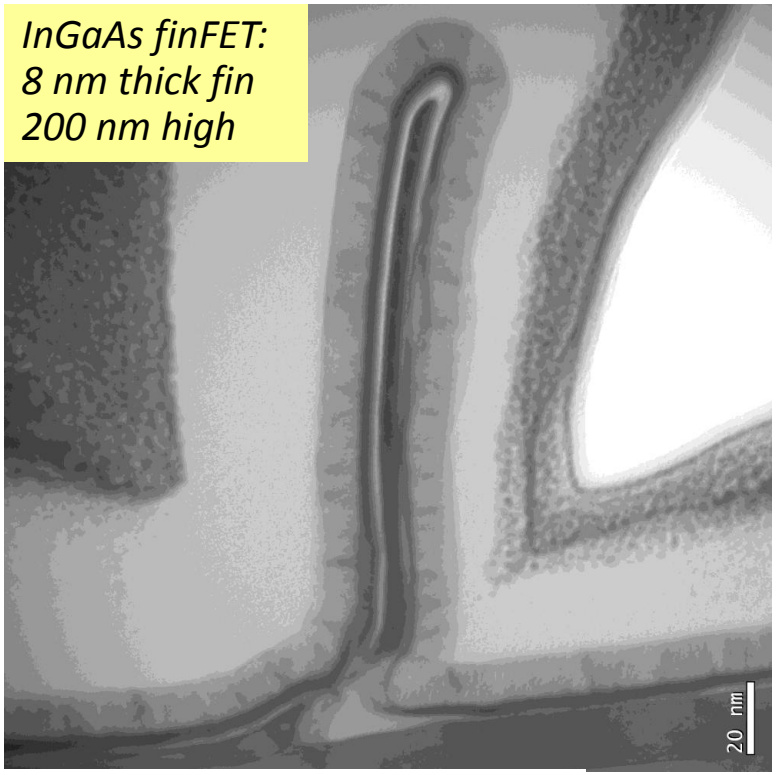
Prospects for High-Aspect-Ratio FinFETs in Low-Power Logic

Mark Rodwell, Doron Elias
University of California, Santa Barbara

High Aspect Ratio Fins for Low-Power Logic

Fin thickness defined by Atomic layer epitaxy
→ nm thickness control
Fin height defined by sidewall growth
→ 200 nm high fins

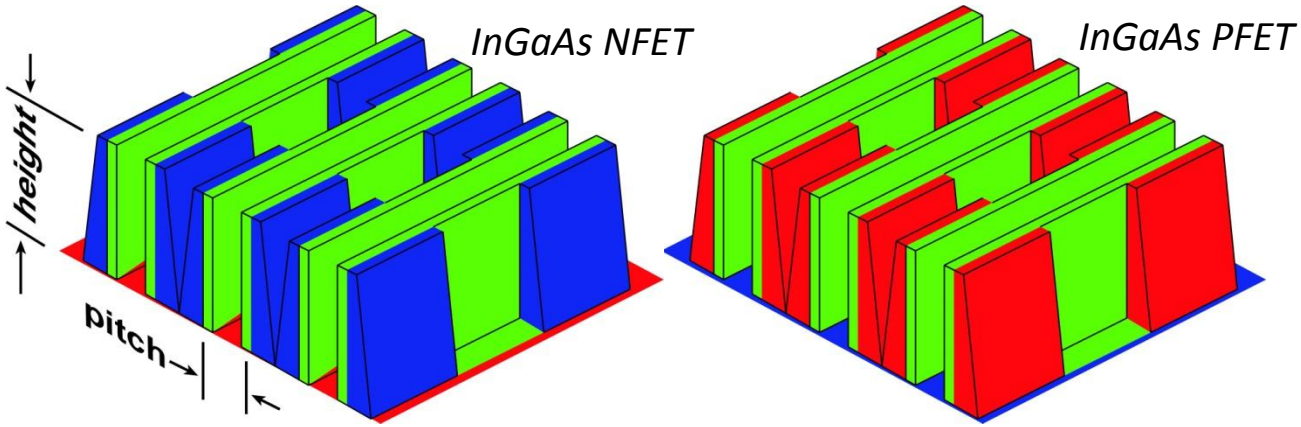
Enables ~4 nm fin bodies→ 8 nm gate length
10:1 more current per unit die area
→ smaller IC die area
complements lithographic scaling
Enables high speed, ultra low-power logic,
V_{dd}~300 mV



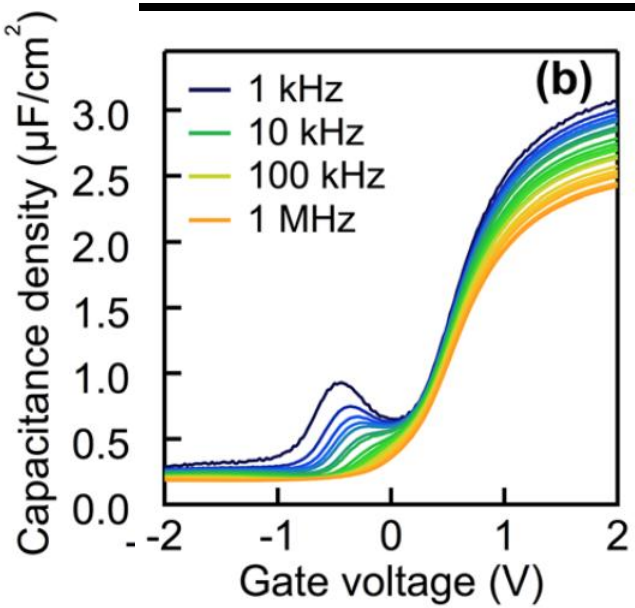
InGaAs finFET:
8 nm thick fin
200 nm high

D. Elias, DRC 2013, June, Notre Dame

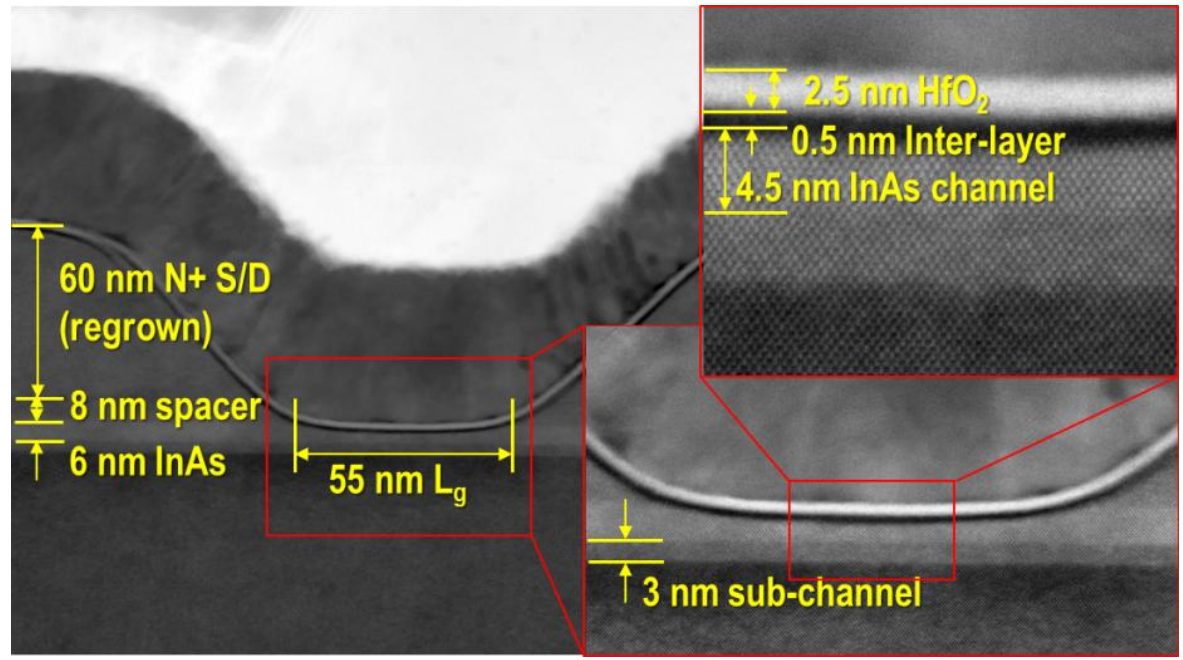
height>> pitch



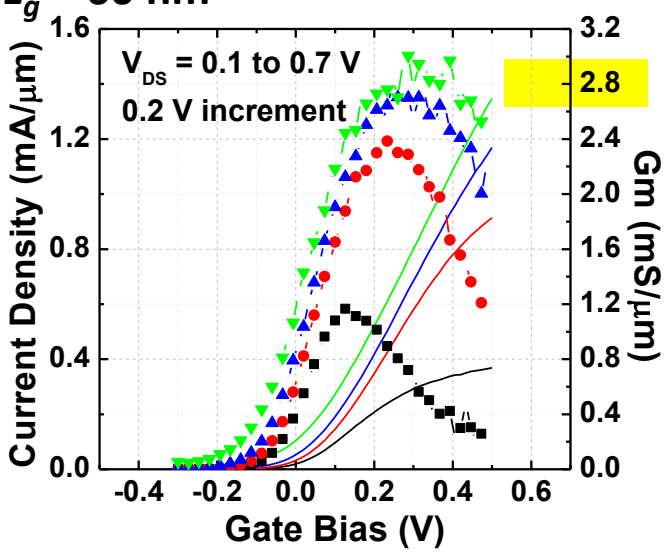
Background: III-V MOS

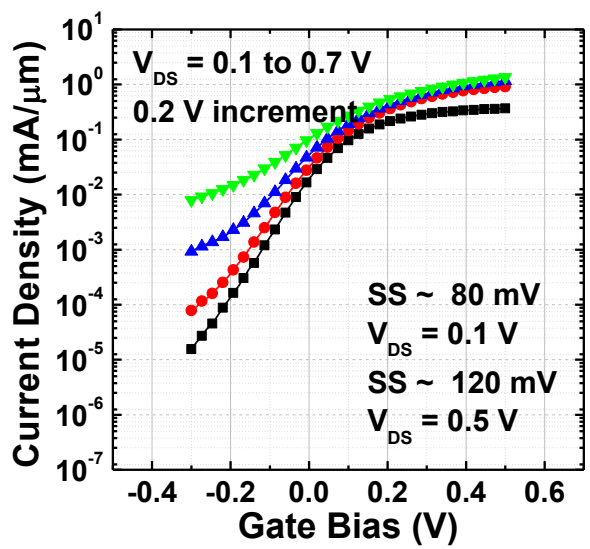


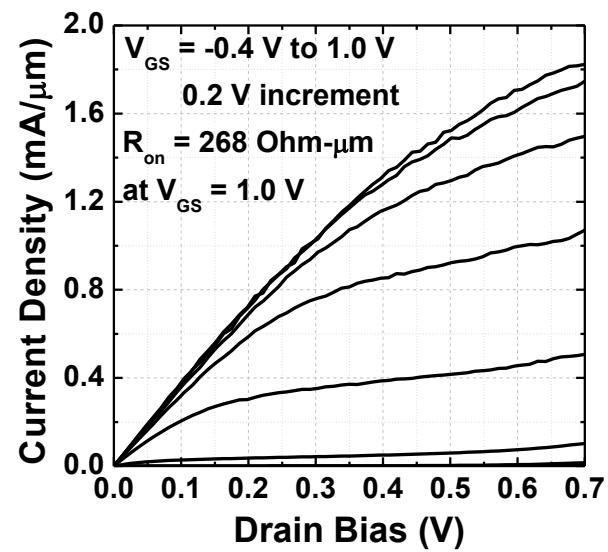
V. Chobpattana et al (Stemmer group),
APPLIED PHYSICS LETTERS 102, 022907 (2013)



