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

Mark Rodwell, Doron Elias,
ECE Department, University of California, Santa Barbara, USA, rodwell@ece.ucsb.edu

As we reduce transistor capacitances, node capacitances are limited by wiring, setting a minimum power dissipation determined by the number of gates, the mean wire length, the mean switching rate, and the supply voltage V_{DD}. With thermally-activated, FETs, the off-state leakage I_{off} and target on-current I_{on} then determine the minimum feasible V_{DD}, and the IC clock frequency can then be increased only at the expense of increased power consumption. Tunnel transistors [1] offer subthreshold characteristics steeper than 60mV/decade, but achieving high I_{on} at low I_{off} and low V_{DD} is challenging. Subthreshold logic [2] operates at low V_{DD}, but is slow because of low I_{on}. Here we propose low-power logic using high-aspect-ratio finFETs, devices we have fabricated with few-nm body thicknesses and 180nm height [3]. If these can fabricated at ~20nm pitch, then the fin surface area can exceed its footprint area —i.e. the area the transistor occupies on the IC—by ~10:1. IC performance can be then improved by maintaining fixed V_{DD}, but with reduced FET footprint area hence reduced die size and therefore reduced wiring capacitance, or can be improved by reducing V_{DD} to ~300mV while maintaining large I_{on} per unit IC die area.

We form high-aspect-ratio finFETs (fig. 1) by first etching ridges InP using H_3PO_4:HCl. The InGaAs channel is then grown by atomic layer epitaxy (ALE). This permits single-monolayer control of fin width, and ~4nm thick fins, both attributes necessary for ~8nm gate length L_g [3]. Because the H_3PO_4:HCl etch is selective to the (011) facet, very tall, ~200nm, yet vertical fins can be formed (fig. 2). The on-current per unit transistor footprint width increases in proportion (fig. 1) to the ratio of fin height to fin pitch.

With V_{DD} held constant, the increased drive current per unit die area reduces the die size (fig. 3). While the majority of FETs in VLSI are drawn at minimum width, for delay-critical gates (clock buffers, interconnect drivers, key logic paths), FET size is selected for a desired minimum I_{on} so as to reduce the $C_{wire}V_{DD}/I_{on}$ delay. For such FETs, a 10:1 fin height/pitch ratio reduces by 10:1 the required FET footprint width and footprint area, reducing the IC die area, hence the mean wire length, mean interconnect capacitance, mean $C_{wire}V_{DD}/I_{on}$ delay, and mean $C_{wire}V_{DD}^2/2$ dissipation. It is desire for increased I_{on} which drives interest in III-V MOS; with high-aspect-ratio finFETs, high I_{on} per unit die area is obtained both through transport and through geometry.

With high aspect ratio finFETs, reducing V_{DD} will reduce power consumption while maintaining low delay. Assuming ballistic [4] *I-V* characteristics (fig. 4), with nominal 0.1A/m I_{off} and 1000A/m I_{on} at 500 mV V_{DD}, with a 10:1 FET height/pitch ratio, after shifting the threshold by 60 mV, the same 0.1A/m I_{off} and 1000A/m I_{on} is obtained at 268mV V_{DD}, *where the currents are now stated per unit transistor footprint width*. In this near-threshold operation [2], $C_{wire}V_{DD}^2/2$ power is reduced 3.5:1. Yet, we must also consider transistor capacitances.

Near-threshold operation increases the FET total channel width required for a given I_{on}, hence increases the transistor capacitances. FET capacitances are dominated (fig. 5) by gate-source and gate-drain fringing capacitances, both ~0.3 fF/μm. Interconnect capacitance is ~0.2 fF/μm. Assuming a (short) 10 μm interconnect, comparing a 200 nm tall finFET (fig. 4c,d) at 268 mV V_{DD} and 20 μA I_{on} to a reference planar FET (fig. 4a,b) at 20 nm W_g, 500 mV V_{DD} and also 20 μA I_{on}, the large finFET fringing capacitances increase the capacitance 1.4:1, hence the reduction in $C_{total}V_{DD}^2/2$ is only 2.5:1, while the delay is reduced ~20% because of the reduced V_{DD}. The benefit is greatest with large ratios of interconnect length to FET I_{on}, i.e. the low-power design case.

