Trends and Requirements of Future FETs Based on a Simple Physical Device Model

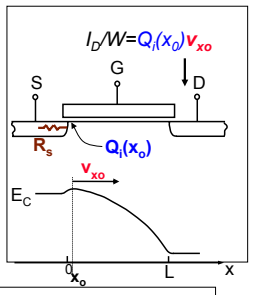
Dimitri A. Antoniadis
Microsystems Technology Labs
MIT

Outline

- Introduction: A 7-parameter physical FET (carrier transport) model for analyzing device performance
- Historical trend of channel carrier velocities and their correlation with carrier mobilities
- CMOS technology scaling trends – aggressive "High Performance" scaling scenario
- Application of model to predict performance and draw conclusions about "roadmap" requirements
- Potential device architectures
- Conclusions

Analytical Physical MOSFET I_D (V_{GS},V_{GS}) Model for Saturation

- $I_D=WQ_i(x_0)v_{xo}$
  v_{xo} : carrier velocity at virtual source
- $Q_i(x_0) \cong C_{ox}^{inv}[V_{GS}^* - V_t^*(V_D)]$
- $V_{GS}^* = V_{GS} - I_D R_s$
  R_s: Source parasitic resistance
- $V_t^* = V_{t0} - (V_{DS} - 2R_s I_D)\delta$ where δ is DIBL in V/V and
  $V_{t0} = V_t^*(V_{DS}=0)$
- Simple model for I_D in saturation:



$$(I_D/W) = C_{ox}^{inv}(V_{GS} - V_t)v$$

where "apparent velocity", v:

$$v = \frac{v_{xo}}{[1 + C_{ox}^{inv} R_s W(1+2\delta)v_{xo}]} \quad \text{and} \quad V_t = V_{t,sat} = V_{t0} - \delta V_{DS}$$

MOSFET Switch Performance: Delay Metric

$$\tau = \frac{\Delta Q_G}{I_{eff}}$$

ΔQ_G **: Charge difference between the ON and OFF states**

I_{eff} **: Effective switching current**

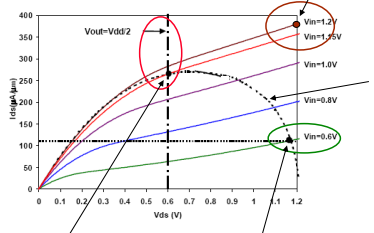
$$\tau = \frac{(1-\delta)V_{dd} - V_t + (C_f^* V_{dd} / C_{ox}^{inv} L_g)}{[(3-\delta)/4]V_{dd} - V_t} \frac{L_g}{v}$$

C_f^* **is the total source and drain side fringing capacitance multiplied by 3/2 in order to include the Miller effect at the drain side.**

D. A. Antoniadis, et al., IBM J. Res. Dev., vol. 50., p. 363, 2006.

Effective Current Drive During Switching

"Traditional"
$I_{on}=I_D(V_D=V_{dd}, V_G=V_{dd})$



Switching I_d-V_d trajectory

$I_{eff}=[I_D(V_D=V_{dd}/2, V_G=V_{dd}) + I_D(V_D=V_{dd}, V_G=V_{dd}/2)]/2$

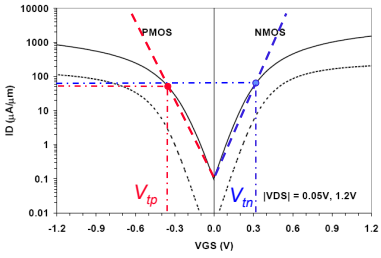
(Based on I_D model and effective gate discharge trajectory theory by Na et al., IEDM '02)