L_g = 60 nm

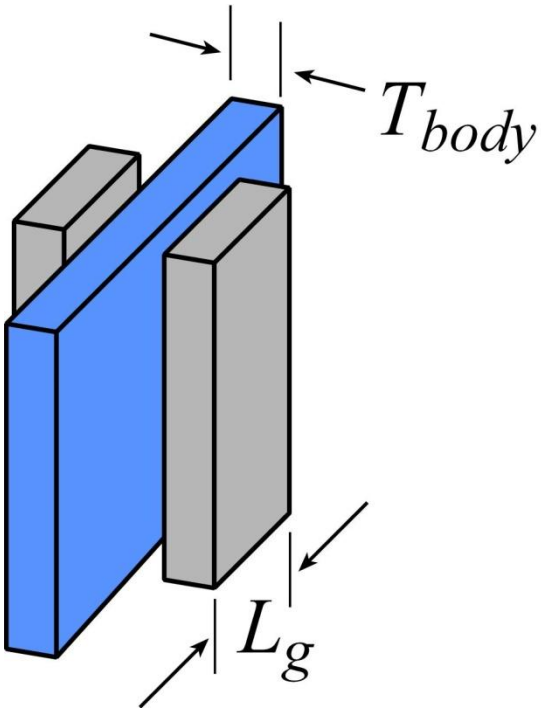






FinFETs by Atomic Layer Epitaxy: Why ?

Electrostatics:
body must be thinner than ~$L_g/2$
→ less than 4 nm thick body for 8 nm L_g



Problem:
threshold becomes sensitive to body thickness

$$\delta V_{th} \propto \delta T_{body} / T_{body}^3$$

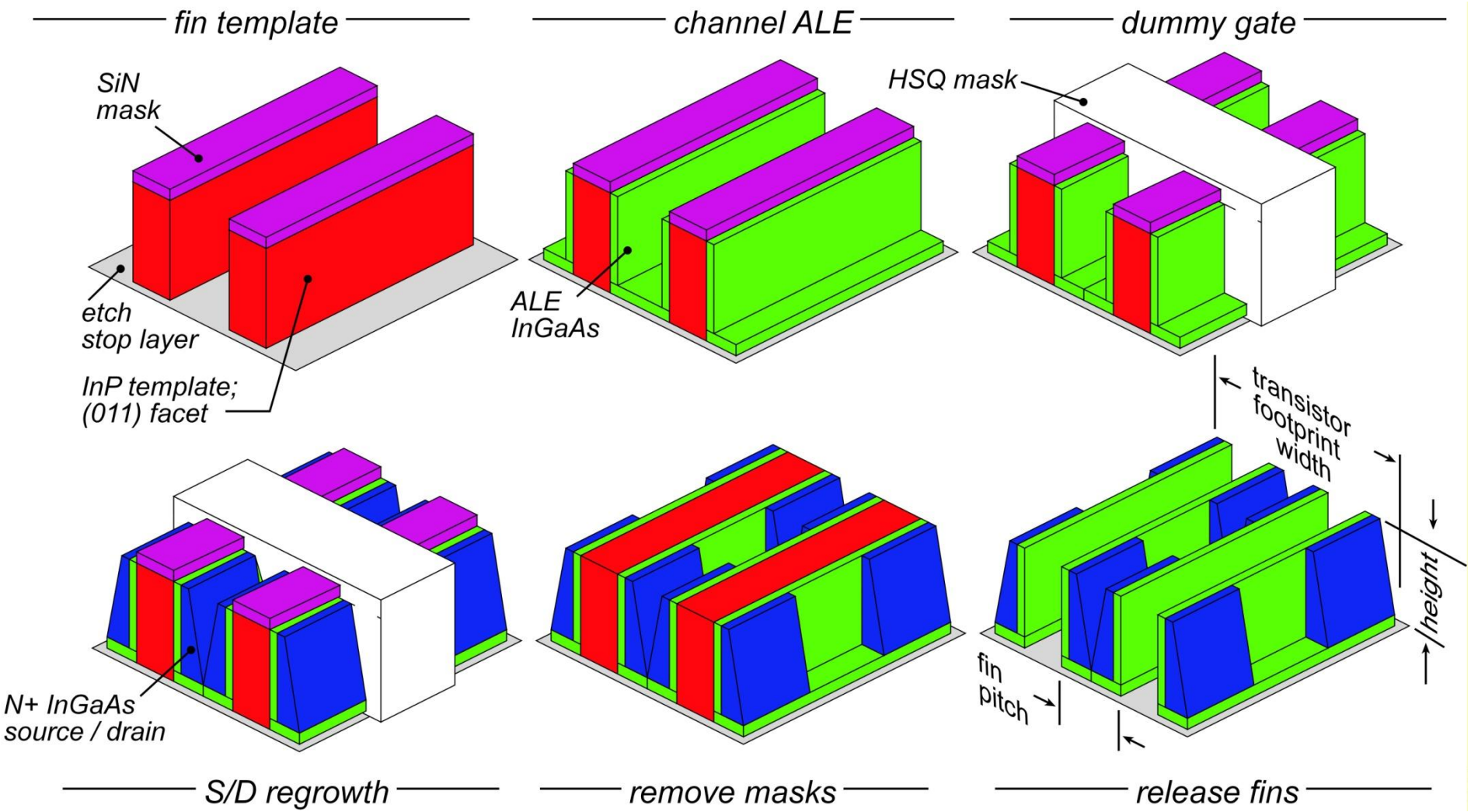
Problem:
low mobility unless surfaces are very smooth

$$\mu \propto T_{body}^6 / \delta T_{body}^2$$

Implication: At sub-8-nm gate length, need :
atomically-smooth interfaces
atomically-precise control of channel thickness

side benefit: high drive current→ low-voltage, low-power logic

ALE-Defined finFET: Process Flow

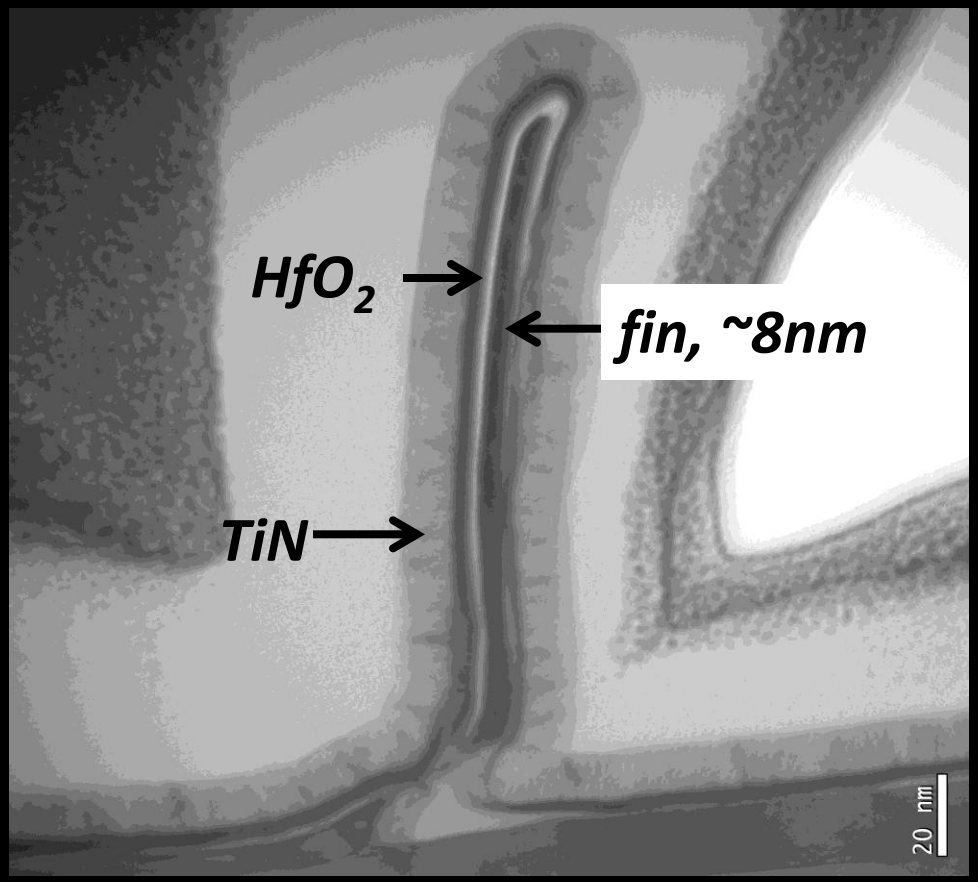


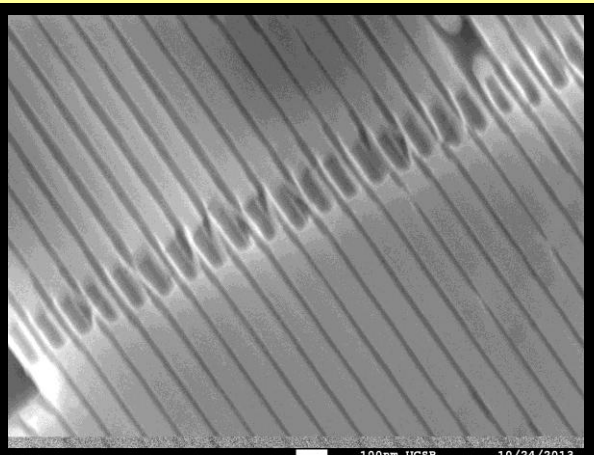
Fin template: formed by {110}-facet-selective etch→ atomically smooth

