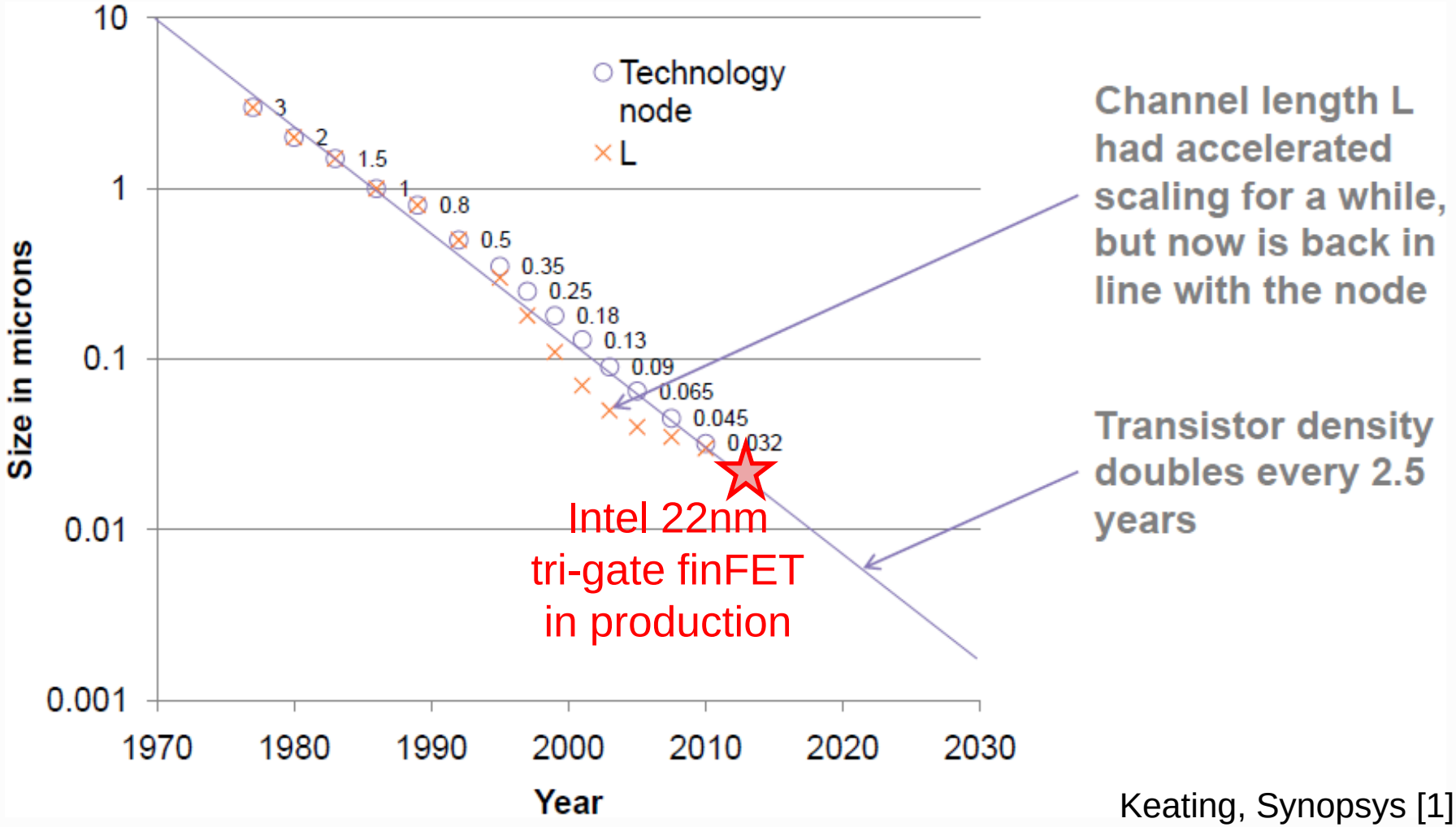
2D to 3D MOS Technology Evolution for Circuit Designers

Alvin Loke[1], Ray Stephany[1], Andy Wei[2], Bich-Yen Nguyen[3], Tin Tin Wee[1], John Faricelli[1], Jung-Suk Goo[2], and Shawn Searles[1]

[1]Advanced Micro Devices
[2]GlobalFoundries
[3]Soitec

CMOS Scaling Still Alive…

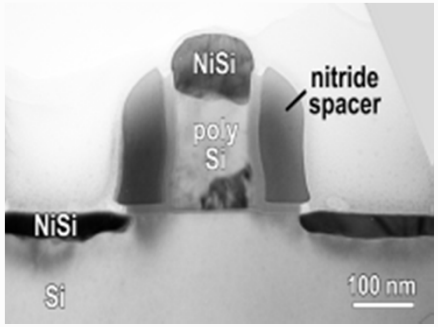


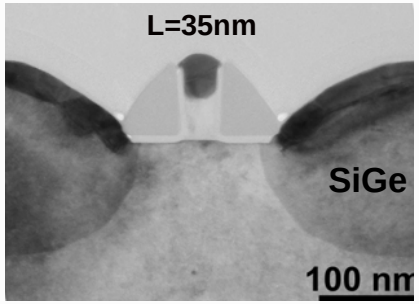
Keating, Synopsys [1]

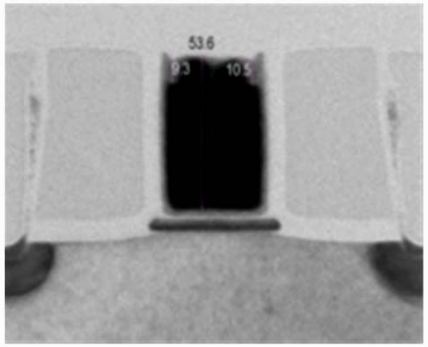
- Leading foundries frantically after manufacturable tri-gate

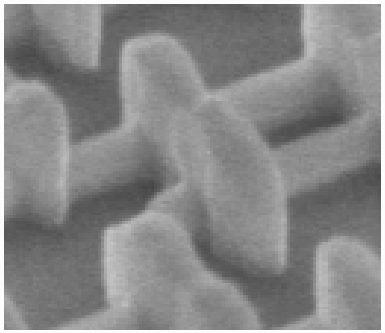
Our Objective

- Understand how MOSFET structure has evolved
- Learn about enabling technologies
- Understand why it has evolved this way





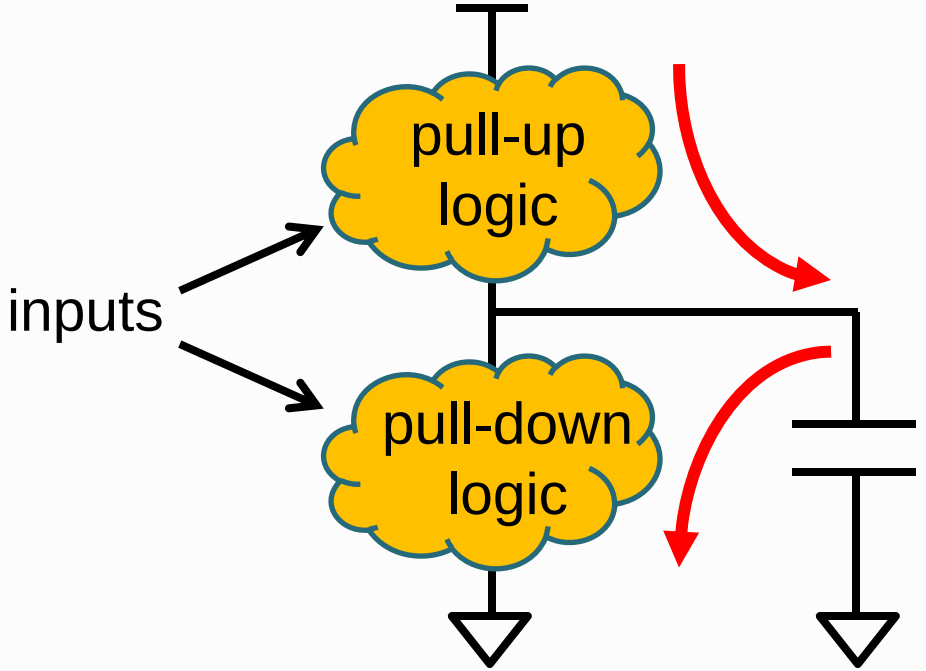




Outline

- Motivation
- MOSFET & Short-Channel Fundamentals
- 130nm MOSFET Fabrication
- Strain Engineering (90nm & Beyond)
- High-K / Metal-Gate (45nm & Beyond)
- Fully-Depleted (22nm & Beyond)
- Tri-Gate FinFETs
- Conclusions

The Basis of All CMOS Digital ICs

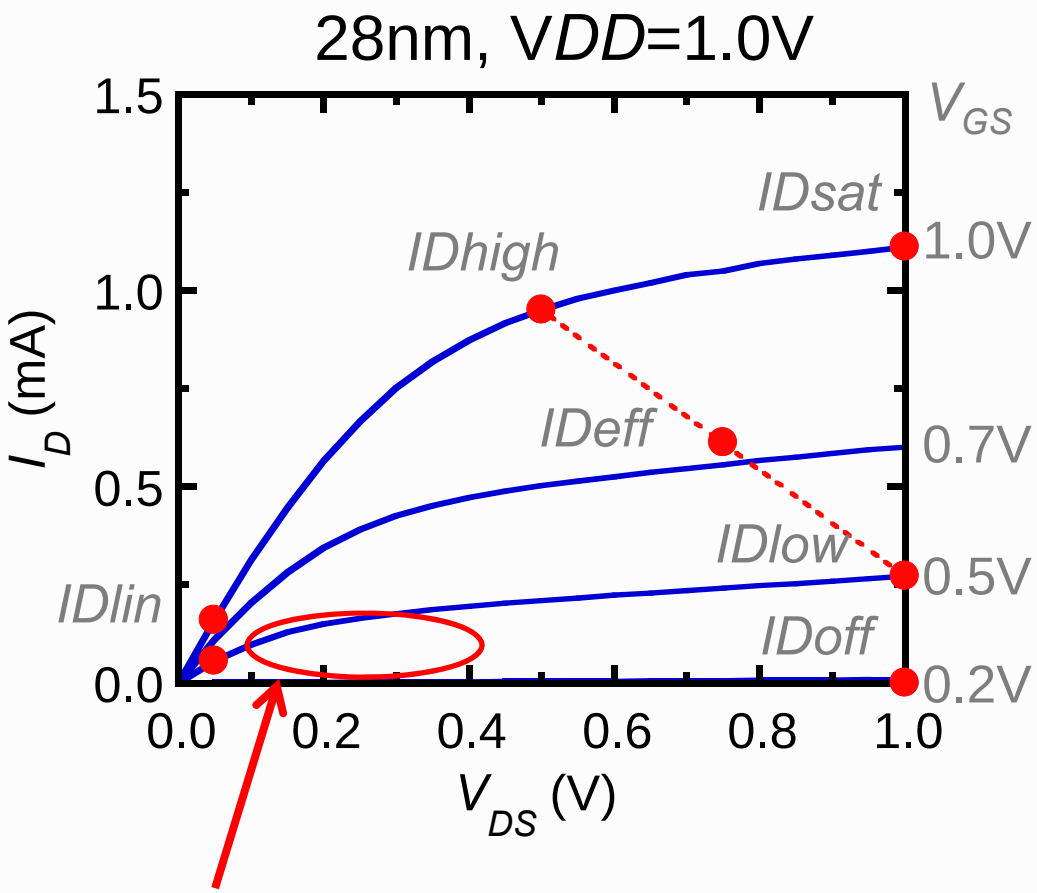


$$t_{delay} \approx \frac{Q_{load}}{I_{eff}} = \frac{C_{load} V_{DD}}{I_{eff}}$$

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

- Charging and discharging a capacitor… very quickly!
- For shorter delay and lower power
  - C_{load} ↓ → reduce parasitics (wires, gates, junctions, …)
  - V_{DD} ↓ → reduce logic swing
  - I_{eff} ↑ → move charge quicker

Important MOSFET Current Parameters



typical analog usage
V_{GS}=V_T to V_T+0.2V

IDeff estimates effective inverter current drawn during switching event, more realistic and way less optimistic than *IDsat*

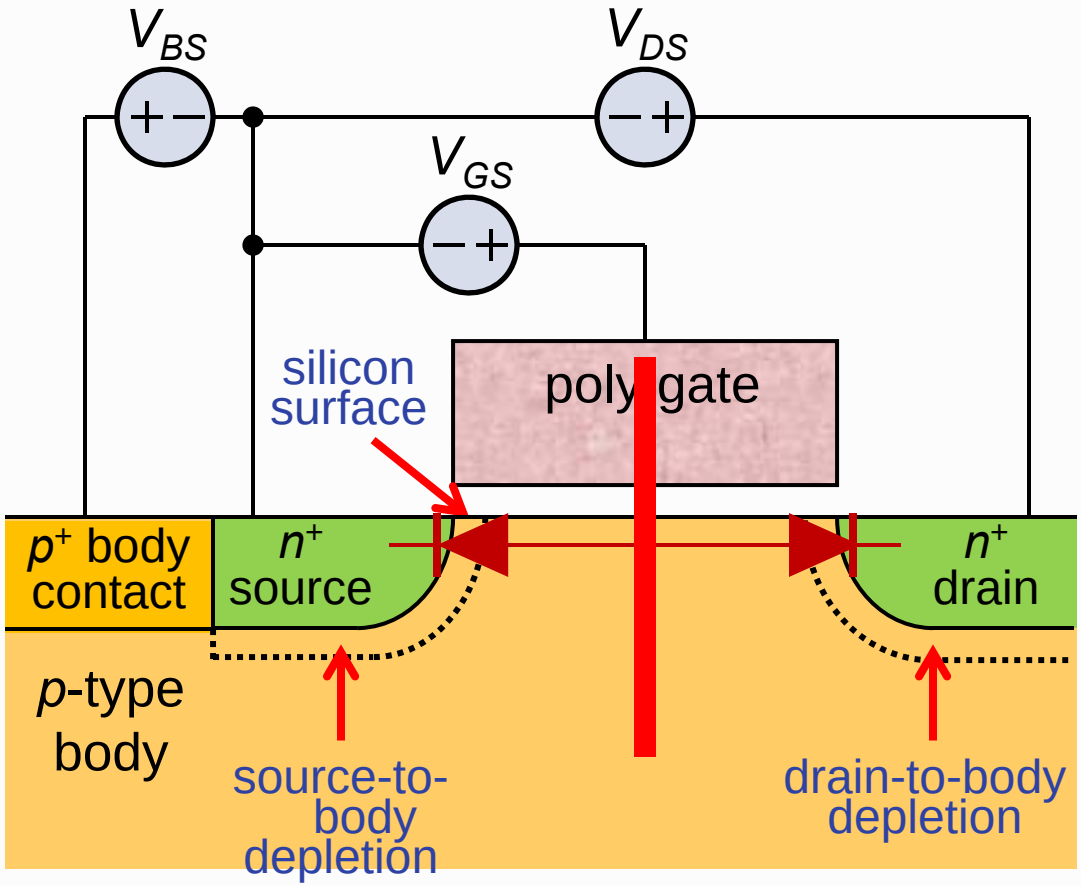
$$IDeff = \frac{IDlow + IDhigh}{2}$$

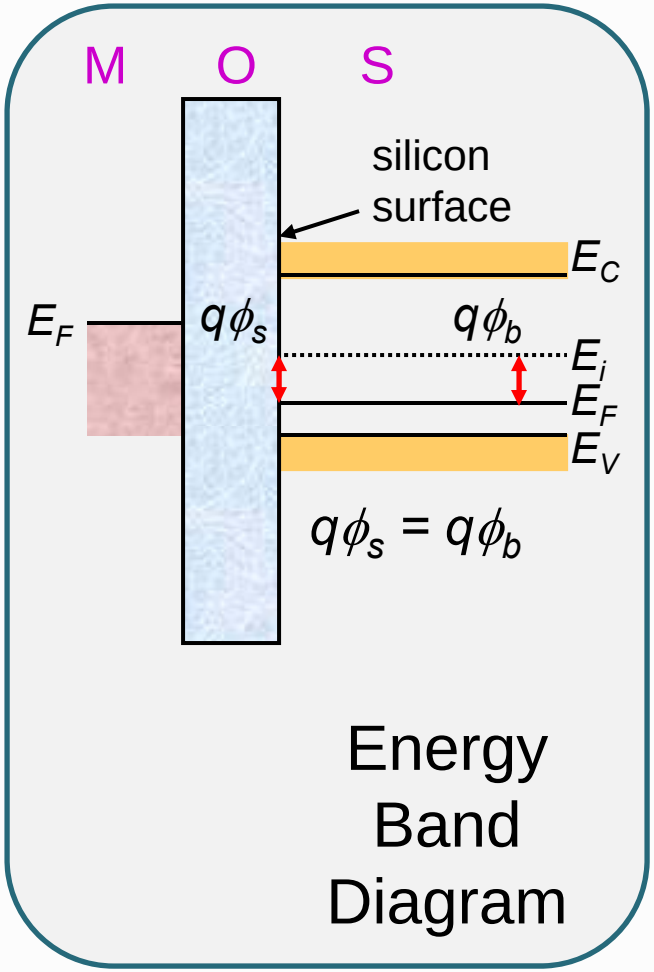
$$IDlow = ID\left(V_{GS} = \frac{V_{DD}}{2}, V_{DS} = V_{DD}\right)$$

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

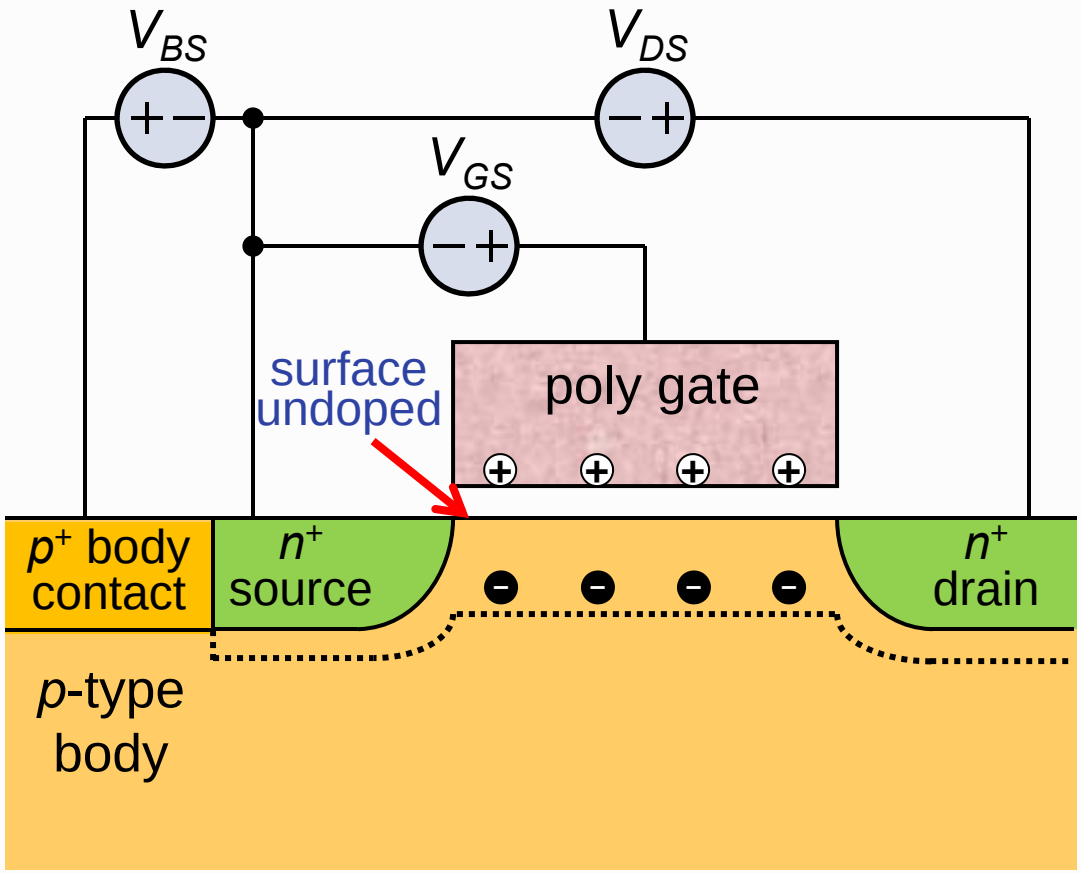
Na *et al.*, IBM [3]

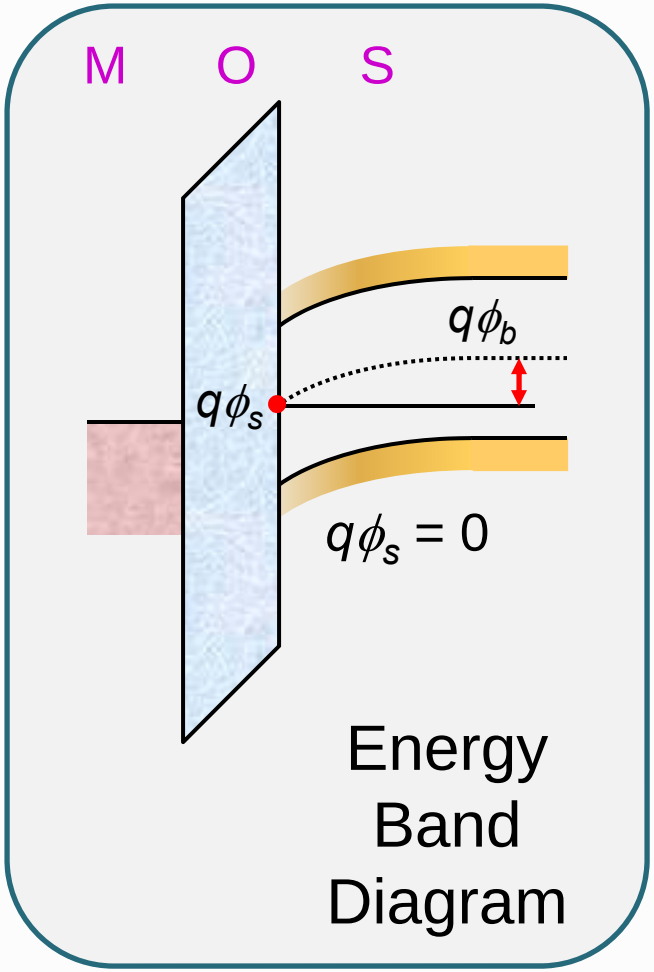
Flatband Condition (V_{GS}=V_{FB})





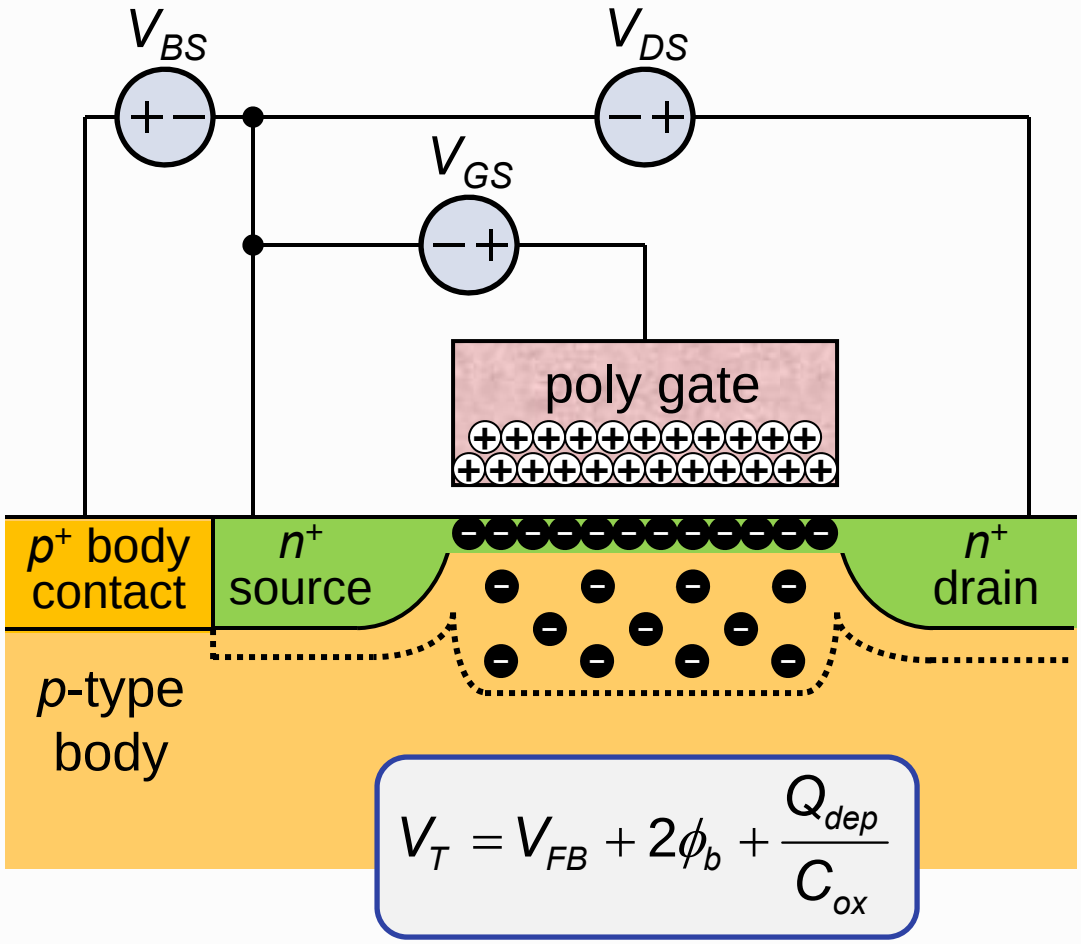
Onset of Surface Inversion (ϕ_s=0)



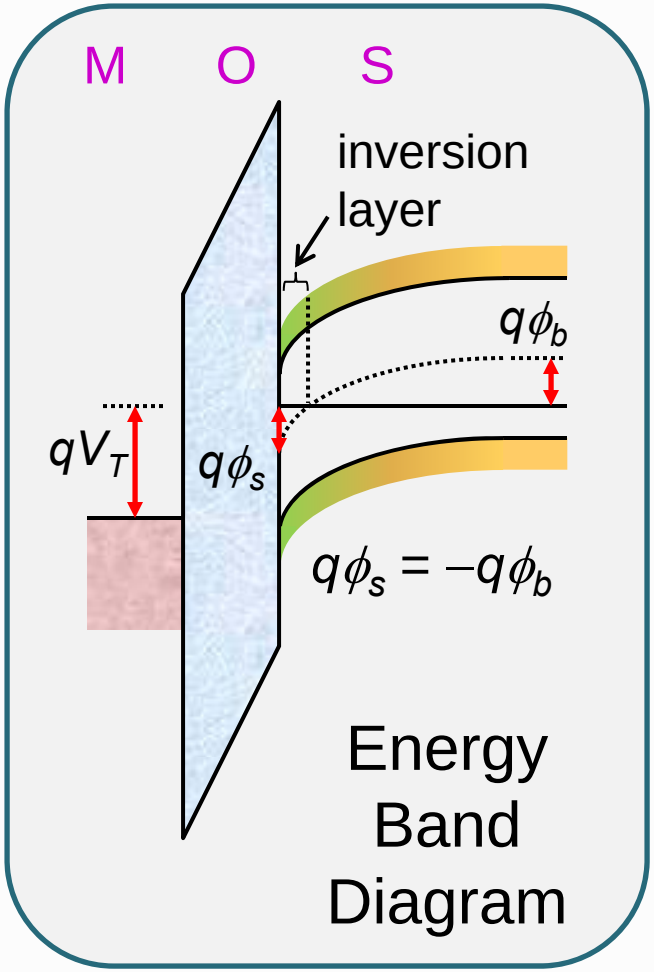


+ charge terminating on – charge

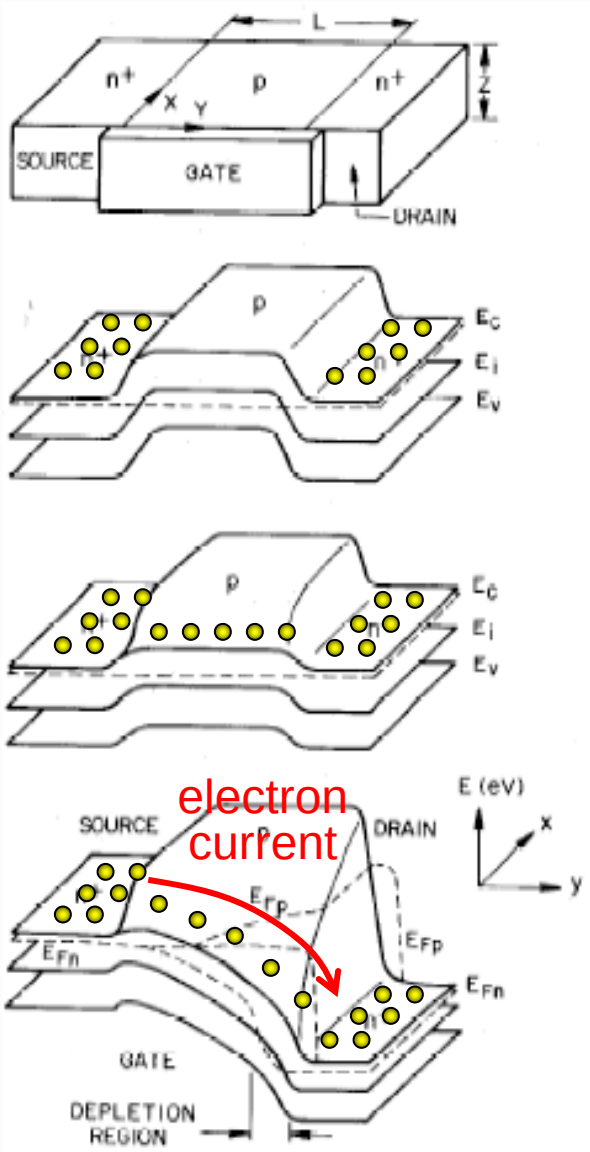
Onset of Strong Surface Inversion ($V_{GS}=V_T$)



