [27]





Conclusions

- Digital needs will continue to drive CMOS scaling but at slower pace
- Expect new learning in 20nm & 14nm as we cope with fin design & layout
- SPICE models will lag to include new effects
- Designers with good technology knowledge are best positioned for silicon success
- Exciting time to be designing

References

[1] M. Keating, “Science fiction or technology roadmap: a look at the future of SoC design,” in *SNUG San Jose Conf.*, Mar. 2010.

[2] M. Na *et al.*, “The effective drive current in CMOS inverters,” in *IEEE Int. Electron Devices Meeting Tech. Dig.*, pp. 121–124, Dec. 2002.

[3] S.M. Sze, *Physics of Semiconductor Devices (2nd ed.)*, John Wiley & Sons, 1981.

[4] A. Wei, “Foundry trends: technology challenges and opportunities beyond 32nm,” in *IEEE Vail Computer Elements Workshop*, Jun. 2010.

[5] www.soitec.com

[6] J. Plummer *et al.*, *Silicon VLSI Technology– Fundamentals, Practice and Modeling*, Prentice-Hall, 2000.

[7] V. Chan *et al.*, “Strain for CMOS performance improvement,” in *Proc. IEEE Custom Integrated Circuits Conf.*, pp. 667–674, Sep.2005.

[8] X. Xi *et al.*, BSIM4.3.0 MOSFET Model – User’s Manual, *The Regents of the University of California at Berkeley*, 2003

[9] J. Faricelli, “Layout-dependent proximity effects in deep nanoscale CMOS,” in *Proc. IEEE Custom Integrated Circuits Conf.*, pp. 1–8, Sep.2010.

[10] J. McPherson, “Reliability trends with advanced CMOS scaling and the implications on design,” in *Proc. IEEE Custom Integrated Circuits Conf.*, pp. 405–412, Sep. 2007.

[11] P. Wong, “Beyond the conventional transistor,” *IBM J. Research & Development*, pp. 133–168, vol. 2-3, no. 46, Mar. 2002.

[12] www.ICKnowledge.com

[13] M. Horstmann *et al.*, “Advanced SOI CMOS transistor technologies for high-performance microprocessor applications,” in *Proc. IEEE Custom Integrated Circuits Conf.*, pp. 149–152, Sep. 2009.

[14] C. Auth, “45nm high-k + metal-gate strain-enhanced CMOS transistors,” in *Proc. IEEE Custom Integrated Circuits Conf.*, pp. 379–386, Sep. 2008.

[15] P. Packan *et al.*, “High performance 32nm logic technology featuring 2nd generation high-k + metal gate transistors,” in *IEEE Int. Electron Devices Meeting Tech. Dig.*, pp. 1–4, Dec. 2009.

[16] R.-H. Yan *et al.*, “Scaling the Si MOSFET: From bulk to SOI to bulk,” *IEEE Trans. Electron Devices*, vol. 39, no. 7, pp. 1704–1710, Jul. 1992.

[17] L. Wei *et al.*, “Exploration of device design space to meet circuit speed targeting 22nm and beyond,” in *Proc. Int. Conf. Solid State Devices and Materials*, pp. 808–809, Sep. 2009.

[18] K. Cheng *et al.*, “Fully depleted extremely thin SOI technology fabricate by a novel integration scheme featuring implant-free, zero-silicon-loss, and faceted raised source/drain,” in *IEEE Symp. VLSI Technology Tech. Dig.*, pp. 212–213, Jun.2009.

[19] T. Skotnicki, “CMOS technologies – trends, scaling and issues ,” in *IEEE Int. Electron Devices Meeting Short Course,* Dec. 2010.

[20] M. Yamaoka *et al.*, “SRAM circuit with expanded operating margin and reduced stand-by leakage current using thin BOX FD-SOI transistors,“ *IEEE J. Solid-State Circuits*, vol. 41, no.11, Nov. 2006.

[21] T. Skotnicki, “Competitive SOC with UTBB SOI,” in *Proc. IEEE SOI Conf.*, Oct. 2011.

[22] K. Fujita *et al.*, “Advanced channel engineering achieving aggressive reduction of V_T variation for ultra-low power applications,” in *IEEE Int. Electron Devices Meeting Tech. Dig.*, pp. 32.3.1–32.3.4, Dec. 2011.

[23] C. Hu, “FinFET 3D transistor and the concept behind it,” in *IEEE Electron Device Soc. Webinar*, Jul. 2011.

[24] M. Bohr, “22 nm tri-gate transistors for industry-leading low power capabilities,” in *Intel Developer Forum*, Sep. 2011.

[25] M. Yang *et al.*, “Hybrid-orientation technology (HOT): opportunities and challenges*,” IEEE Trans. Electron Devices*, vol. 53, no. 5, pp. 965–978, May 2006.

[26] S. Paul, “FinFET vs. trigate: parasitic capacitance and resistance”, *AMD Internal Presentation*, Aug. 2011.

[27] C. Auth *et al.*, “A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors,” in *IEEE Symp. VLSI Technology Tech. Dig.*, pp. 131–132, Jun. 2012.