V_t, I_{off}, Subthreshold Swing, and DIBL

- V_t and I_{off} determine approximate value of S:

$$V_t = S \log \frac{I_{ref}/L_g}{(I_{off}/W)}$$

where I_{ref} typically about 3.0 μA for nFETs and 2 μA for pFETs

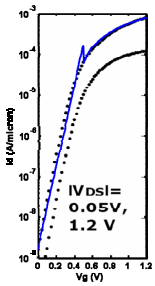


e.g. L=35 nm :
I_{refn}/L=85 μA/μm
I_{refp}/L=55 μA/μm

S_n=120 mV/dec
S_p=130 mV/dec

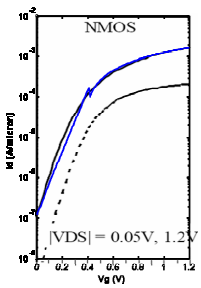
Data from Bai et a. IEDM '04

nFET Model-Data comparison: Transfer I-V



Intel (LP) L_g=55 nm CMOS
(Jan, IEDM '05)

t_{oxinv}= 2.4 nm (EOT=1.39 nm)
V_t = 0.45 V
R_{SD} = 320 Ohm-micron
δ = 105 mV/V
S= 100 mV/decade
v_{xo}= 9.15 x 10^6 cm/s

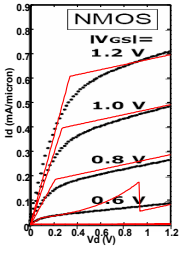


Intel (H-P) L_g=35 nm CMOS
(Tyagi, IEDM '05)

t_{oxinv}= 1.9 nm (EOT=1.2 nm)
V_t = 0.37 V
R_{SD} = 160 Ohm-micron
δ = 130 mV/V
S = 130 mV/decade
v_{xo}= 1.35 x 10^7 cm/s

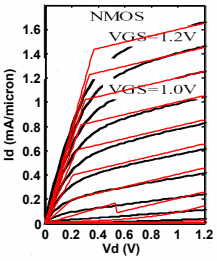
Model reproduces I_D-V_{GS} reasonably well with just six parameters

nFET Model-Data comparison: Output I-V



Intel (LP) L_g=55 nm CMOS
(Jan, IEDM '05)

t_{oxinv}= 2.4 nm (EOT=1.39 nm)
V_t = 0.45 V
R_{SD} = 320 Ohm-micron
δ = 105 mV/V
S= 100 mV/decade
v_{xo}= 9.15 x 10^6 cm/s



Intel (H-P) L_g=35 nm CMOS
(Tyagi, IEDM '05)

t_{oxinv}= 1.9 nm (EOT=1.2 nm)
V_t = 0.37 V
R_{SD} = 160 Ohm-micron
δ = 130 mV/V
S = 130 mV/decade
v_{xo}= 1.35 x 10^7 cm/s

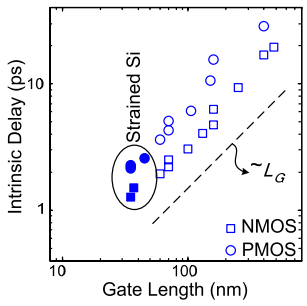
Model reproduces I_D-V_{DS} reasonably well with just six parameters

Outline

- A 5-parameter physical FET (carrier transport) model for analyzing device performance
- Historical trend of channel carrier velocities and their correlation with carrier mobilities
- CMOS technology scaling trends – aggressive "High Performance" scaling scenario
- Application of model to predict performance and draw conclusions about "roadmap" requirements
- Potential device architectures
- Conclusions

Historical MOSFET Switching Delay

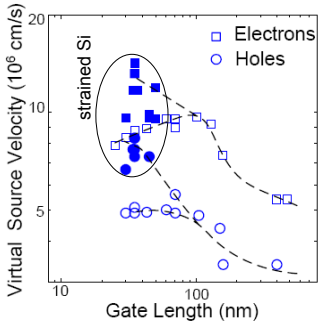
- Historical performance has followed gate length scaling remarkably well.