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



Lower the Surface Barrier



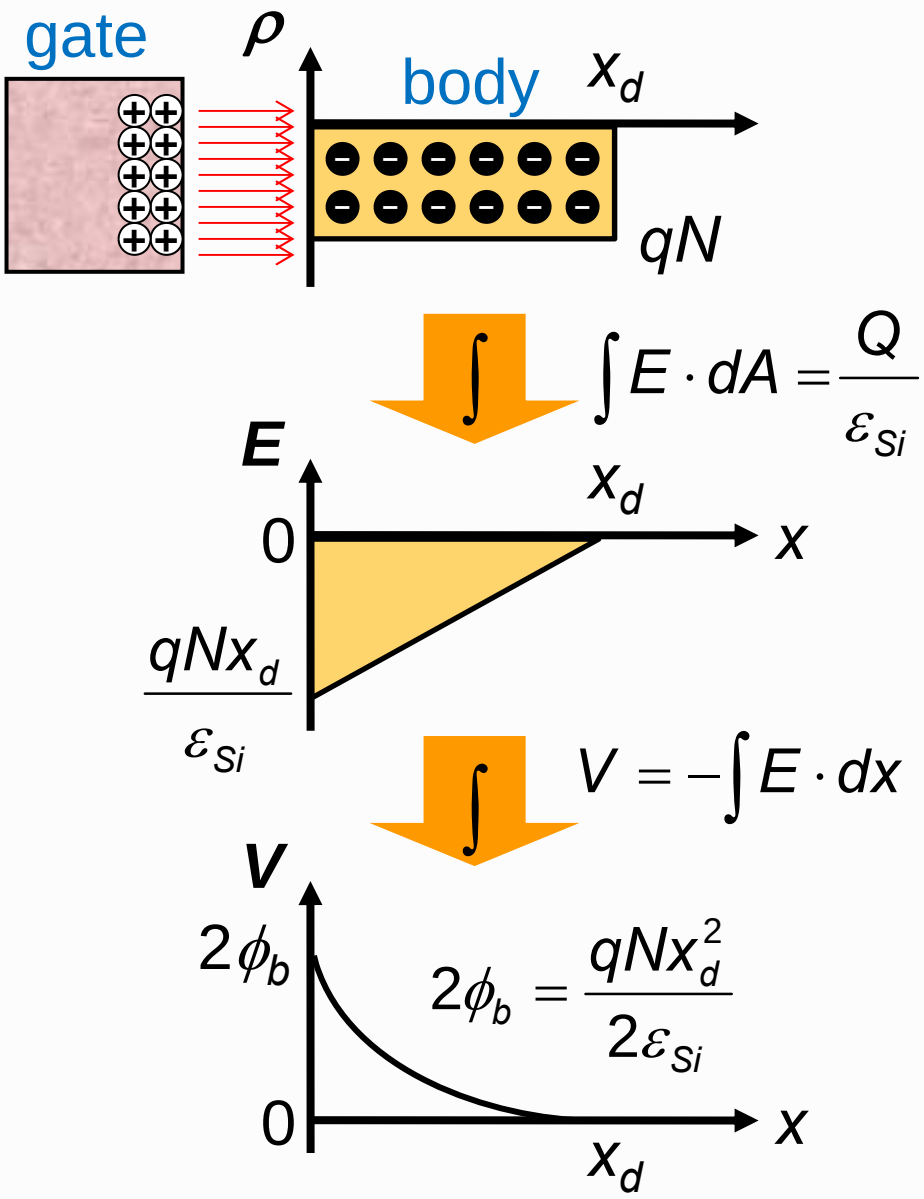
$V_{GS} = 0$
$V_{DS} = 0$ (no current)
Large source barrier
(back-to-back diodes)

$V_{GS} \approx V_T$
$V_{DS} = 0$ (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 [4]

Quantifying Charge to Move ϕ_s by $2\phi_b$

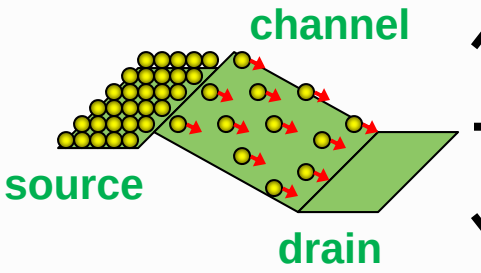


- Assume *uniformly doped* p-type body
- How much body must be depleted to reach strong inversion?

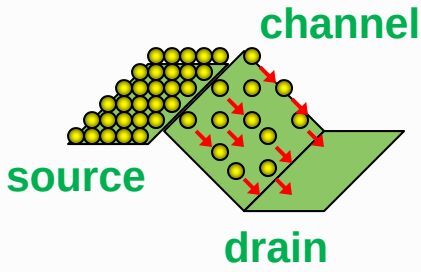
$$x_d = \sqrt{\frac{2\varepsilon_{Si} \cdot 2\phi_b}{qN}} \propto \frac{1}{\sqrt{N}}$$

$$Q_{dep} = qNx_d$$

The Roads to Higher Performance

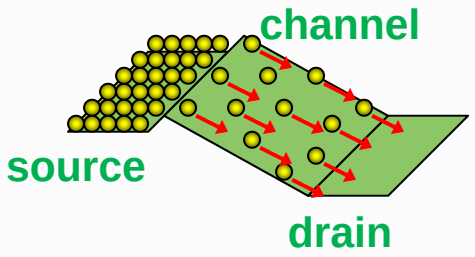


Decrease L – steepen the hill



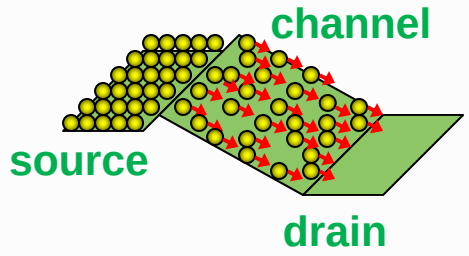
lithography
scaling

Increase μ – move carriers faster



strain engineering

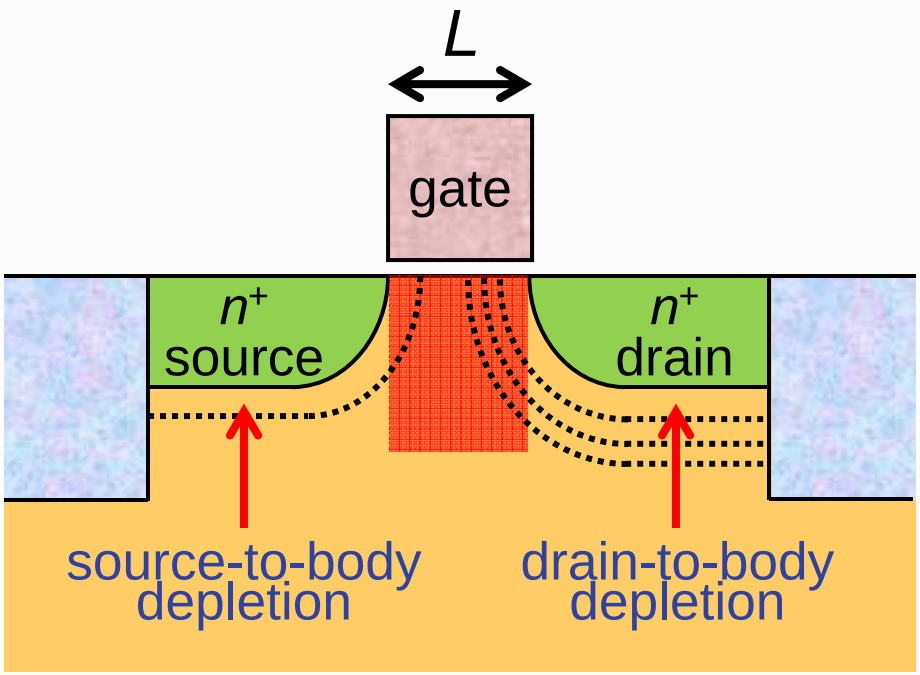
Increase C_{ox} – move more carriers

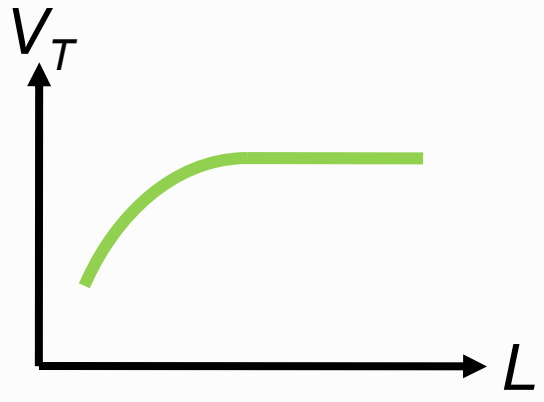


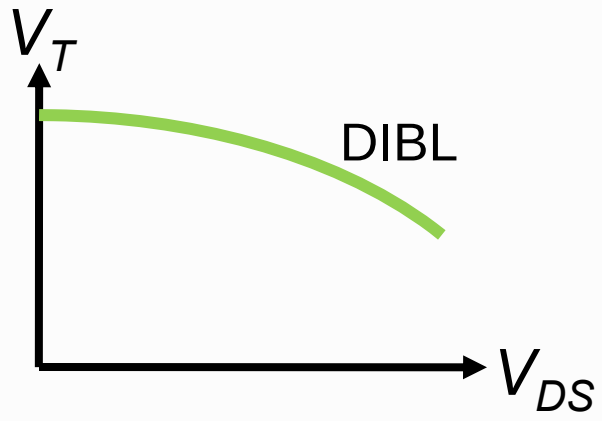
high-K dielectric
metal gate

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

Short-Channel Effects (SCEs)







V_{DD} not scaling as aggressively as L

→ Higher channel electric fields

- Velocity saturation
- Mobility degradation

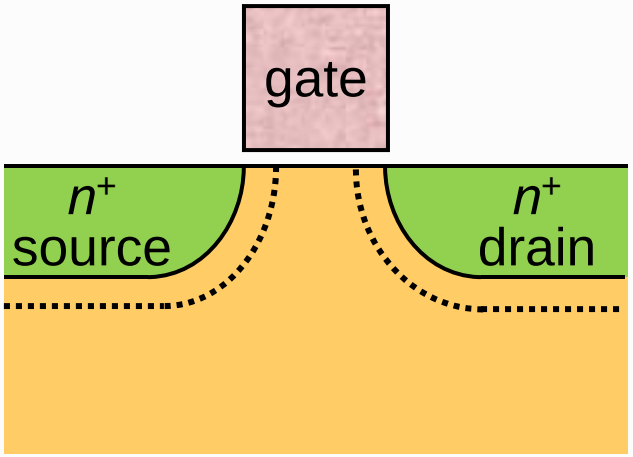
Overcoming Short-Channel Effects

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)

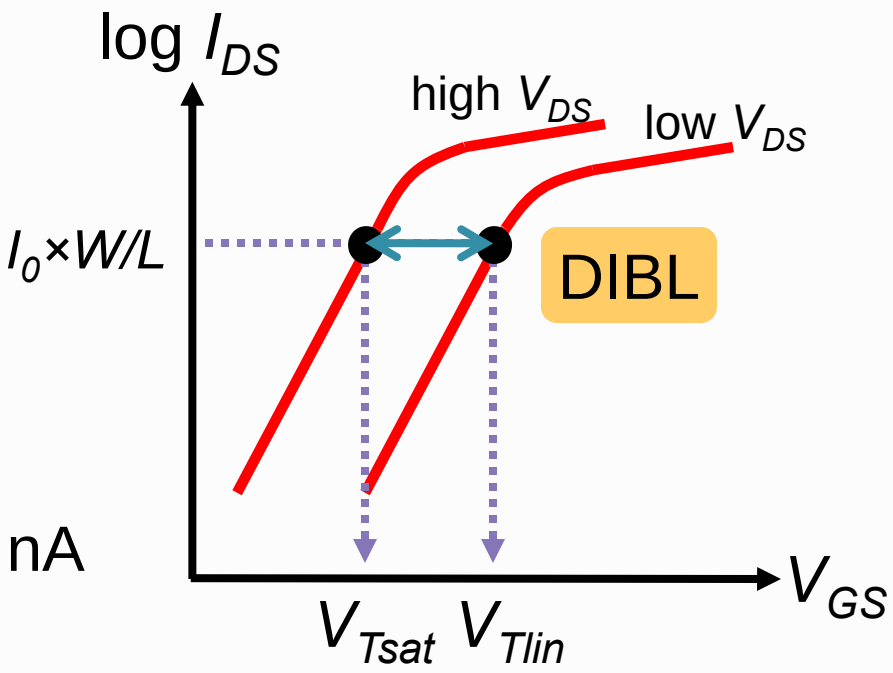
Stressors for mobility enhancement

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



Constant-Current V_T Measurement

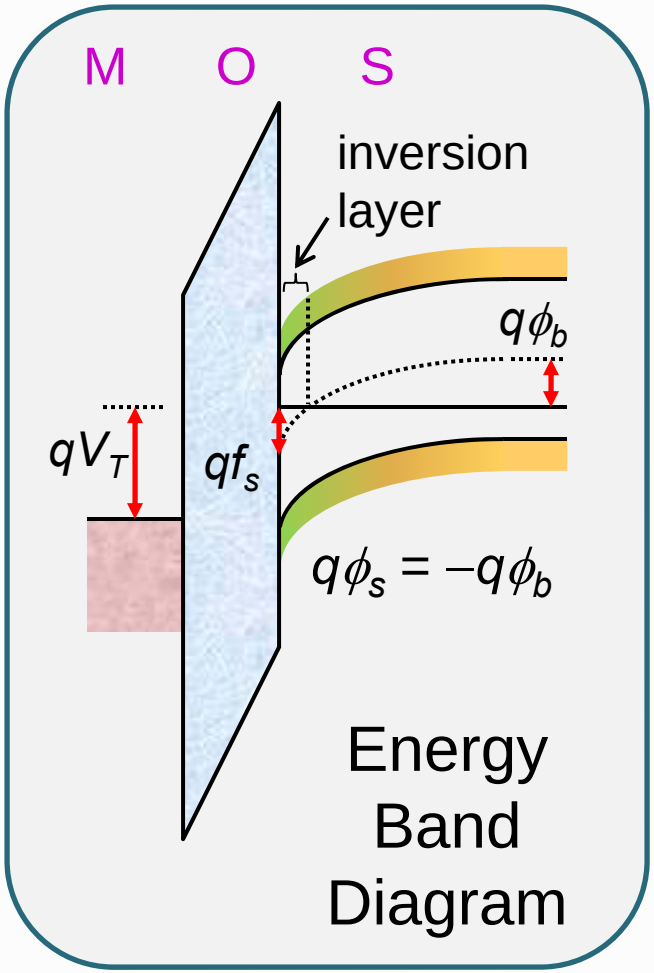
- Onset of strong inversion near impossible to measure
- Sweep log I_{DS} vs. V_{GS}
- Find V_{GS} when I_{DS} crosses user-specified threshold I_0 normalized to W/L
- Foundry-specific I_0 ~ 10 to 500 nA
- No physical connection to "fundamental" V_T definition



$$V_T = V_{GS}\Big|_{I_{DS}=I_0\times\frac{W_{drawn}}{L_{drawn}}}$$

Loke *et al.*, AMD [5]

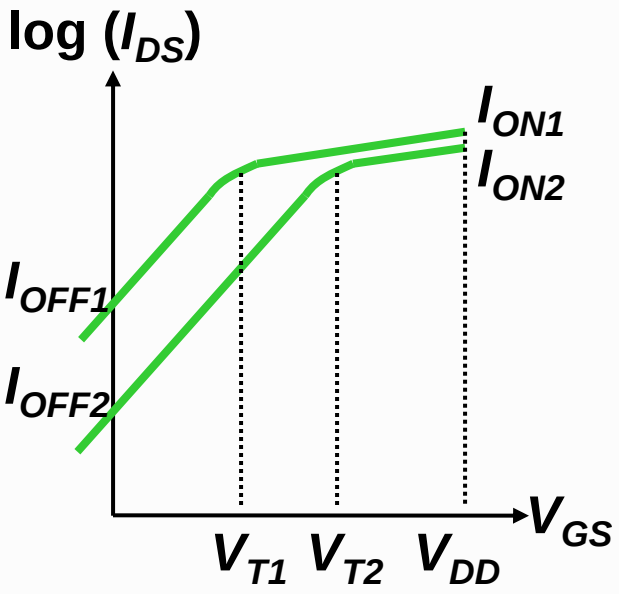
Not So Fundamental After All

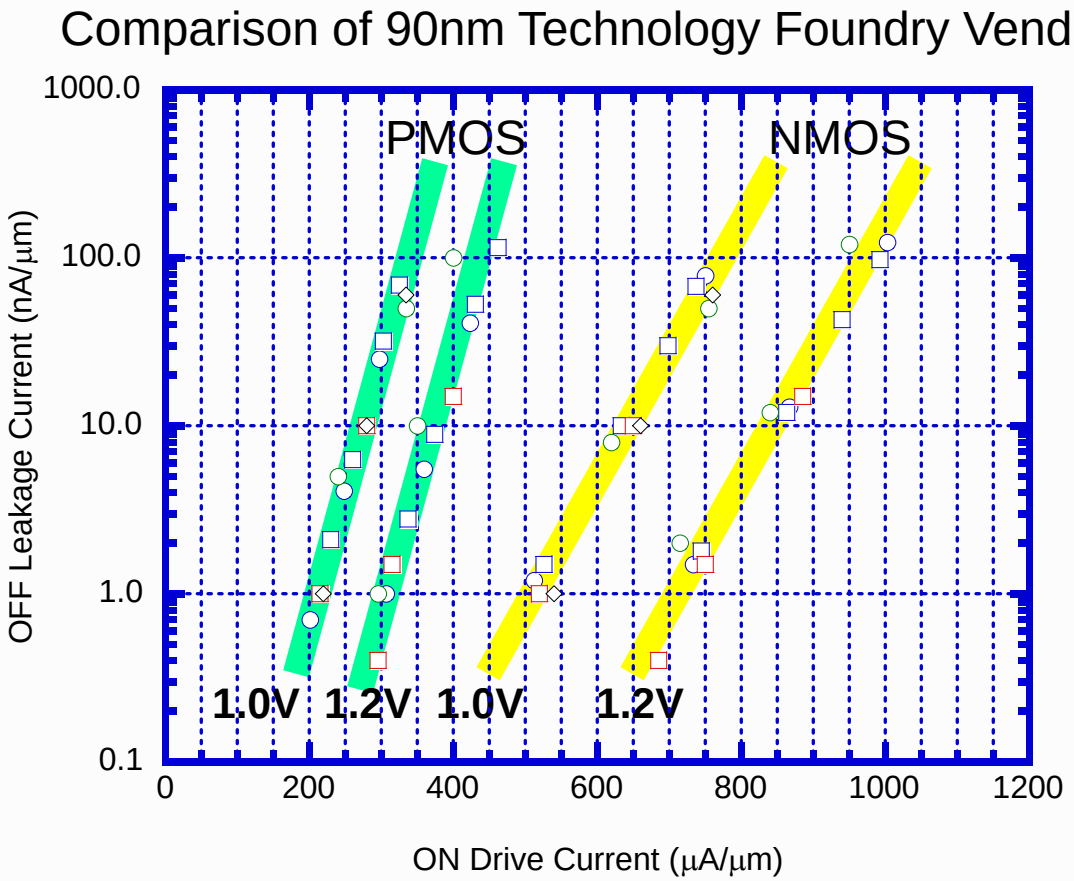


$$V_T = V_{FB} + 2\phi_b + \frac{Q_{dep}}{C_{ox}}$$

- 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

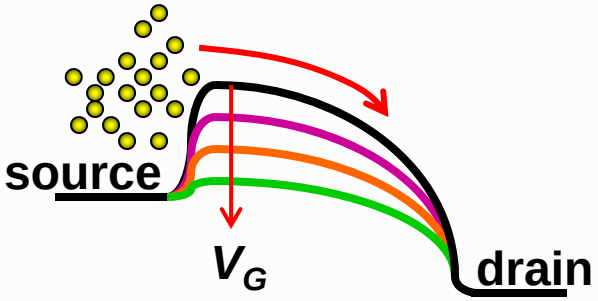




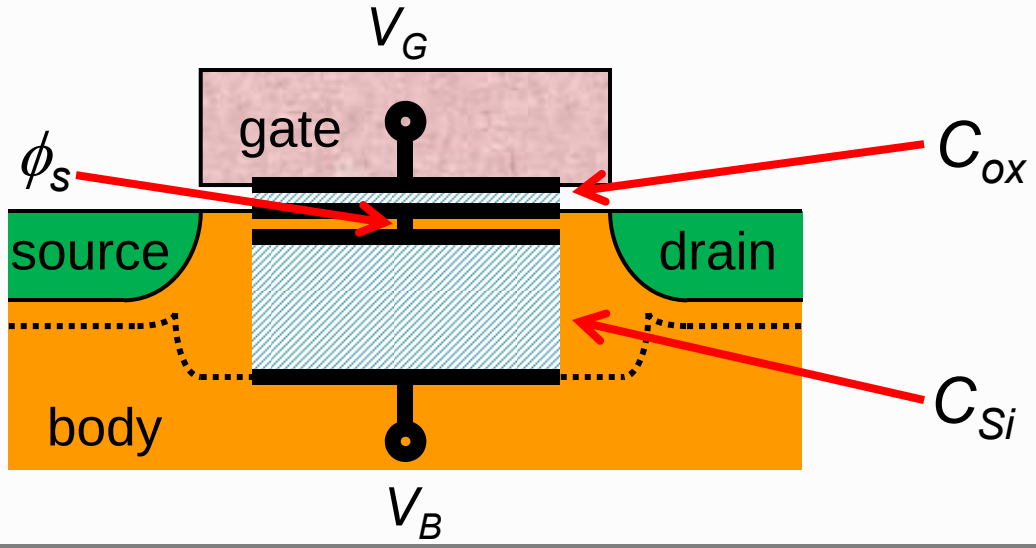
- High I_{ON} → high I_{OFF} & low I_{ON} → low I_{OFF}
- OFF leakage prevents V_T from scaling with gate length
- Several V_T's enable trade-off between high speed vs. low leakage

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



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

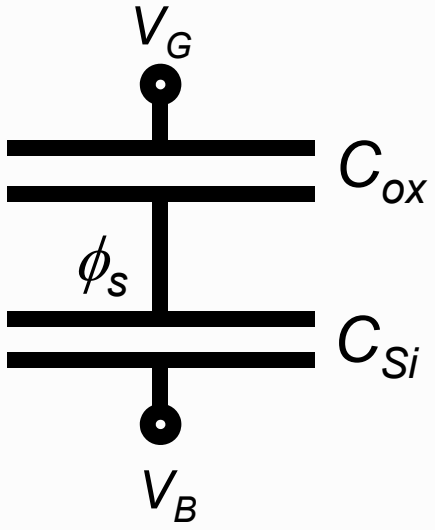


Subthreshold Slope

- V_G needed for 10× change in current

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

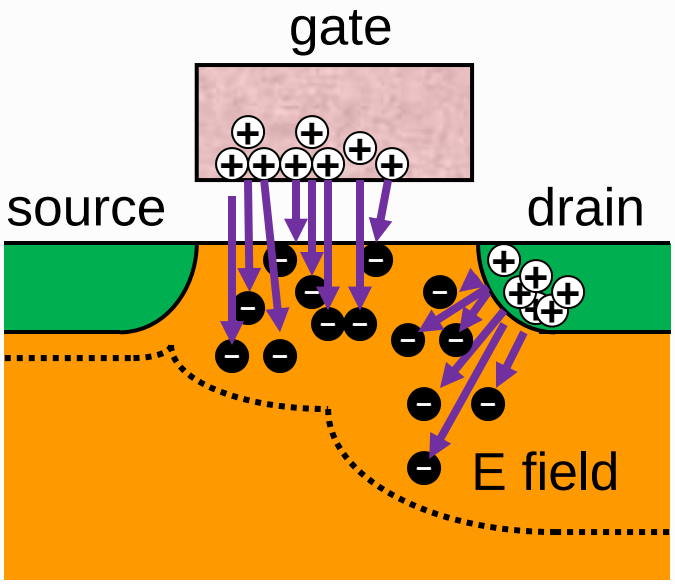
$$S = (60mV/dec) \cdot \frac{C_{ox} + C_{Si}}{C_{ox}} \quad \text{at } 25°C$$

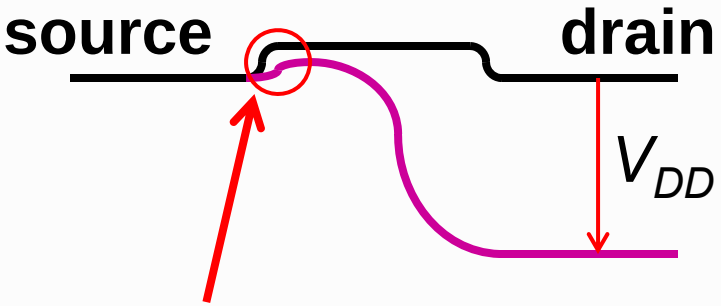


- **Planar 28nm: *S* = 100–110mV/dec at 25°C**
- Want tight coupling of V_G to ϕ_s but have to overcome C_{Si}
  - Large C_{ox} → thinner gate oxide, HKMG
  - Small C_{Si} → lower body doping, FD-SOI, finFET
  - Get diode limit when $C_{ox} \rightarrow \infty$ & $C_{Si} \rightarrow 0$ ($\eta = 1$)
-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

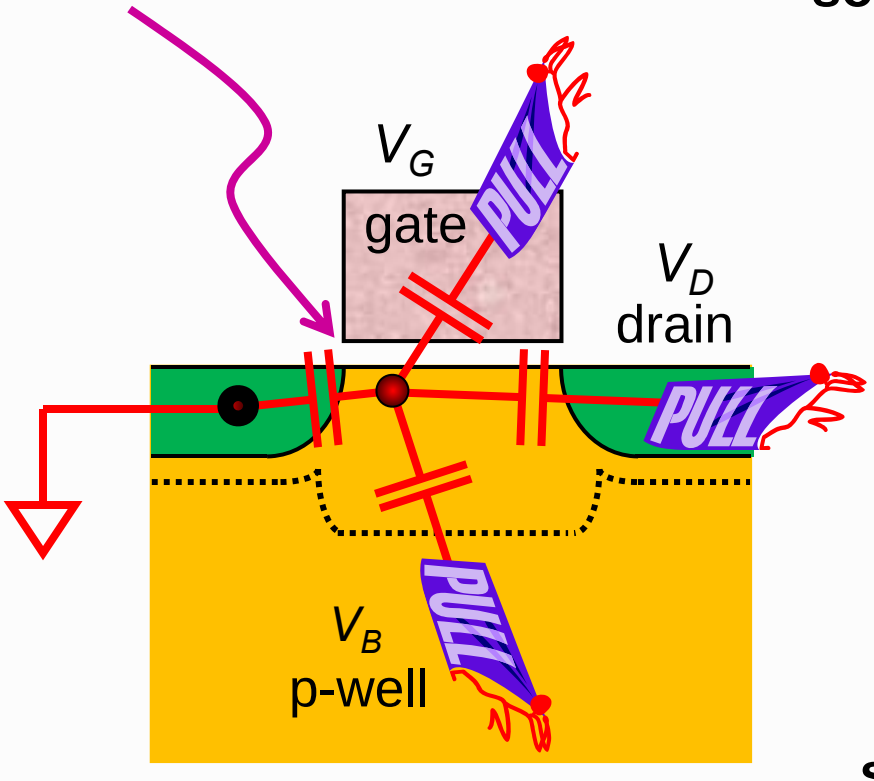


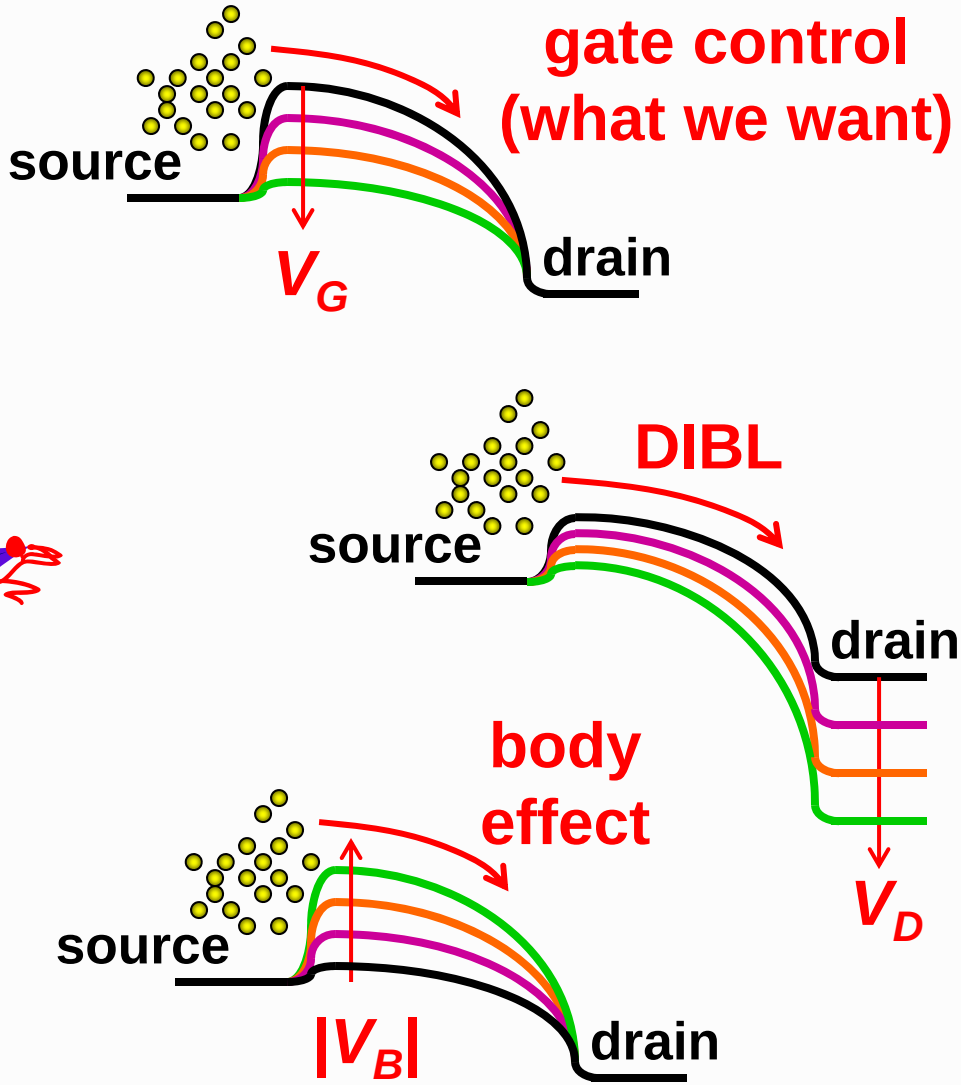


reduction of barrier height
at edge of source

3-Way Competition for Body Charge

What's happening to surface potential?