Tall finFETs in near-threshold operation differ from tunnel FETs and from planar FETs in subthreshold operation. In all cases, low I_{on} per unit transistor width increases the transistor capacitive loading if the FET width is increased to maintain drive current. However, only with the tall finFETs is large drive current per unit IC die area maintained; without this feature, a 10:1 decrease in I_{on} per unit die area forces either a direct 10:1 increase in gate delay or, if FET areas are correspondingly increased, increased die area, increased mean interconnect length, increased mean interconnect capacitance, and correspondingly increased $C_{wire}V_{DD}^2/2$ dissipation.

Work supported by the NSF-NEB Program. Work done in the UCSB Nanofabrication Facility, part of the NSF-funded NNIN network, and in the MRL Central Facilities supported by the MRSEC Program of the NSF under Award MR05-20415. We thank Saumitra Mehrotra and Michael Povolotskyi (Purdue) for FET simulation (fig. 4ab).

References:

[1] G. Zhou *et al*, 2012 IEDM, December 10-12, San Francisco.
[2] S. Vitale *et al*, IEEE Proceedings, Vol. 98, No. 2, February 2010, pp. 333-342.
[3] D. Elias *et al*, 2013 IEEE DRC, June 23-26, Notre Dame. See also J. J. Gu *et al*, 2012 IEDM, Dec. 10-12, San Francisco.
[4] A. Rahman *et al*, IEEE TED, vol.50, no.9, pp.1853-1864, Sept. 2003.

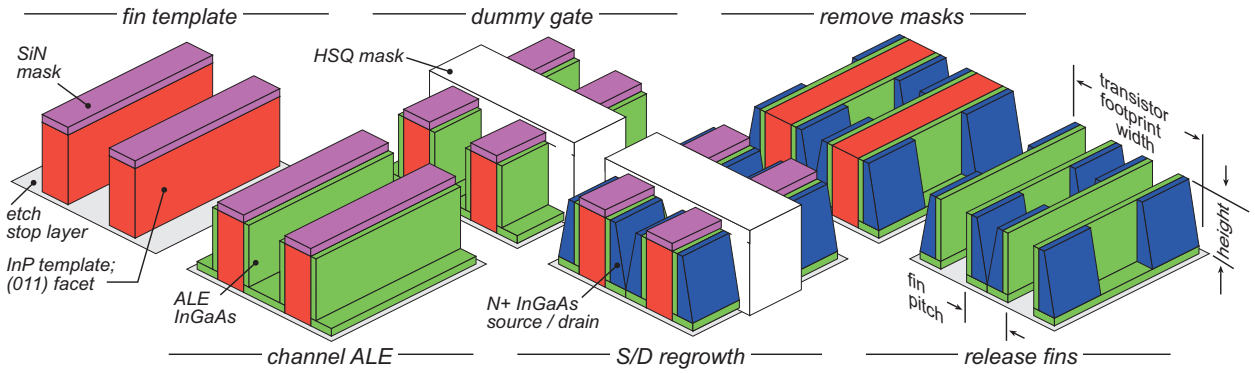


Figure 1: Process flow for high aspect ratio finFETs. The InP template is defined by an wet etch selective to the vertical (011) facet. ALE growth can provide monolayer control of the channel thickness. Not shown: ALD gate dielectric, gate & S/D metals.