For details see Antoniadis et al. IBM JRD 2006 p. 363

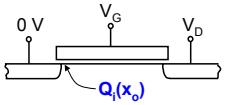
Historical Carrier Virtual Source Velocity

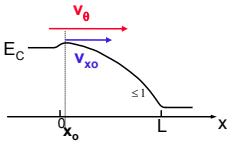
- *Virtual source velocity*, v_{xo}, increased significantly to compensate the increase in the *time delay prefactor*.
- Strain has been essential for continuous v_{xo} increase



Khakifirooz et al. IEDM 2006

What Determines the Virtual-Source Velocity?





v_{xo}: virtual source velocity
v_θ: unidirectional thermal velocity (Ballistic)

$$v_{xo} = B v_\theta$$

where B is the ballistic efficiency factor (B=1 means fully ballistic transport)

Since $B \le 1$, $\quad v_\theta \ge v_{xo}$

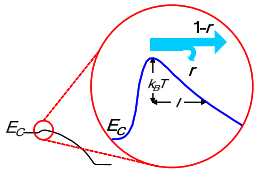
- Can practical nFETs become ballistic (B=1) ?
- What is the limit of v_θ (electrons) in practical Si devices?

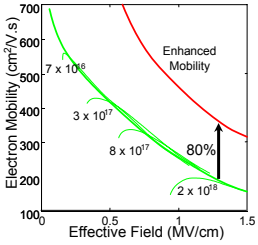
What Determines v_{xo}? [continued]

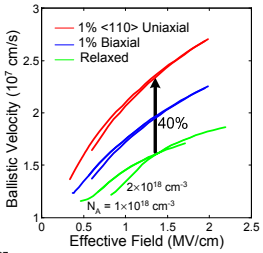
$$v_{xo} = Bv_\theta$$

$$B = \frac{1-r}{1+r} = \frac{\lambda/l}{2+\lambda/l}$$

- v_θ, mobility, and mean free path, λ are closely related







Mobility-Velocity Relationship (at fixed Q_i)

1. $\mu \propto \frac{1}{m_C m_D}$, $v_\theta \propto \frac{1}{\sqrt{m_C m_D}} \Rightarrow v_\theta \propto \mu^\alpha$ with $\alpha \approx 0.5$ for uniaxially strained Si.
2. $l \propto \mu^{-\beta}$ from NEGF simulation.
3. $\lambda = \frac{2\mu}{v_\theta}\frac{k_B T}{q}\frac{F_0(\eta_F)}{F_{-1}(\eta_F)} \propto \mu^{1-\alpha}$

$$v_{x0} = v_\theta \frac{\lambda}{2l+\lambda} = v_\theta B$$

$$\frac{\partial v_{x0}}{v_{x0}} = \frac{\partial v_\theta}{v_\theta} + (1-B)\frac{\partial \lambda}{\lambda} + (1-B)\frac{\partial l}{l}$$

$$\boxed{\frac{\partial v_{x0}}{v_{x0}} = [\alpha + (1-B)(1-\alpha+\beta)]\frac{\partial \mu}{\mu}}$$

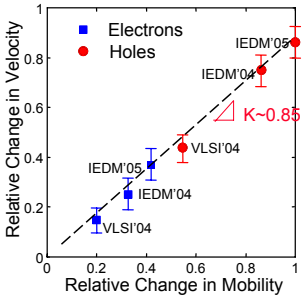
How close to the ballistic limit?

- The incremental velocity-mobility correlation factor, K, can be determined from experimental data

$$\frac{\partial v_{xo}}{v_{xo}} = [\alpha + (1-B)(1-\alpha+\beta)]\frac{\partial \mu}{\mu} = K\frac{\partial \mu}{\mu}$$

- Then, applying theory we can estimate the transmission ratio and therefore how close to ballistic limit devices operate

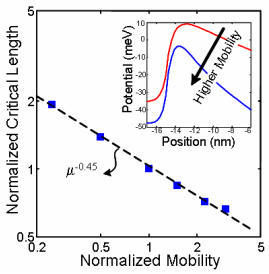
Carrier Velocity-Mobility Correlation



- The ratio of velocity change to that of mobility in modern technologies is higher than the commonly accepted value of 0.5