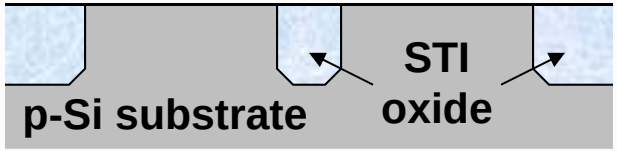


Outline

- Motivation
- MOSFET & Short-Channel Fundamentals
- **130nm MOSFET Fabrication**
- Strain Engineering (90nm & Beyond)
- High-K / Metal-Gate (45nm & Beyond)
- Fully-Depleted (22nm & Beyond)
- Tri-Gate FinFETs
- Conclusions

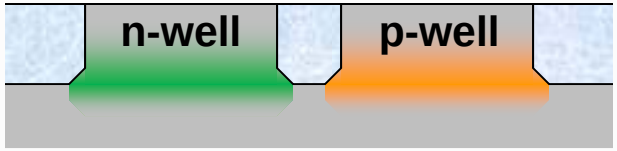
130nm MOSFET Fabrication

1



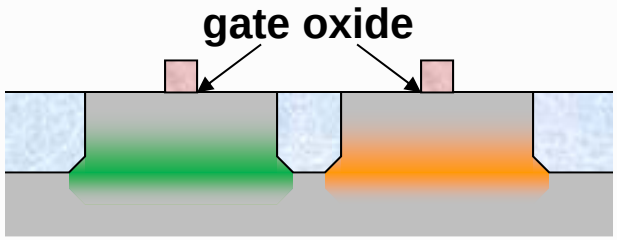
Shallow Trench Isolation

2



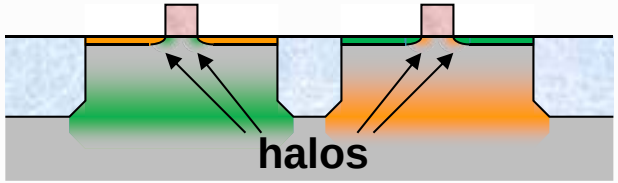
Well Implantation

3



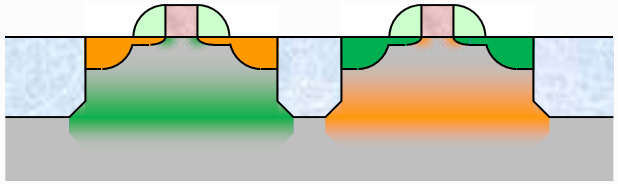
Gate Oxidation &
Poly Definition

4



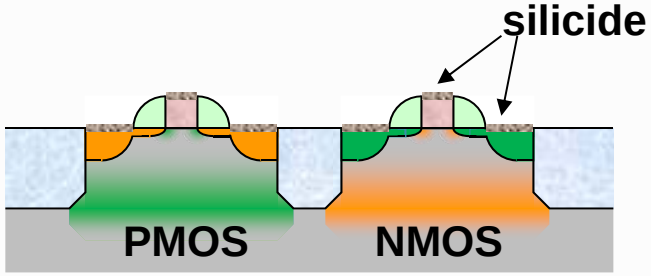
Source/Drain Extension
& Halo Implantation

5



Spacer Formation &
Source/Drain Implantation

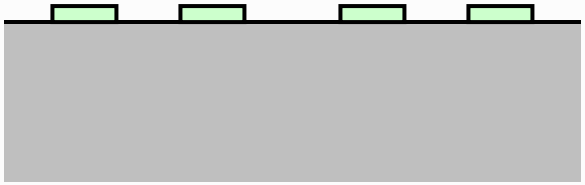
6



Salicidation

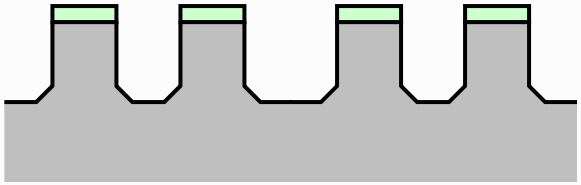
Shallow Trench Isolation

1



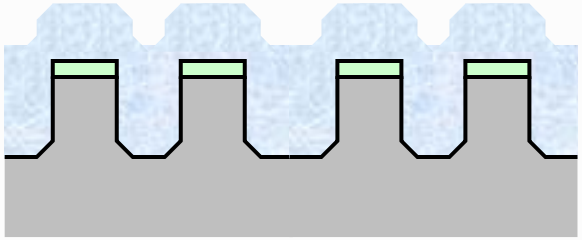
Deposit & pattern thin Si_3N_4 etch mask & polish stop

2



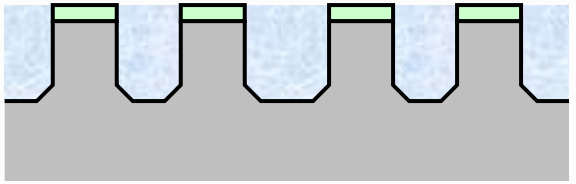
Etch silicon around active area – profile critical to minimize stress

3



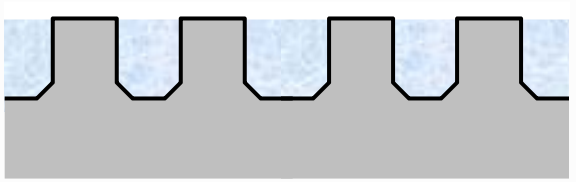
Grow liner SiO_2, then deposit conformal SiO_2 – void-free deposition is critical

4



CMP excess SiO_2

5



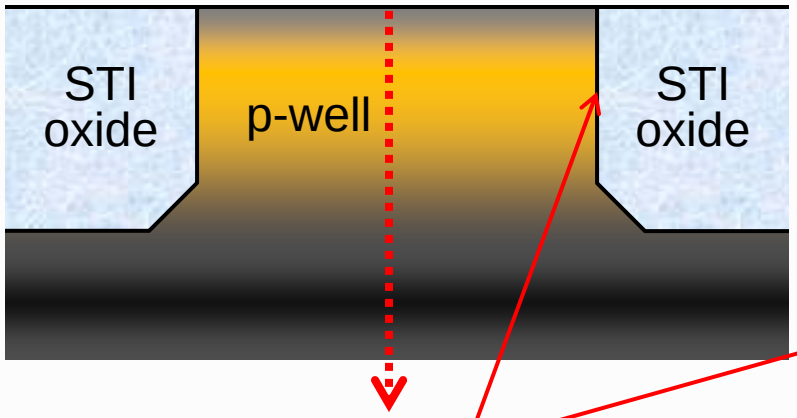
Recess SiO_2
Strip Si_3N_4 polish stop

Advantages over LOCOS

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

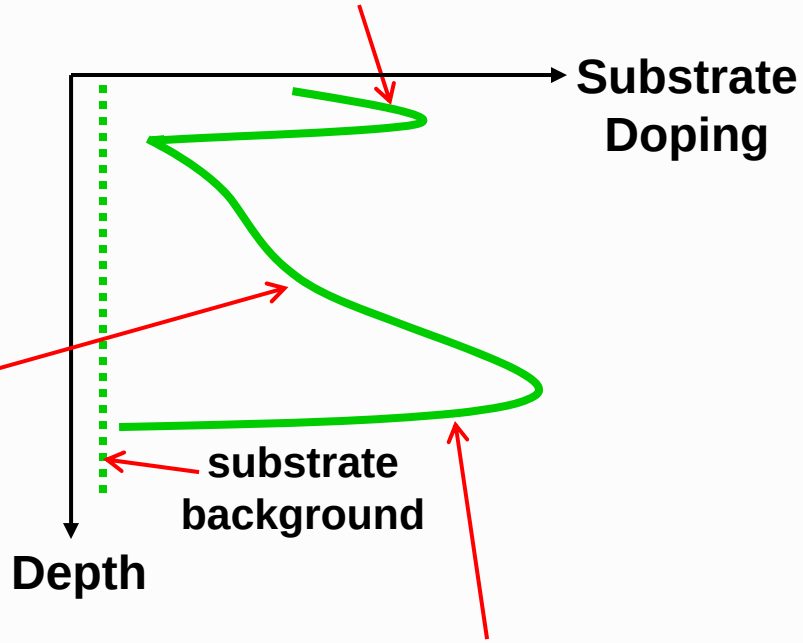
Well Implant Engineering

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)

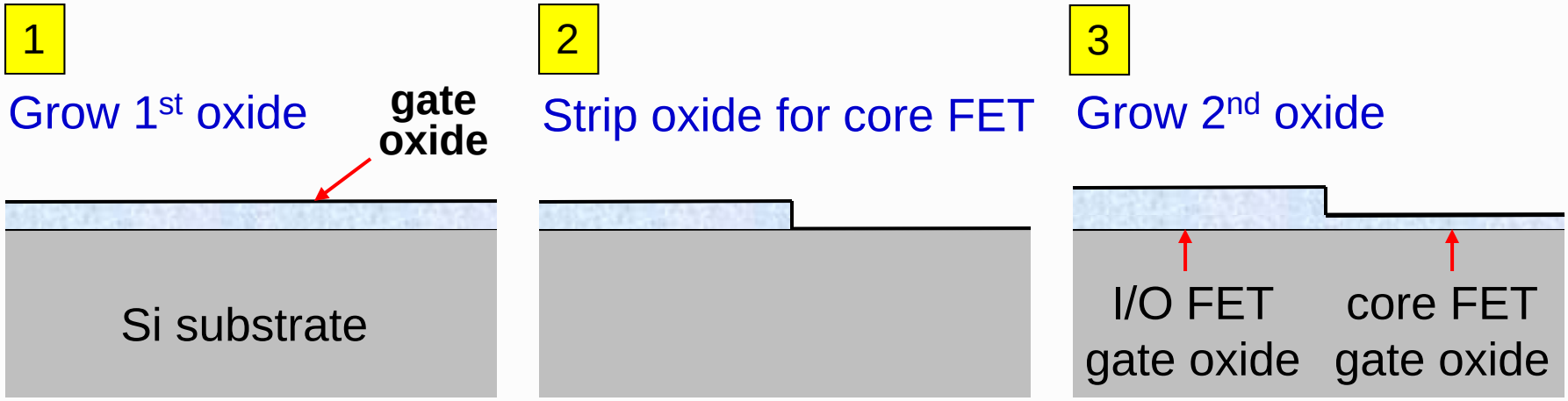
Very deep high-dose implant

- Latchup prevention
- Noise immunity
- Faster diffusers (B, As/P)

Sequence implant to reduce ion channeling, especially for shallow implant

Gate Oxide Growth

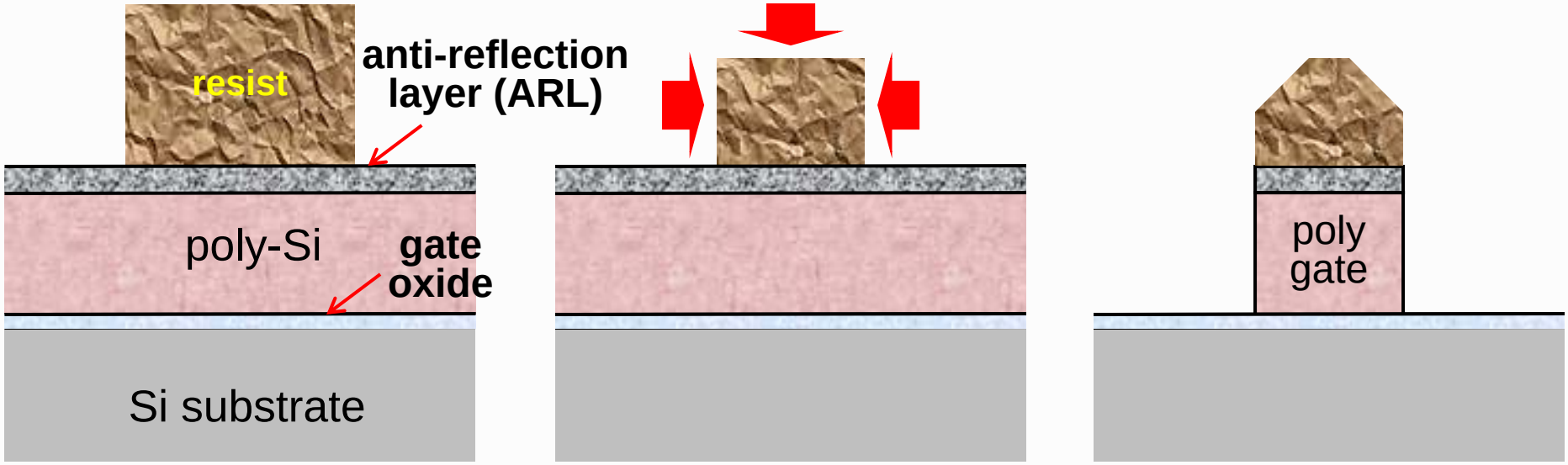
- Need two gate oxide t_{ox}'s – thin for core FET & thick for I/O FET



- Oxide is grown, not deposited
  - Need high-quality Si-SiO_2 interface with low Q_f & D_{it}
- Gate oxide is really made of silicon oxynitride (SiO_xN_y)
  - Nitrogen prevents boron diffusion from p+ poly to channel
  - Improves GOI (gate oxide integrity) reliability
  - Side benefit – increased ε_{ox}

Poly Gate Definition

- Gate CD way smaller than lithography capability



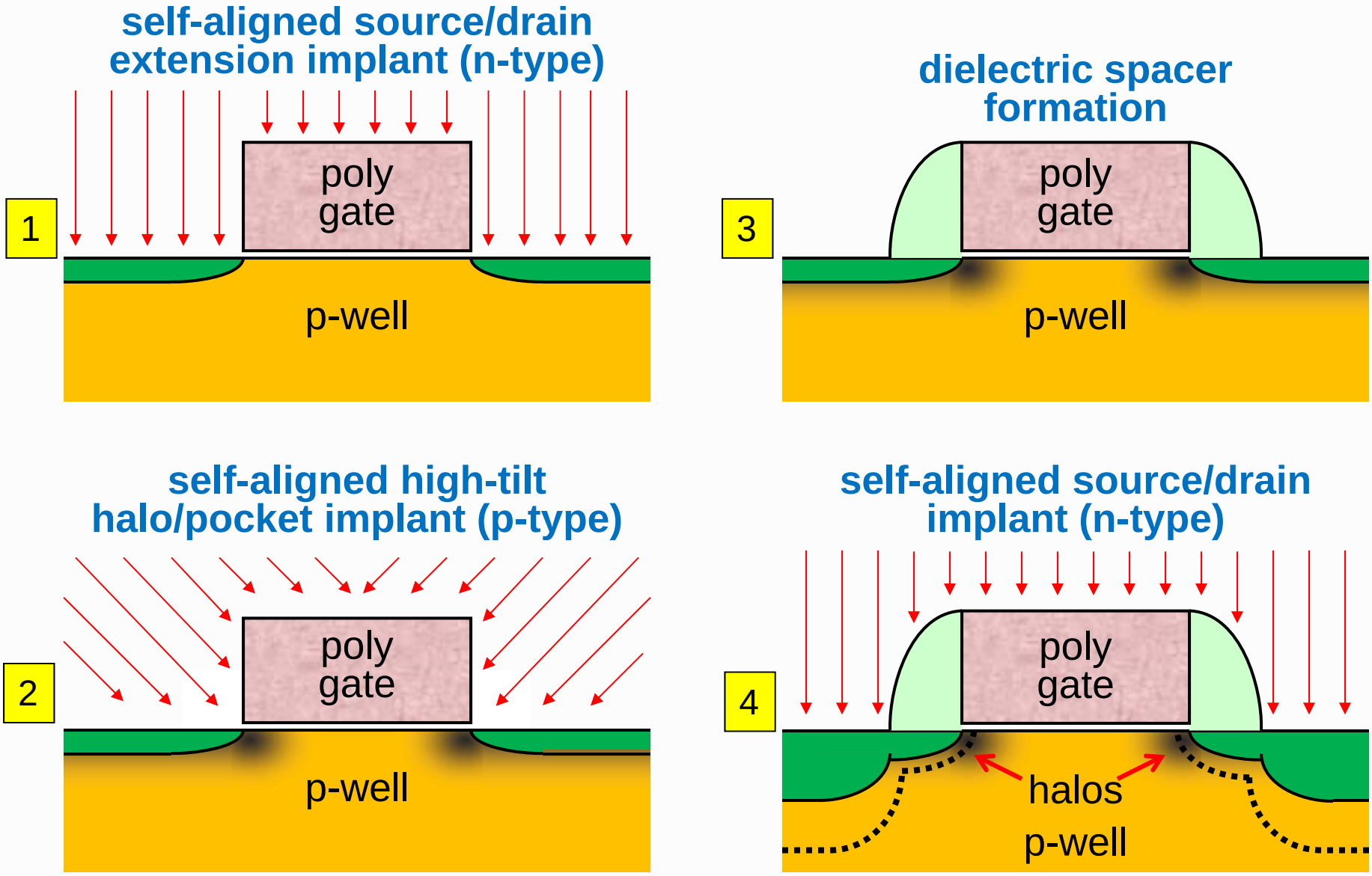
1 Pattern resist

2 Trim resist (oxygen ash)

3 Etch gate stack

- *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 → immersion litho

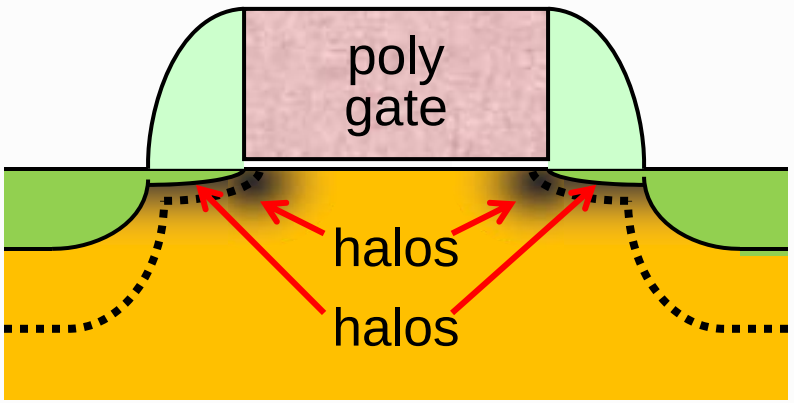
Channel & Source/Drain Engineering



Benefits of Halo and Extension

Resulting structure

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



Not to be confused with LDD in I/O FET

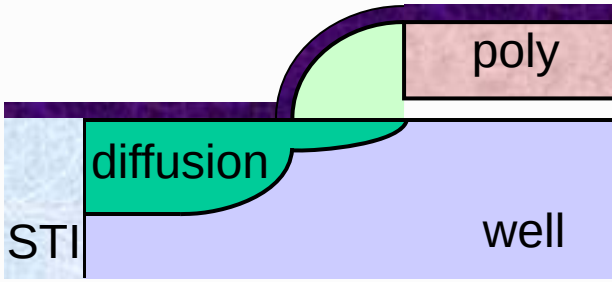
- Same process with spacers but lightly doped drain (LDD) is used for minimizing peak electric fields that cause hot carriers & breakdown
- Extensions need to be heavily doped to minimize series resistance

Different halo & extension/LDD implants for each FET variant

Self-Aligned Silicidation (Salicidation)

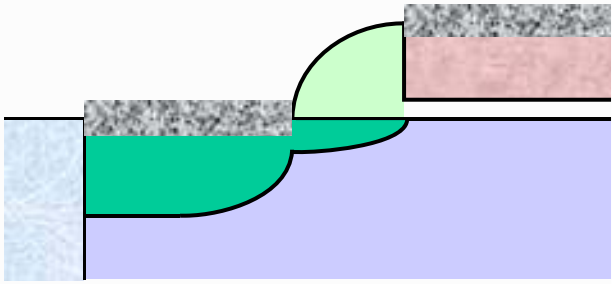
- Need to reduce poly & diffusion R_s, or get severe I_{FET} degradation

1



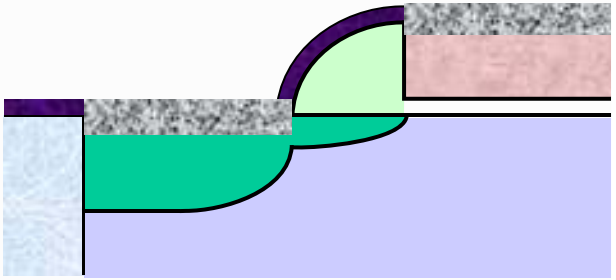
Deposit sicilide metal (Ti, Co, Ni)

3



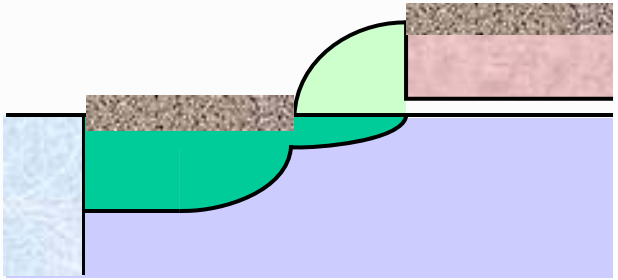
Strip unreacted metal

2



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

4



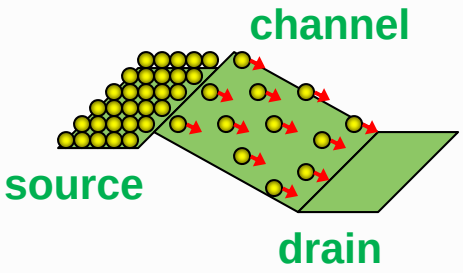
RTA2 (high temperature)
Transforms silicide into low-ρ phase by consuming more Si

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

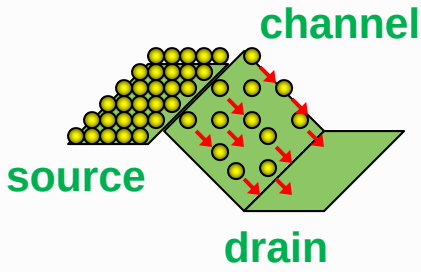
Outline

- Motivation
- MOSFET & Short-Channel Fundamentals
- 130nm MOSFET Fabrication
- **Strain Engineering (90nm & Beyond)**
- High-K / Metal-Gate (45nm & Beyond)
- Fully-Depleted (22nm & Beyond)
- Tri-Gate FinFETs
- Conclusions

The Roads to Higher Performance

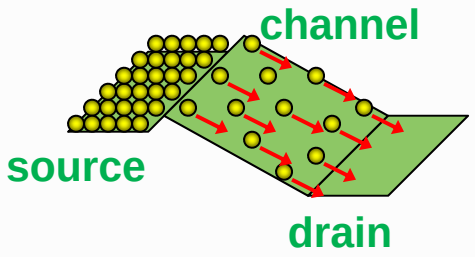


Decrease L – steepen the hill



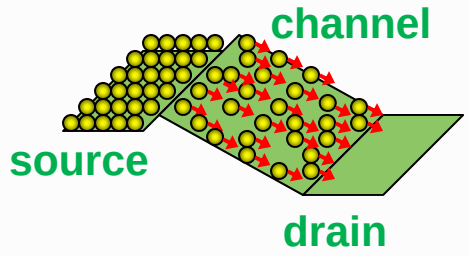
lithography
scaling

Increase μ – move carriers faster



strain engineering

Increase C_{ox} – move more carriers



high-K dielectric
metal gate

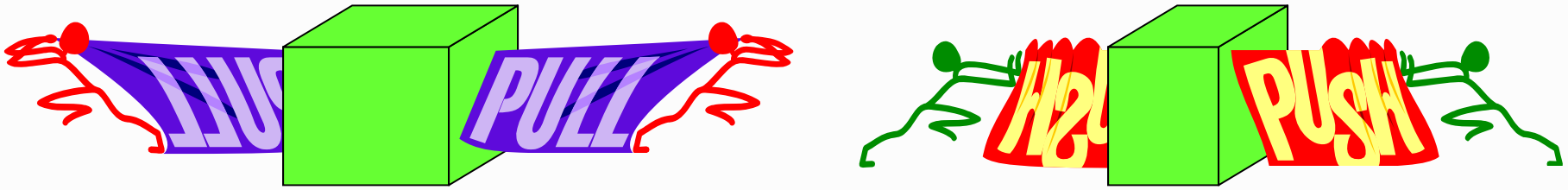
Mechanical Stresses & Strains

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

Tension
(positive stress)

vs.