Channel thickness set by ALE growth→ atomically precise

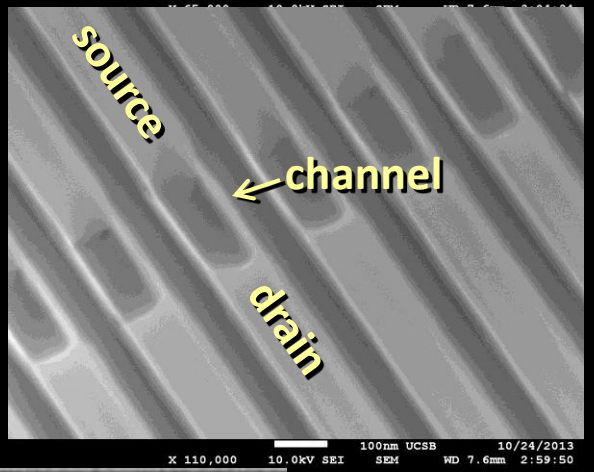
Not shown: gate dielectric, gate metal, S/D metal

Images



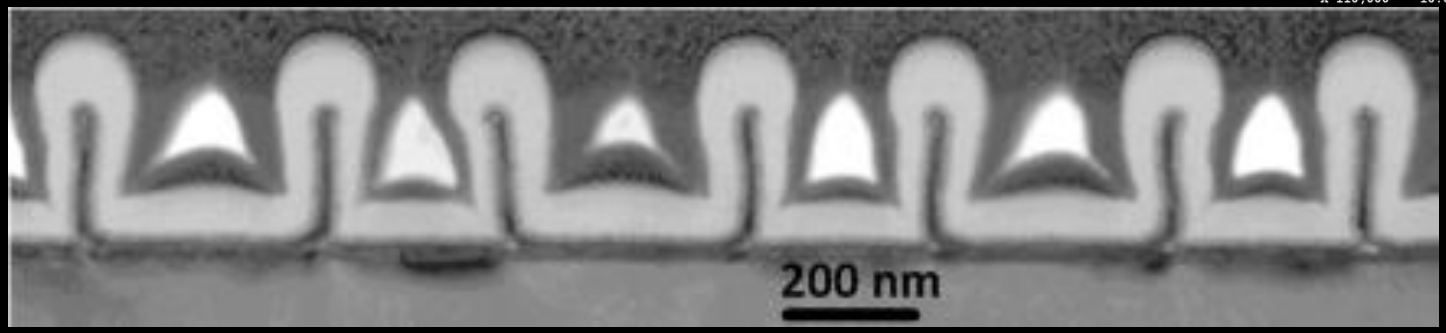


50 nm fin pitch



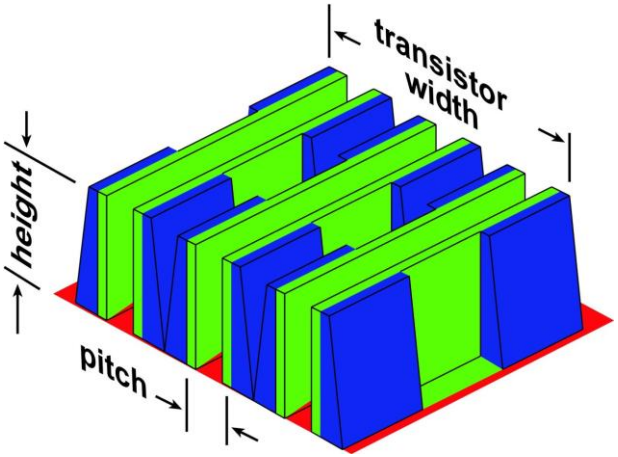
100 nm fin pitch

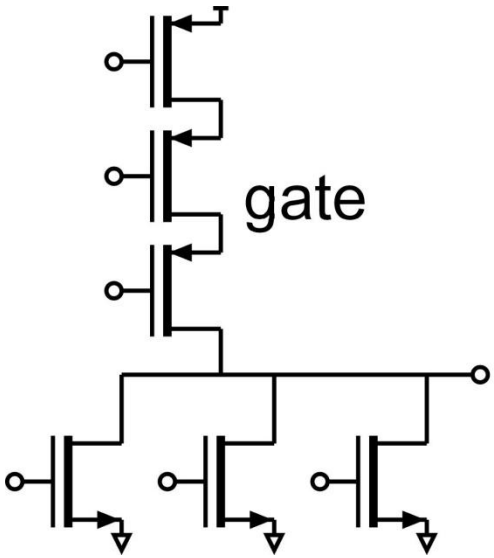
10 nm thick fins, 100 nm tall



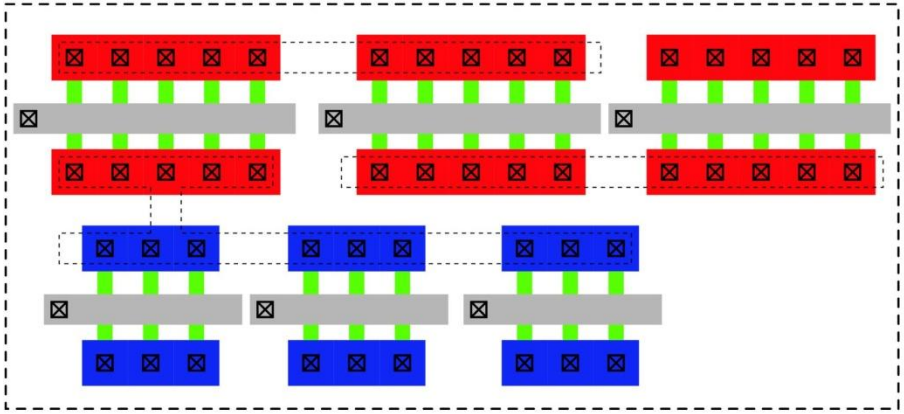
Goal: Tall Fins for High Drive Current

$$\frac{current}{transistor\ width} = J_{surface} \cdot \frac{fin\ height}{fin\ pitch}$$

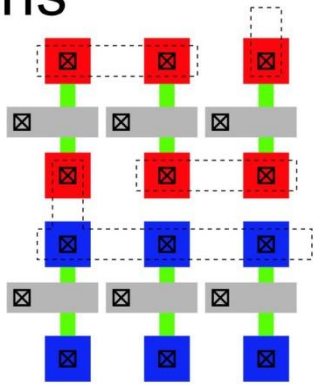




low-current fins



high-current fins

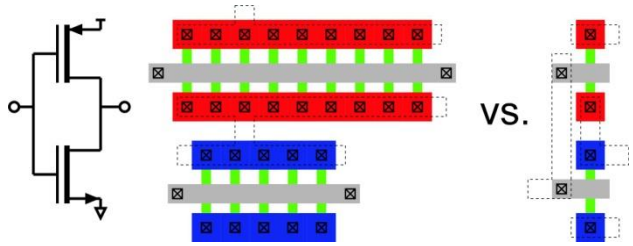


Goal: fin height >> fin pitch (spacing)→ more current per fin → less fins needed → higher integration density

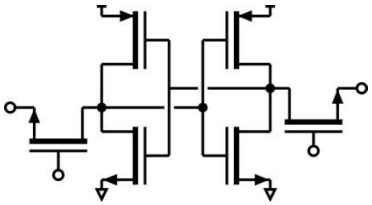
Higher density→ shorter wires→ smaller $C_{wire}V_{dd}/I$, $C_{wire}V_{dd}^2/2$

Is the IC Area Reduction Significant ?

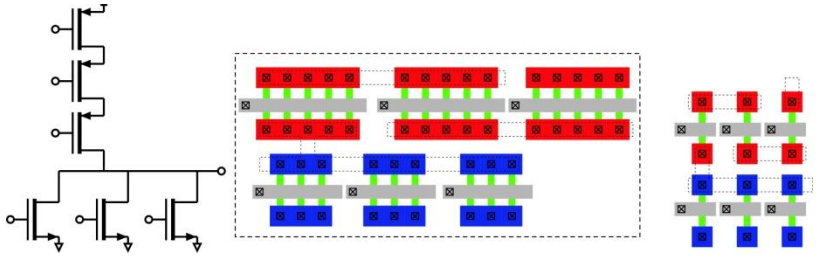
Clock/interconnect drivers need large drive currents.
Area reduction for these is likely substantial.



FETs in **Cache Memory & Registers** are drawn at minimum width
No area reduction for these.



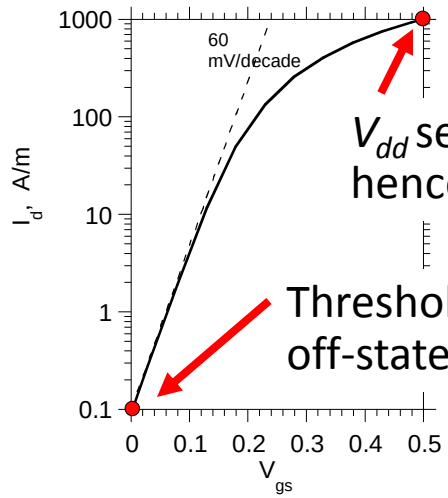
Most, but not all, **Logic Gates** will be drawn at minimum width.



Benefit must be evaluated by VLSI architect, not by device physicist.

300 mV Logic: Can We Address The $CV^2/2$ Limit ?

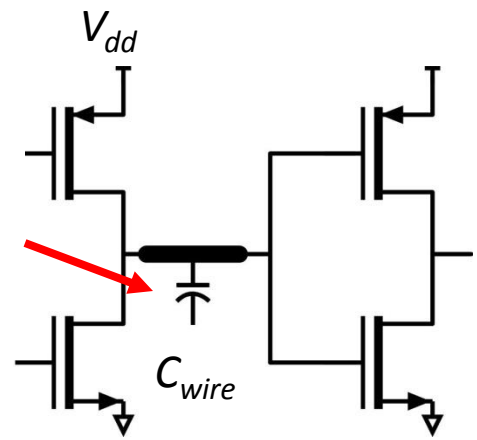
The $CV^2/2$ dissipation limit



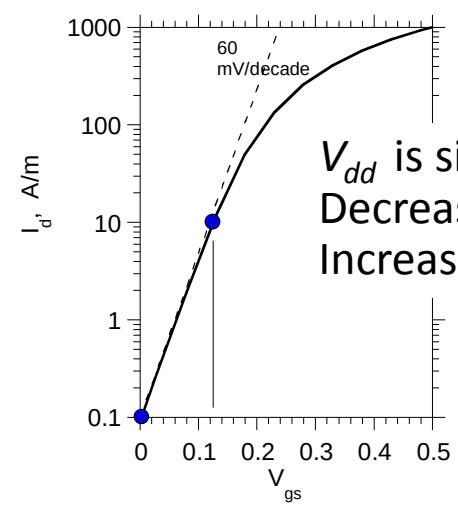
V_{dd} set for target I_{on}, hence acceptable CV_{dd}/I_{on}.