Theory: Critical Channel Length vs. Mobility

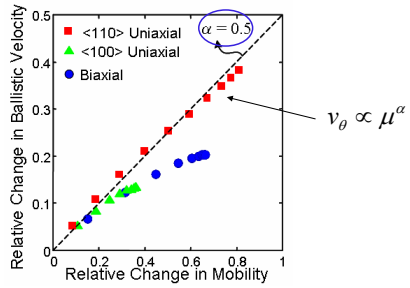
NFET Simulation by NanoMOS (NEGF + scattering)
Used for both electrons and holes



$$\frac{\partial v_{xo}}{v_{xo}} = [\alpha + (1-B)(1-\alpha+\beta)]\frac{\partial \mu}{\mu} = K\frac{\partial \mu}{\mu}$$

0.45

Theory: Relative Change in Ballistic Velocity vs. Relative Change in Mobility



$$\frac{\partial v_{xo}}{v_{xo}} = [\alpha + (1-B)(1-\alpha+\beta)]\frac{\partial \mu}{\mu} = K\frac{\partial \mu}{\mu}$$

From Uchida IEDM '05

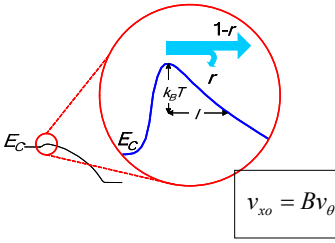
Estimation of Transmission Coefficient

- For the IEDM '05 Intel 65 nm CMOS technology, which uses uniaxial strain for electron and hole mobility enhancement we can estimate how close to the ballistic limit L_G=35 nm FETs are:

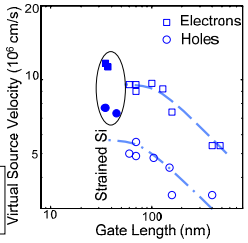
$$B = 1 - \frac{K - \alpha}{1 - \alpha + \beta} \approx 0.65$$

- We can conclude that state-of-the-art FETs operate at ~65% ballistic efficiency

Historical Evolution of v_{xo}

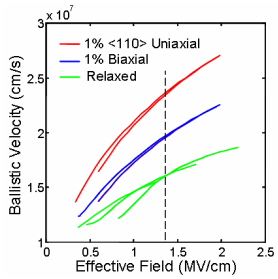


$$v_{xo} = Bv_\theta = \left(\frac{1-r}{1+r}\right)v_\theta$$



- For unstrained Si, v_{xo} increases with scaling primarily because r decreases, while v_θ nearly constant (modest increase with E_{eff}).
- B is estimated to have reached 0.6 at L_g~ 65 nm (unstrained) for both electrons and holes.
- Introducing strain, v_{xo} increases primarily because v_θ increases while B (~0.6) nearly constant with further scaling.
- B unlikely to increase much above 0.6-07 in practical "32 nm" FETs

Si Ballistic Electron Velocity, v_θ



Calculations based on assumption that only Δ_2 valleys are populated under strain.

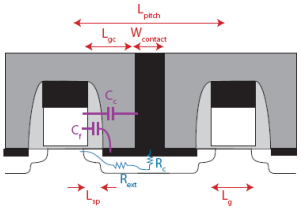
1% Biaxial strain: Increased degeneracy and smaller effective mass

1% <110> Uniaxial Strain: More effective mass change (Uchida IEDM 05)

- Unlikely to have v_θ (Si-electrons) larger than ~2.5x10^7 cm/s i.e. at ~1% uniaxial tensile strain

Scaling scenario, 130 to 15 nm CMOS

- **Contacted pitch, L_{pitch} is the main driver**: 30% per generation
- Performance scaling both historical and projected based on feature dimensions



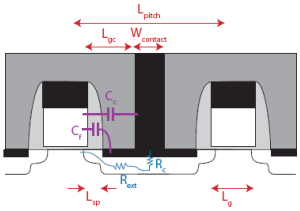
"Aggressive (HP) Scaling Path"