Compression
(negative stress)



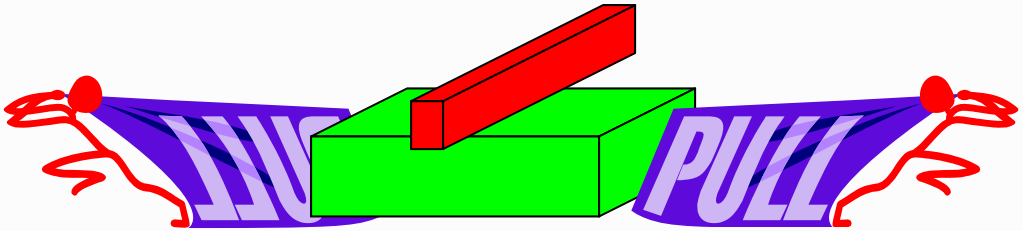
atomic spacing > equilibrium spacing

atomic spacing < equilibrium spacing

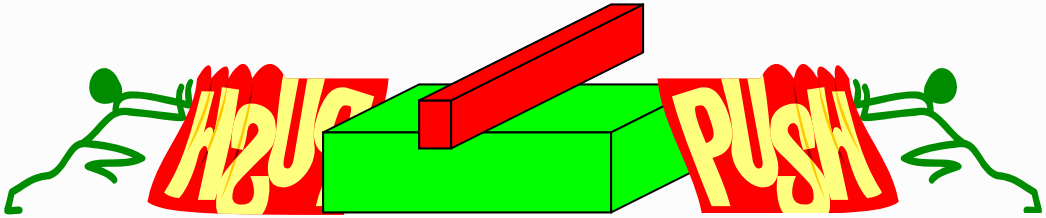
- Stretching / compressing FET channel atoms by *as little as 1%* can improve electron / hole mobilities by *several times*
- Strain perturbs crystal structure (energy bands, density of states, etc.) → changes effective mass of electrons & holes
- Increase I_{ON} for the same I_{OFF} without increasing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

Longitudinal Uni-Axial Strain

tension (stretch atoms apart) → faster NMOS

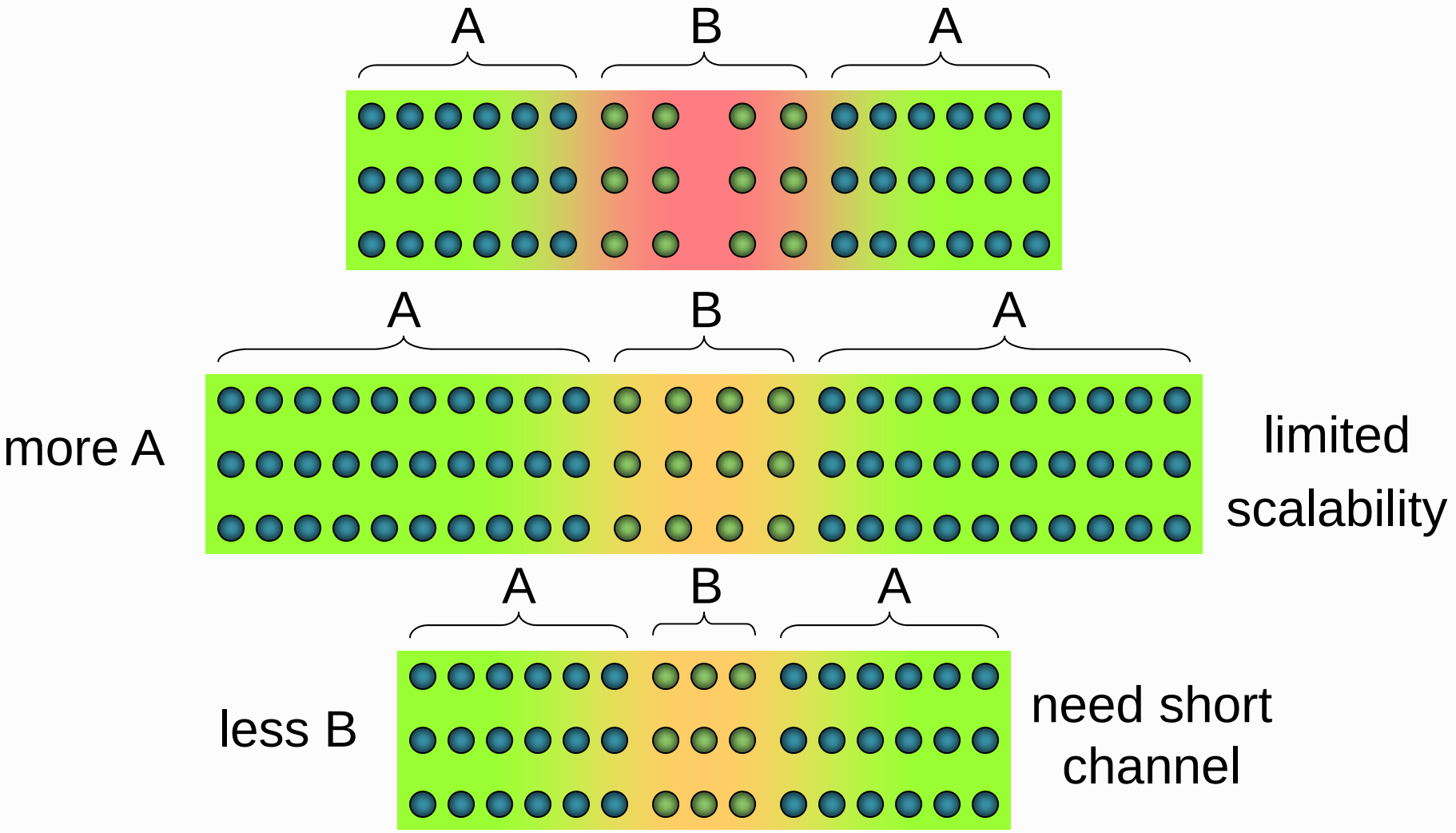


compression (squeeze atoms together) → faster PMOS



- 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



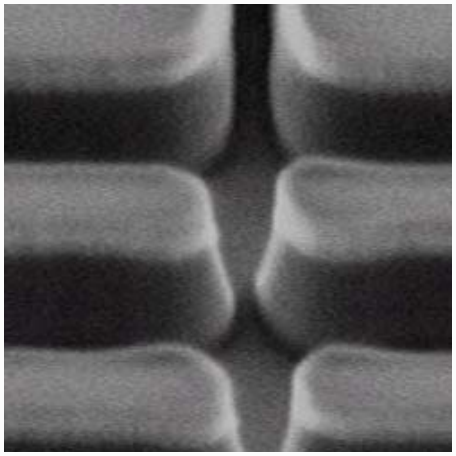
Ways to Incorporate Uni-Axial Strain

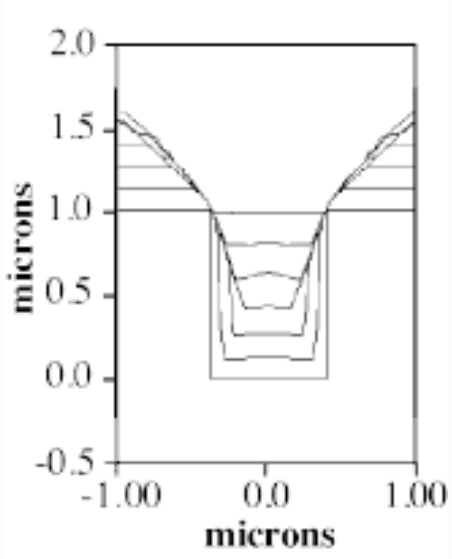
- NMOS wants tension, PMOS wants compression
- Un-Intentional (comes for *free*)
  - Shallow Trench Isolation – **NMOS** ☹ / **PMOS** ☺
- Intentional (requires extra processing)
  - Stress Memorization Technique – **NMOS** ☺
  - Embedded-SiGe Source/Drain – **PMOS** ☺
  - Embedded-SiC Source/Drain –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

Shallow Trench Isolation (STI)

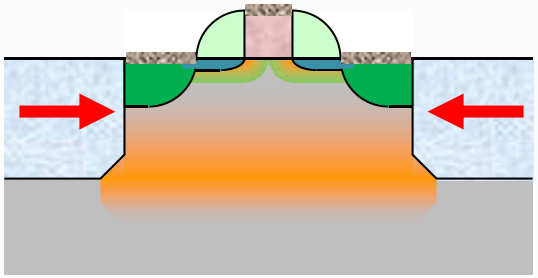
NMOS ☹ & PMOS ☺

- 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





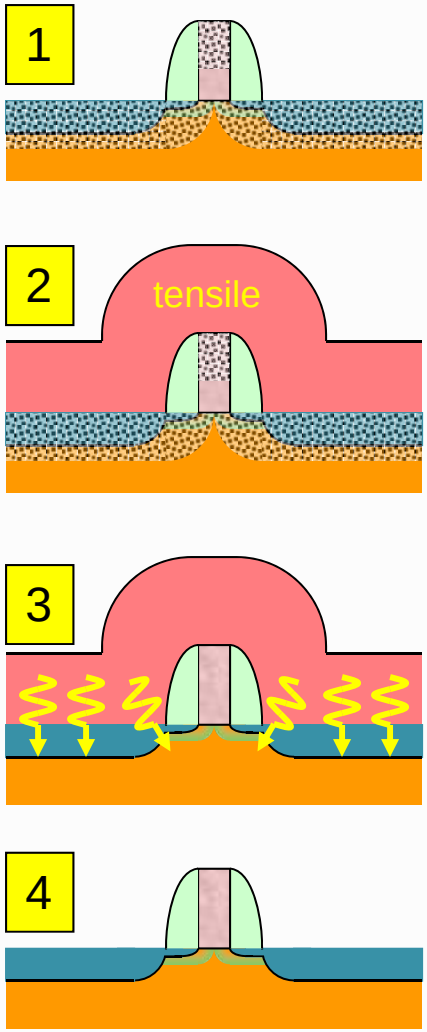
Plummer *et al.*, Stanford [6]



Bianchi *et al.*, AMD [7]

Stress Memorization Technique (SMT)

NMOS ☺

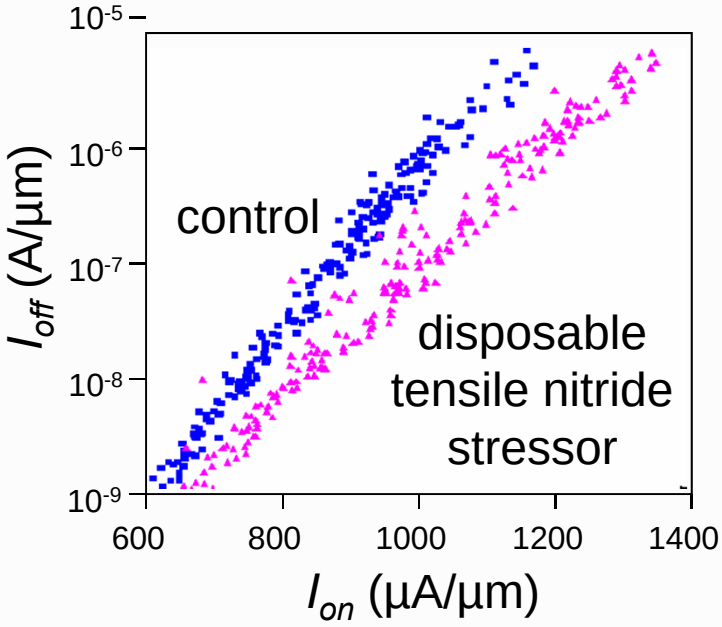


Amorphize poly & diffusion with silicon implant

Deposit tensile nitride

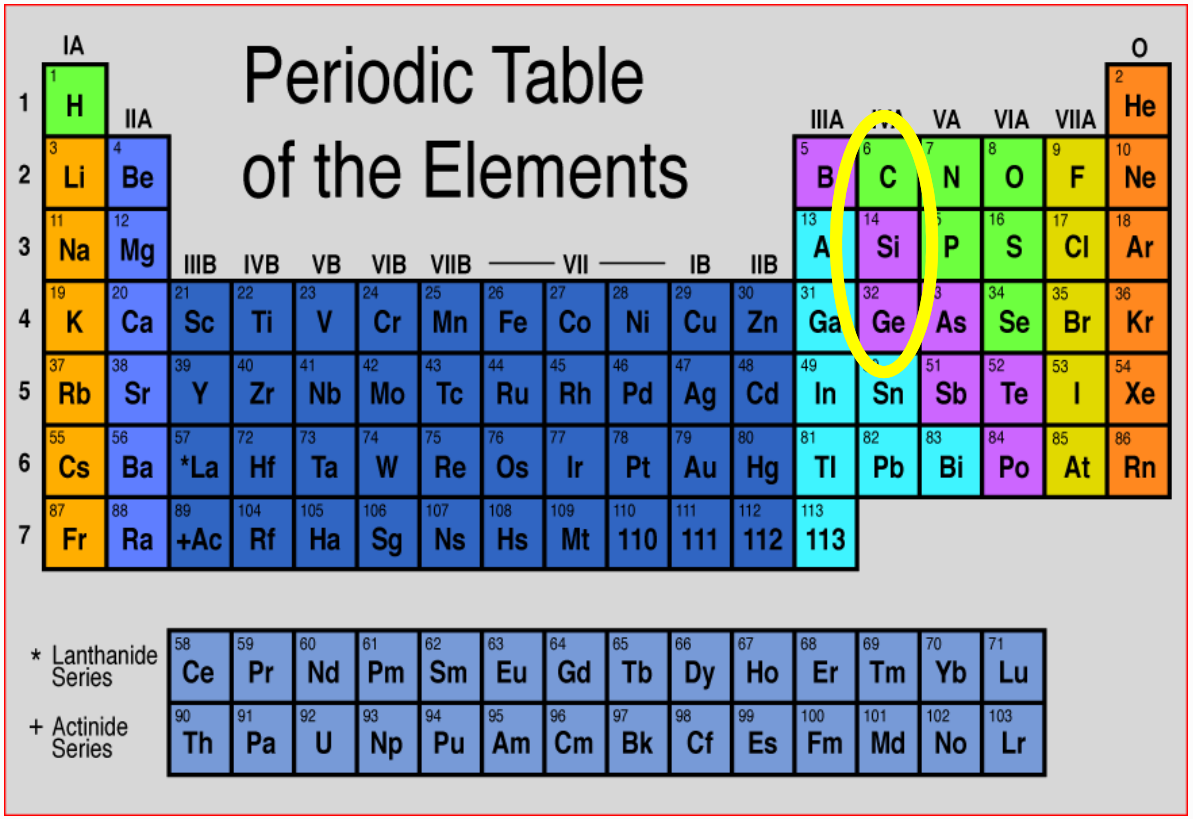
Anneal to *make nitride more tensile* and transfer nitride tension to crystallizing amorphous channel

Remove nitride stressor (tension now frozen in diffusion)



Chan *et al.*, IBM [8]

Periodic Table Trends



lattice spacing ↑

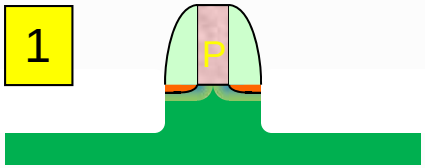
bandgap ↓

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

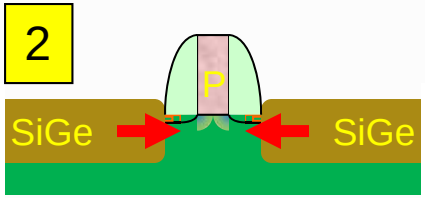
Embedded-SiGe Source/Drain (e-SiGe)

PMOS ☺

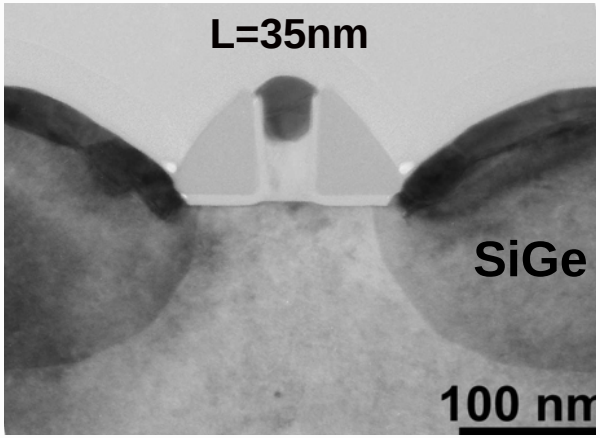
- SiGe constrained to Si lattice will be in compression
- Compressive SiGe source/drain transfers compression to Si channel



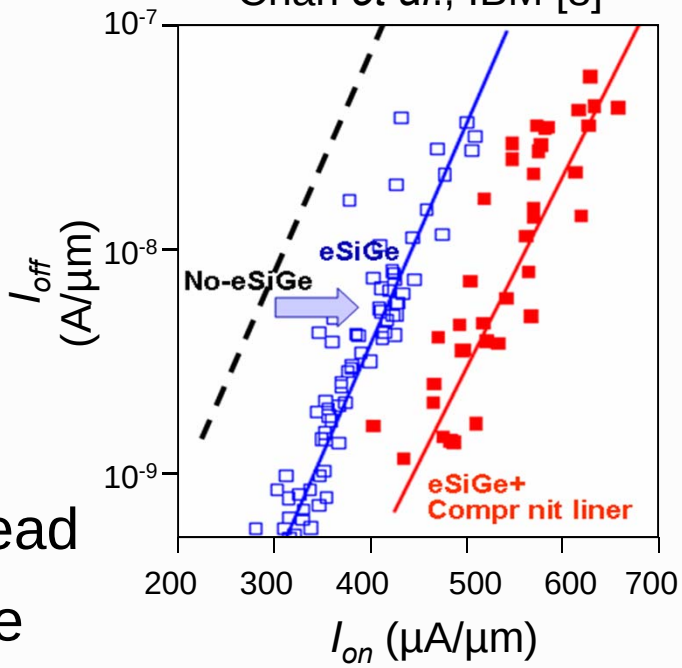
Etch source/drain recess



Grow SiGe epitaxially in recessed regions



Chan *et al.*, IBM [8]



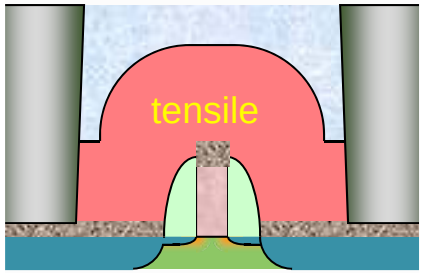
- e-SiC is similar but introduces tension instead
- Epitaxial SiC much tougher to do than SiGe

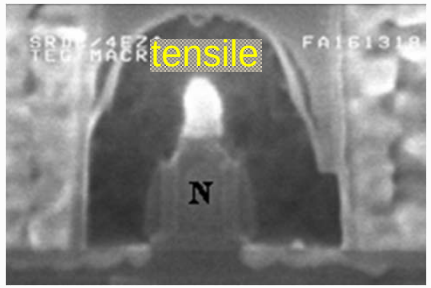
Dual-Stress Liners

NMOS ☺ & PMOS ☺

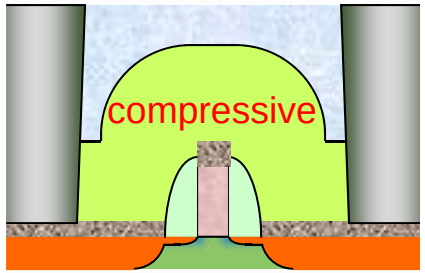
- Deposit tensile/compressive PECVD SiN (PEN) liners on N/PMOS
- Liner stress is dialed in by liner deposition conditions (gas flow, pressure, temperature, etc.)

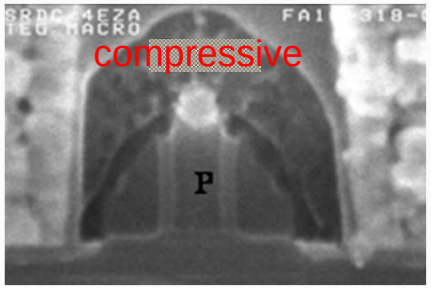
TPEN for NMOS



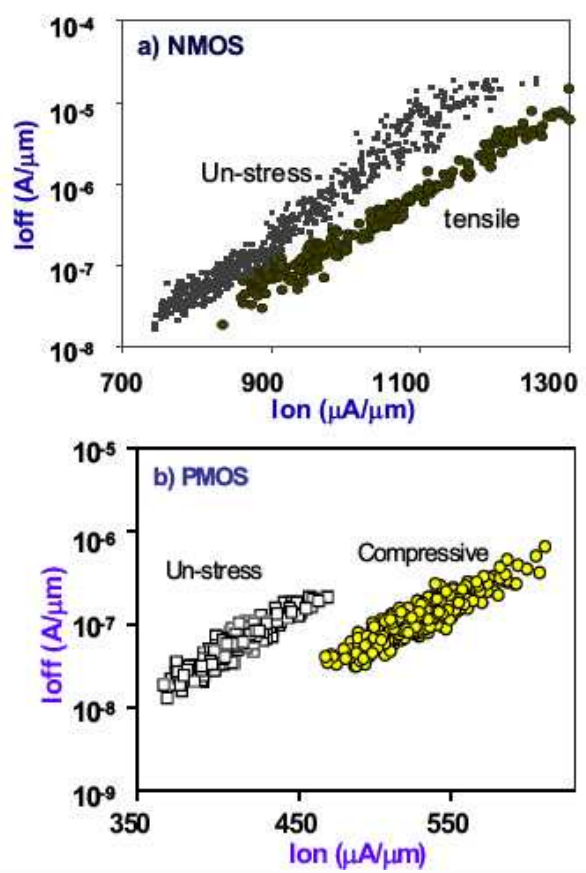


CPEN for PMOS



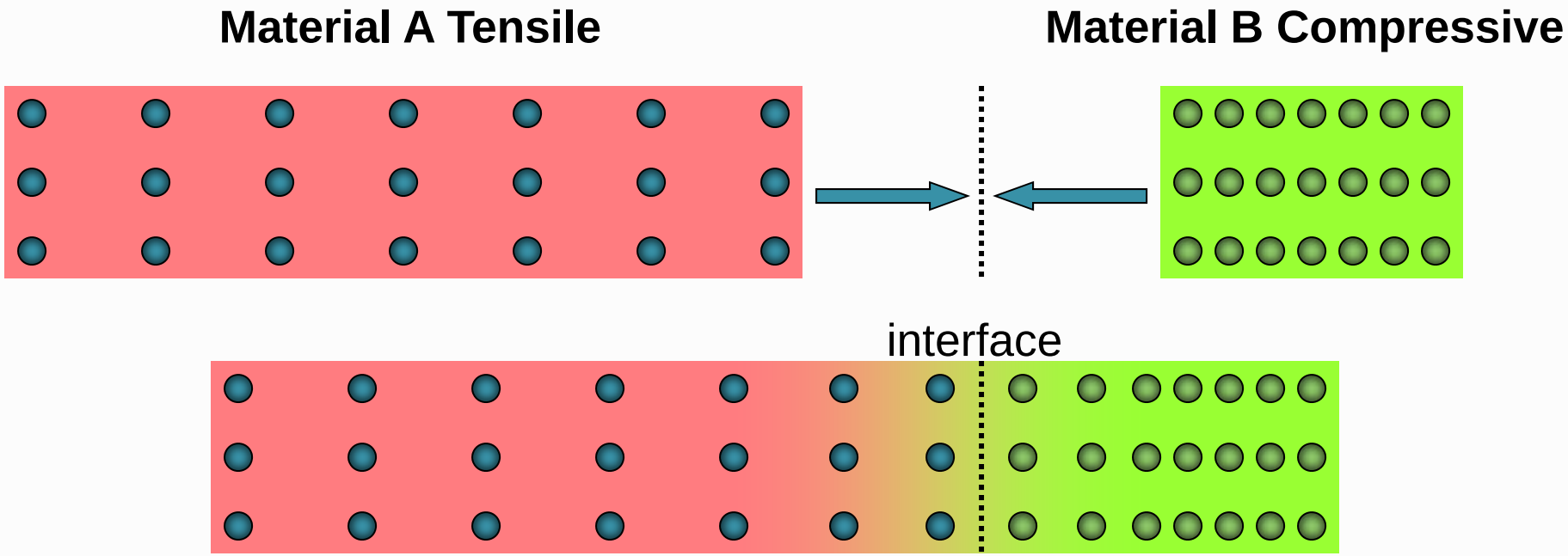


Chan *et al.*, IBM [8]



Strain Relaxation

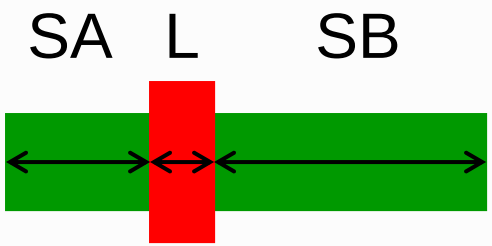
When materials of different strain come together…



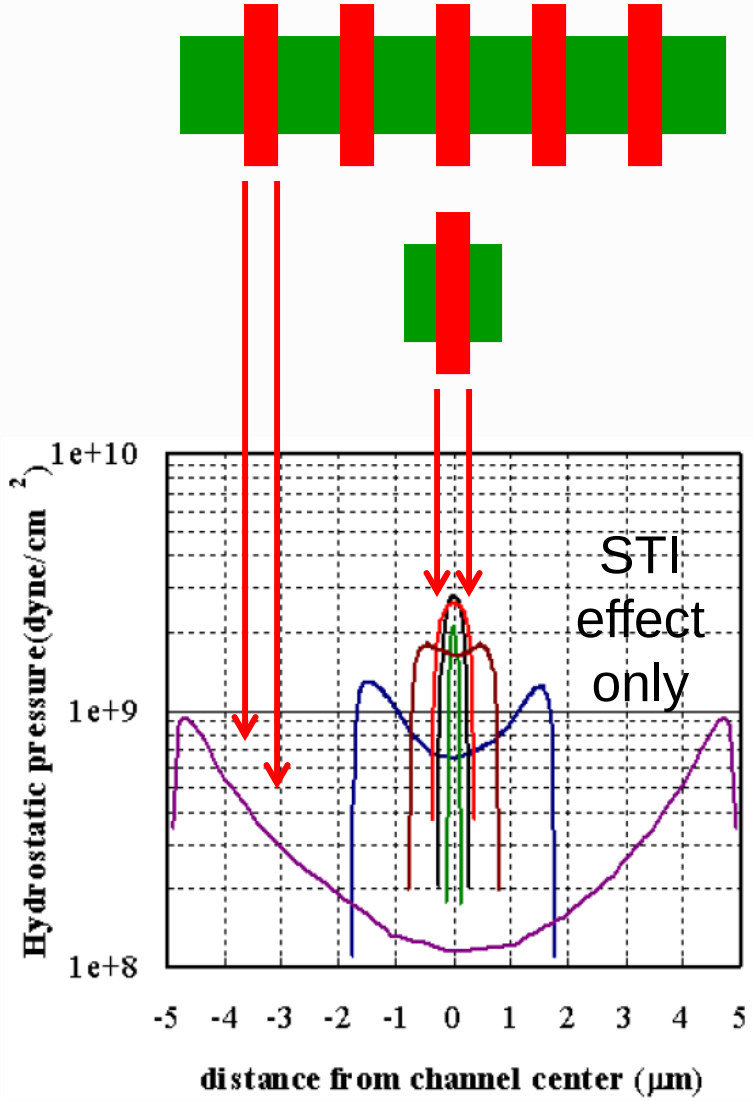
- Both materials will relax at the interface
- Extent of relaxation is gradual, depends on distance from interface
- No relaxation far away from interface