Threshold set for acceptable off-state dissipation $I_{off}V_{dd}$.

With minimum C_{wire}, a minimum switching energy $C_{wire}V_{dd}^2/2$ is set

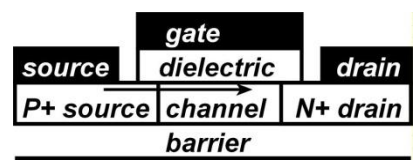


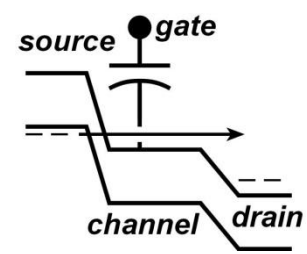
Subthreshold logic



V_{dd} is simply reduced.
Decreases energy $CV_{dd}^2/2$.
Increases delay CV_{dd}/I_{on}.

Tunnel FETs



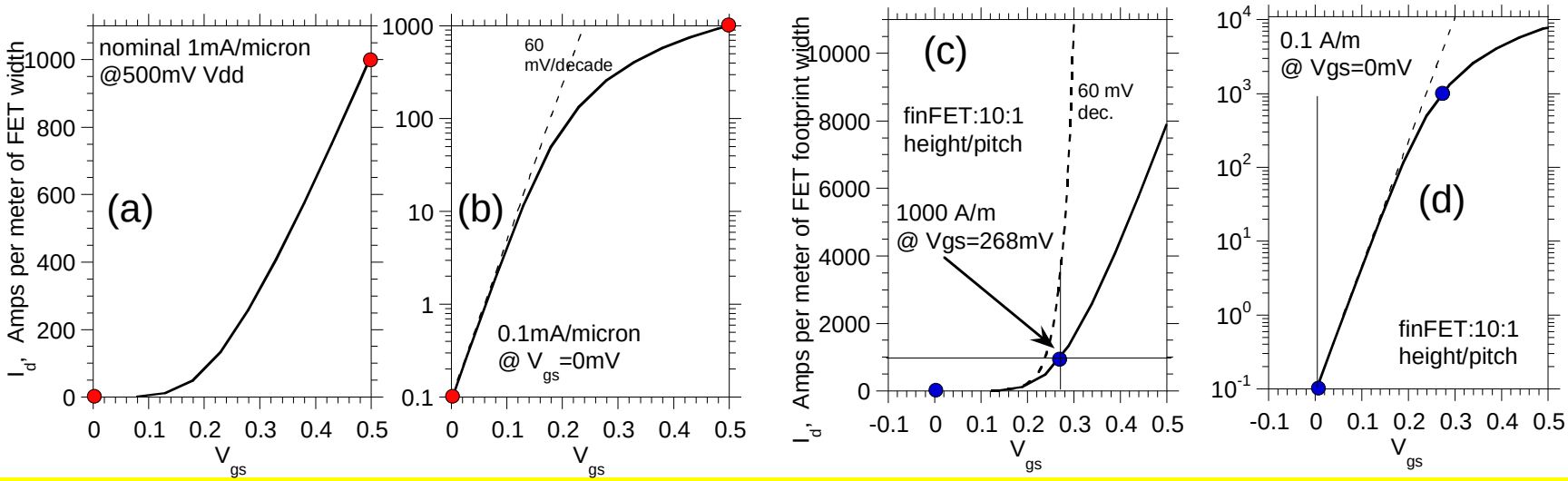


Bandgap of P+ source truncates thermal distribution.

Potential for low I_{off} at low V_{dd}.

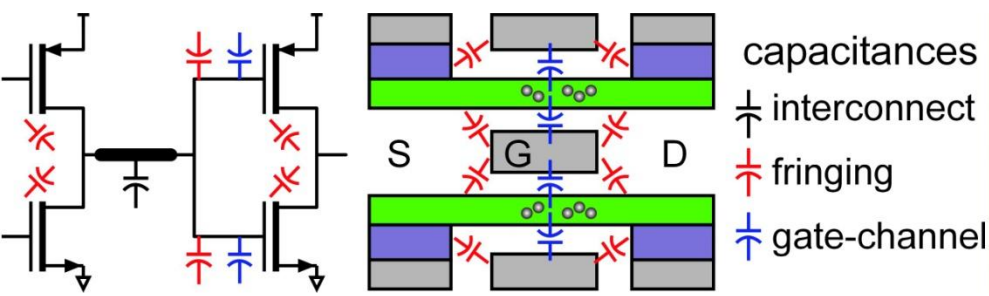
Obtaining high I_{on}/V_{dd} is the challenge.

Goal: Tall Fins for Low-Power, Low-Voltage Logic



Supply reduced from 500mV to 268 mV while maintaining high speed.

3.5:1 power savings ? Must consider FET capacitances.



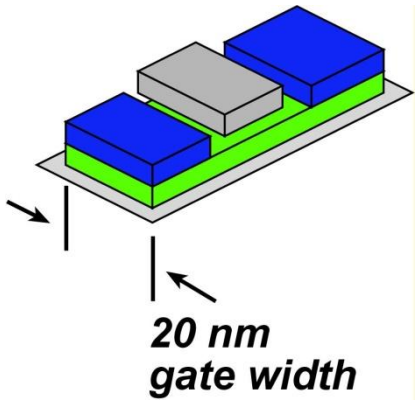
$$C_{total} \approx 0.2\,\text{fF}/\mu\text{m} \cdot L_{wire} \quad \text{interconnect}$$
$$+15 \cdot 0.3\,\text{fF}/\mu\text{m} \cdot W_g \quad \text{fringing}$$
$$+3 \cdot I_{on} L_g / v_{injection} V_{DD} \quad \text{gate-channel}$$

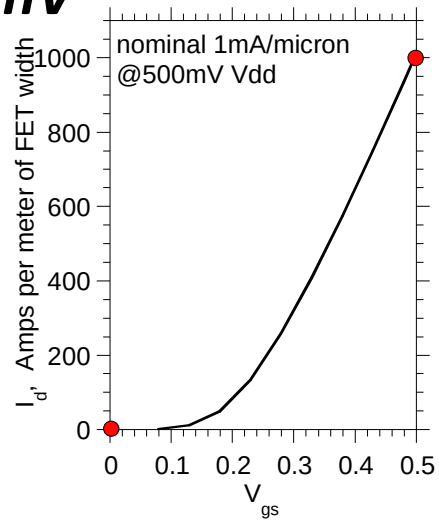
$$\tau_{gate} \approx C_{total} V_{DD} / 2 I_{on} \quad \text{delay}$$
$$E_{sw} \approx C_{total} V_{DD}^2 / 2 \quad \text{energy}$$

Assumes (Hodges & Jackson, 2003): (1) Charge-control analysis (2) $I_{on,PFET} / W_g = 0.5 * I_{on,NFET} / W_g$ (3) FO=FI=1

Power and Delay Comparison

Planar FET, V_{dd}=500 mV





I_{on}=20 μA, I_{off}=2nA

$C_{g\text{-}ch}=I_{on}L_g/v_{inj}V_{dd}$=1.3 aF

$C_{gd\text{-}f} = C_{gs\text{-}f}$ = 6 aF

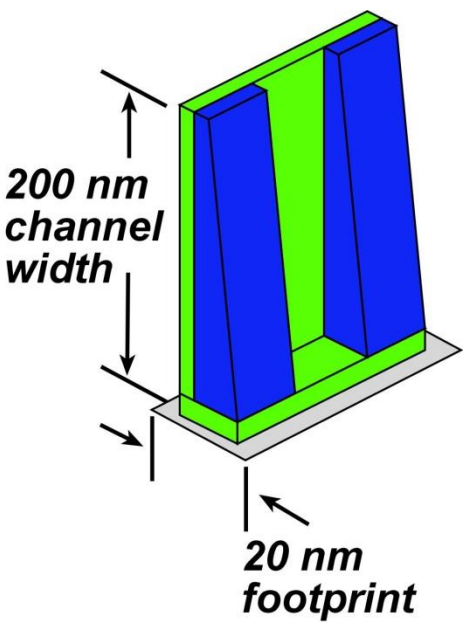
C_{wire}= 2 fF (10 μm length)

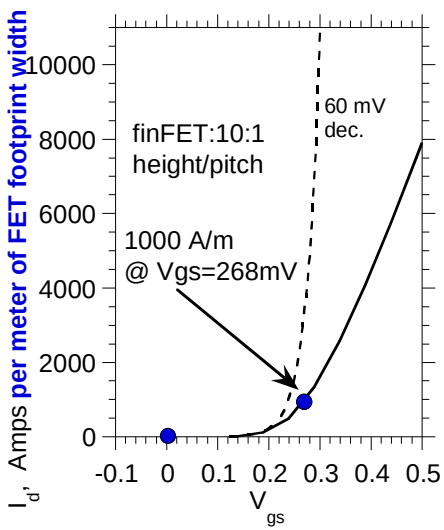
C_{total} = 2.1 fF (various multipliers)

delay= 52 ps

$C_{total}V_{DD}^2$= 0.26 fJ

tall finFET, V_{dd}=268 mV





I_{on}=20 μA, I_{off}=2nA

$C_{g\text{-}ch}=I_{on}L_g/v_{inj}V_{dd}$=3.7 aF

$C_{gd\text{-}f} = C_{gs\text{-}f}$ = 60 aF

C_{wire}= 2 fF (10 μm length)

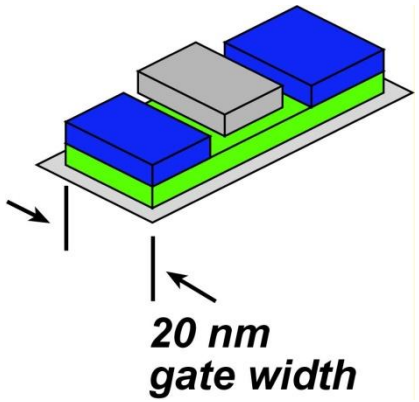
C_{total} = 2.9 fF (various multipliers)

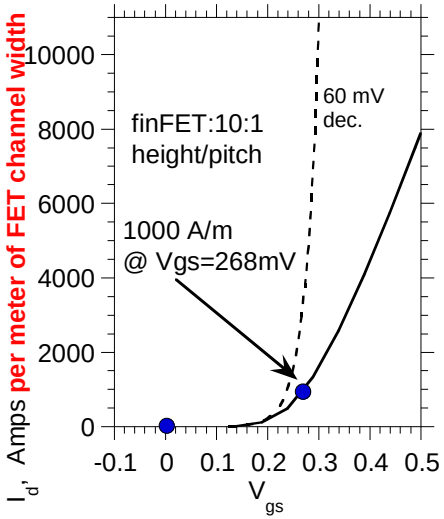
delay= 39 ps

$C_{total}V_{DD}^2$= 0.11 fJ

Why tall finFETs ? Why Not Just Subthreshold Logic ?