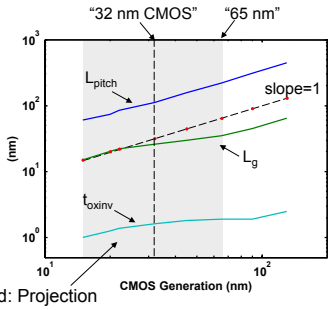
Generation (nm)	130	90	65	45	32	22	15
L_{pitch} (nm)	445	310	220	155	110	84	60
L_g (nm)	65	45	35	30	26	22	15
t_{oxinv} (nm)	2.5	1.9	1.9	1.8	1.63	1.38	1
L_{sp} (nm)	100	50	30	21	14	10	5
L_{gc} (nm)	125	87	60	40	26	20	15
V_{dd} (V)	1.4	1.2	1.2	1	0.9	0.8	0.7
DIBL (δ)	0.10	0.12	0.15	0.15	0.15	0.15	0.15
S (V/dec)	0.12	0.12	0.12	0.12	0.12	0.12	0.12
I_{on}/W (nA/μm)	100	100	100	200	300	300	300
ρ_c (10^{-8} Ω-cm^2)	6	5	4	3	2.5	2	2
R_{sheet} (Ω)	250	200	200	200	200	200	200
V_{xo} (10^7 cm/s)	0.95	1.08	1.38	1.65	1.65	1.65	1.65

Scaling scenario, 130 to 15 nm CMOS

("130 nm" to "15 nm" CMOS generations)

- Key feature sizes

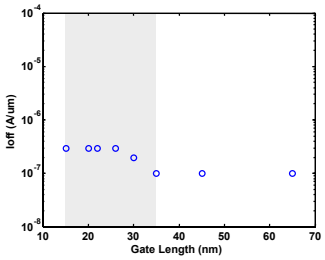


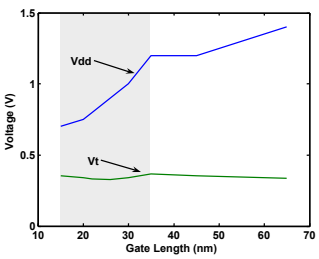


Shaded: Projection

Scaling scenario, 130 to 15 nm CMOS

- I_{off} and V_{dd} given
- V_t calculated to meet I_{off} target, given SS (S) and DIBL (δ)
- No Band-to-band tunneling leakage is included in I_{off}



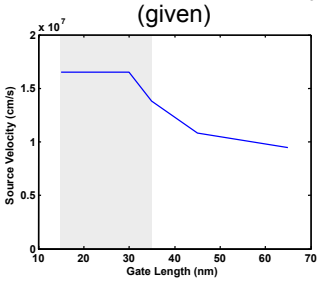


$$V_t = S / \log(WI_{ref} / L_g I_{off})$$

where: I_{ref}= 3 μA

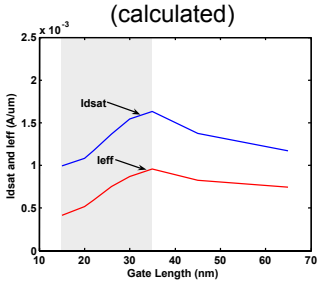
Scaling scenario, 130 to 15 nm CMOS

Virtual Source Velocity, v_{xo} (given)



It is assumed that v_{xo} will peak at "45-nm" somewhat above "65-nm" and stay constant with scaling

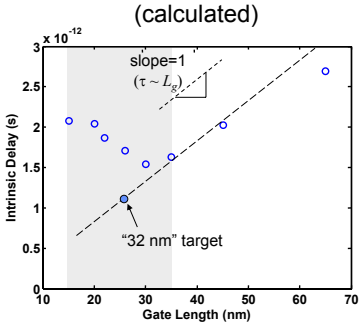
I_{dsat} and I_{eff} (calculated)