Strain Depends on Channel Location

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



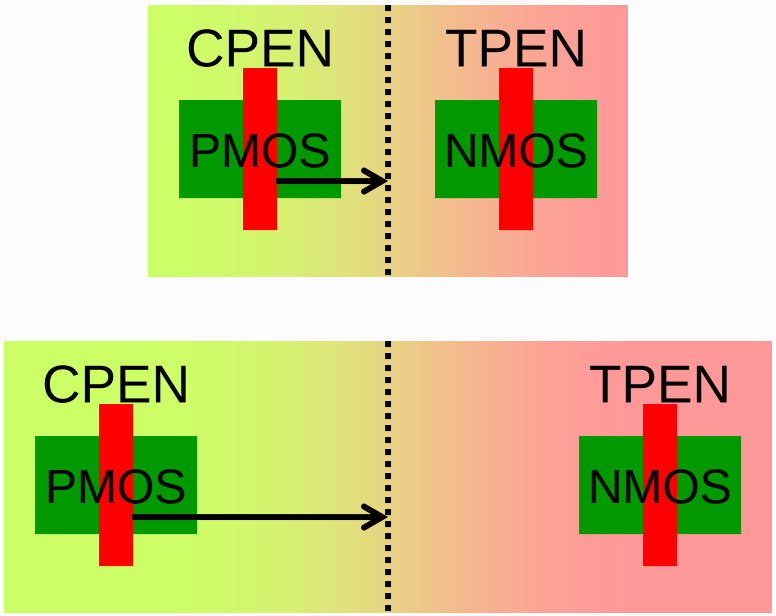
- Critical for modeling device mobility change due to STI, SMT & e-SiGe/e-SiC
- Strain at source & drain ends of channel may be different
- Important consideration for matching, e.g., current mirrors



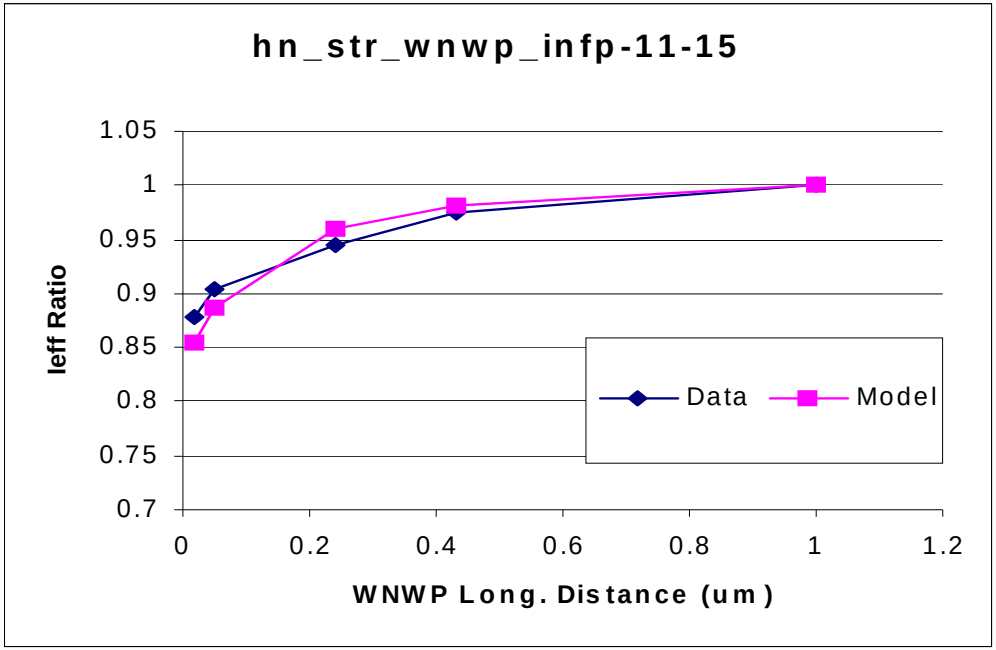
Xi *et al.*, UC Berkeley [9]

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



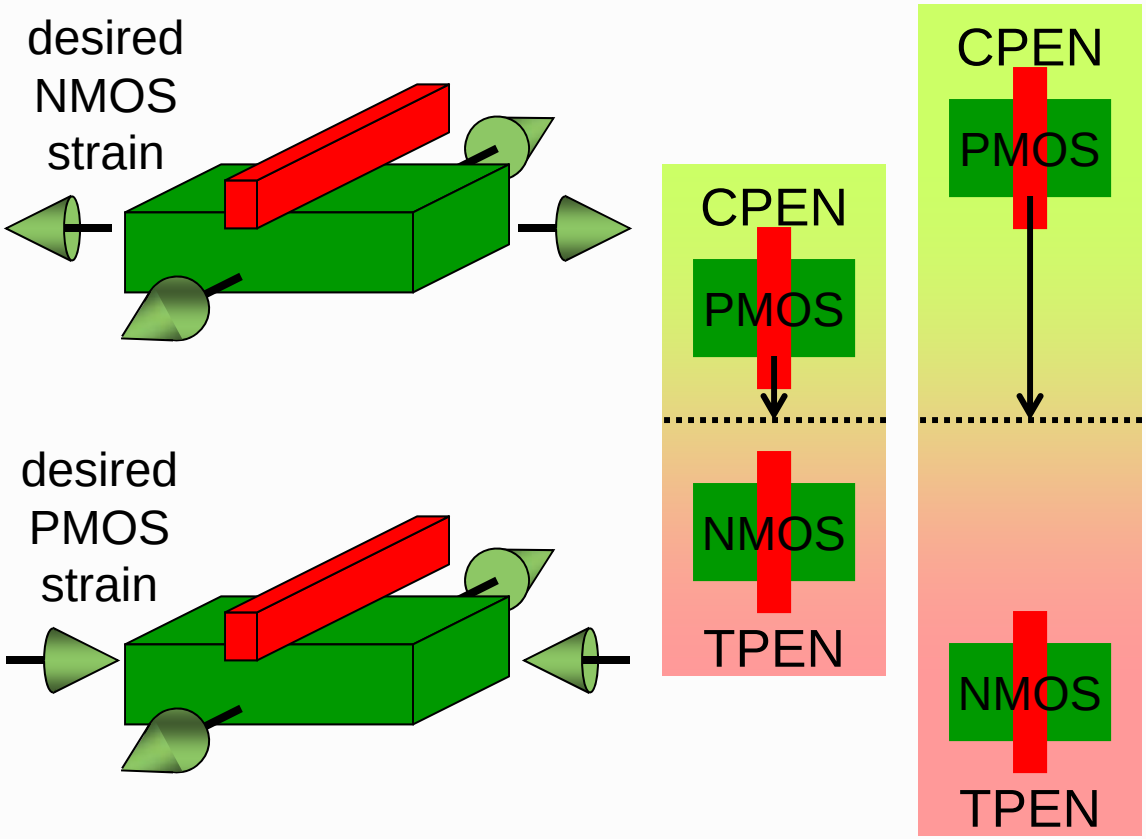
PMOS Longitudinal Proximity



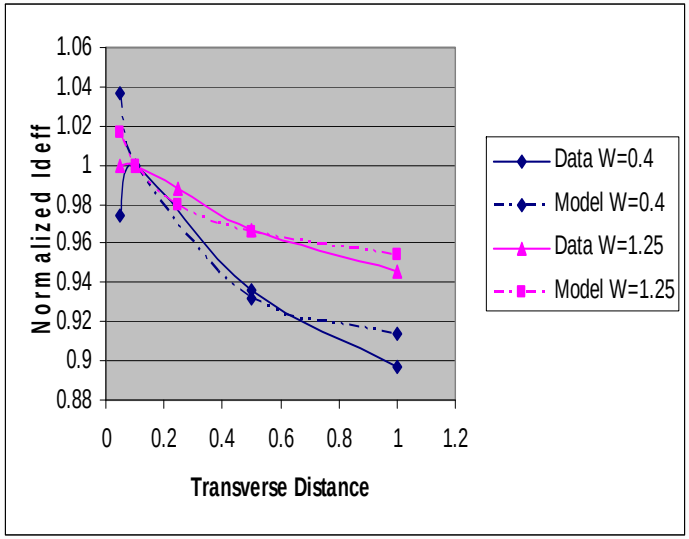
Faricelli, AMD [10]

Transverse DSL Proximity

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



PMOS Transverse Proximity

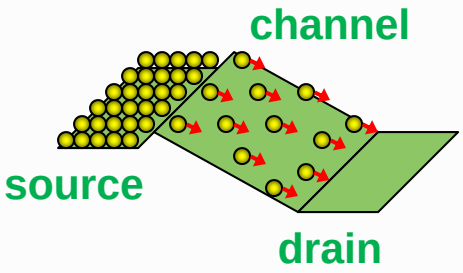


Faricelli, AMD [10]

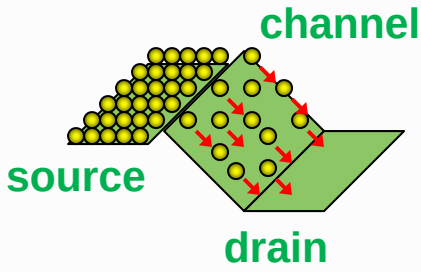
Outline

- Motivation
- MOSFET & Short-Channel Fundamentals
- 130nm MOSFET Fabrication
- Strain Engineering (90nm & Beyond)
- **High-K / Metal-Gate (45nm & Beyond)**
- Fully-Depleted (22nm & Beyond)
- Tri-Gate FinFETs
- Conclusions

The Roads to Higher Performance

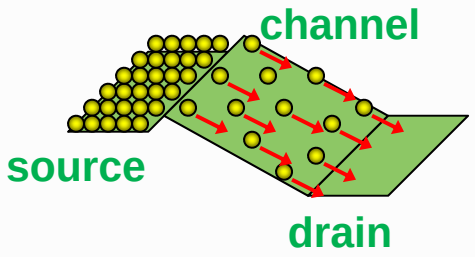


Decrease L – steepen the hill



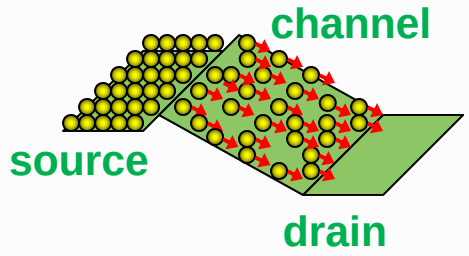
lithography
scaling

Increase μ – move carriers faster



strain engineering

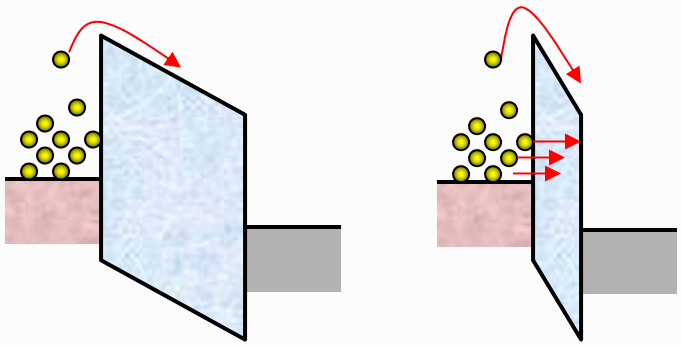
Increase C_{ox} – move more carriers



high-*K* dielectric
metal gate

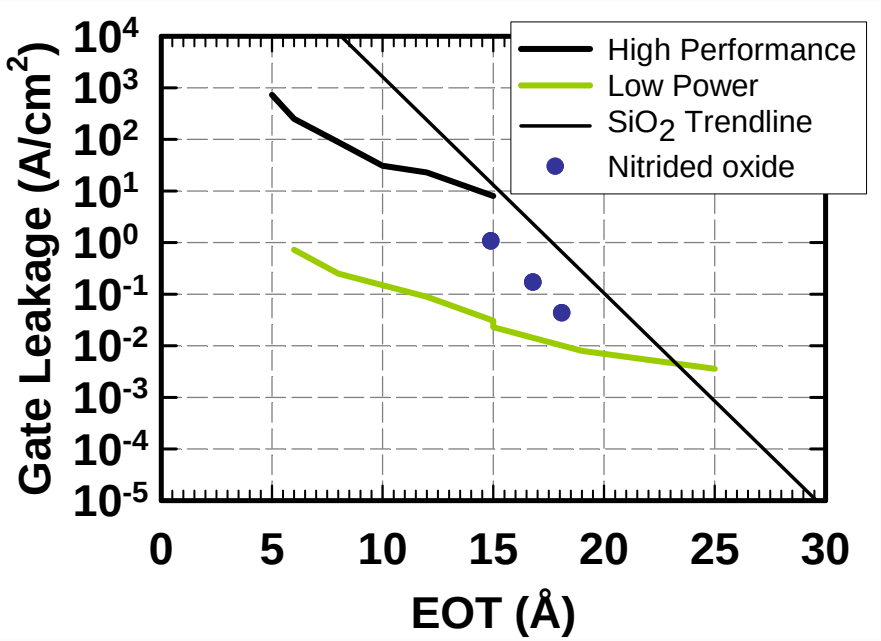
Direct Tunneling Gate Leakage

- 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 [11]

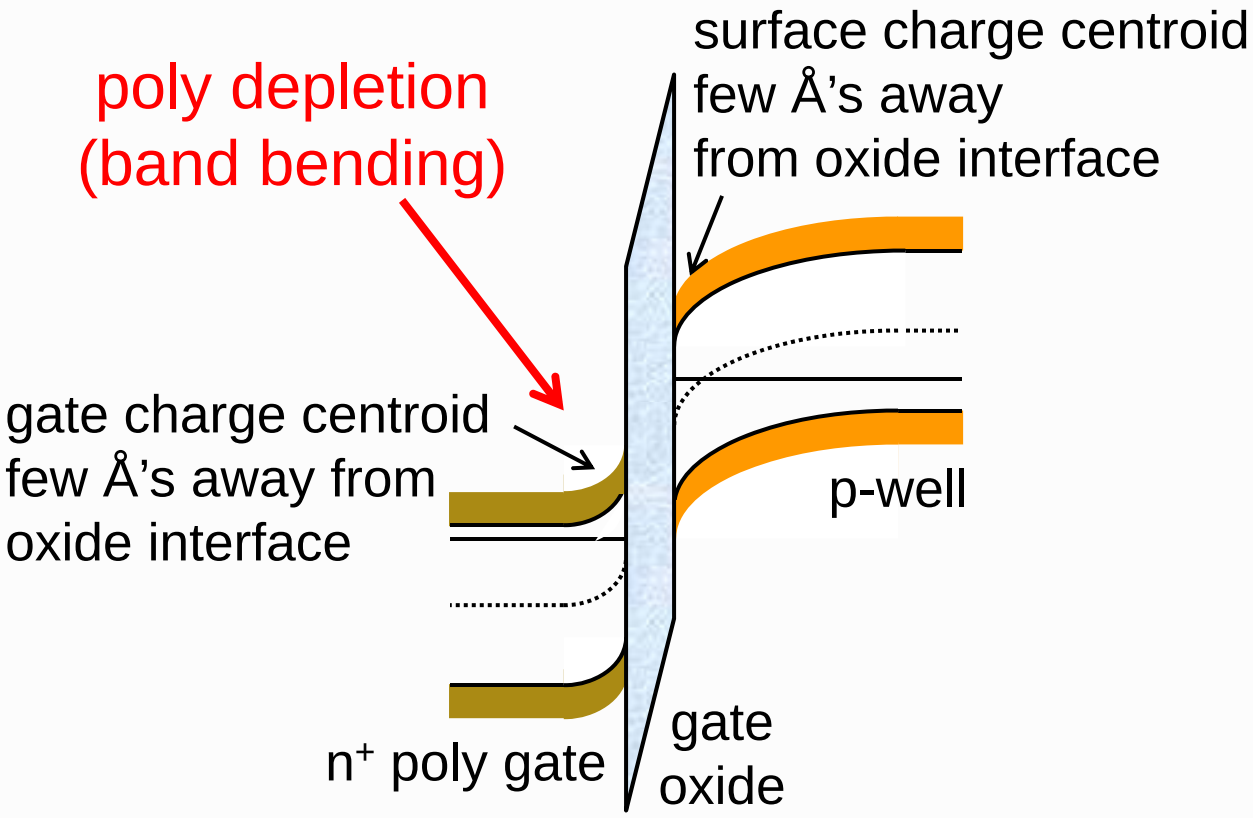


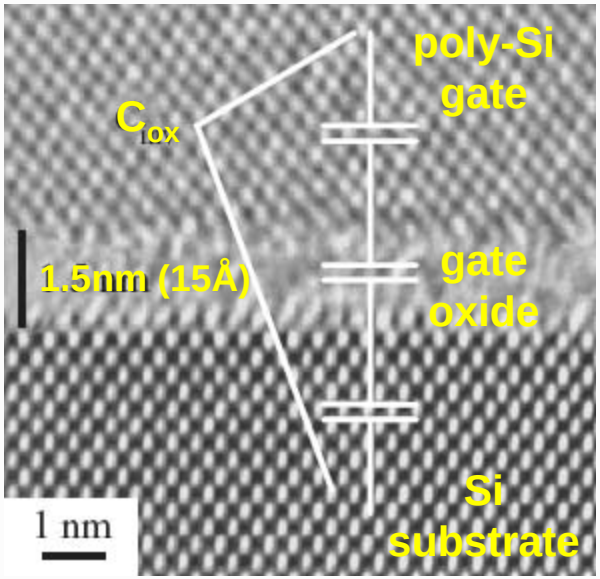
EOT = Equivalent Oxide Thickness

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

Poly Depletion & Charge Centroid

Dielectric Only Half the Story





Wong, IBM [12]

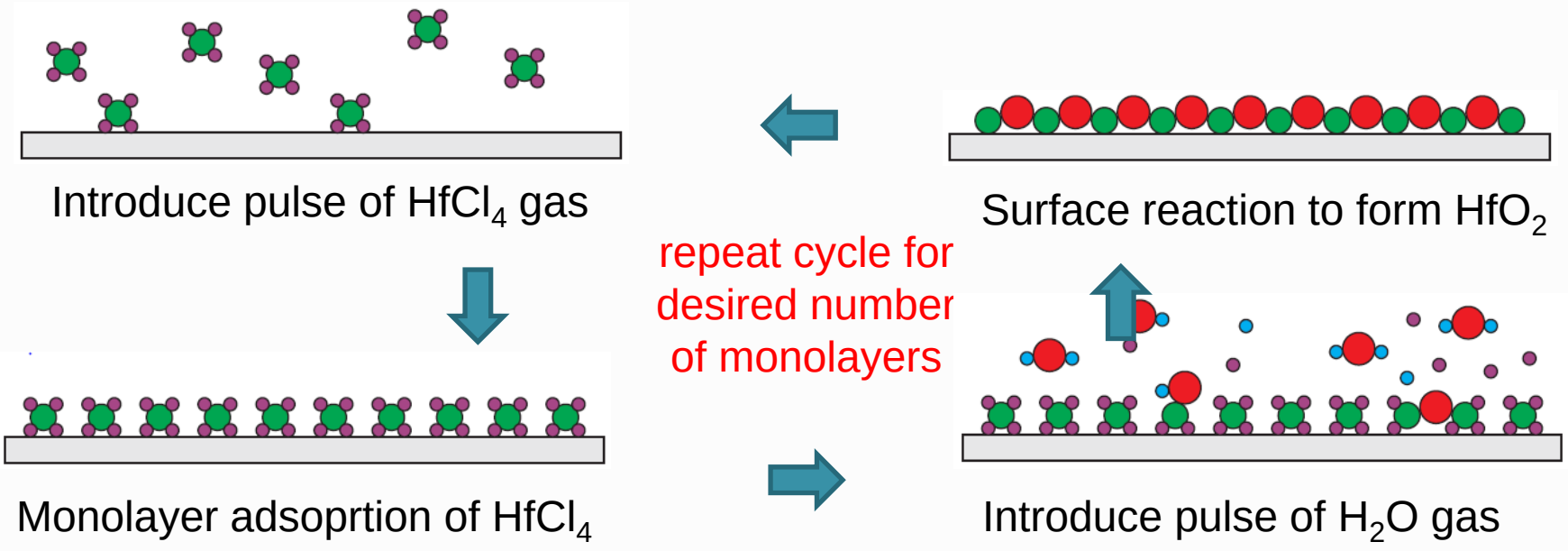
- Even heavily-doped poly is a limited conductor
- Discrepancy between electrical & physical thicknesses since charge is not intimately in contact with oxide interface

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 $\Phi_M \rightarrow V_{FB} \rightarrow V_T$
  - 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 [13]

HK-First / MG-First Integration

1

Deposit HK
Deposit MG1

2

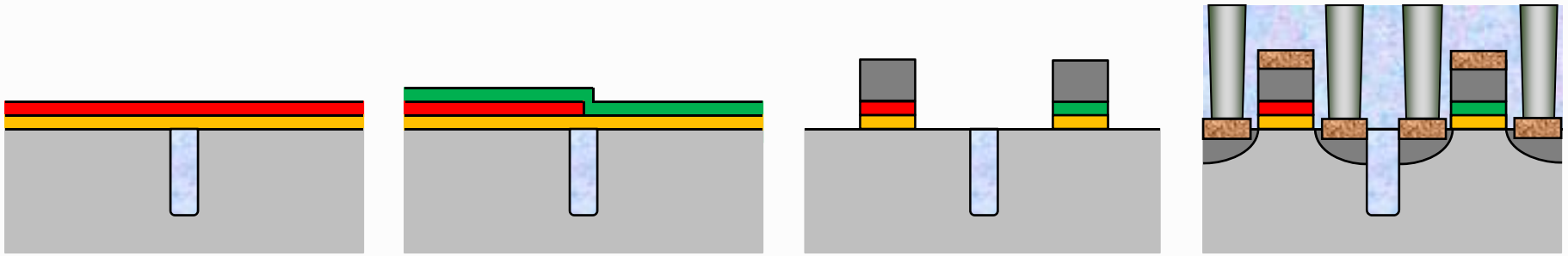
Pattern MG1
Deposit MG2

3

Pattern MG2
Deposit gate
Pattern gates / MGs / HK

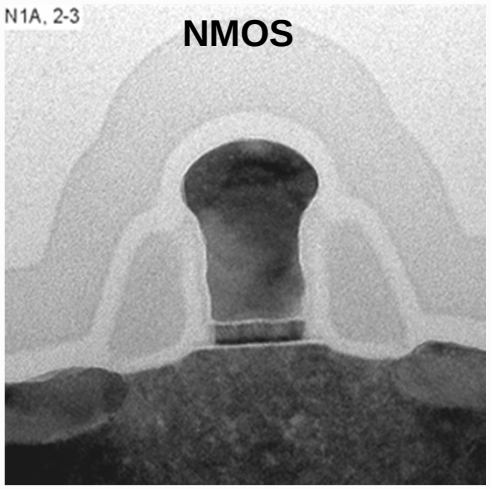
4

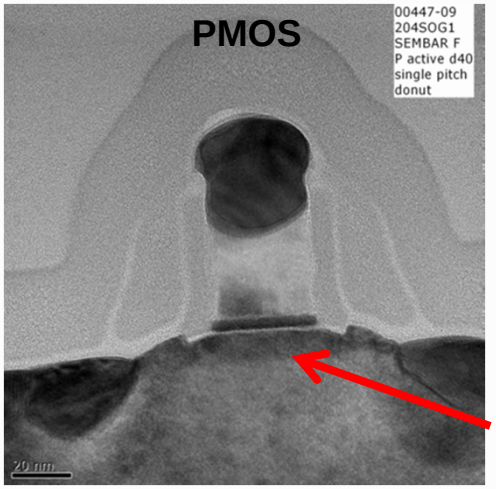
Implant/anneal S/ D
Form silicide
Deposit/CMP ILD0
Form contacts



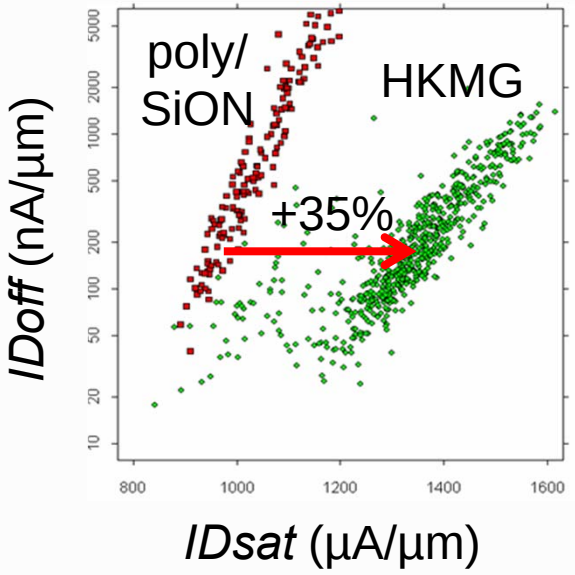
- Obvious extension of poly-Si gate integration
- Seems obvious & “easy” at first but plagued with unstable work function when HKMG is exposed to activation anneals
- Especially problematic with PMOS V_T coming out too high

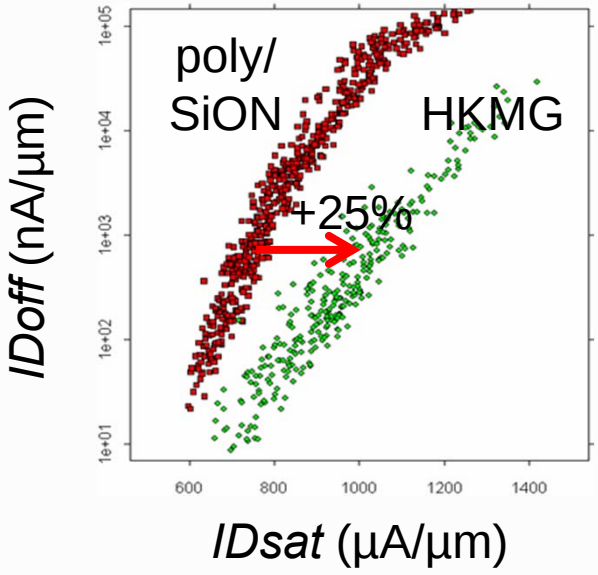
GlobalFoundries 32nm-SOI





epi-cSiGe to set channel Φ_M



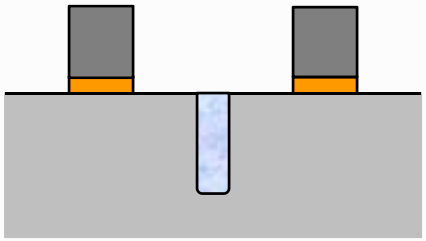


Horstmann *et al.*, GlobalFoundries [14]

HK-First / MG-Last Integration

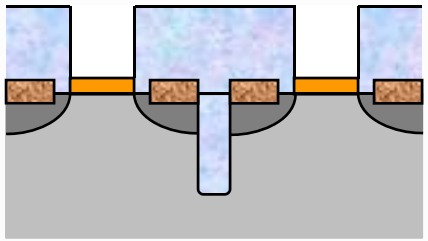
1

Deposit HK / gate
Pattern gate / HK



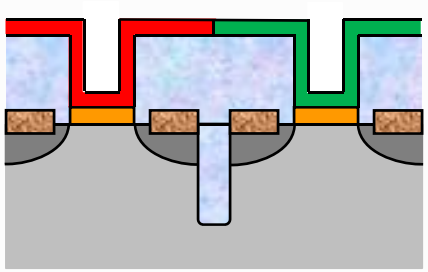
2

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



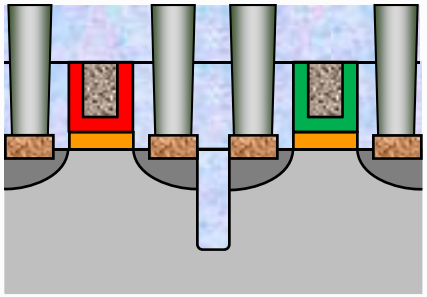
3

Deposit MG1
Pattern MG1
Deposit MG2
Pattern MG2



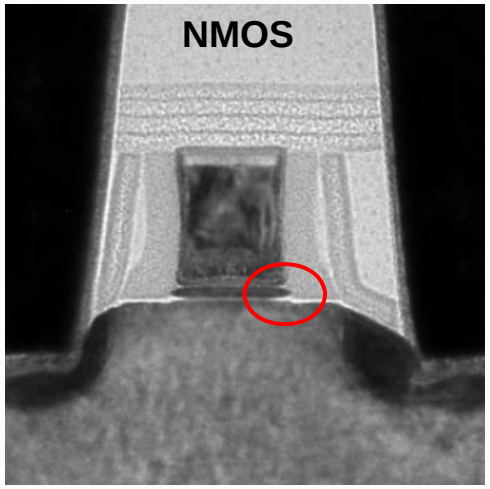
4

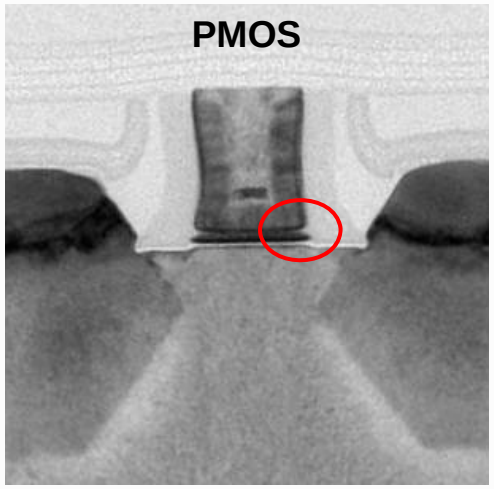
Deposit gate-fill
CMP gate-fill / MGs
Deposit more ILD0
Form contacts