Planar FET, V_{dd}=268 mV

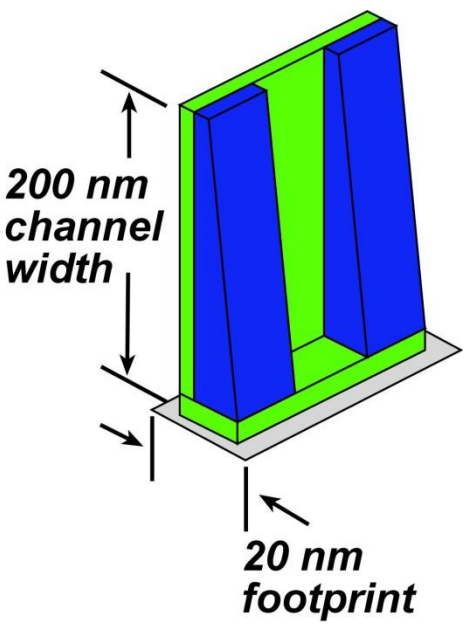


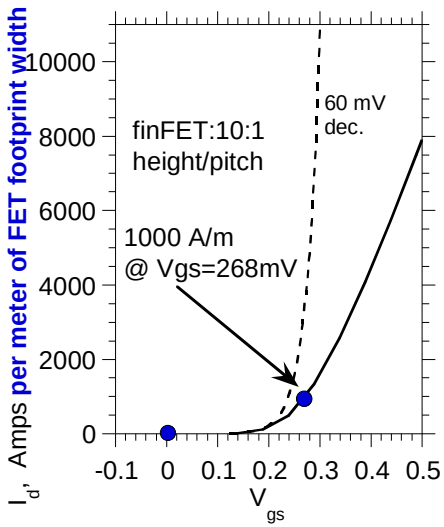


I_{on}=2.0 μA

Low I_{on}→ large CV_{DD}/I_{on} delay, subthreshold logic is slow.

tall finFET, V_{dd}=268 mV

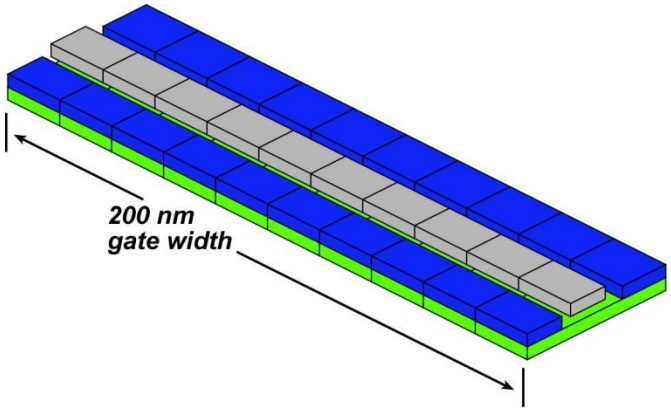


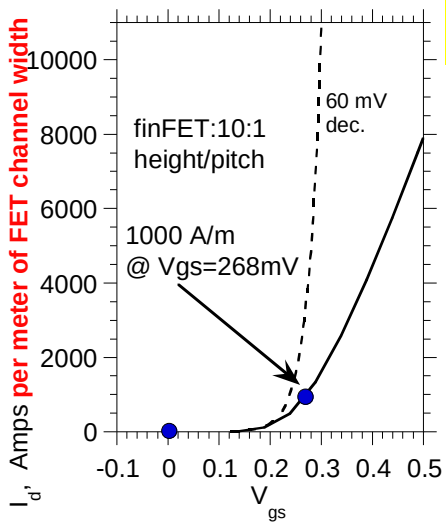


I_{on}=20 μA

Why tall finFETs ? Why Not Just Subthreshold Logic ?

Planar FET, V_{dd}=268 mV



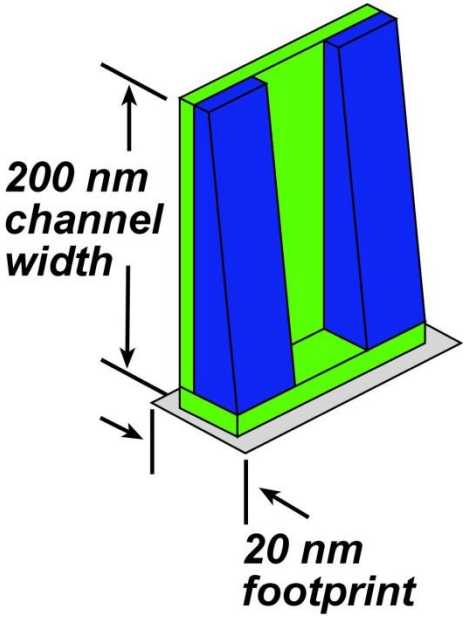


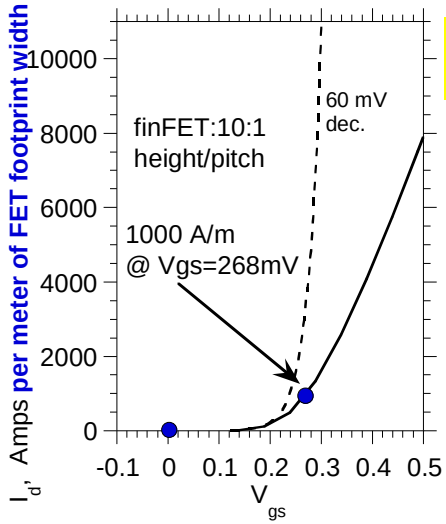
I_{on}=20 μA

Die size increased 10:1

(also: longer interconnects, etc)

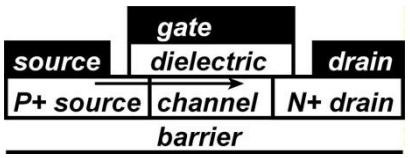
tall finFET, V_{dd}=268 mV

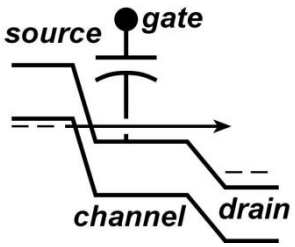




I_{on}=20 μA

Tunnel FETs & High-Aspect-Ratio Fins





Quick performance estimate:

Assume, for a moment, that P/N tunneling probability is 10%*.

Typical of the best reported ohmic contacts .*

Then on-currents for tunnel FETs are ~10:1 smaller than that of normal FETs.

Unless I_{on}/W_g is high, tunnel FETs will suffer from either large $C_{wire}V/I$ gate delays or (increasing FET widths) large die areas.

Using high-aspect ratio fin structures, tunnel FET drive currents can be increased. Parasitic fringing capacitance will then also contribute to CV/I & CV^2.

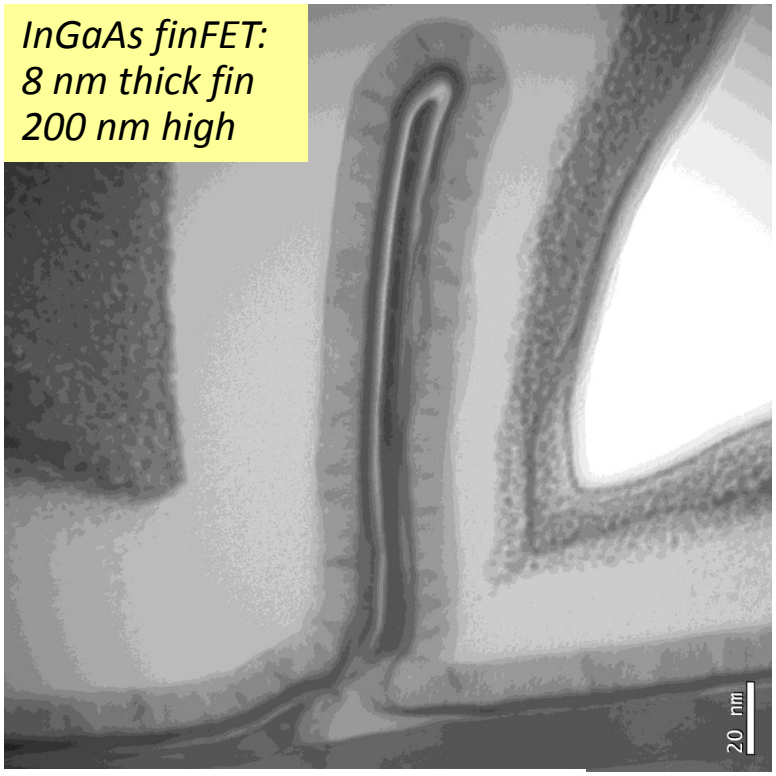
*Contact to N-InGaAs @ 6E19/cm^3 doping: $m^*=0.1m_0$, 0.2 eV, 0.5 nm barrier
Baraskar, et al: Journal of Applied Physics, 114, 154516 (2013)

finFETs Defined by Atomic Layer Epitaxy

Fin thickness defined by Atomic layer epitaxy
→ nm thickness control
Fin height defined by sidewall growth
→ 200 nm high fins