Both I_{dsat} and I_{eff} peak at "65-nm" with assumed V_{dd}=1.2 V and degrade sharply with assumed V_{dd} and EOT scaling

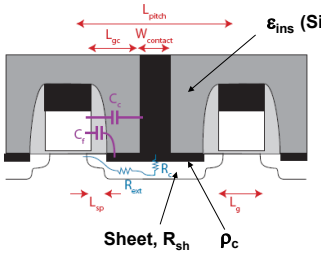
Scaling scenario, 130 to 15 nm CMOS

nFET delay vs. Gate Length (calculated)



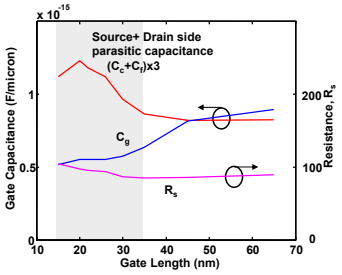
Much of the "counter-scaling" of delay is due to reduction of I_{eff} scaling beyond "65-nm" generation

Scaling scenario, 130 to 15 nm CMOS



$$R_S = R_{ext} + R_c = R_{ext} + \sqrt{R_{sh}\rho_c}\,\coth(L_{silcd}/l_{tr})$$

$$where,\ l_{tr} = \sqrt{\rho_c / R_{sh}}$$



Given the geometry, ε_{ins}, ρ_c, and R_{sh} : Calculate C_c and R_s

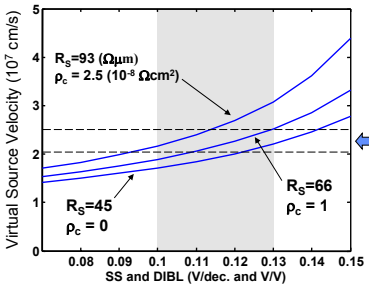
R_{ext} assumed constant, 45 Ω–μm
C_f assumed constant, 0.5 fF/μm
(To include both S/D and Miller effect multiply C_c and C_f by 3)

Required Electron Velocity, v_{xo}, to Meet Target "H-P 32 nm" τ

- Trade-off between virtual source velocity and electrostatic integrity for different source resistance values

32 nm Generation
L_g = 26 nm
t_{oxinv}=1.63 nm
V_{dd} = 0.9 V
I_{off} = 300 nA/μm

Target:
τ = 1.1 ps



Range of ballistic electron velocity in uniaxially strained Si

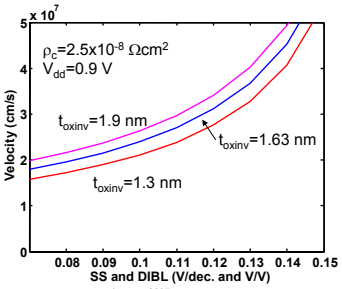
Target delay here is the nFET delay metric.
For overall CMOS delay scaling it is assumed that pFET will have to scale proportionally to nFET

Required Electron Velocity, v_{xo}, to Meet Target "H-P 32 nm" τ

- Trade-off between virtual source velocity and electrostatic integrity for different values of t_{oxinv}
- Direct sensitivity to t_{oxinv} is rather modest, particularly in the presence of R_S and DIBL
- However, reduced t_{oxinv} would somewhat reduce SS and DIBL

32 nm Generation
L_g = 26 nm
t_{oxinv}=1.63 nm
V_{dd} = 0.9 V
I_{off} = 300 nA/mm

Target:
τ = 1.1 ps

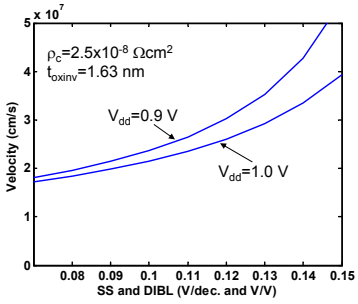


Required Electron Velocity, v_{xo}, to Meet Target "H-P 32 nm" τ

- Effect of increasing V_{dd} by 10%
- Increasingly beneficial as electrostatic integrity diminishes, i.e. SS and DIBL increase