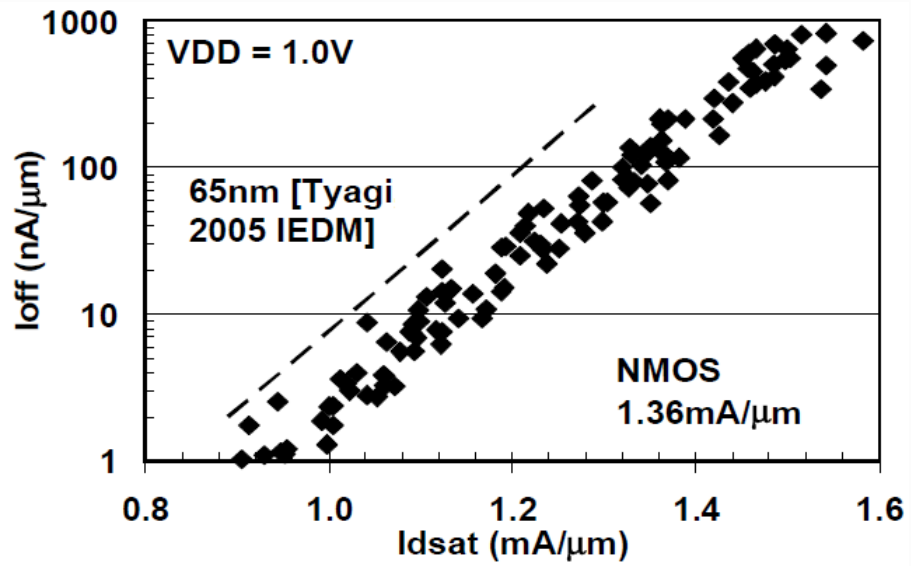


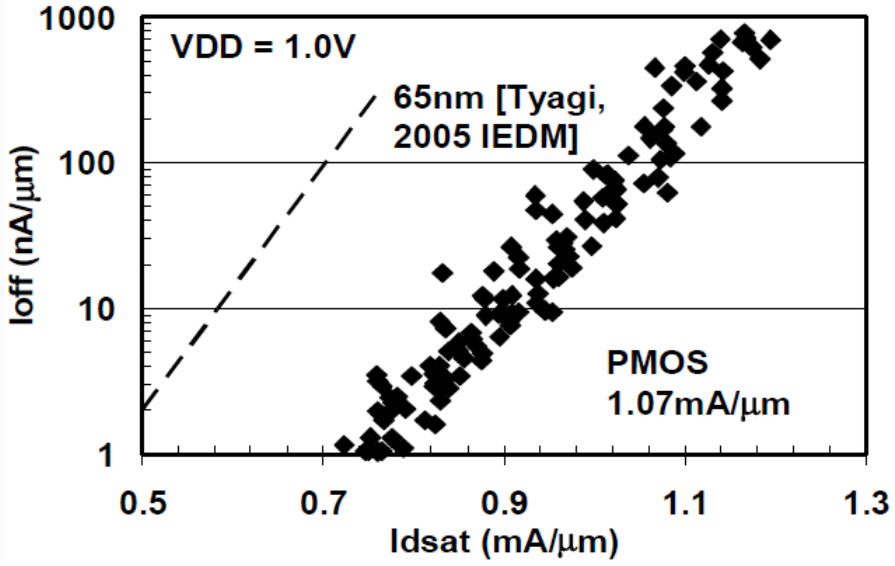
- High thermal budget available for middle-of-line
- Low thermal budget for metal gate → more gate metal choices
- Enhanced strain when sacrificial poly is removed & resulting trench is filled with gate fill metal

Intel 45nm







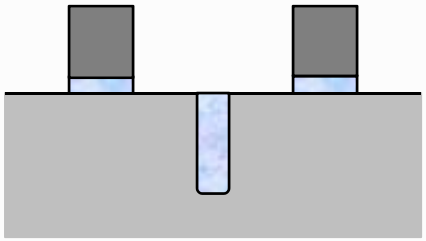


Auth *et al*, Intel [15]

HK-Last / MG-Last Integration

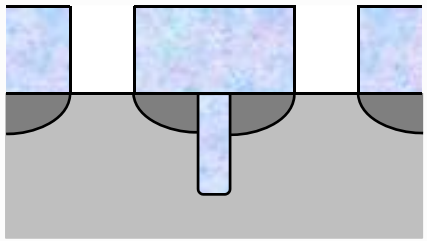
1

Deposit oxide / gate
Pattern gate / oxide



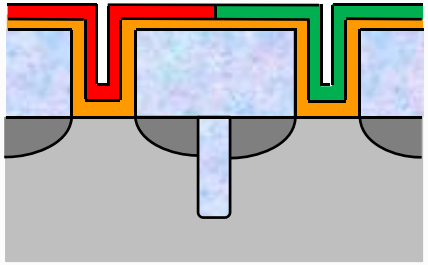
2

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



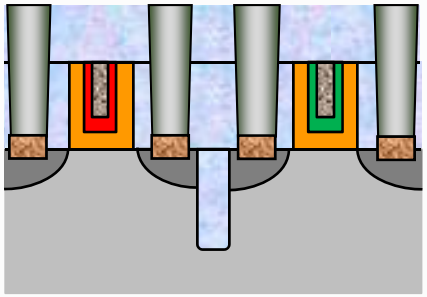
3

Deposit HK
Deposit MG1
Pattern MG1
Deposit MG2
Pattern MG2



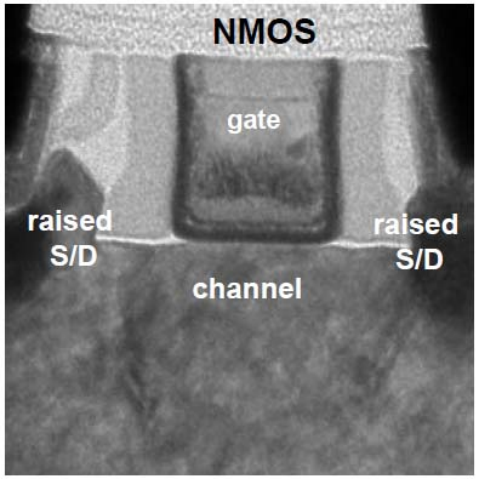
4

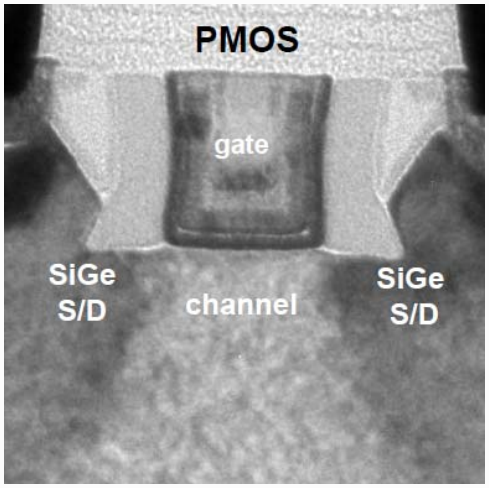
Deposit gate-fill
CMP gate-fill / MGs
Deposit more ILD0
Pattern contacts
Form silicide
Form contacts

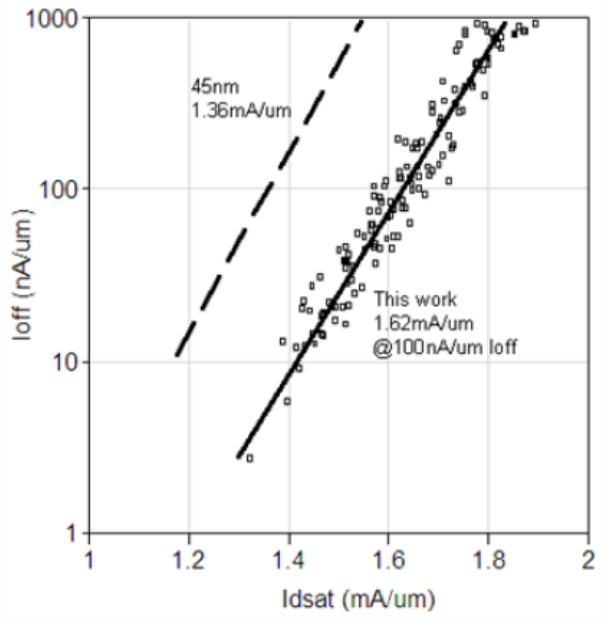


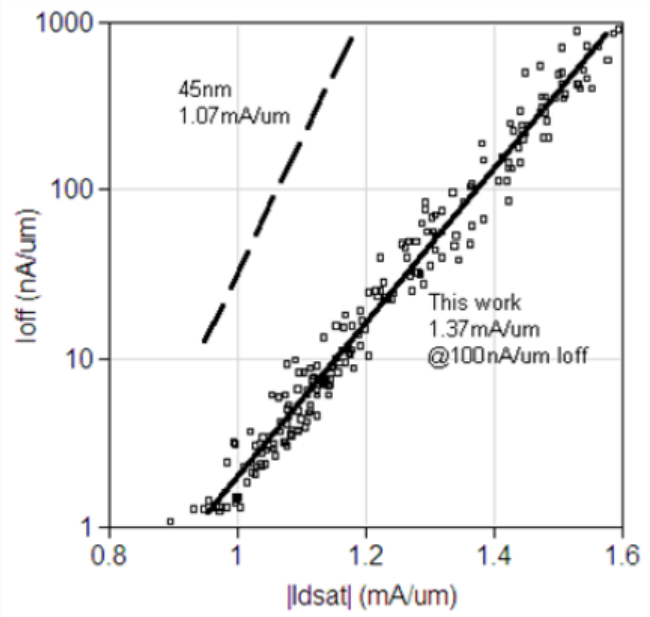
- Same advantages as HK-first / MG-last integration
- Overcomes EOT scaling limitations in HK-first / MG-last
- Need to postpone silicidation to after contact etch
- Need trench contacts to accommodate silicide metal deposition
- DSL relax & no longer useful since contacts cut through FET width

Intel 32nm









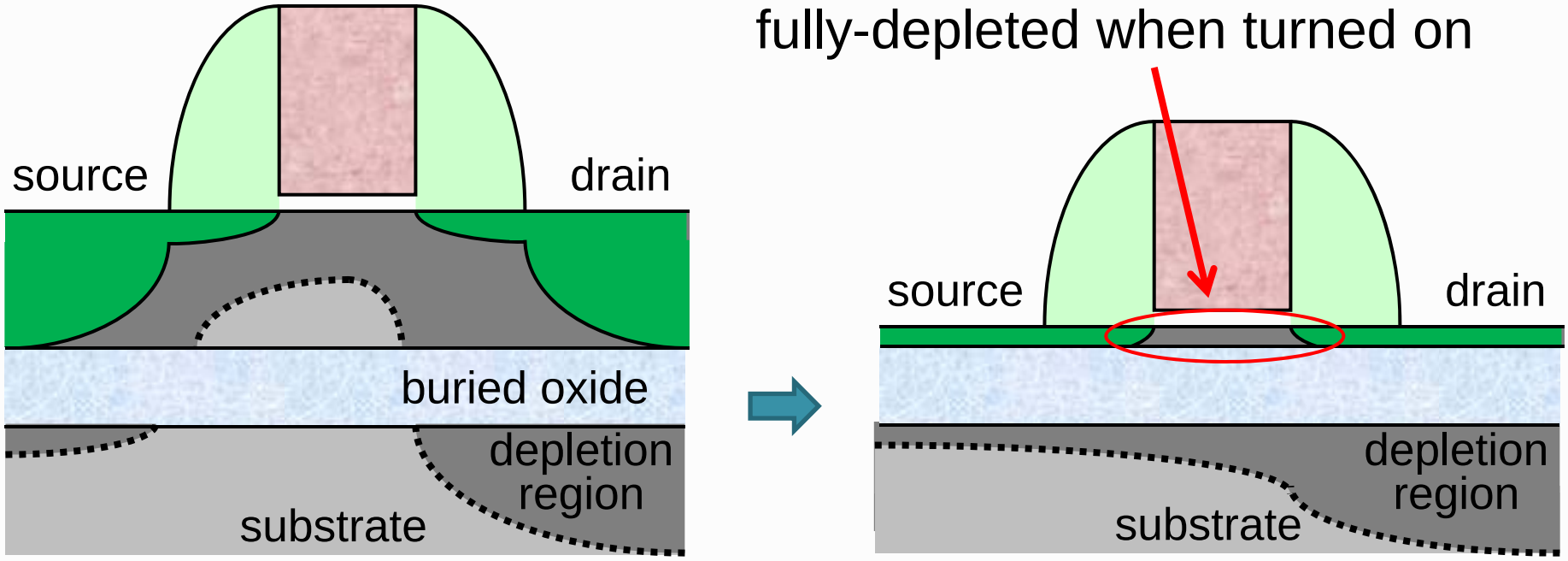
Packan *et al*, Intel [16]

Outline

- Motivation
- MOSFET & Short-Channel Fundamentals
- 130nm MOSFET Fabrication
- Strain Engineering (90nm & Beyond)
- High-K / Metal-Gate (45nm & Beyond)
- **Fully-Depleted (22nm & Beyond)**
- Tri-Gate FinFETs
- Conclusions

What Does Fully-Depleted Really Mean?

- Consider what happens when SOI layer thins down



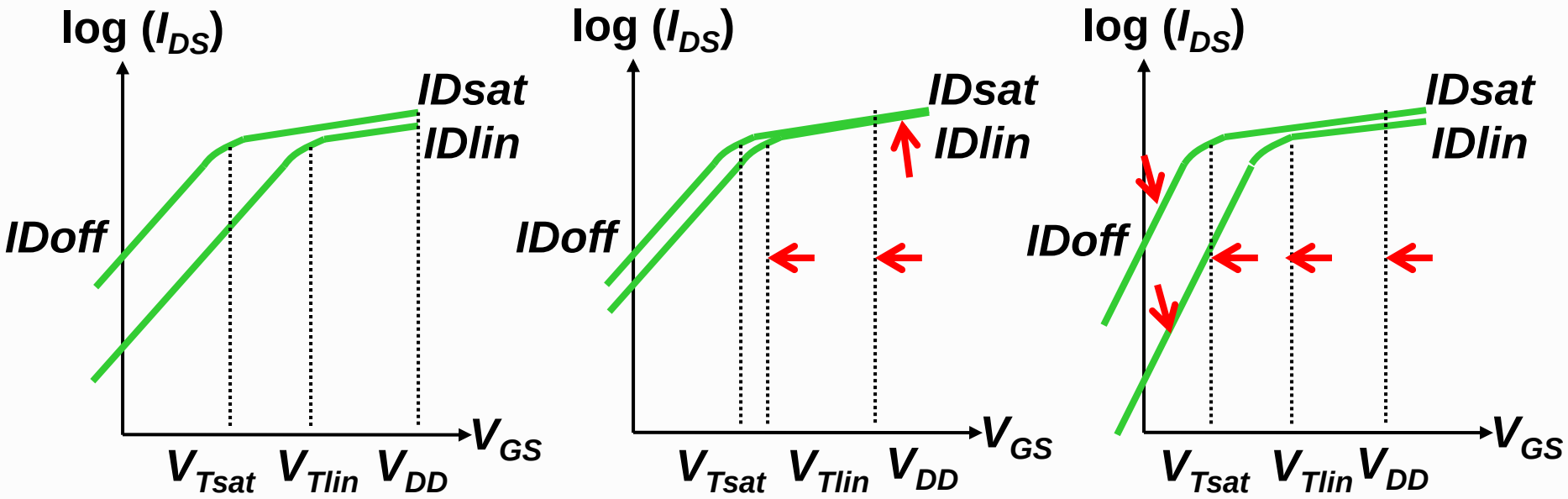
- Conservation of charge cannot be violated
- So once body is fully depleted, extra gate charge must be balanced by charge elsewhere, e.g., beneath buried oxide
- If substrate is insulator, then charge must come from source/drain

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

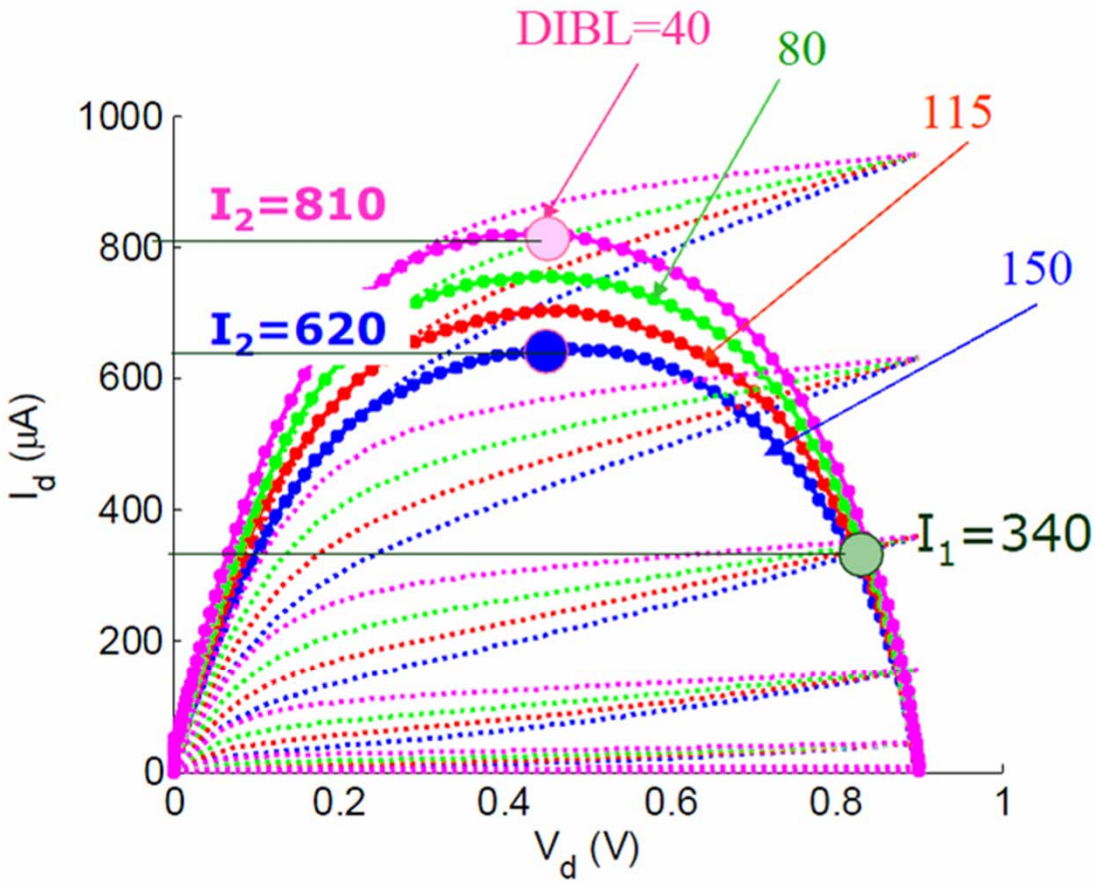
The Big Deal with Lower DIBL

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

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

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

Lower DIBL
=
Higher Performance

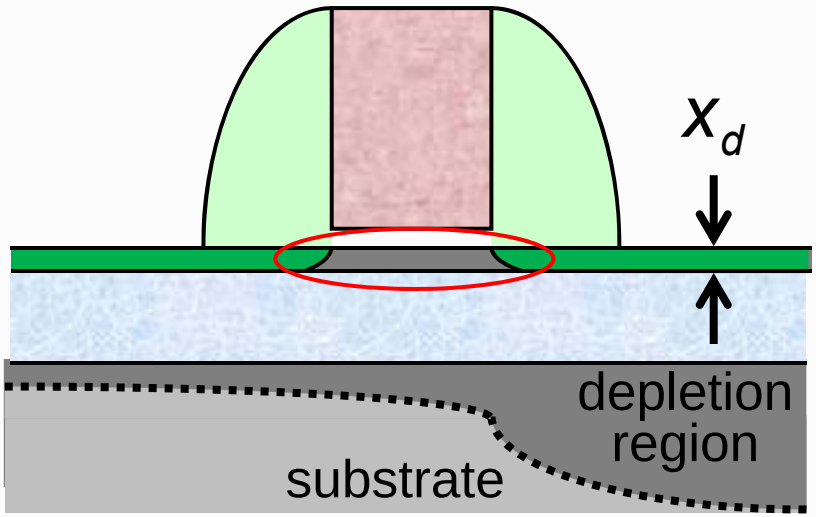


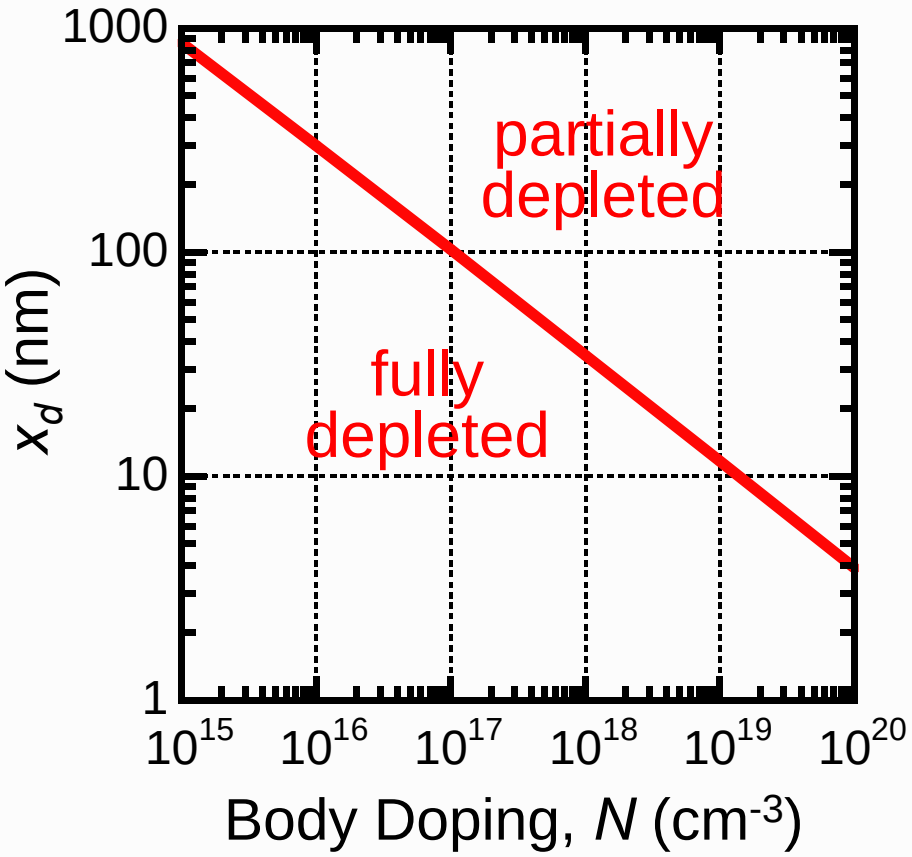
Higher performance for the same *IDsat* & *IDoff*

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

Body Thickness for Fully-Depleted

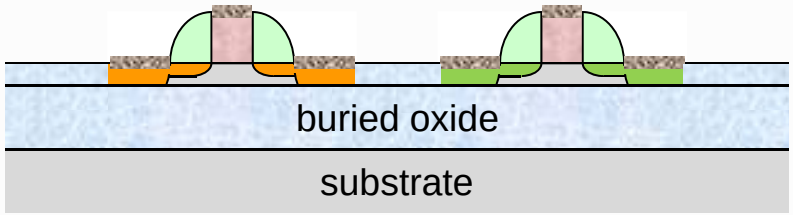
$$x_d < \sqrt{\frac{2\varepsilon_{Si} \cdot 2\phi_b}{qN}}$$