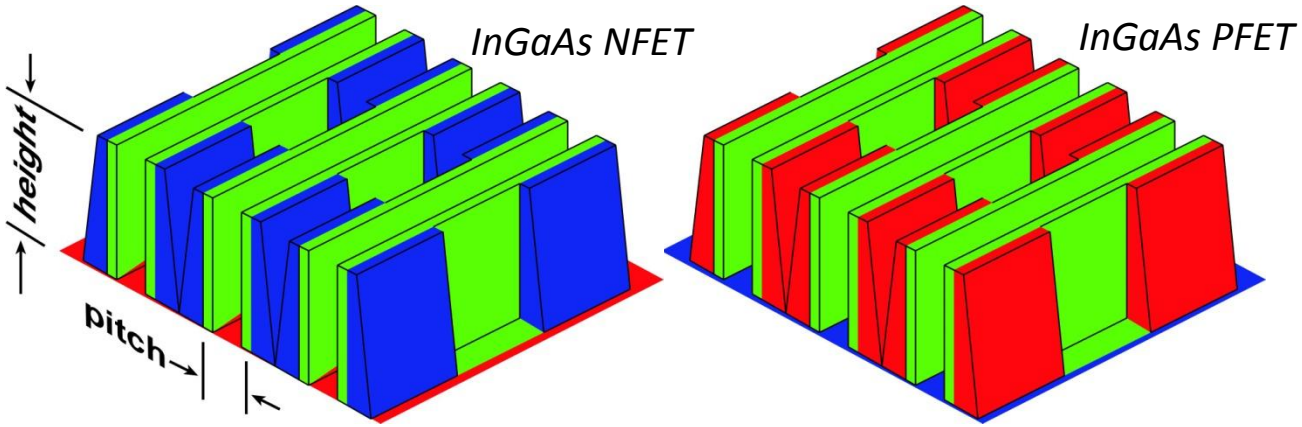
Benefits:
Enables ~4 nm fin bodies→ 8 nm gate length
10:1 more current per unit die area
→ smaller IC die area
Enables high speed, ultra low-power logic,
V_{dd}~300 mV

InGaAs finFET:
8 nm thick fin
200 nm high



D. Elias, DRC 2013, June, Notre Dame

height>> pitch



(end)

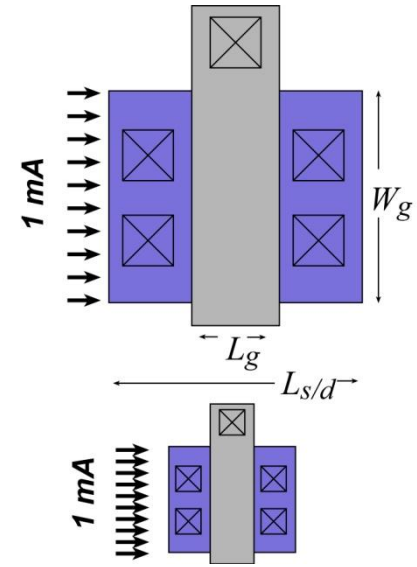
Backups

Lithographic Scaling vs. 3-D for High-Density Logic

Past VLSI Scaling: more FETs per IC because of
(1) shorter gate & contact lengths
(2) increased mA/micron→ less gate width W_g.

Today, I_{on}/W_g (mA/micron) is not increasing soon, once S/D tunneling dominates, I_{on}/W_g will start to ***decrease***

Sub-16nm lithography is also difficult.
Further reductions in feature size are expensive.

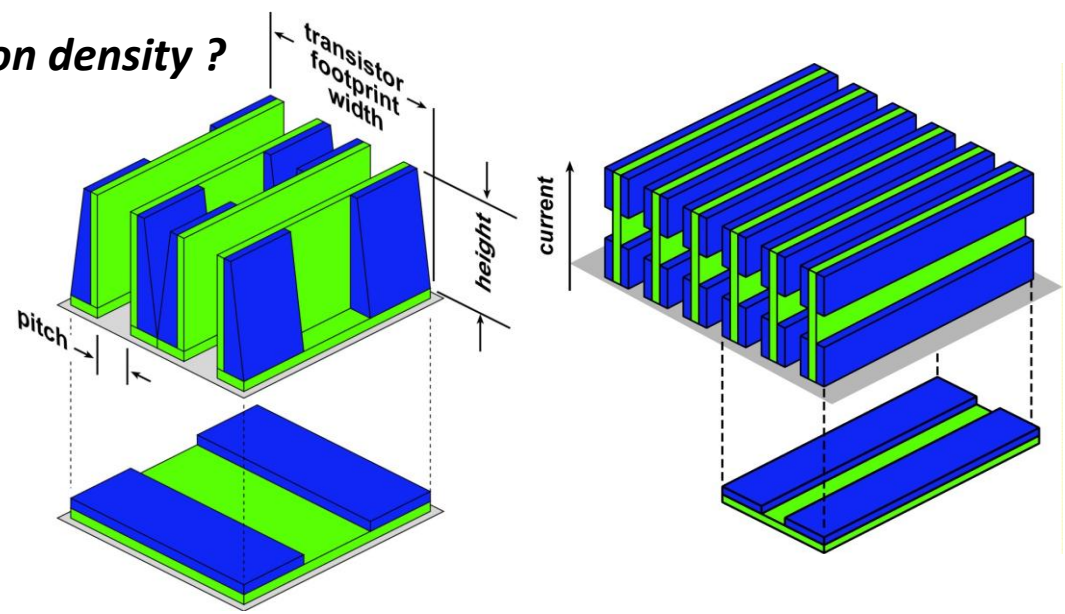


Can 3-D transistors increase integration density ?

Clear: increased I_{on}
for a given FET footprint size.

Clear: can suppress S/D tunneling:
tunneling distance >>
lithographic distance.

Less clear: decreased size
of minimum-geometry FET.

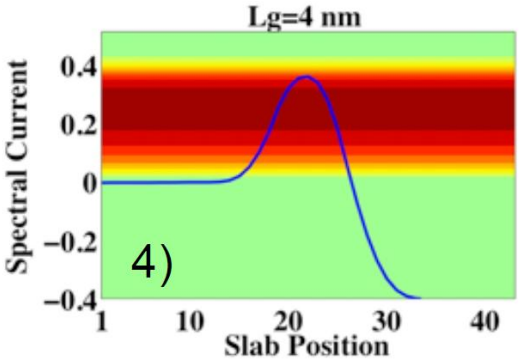


Scaling in the S/D tunneling limit

At 4-8 nm Gate Lengths,
high leakage from source/drain tunneling

$$J_{S/D\,tunnel} \propto \exp\left(-(2m^* qV_{th})^{1/2} L_g / \hbar\right)$$

increases exponentially as gate length is reduced.

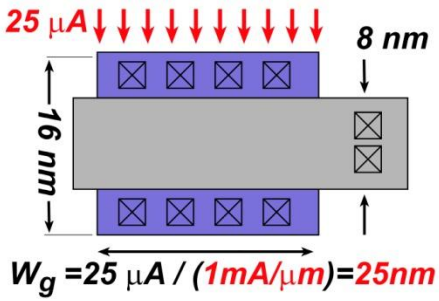


M. Luisier, ETH Zurich / Purdue

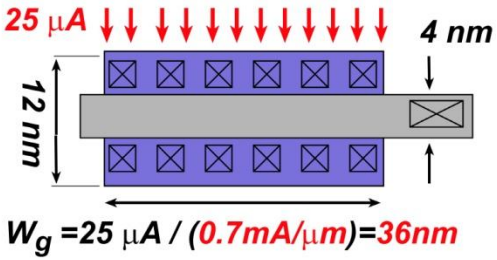
Reducing tunneling through increased mass can be counterproductive

increasing m* → less tunneling, but lower FET on-current→ need more die area

Before:
25 μA in
400 nm²

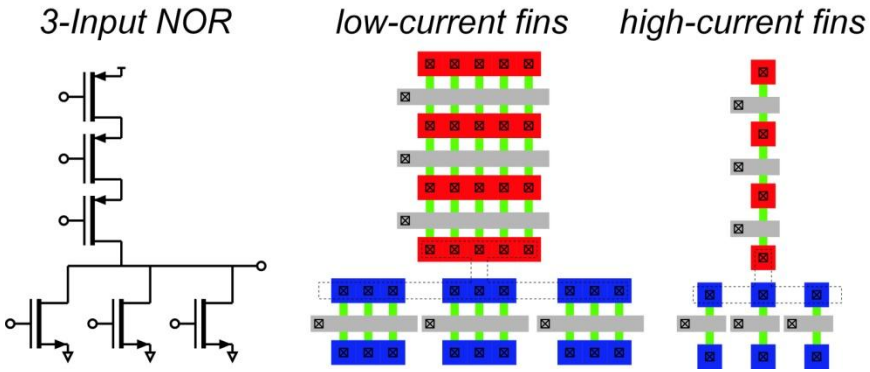


After:
25 μA in
432 nm²



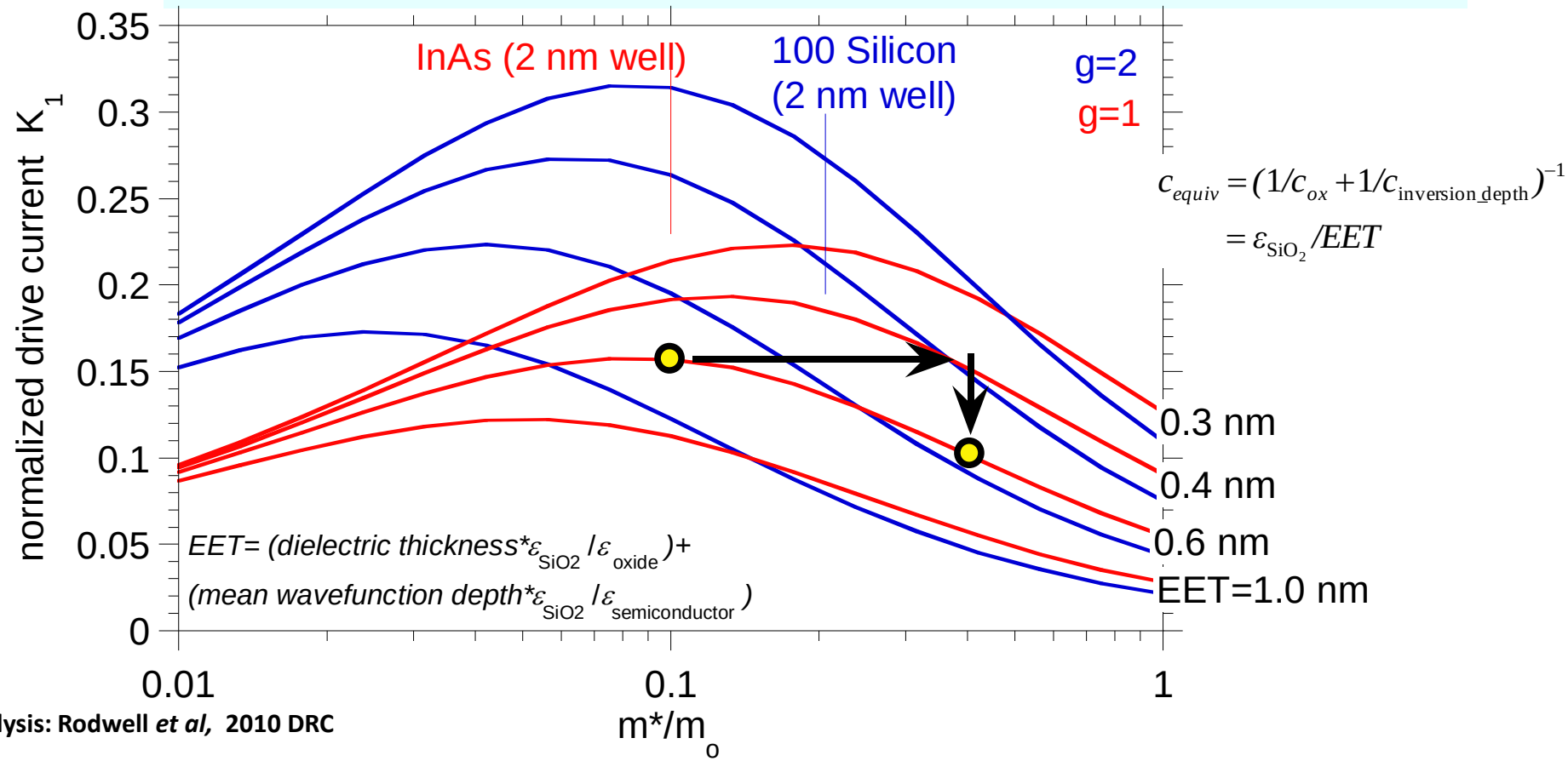
Instead: Ultra-tall fins to increase
the integration density