32 nm Generation
L_g = 26 nm
t_{oxinv}=1.63 nm
V_{dd} = 0.9 V
I_{off} = 300 nA/mm

Target:
τ = 1.1 ps

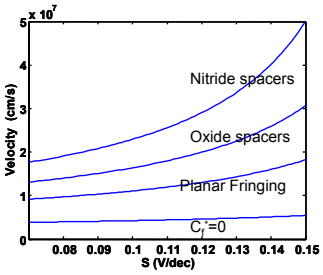


Required Electron Velocity, v_{xo}, to Meet Target "H-P 32 nm" τ

- Fringing capacitance has a very significant effect at short L_g because of the $C_f^*/C_{oxinv}L_g$ term in delay.

32 nm Generation
L_g = 26 nm
t_{oxinv}=1.63 nm
V_{dd} = 0.9 V
I_{off} = 300 nA/mm

Target:
τ = 1.1 ps



- Assumed: Gate height = $3xL_g$
  - "Nitride spacers": gate fully embedded in SiN (k=7.8)
  - "Oxide spacers" : gate fully embedded in SiO (k=3.9)
  - "Planar fringing" ignores non-planar caps: $C_f^*=C_f$=0.5 fF/μm

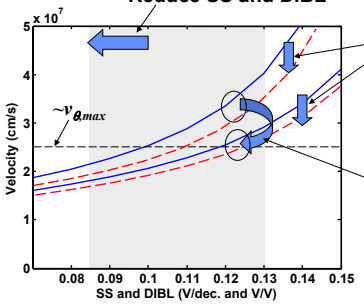
Can "H-P 32-nm" performance be improved?

- Can v_{xo} (electrons) in practical Si exceed 2.5x10^7 cm/s?
- Can practical nFETs become ballistic at L_g~26 nm (B=1) ?

These goals appear out of reach of Si channel

- What then?

Reduce SS and DIBL

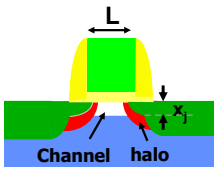


Reduce EOT
(t_{oxinv} :1.6 →1.3 nm)

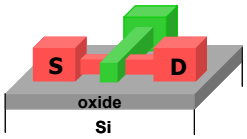
Reduce R_{sd}
(ρ_c: 3 →1x10^{-8} Ωcm^2)
or reduce C_f^*
(SiN→SiO)
or increase V_{dd}
(0.9 V→1.05 V)

Electrostatics: Keep the MOSFET Well Tempered

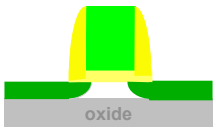
- As L is reduced, gate loses control of channel charge
  - Short channel effects
  - Bulk OK for L> ~30 nm
- Technology solutions for L<~25 nm:
  - Ultra-thin Si channel body
  - Multi-gate channel
- "32 nm CMOS" is on the cusp!



Double or triple-gate structures (non-planar)



Thin body MOSFET



Concluding Remarks

- Channel carrier velocity is the key performance parameter
- Revised parametrized [H-P] device "roadmap"
- Uniaxial strain increase has been effective in increasing velocity though 65 nm node. There much not much left for 45 and 32 nm nodes
- 32 nm performance will slip off trend unless electrostatic integrity is improved, and EOT and parasitic resistance are reduced
- Non-planar device architectures offer improved scalability but…

Concluding Remarks [continued]

- Potentially increased R_{SD} and C_f^*
- Potentially increased device variations
  - Trade Random-Dopant-Fluctuation for T_{si} and/or W_{si} variations
  - Lose halo V_t stabilization against LER and ACLV
  - Workfunction variability if adjustable for V_t tuning or,
  - Trade tuned- and multi-V_t ability for fixed gate workfunctions