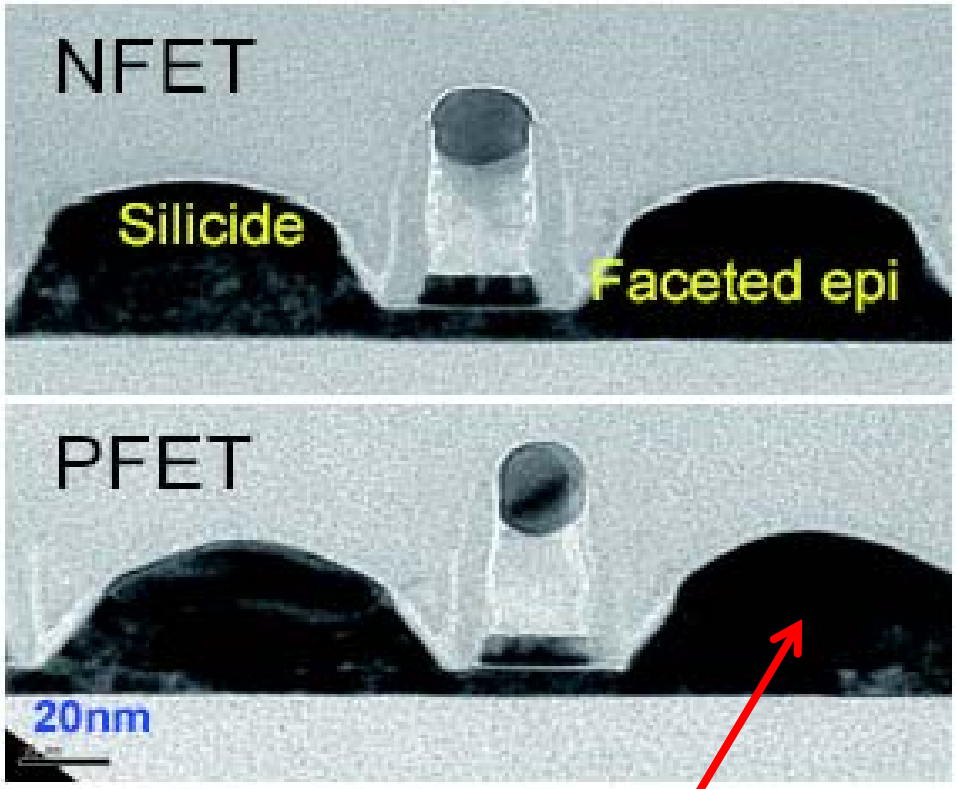




Fully-Depleted Planar on SOI

- 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

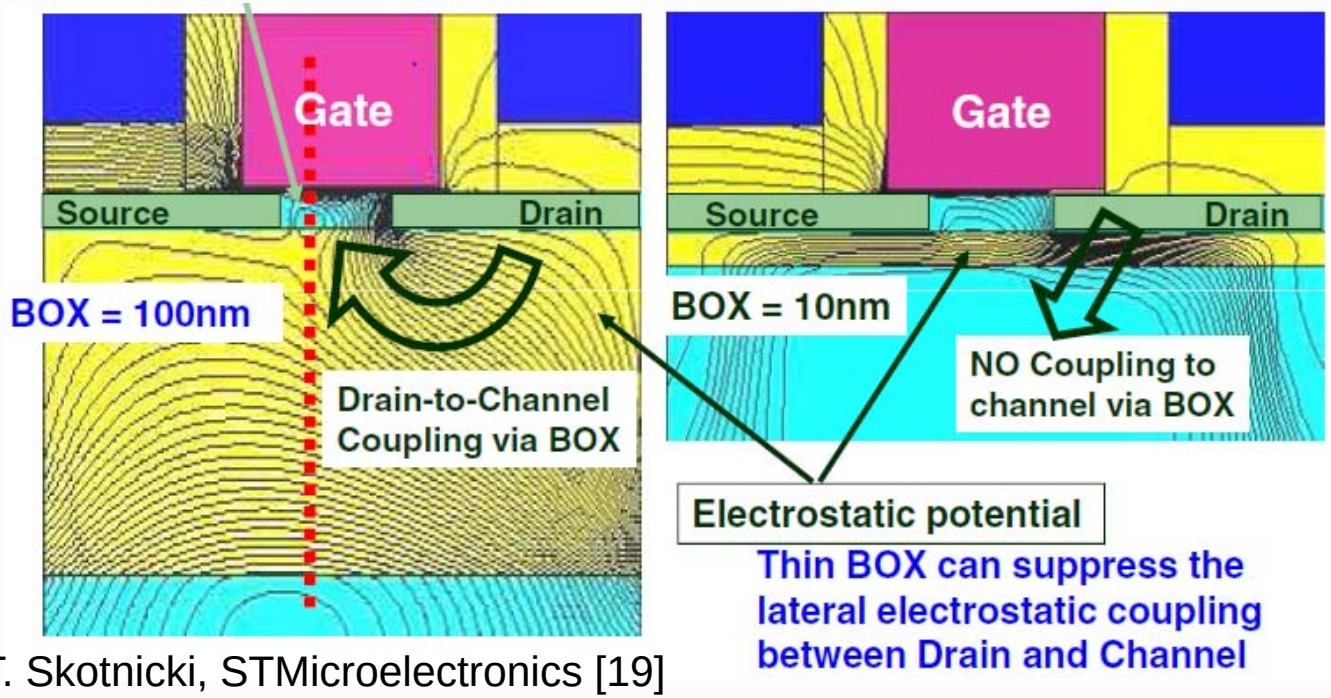




K. Cheng *et al*, IBM [18]

thick to reduce series resistance & apply stress

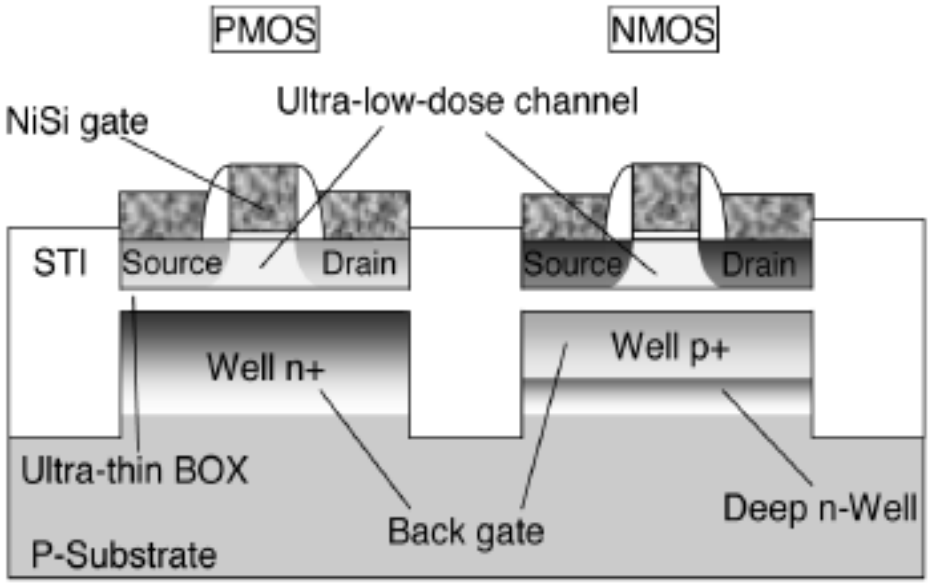
Thin BOX to Suppress SCE



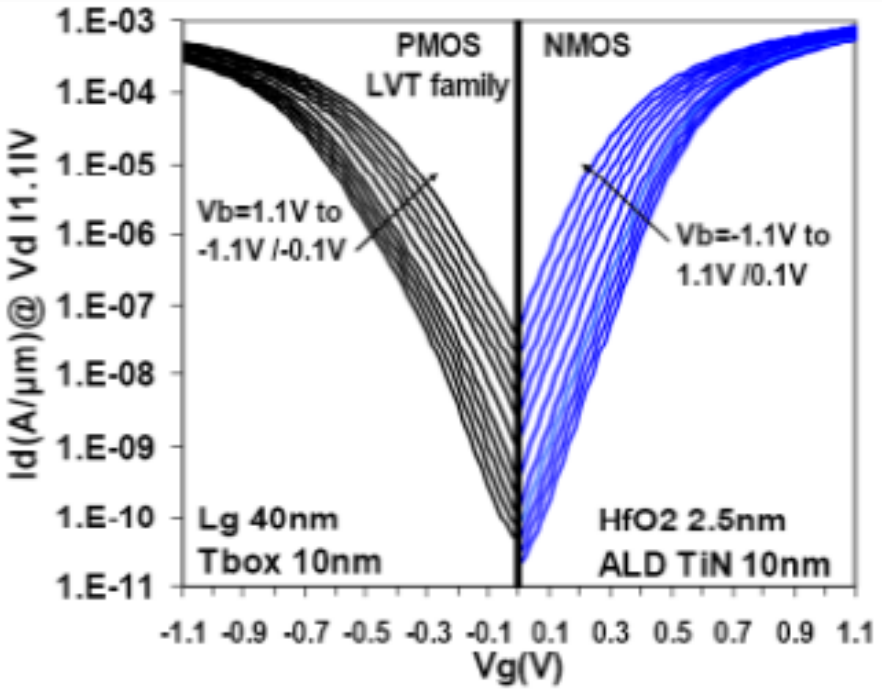
T. Skotnicki, STMicroelectronics [19]

- Fully-depleted alone does not eliminate SCE
- Field lines from drain are still competing for body charge
- If body is fully depleted, these field lines 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 ☹

Performance Tuning with Backgate Bias



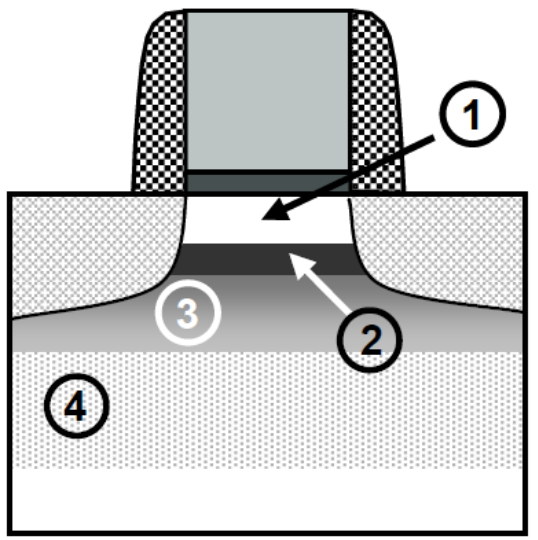
Yamaoka *et al*., Hitachi [20]

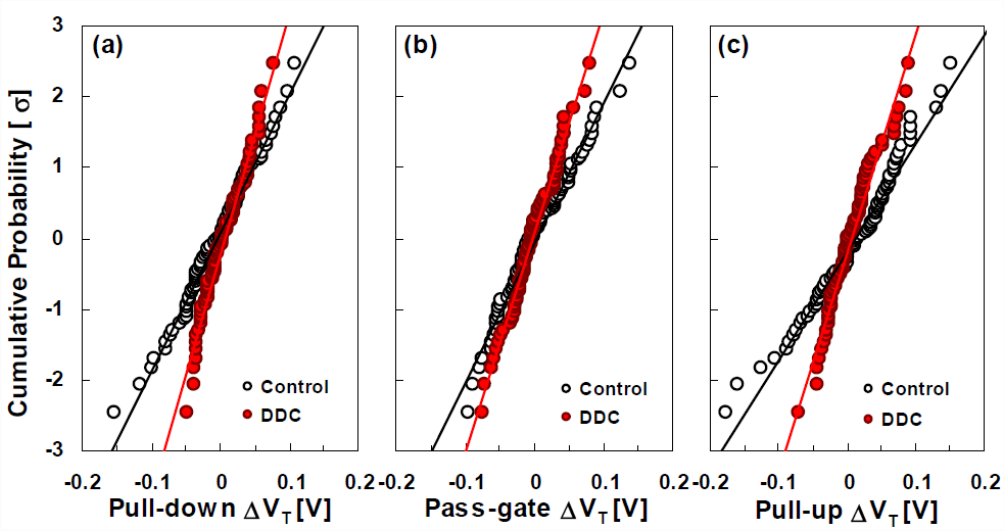


T. Skotnicki, STMicroelectronics [21]

- 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





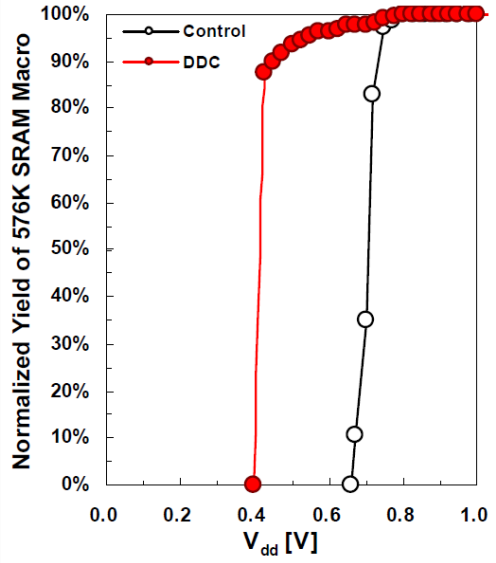
1 Low-doped layer for RDF reduction (fully depleted)

2 V_T setting layer for multiple V_T devices

3 Highly-doped screening layer to terminate depletion

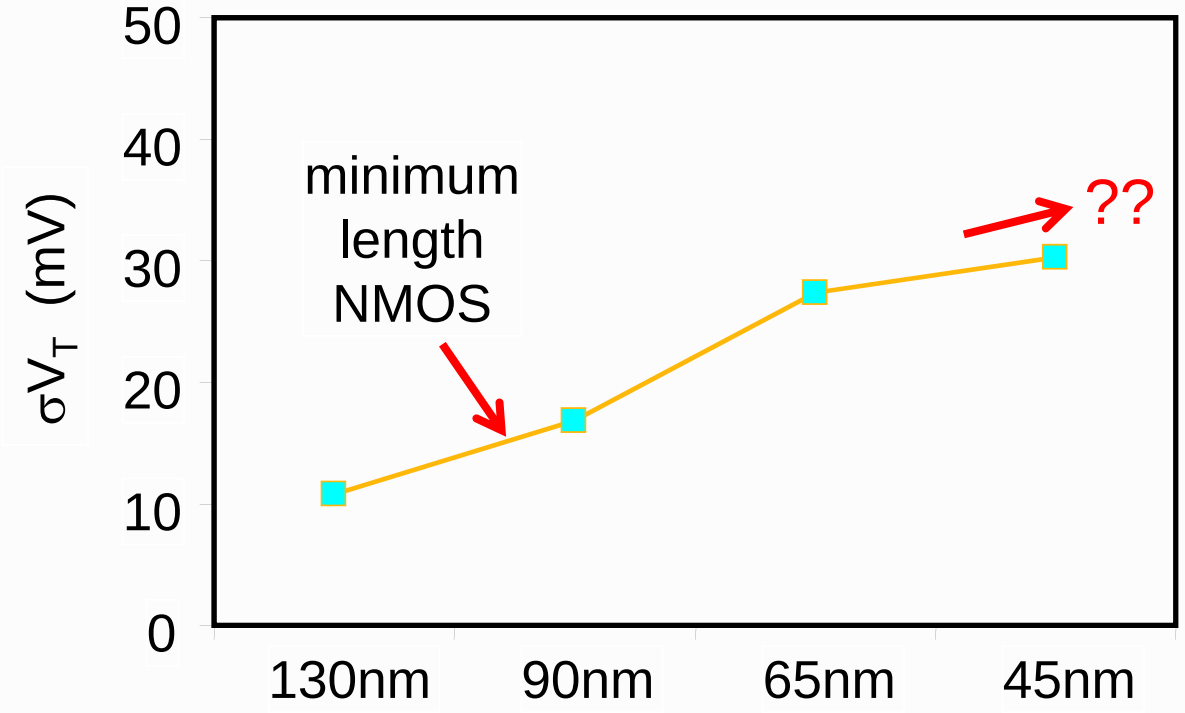
4 Sub-surface punchthrough prevention

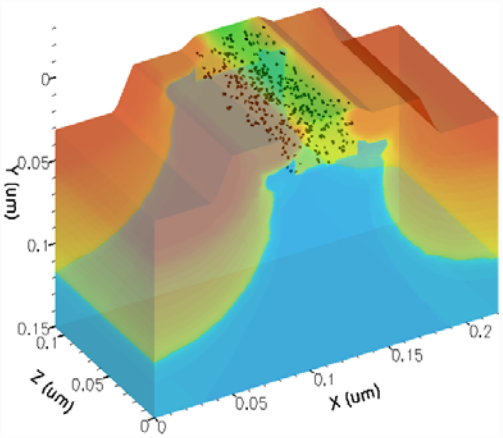
Reduced RDF for tighter V_T control & lower SRAM V_{DDmin}



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

Random Dopant Fluctuation (RDF)





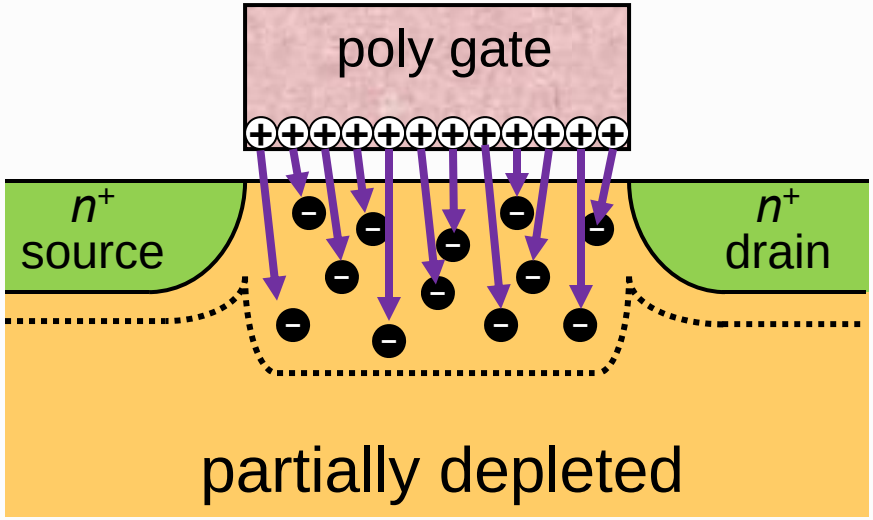
- RDF more prevalent with scaling since number of dopants is decreasing with each MOS generation
- Why does RDF impact magically disappear in fully-depleted?

Auth, Intel [15]

RDF in Conventional MOS

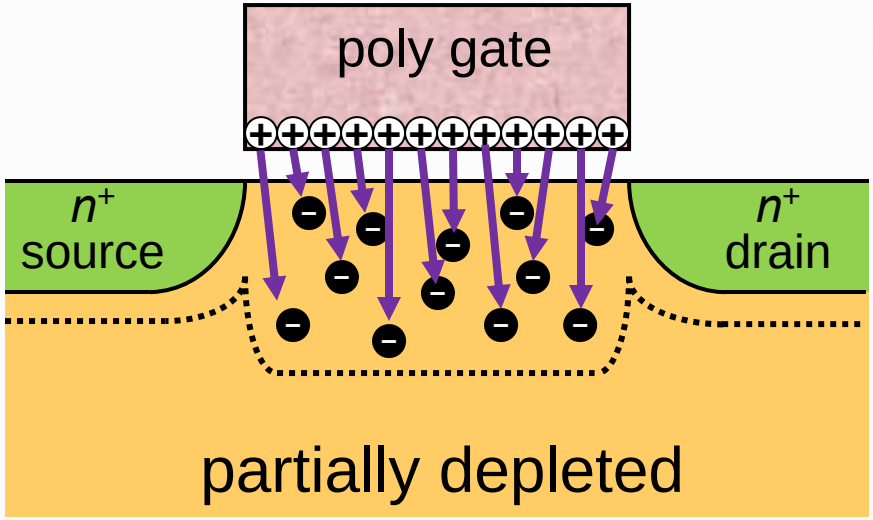
- Back to basics
  - Conservation of charge
  - Electric field lines start at +ve charge & end at –ve charge
- Random locations of dopant atoms
  - *Lengths* of field lines exhibit variation
  - Integrated field (voltage or band bending) has V_T variation

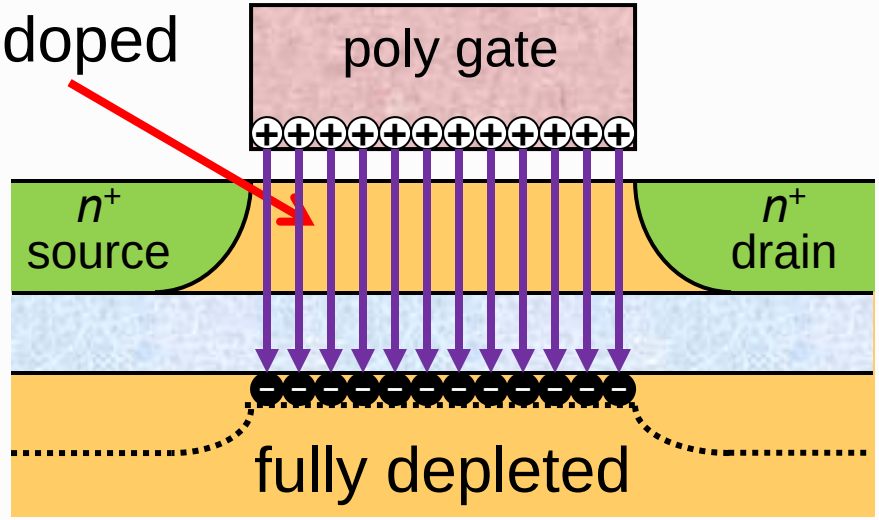
$$V = -\int E \cdot dx$$



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 very tight distribution
  - V_T variation is small
  - However, V_T now very sensitive to dimensional variation, e.g., SOI and BOX thickness



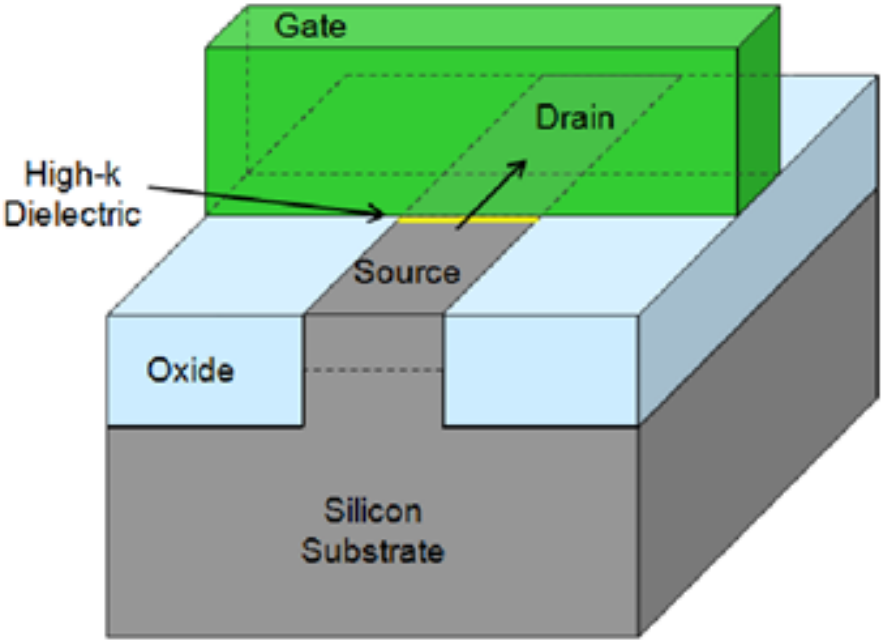


Outline

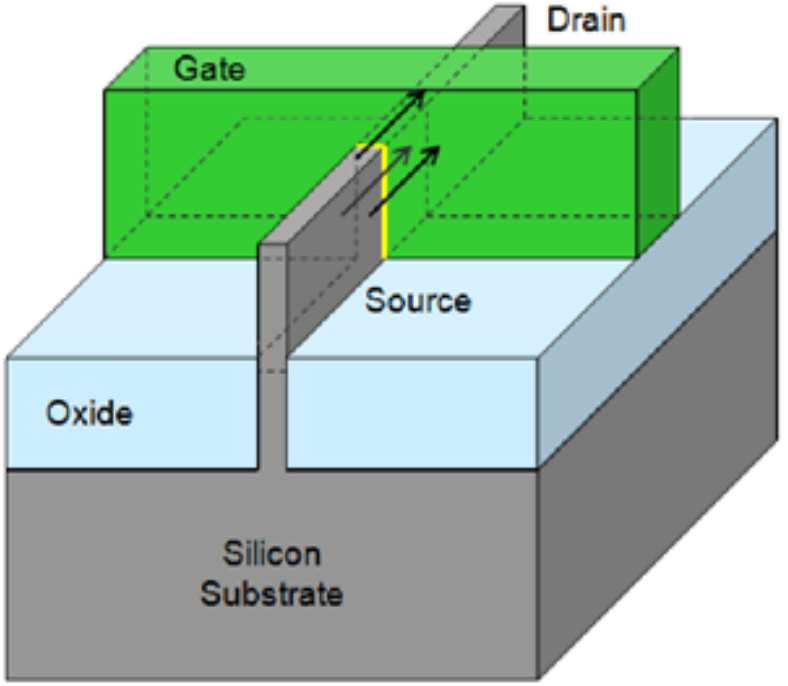
- Motivation
- MOSFET & Short-Channel Fundamentals
- 130nm MOSFET Fabrication
- Strain Engineering (90nm & Beyond)
- High-K / Metal-Gate (45nm & Beyond)
- Fully-Depleted (22nm & Beyond)
- **Tri-Gate FinFETs**
- Conclusions

What is Fully-Depleted Tri-Gate?

32nm planar



22nm tri-gate



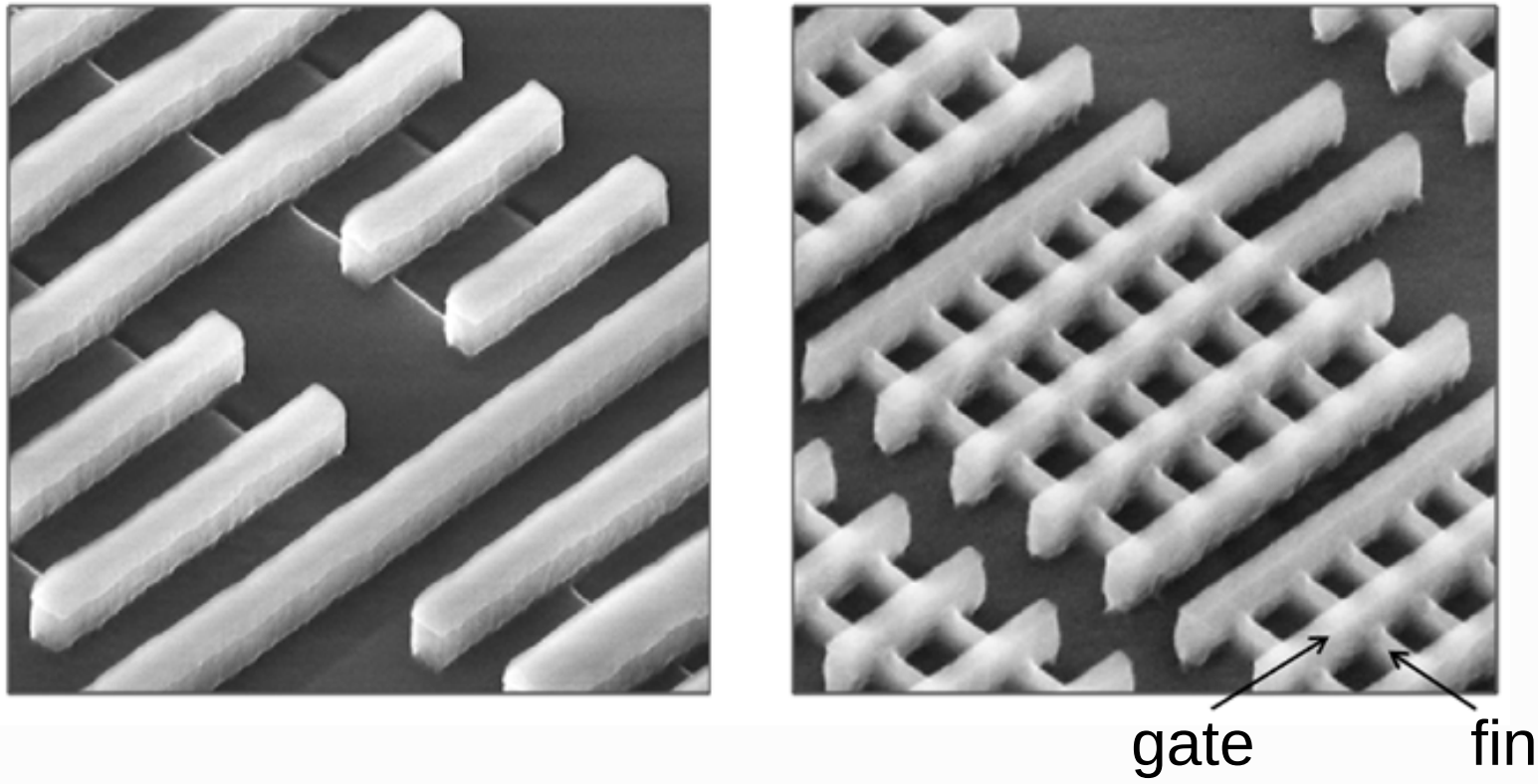
- Channel on 3 sides
- Fin width is *quantized* (SRAM & logic implications)

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

Tri-Gate FinFETs in Production

32nm planar

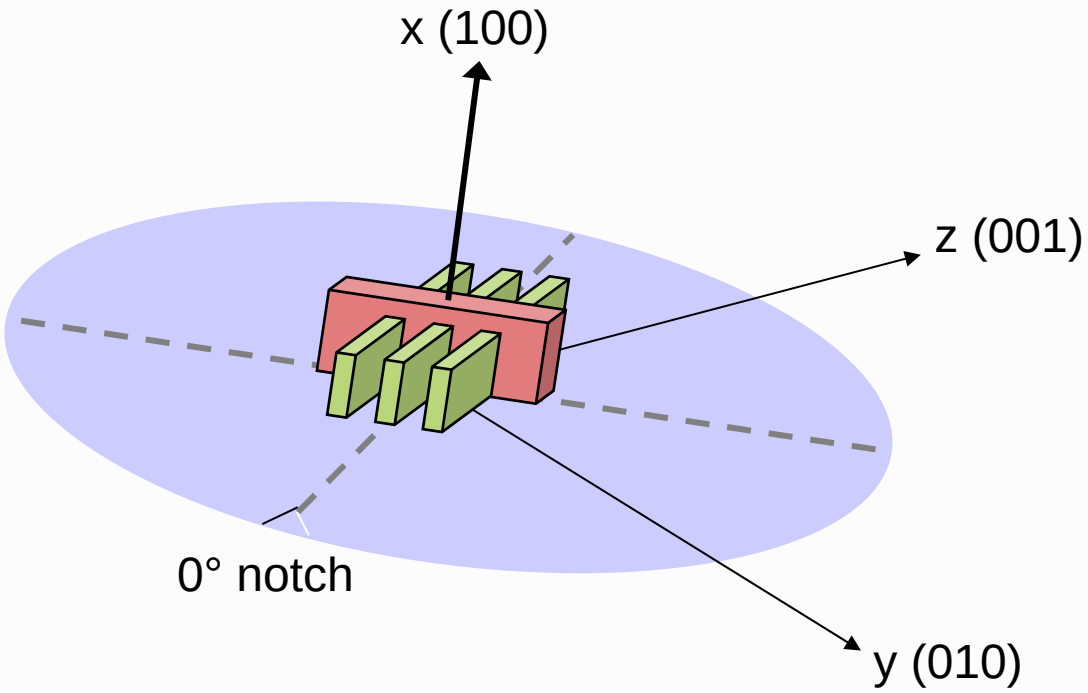
22nm tri-gate

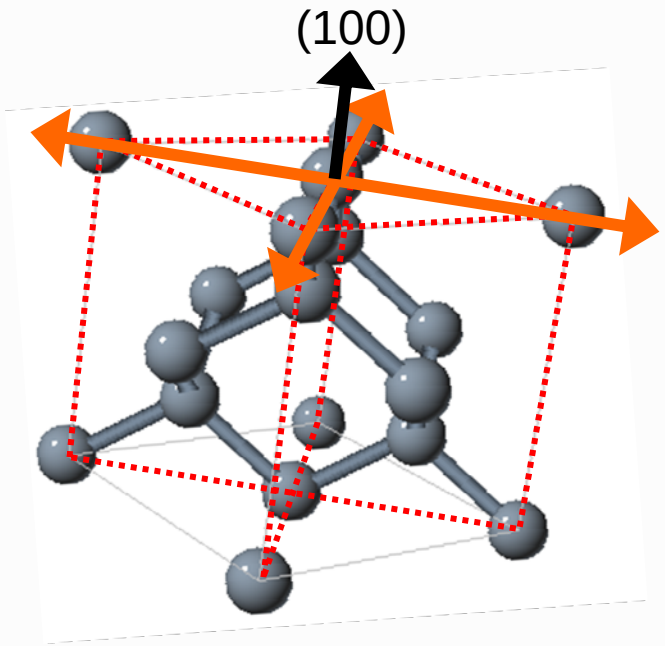


Truly impressive!!!