example: 3-input NOR gate.
other cases: clock & interconnect drivers



Minimum gate length: source-drain tunneling (2)

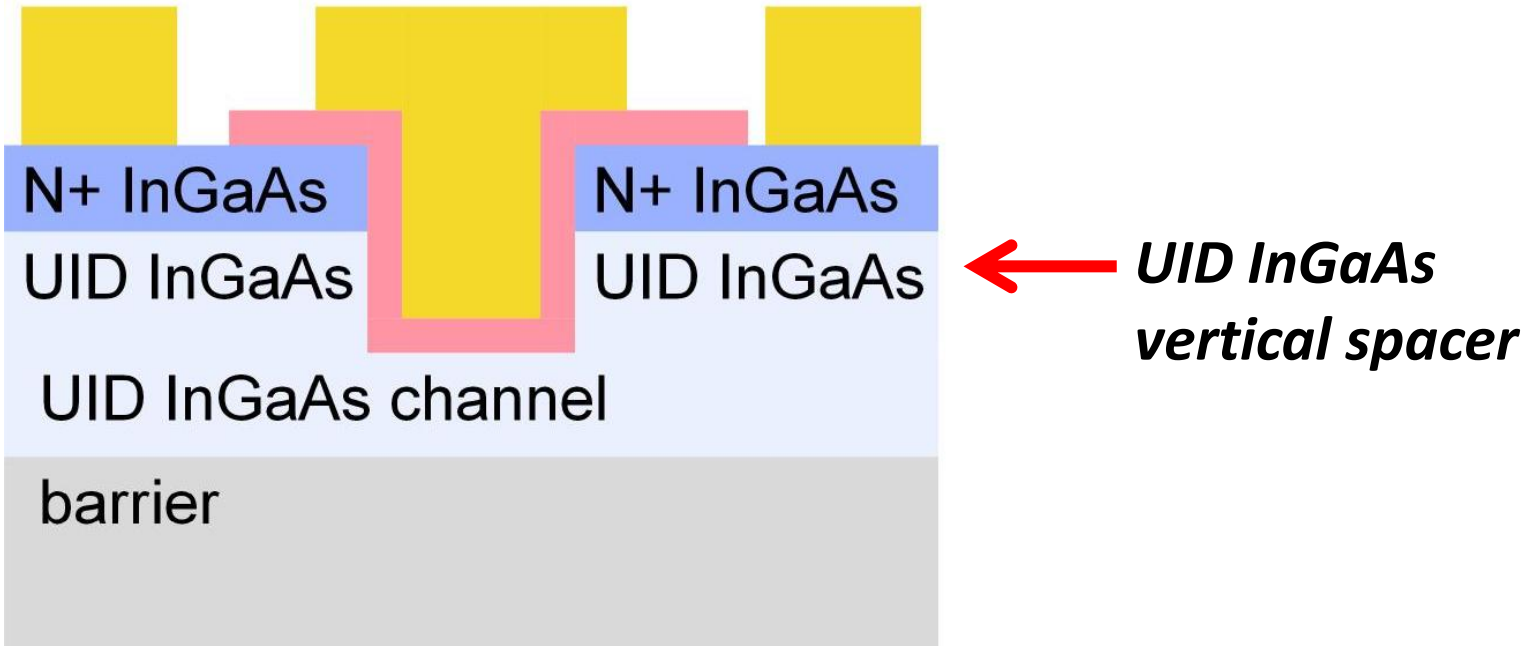
$$J = K_1 \cdot \left(84\frac{\text{mA}}{\mu\text{m}}\right) \cdot \left(\frac{V_{gs}-V_{th}}{1\,\text{V}}\right)^{3/2}, \quad \text{where } K_1 = \frac{g \cdot \left(m^*/m_o\right)^{1/2}}{\left(1+(c_{dos,o}/c_{equiv}) \cdot g \cdot (m^*/m_o)\right)^{3/2}}$$



analysis: Rodwell *et al*, 2010 DRC

increased m* → decreased I_{on}/W_g → decreases packing density

Geometric Solutions to S/D Tunneling

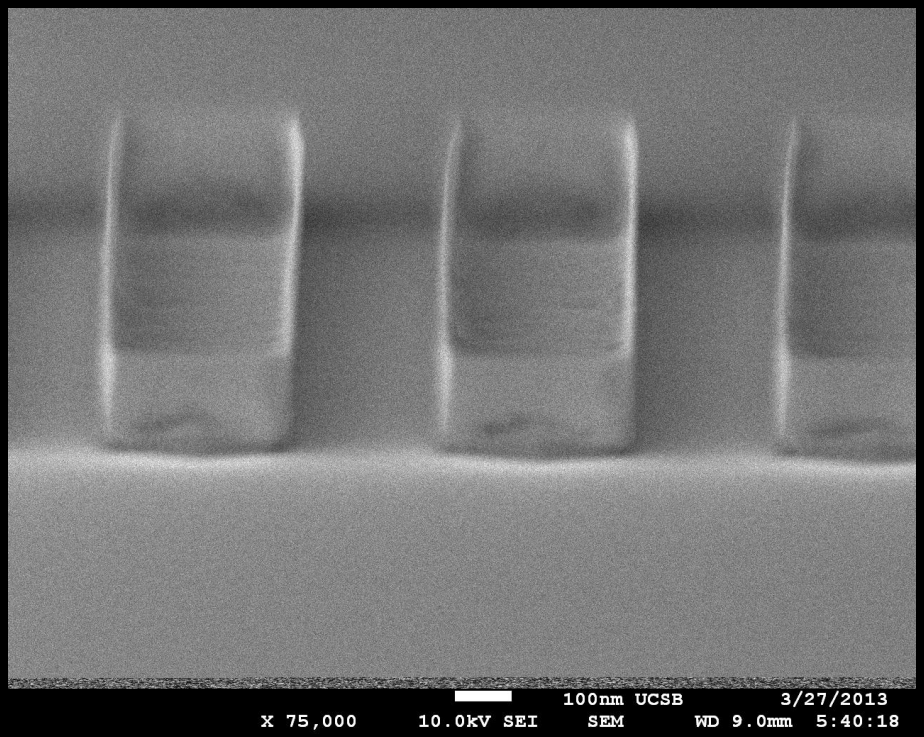


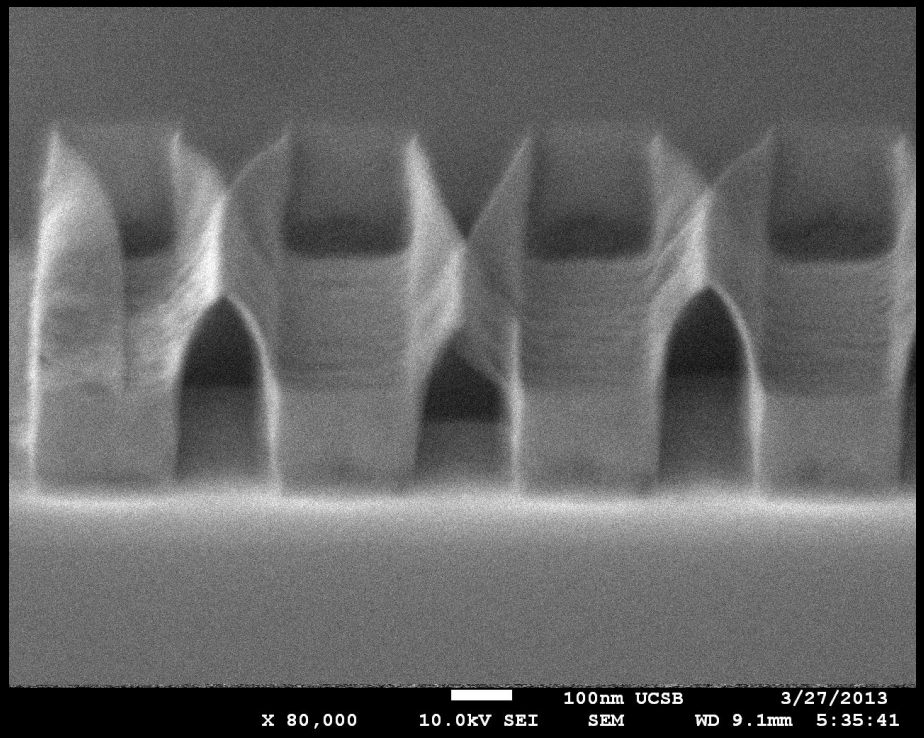
Transport (& tunneling) distance larger than lithographic gate length.
Feature used NOW in our current planar FETs.
Can be incorporated in high-aspect-ratio finFETs

Why Not Release Fins Before S/D Regrowth ?