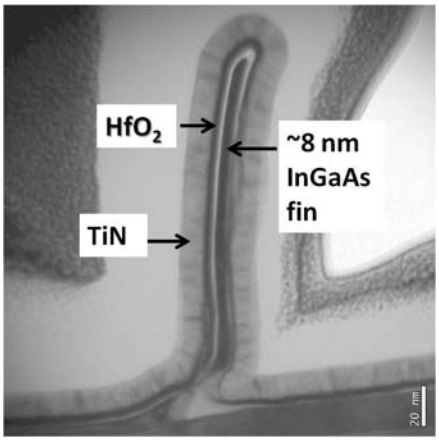


Figure 2: TEM image of a single fin of ~180nm height

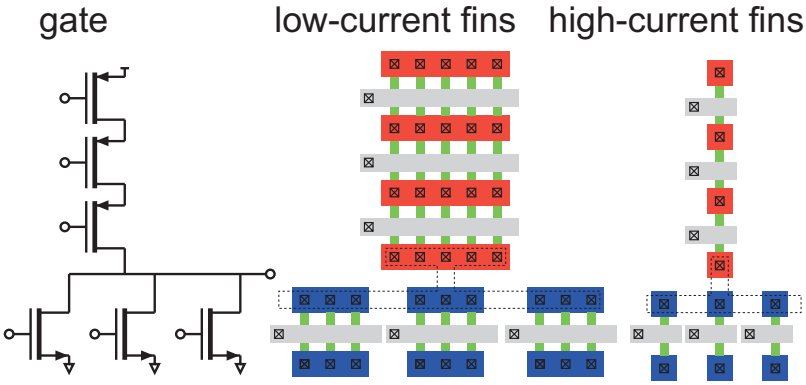


Figure 3: At fixed V_{dd}, increasing the fin height reduces the number of fins, and the gate footprint area, of delay-sensitive gates and of clock & interconnect drivers. This decreases the IC size, hence the mean wire length and consequently the $C_{wire}V^2/2$ switching energy.

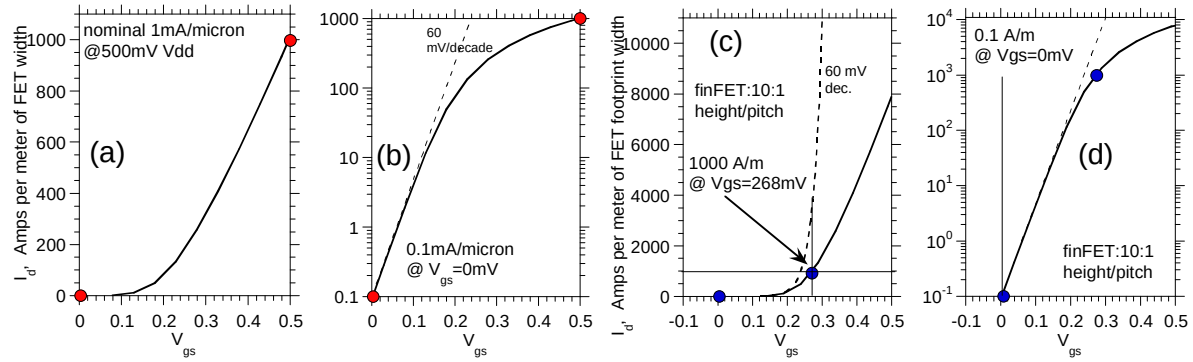
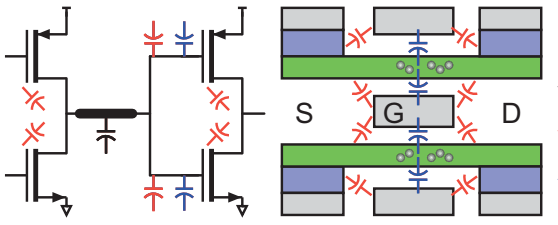


Figure 4: Simulated (a,b) characteristics of a planar ballistic FET with I_{on}=1000A/m and I_{off}=0.1A/m for a 500 mV supply. With a 10:1 height/pitch ratio, the same on- and off-current per unit *footprint width* (c,d) can obtained with a 268 mV supply.



capacitances
interconnect
fringing
gate-channel

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

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

Figure 5: Inverter capacitances include interconnect capacitance C_w, gate-source and gate-drain fringing capacitances C_f, and the capacitance associated with the inversion layer C_{inv}. In modern VLSI, C_{inv} is much smaller than C_f. In turn, in many switching nodes, C_f is smaller than C_w and will only moderately increase the node capacitance. The total capacitance expression, from charge-control analysis, assumes FO=FI=1 and $I_{on,PFET}$= $I_{on,NFET}$/2.