M. Bohr, Intel [24]

Conventional Wafer Surface Orientation & Channel Direction

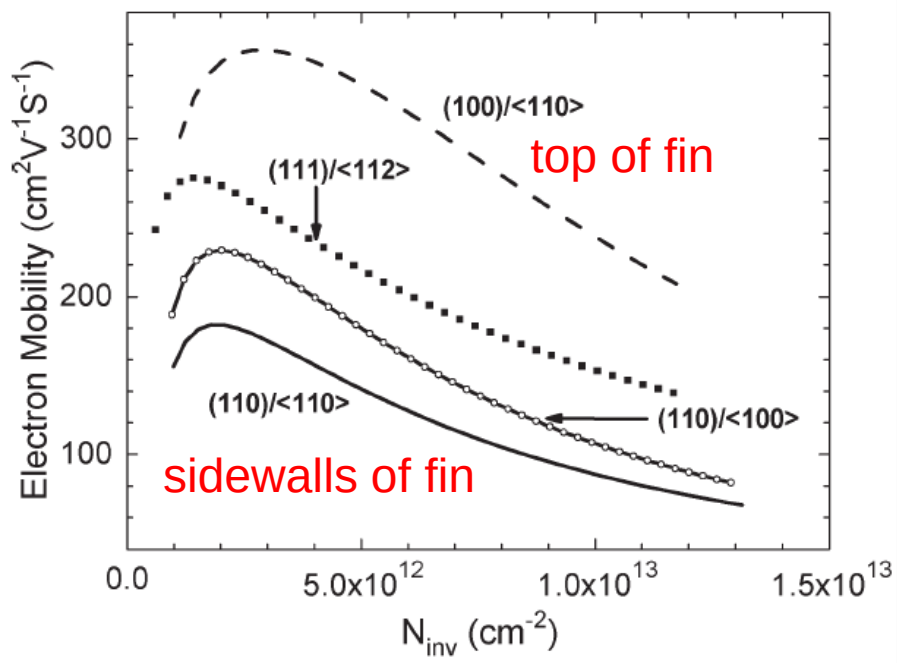




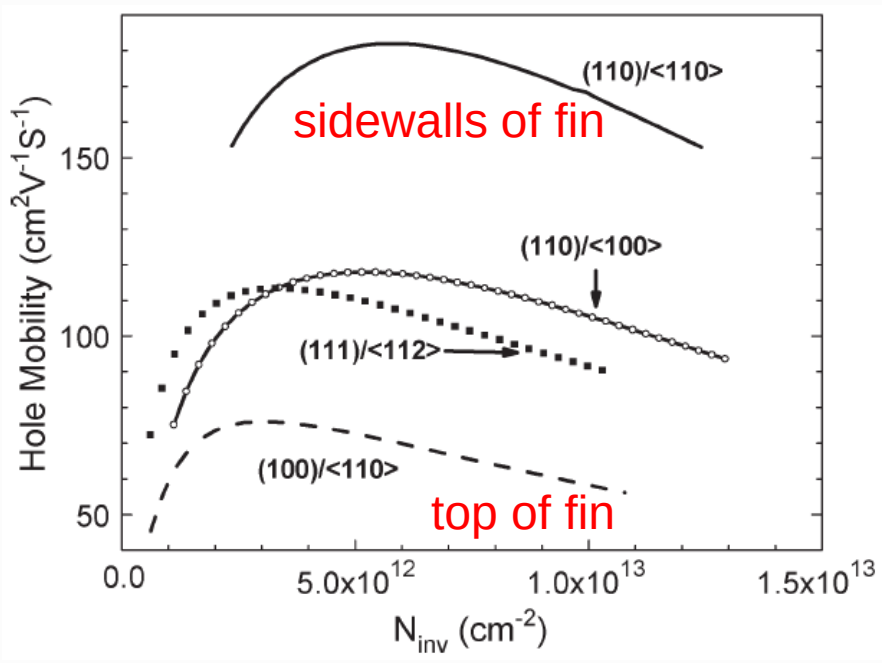
- Wafer normal is (100), current flows in <110> direction
- Tri-Gate FinFET: top surface (100), sidewall surfaces (110)

Mobility Dependence on Surface Orientation & Direction of Current

NMOS



PMOS

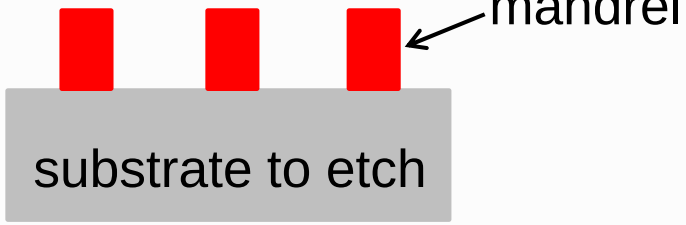


Yang *et al.*, IBM [25]

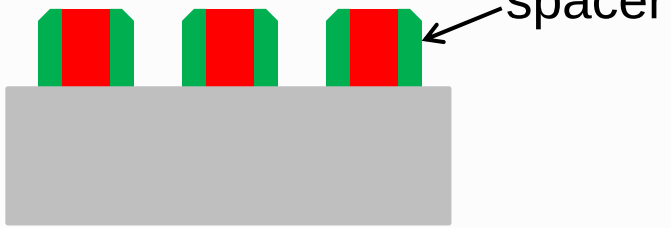
- Strain-induced mobility boost also depends on surface orientation & channel direction – not as strong for current along sidewalls vs. top of fin

Fin Patterning – Sidewall Image Transfer

1 **Deposit & pattern sacrificial mandrel**

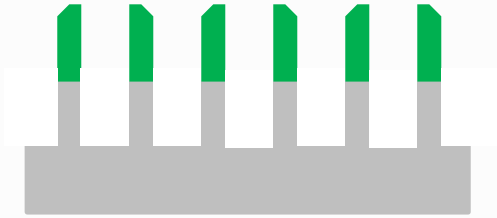


2 **Deposit & etch spacer**



3 **Remove mandrel**

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

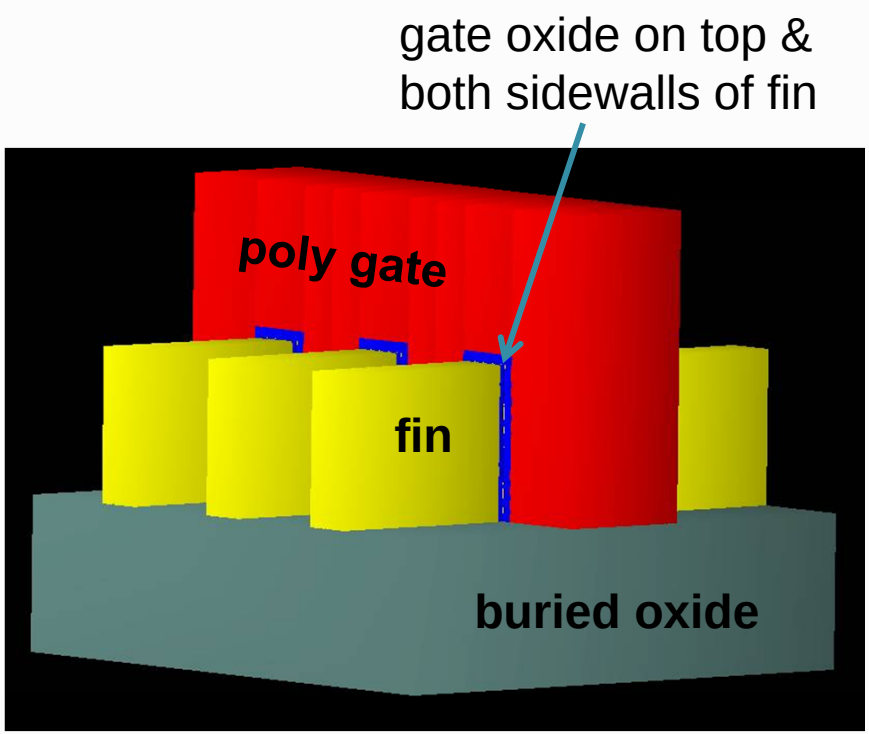


5 **Remove spacer mask**

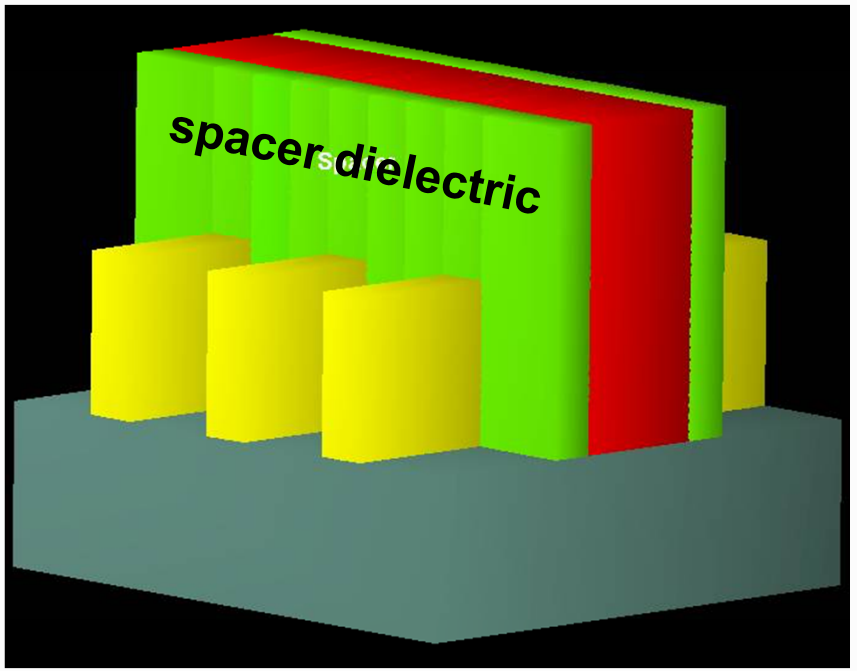
- Standard approach for patterning fins down to 60nm pitch (Intel 22nm)
- In principle, pitch can go down to ~40nm without double patterning

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



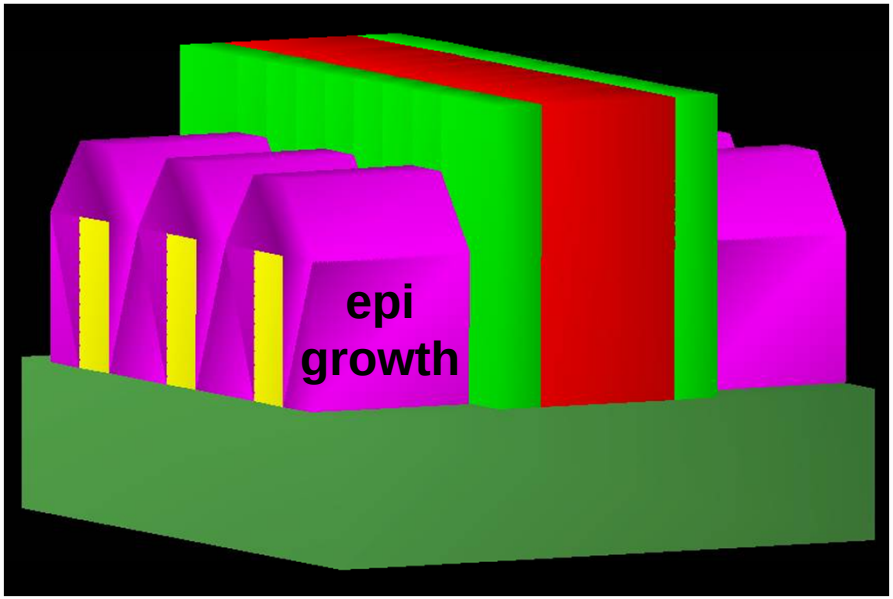
- Deposit spacer dielectric & etch, leaving spacer on gate sidewalls
- Spacer must be removed on fin sidewall



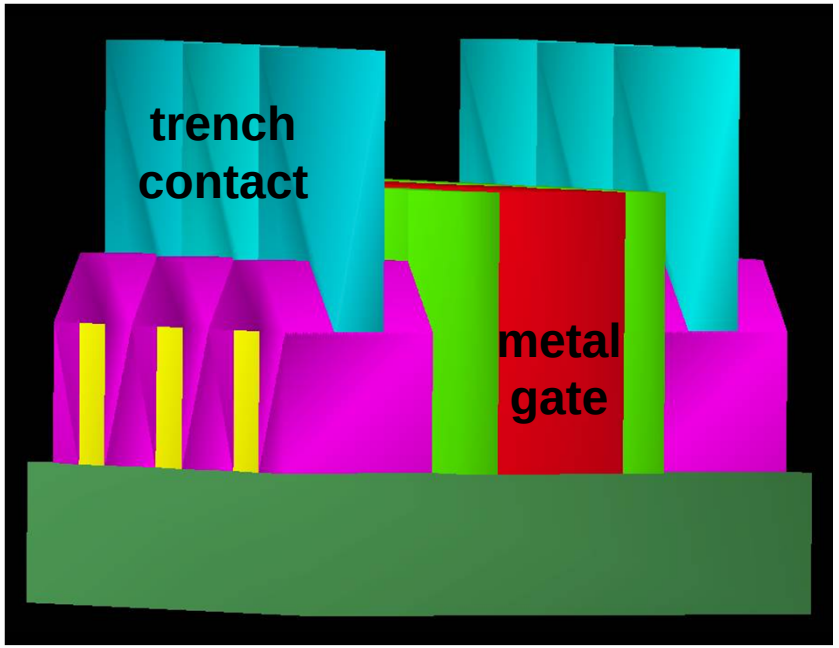
Paul, AMD [26]

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
- Source/drain dopants come from *in situ* doping during epi



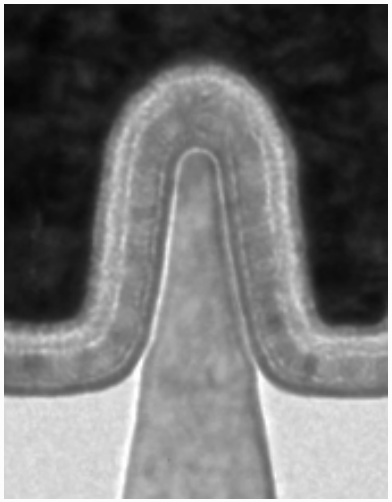
- Deposit ILD0 & CMP to top of poly
- Do replacement-gate HKMG module
- Deposit & pattern contact dielectric
- Form trench contacts (note overlap capacitance to gate)

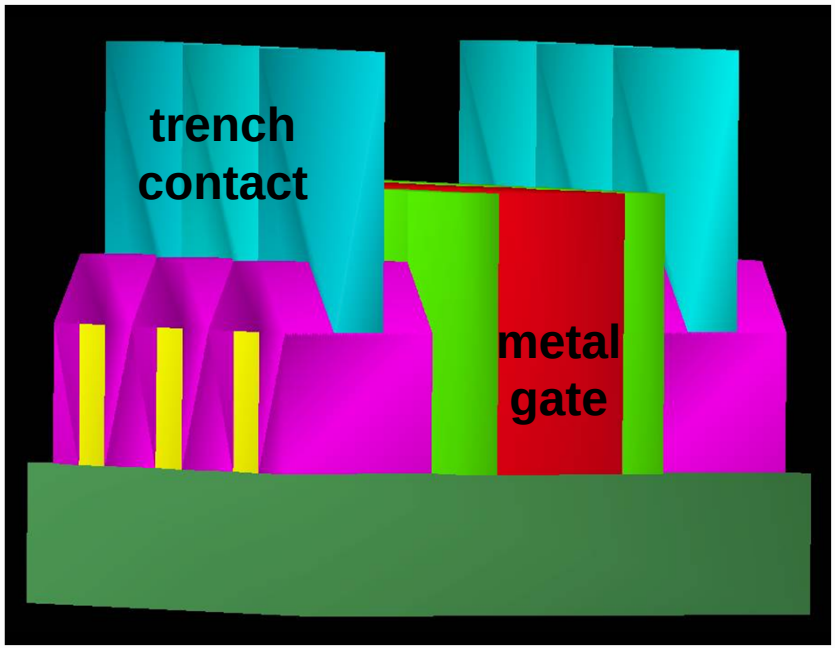


Paul, AMD [26]

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



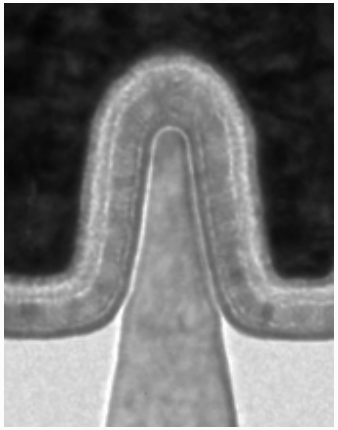


Pacifying The Multi-V_T Addiction

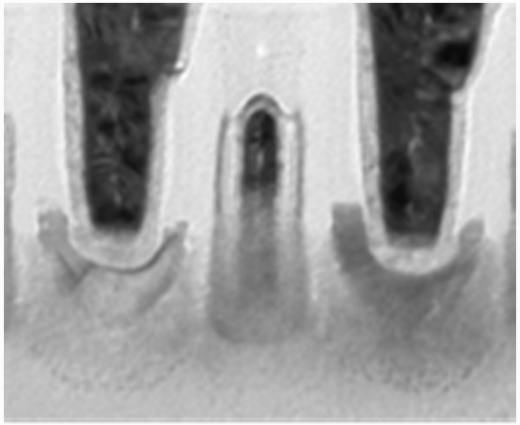
- 8 V_T's typical in 28nm (NMOS vs. PMOS, thick vs. thin oxide)
- Methods of achieving multiple V_T
  1. Bias channel length
     - Exploit SCE (V_T rolloff with shorter L)
     - Increase L for lower I_{ON} & I_{OFF}
  2. Implant fin body with different dose
     - Field lines from gate must terminate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

Intel 22nm TEM Cross-Sections

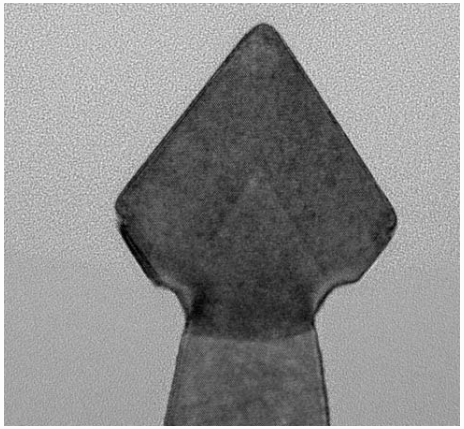
Single fin (along W)



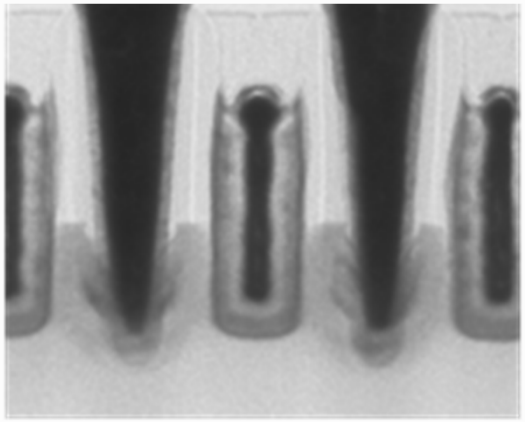
NMOS (along L)



Epi merge (along W)

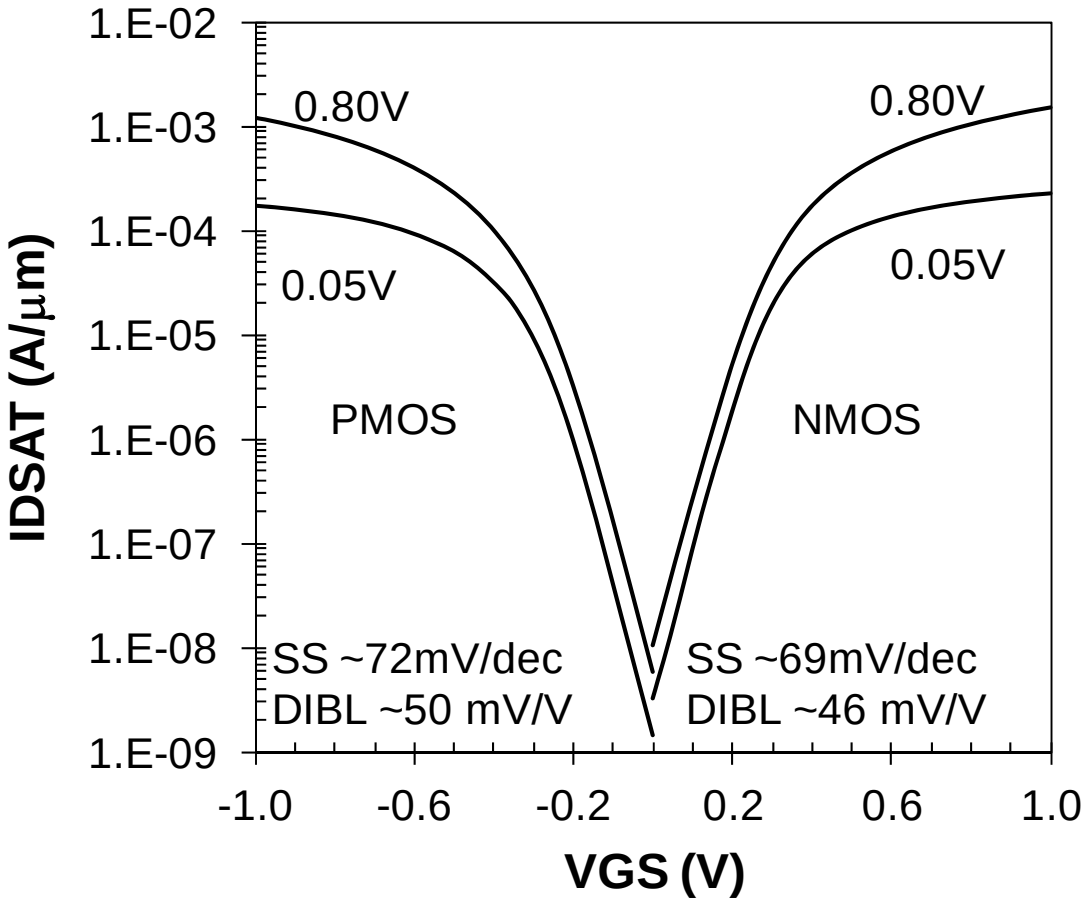


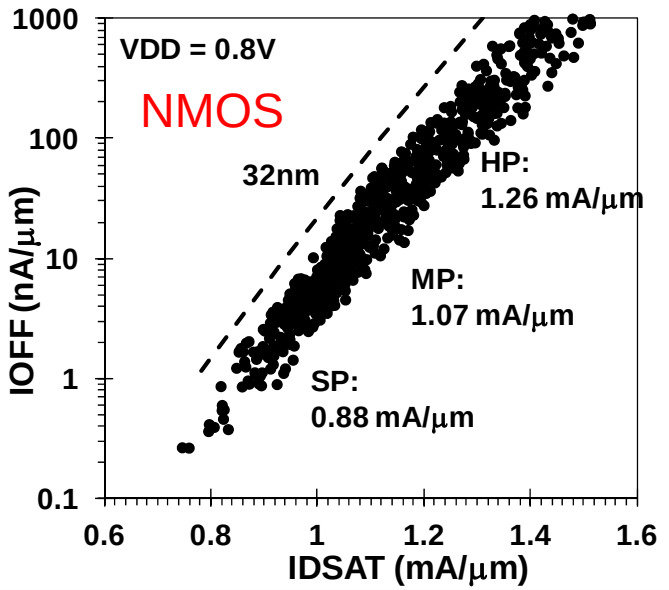
PMOS (along L)

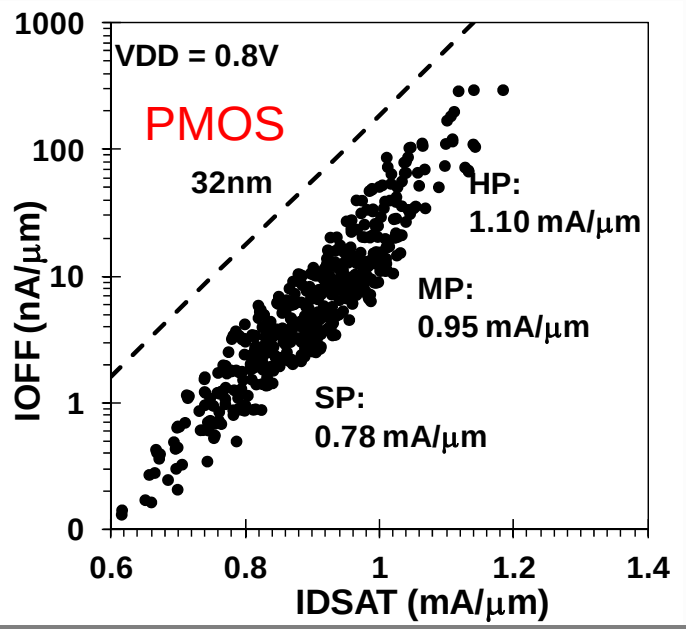


Auth, Intel [27]

Intel 22nm Performance at 0.8V







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
- Exciting time to be designing

References I

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

[2] L. Bair, “Process/product interactions in a concurrent design environment,” in *Proc. IEEE Custom Integrated Circuits Conf.*, pp. 779–782, Sep. 2007.

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

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

[5] A. Loke *et al.*, “Constant-current threshold voltage extraction in HSPICE for nanoscale CMOS analog design,” in *SNUG San Jose Conf.*, Mar. 2010.

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

[7] R. Bianchi *et al.*, “Accurate modeling of trench isolation Induced mechanical stress effects on MOSFET electrical performance,” in *IEEE Int. Electron Devices Meeting Tech. Dig.*, pp. 117-120, Dec. 2002.

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

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

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

[11] 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.

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

[13] www.ICKnowledge.com

[14] 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.

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

[16] 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.

References II

[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.

Acknowledgments

Thanks to the authors of the countless published material used as illustrations in this presentation

The Shoulders I Humbly Stand On


Bob Barnes
Agilent


Larry Bair
AMD


John Bravman
Bucknell


Tom Cynkar
Avago


Dick Dowell
Avago


Bruce Doyle
AMD


Emerson Fang
Apple


John Faricelli
AMD


Dennis Fischette
AMD


Phil Fisher
Avago


Mike Gilsdorf
Avago


Jung-Suk Goo
GlobalFoundries


Rick Hernandez
PMC-Sierra


Mark Horowitz
Stanford


Ron Kennedy
Avago


Takamaro Kikkawa
Hiroshima Univ.


Greg Kovacs
Stanford


Steve Kuehne
LSI


Tom Lee
Stanford


Justin Leung
Intel


Ying-Keung Leung
GlobalFoundries


Tom Lii
TI


Joe McPherson
TI


Bich-Yen Nguyen
Soitec


Charles Moore
Avago


Michael Oshima
AMD

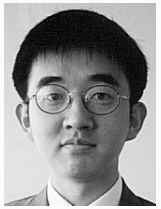

Chintamani Palsule
Sionyx


Jim Pfiester
Avago


Jim Plummer
Stanford


Dave Pulfrey
UBC


Gary Ray
Intel


Changsup Ryu
Samsung


Krishna Saraswat
Stanford


Shawn Searles
AMD


Ray Stephany
AMD


Gerry Talbot
AMD


Tom Tiedje
Univ. Victoria


Paul Townsend
SBA Materials


Jeff Wetzel
SVTC


Martin Wedepohl
UBC

Tin Tin Wee
AMD

Simon Wong
Stanford

Bruce Wooley
Stanford

Matt Angyal
IBM

Qi-Zhong Hong
TI

Wei-Yung Hsu
Applied Materials

Andy Wei
GlobalFoundries

THANK YOU