D. Elias, DRC 2013, June, Notre Dame

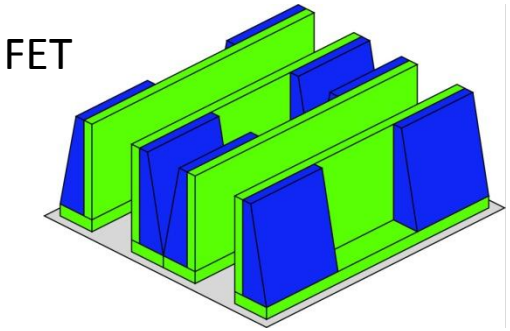
Images of released ~10 nm fins:

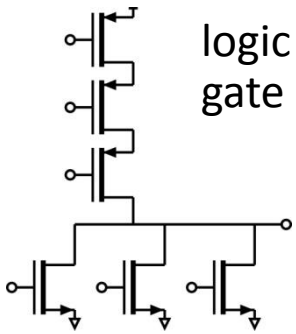


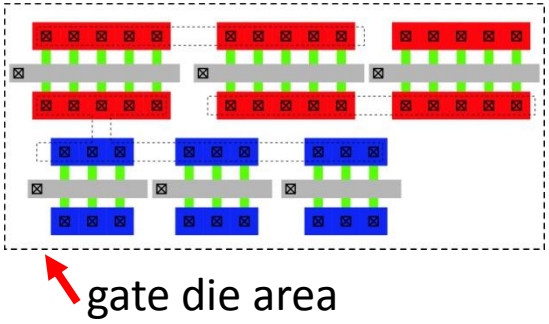


S/D regrowth provides mechanical support

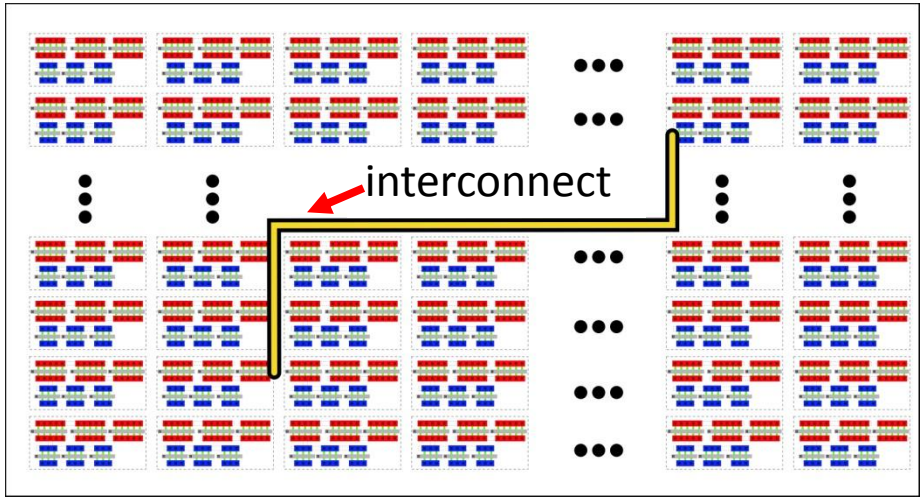
Wire Lengths & Wire Capacitances in VLSI







Integrated Circuit Layout



$$\text{IC area} \propto \text{logic gate area}$$
$$\propto \text{transistor footprint area}$$

$$\text{mean wire lenght} \propto (\text{transistor footprint area})^{1/2}$$

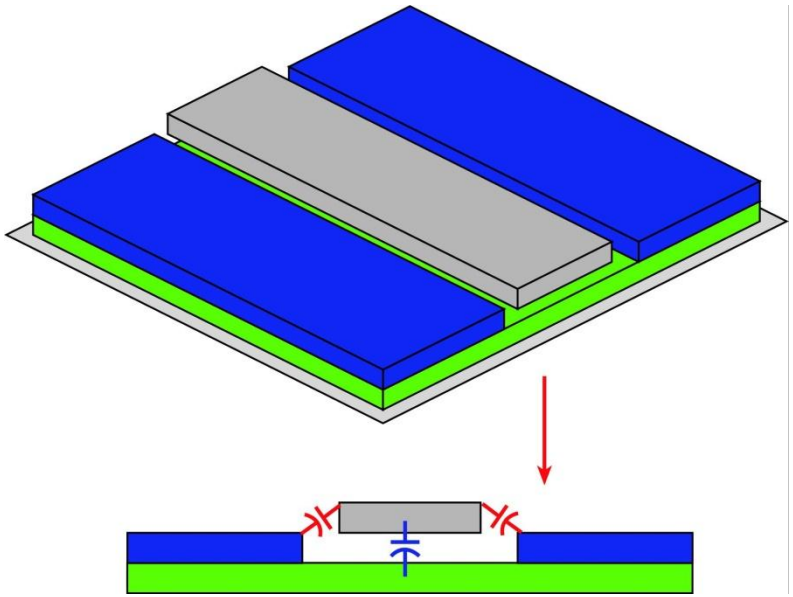
$$\text{wiring capacitance} \approx 0.2\ \text{fF}/\mu\text{m}$$

$$\text{mean wiring capacitance}$$
$$\propto (\text{transistor footprint area})^{1/2}$$

more current per fin→ less fins needed → higher integration density

more current per fin→ shorter wires→ smaller $C_{wire}V_{dd}/I$, $C_{wire}V_{dd}^2/2$

FET Capacitances, Interconnect Capacitances



fringing capacitance

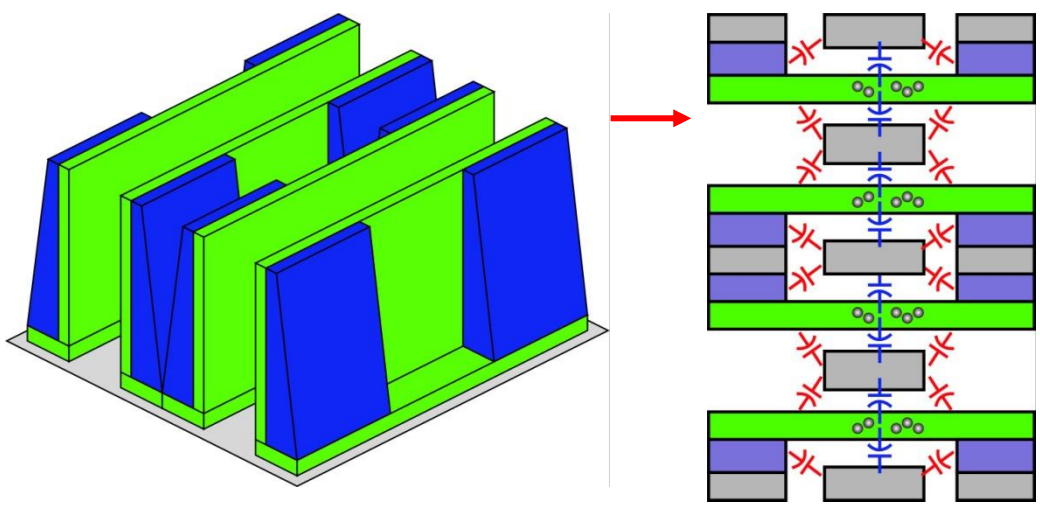
$$C_{gd,f} = C_{gs,f} \approx 0.3\,\mathrm{fF}/\mu\mathrm{m} \cdot W_g$$

gate - channel capacitance

$$C_{g-ch} \approx I_{\mathrm{on}} L_g / v_{\mathrm{injection}} V_{DD}$$

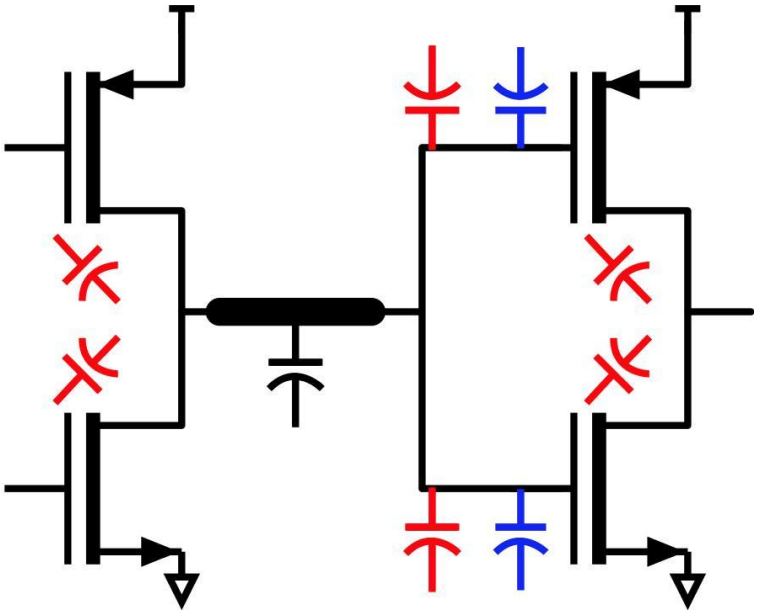
interconnect capacitance

$$C_{wire} \approx 0.2\,\mathrm{fF}/\mu\mathrm{m} \cdot L_{wire}$$



Similar capacitances in finFET

Gate Capacitance, Energy, and Delay



fringing capacitance

$$C_{gd,f} = C_{gs,f} \approx 0.3\,\text{fF}/\mu\text{m} \cdot W_g$$

gate-channel capacitance

$$C_{g-ch} \approx I_{\text{on}} L_g / v_{\text{injection}} V_{DD}$$

interconnect capacitance

$$C_{wire} \approx 0.2\,\text{fF}/\mu\text{m} \cdot L_{wire}$$

$$\begin{aligned} C_{total} \approx\ & 0.2\,\text{fF}/\mu\text{m} \cdot L_{wire} && \text{interconnect} \\ & +15 \cdot 0.3\,\text{fF}/\mu\text{m} \cdot W_g && \text{fringing} \\ & +3 \cdot I_{\text{on}} L_g / v_{\text{injection}} V_{DD} && \text{gate-channel} \end{aligned}$$

$$\tau \approx C_{total} V_{DD} / 2 I_{on} \quad \text{delay}$$

$$E_{sw} \approx C_{total} V_{DD}^2 / 2 \quad \text{energy}$$

Assumes:

(1) Charge-control analysis*

(2) $I_{on,PFET} / W_g = 0.5 * I_{on,NFET} / W_g$

(3) FO=FI=1

*Hodges & Jackson, 2003