Invited

Experimental Technology and Performance of 0.1 μm Gate-Length Low Temperature Operation MOSFETs

G. A. Sai-Halasz

IBM Research Division, T. J. Watson Research Center,
Yorktown Heights, N.Y. 10598

Results are presented from work aimed at demonstrating Si MOSFET technology feasibility at the 0.1μm gate-length regime. Primary objective was low power consumption and reliability. Accordingly, a liquid nitrogen temperature design was chosen as the primary approach because it allows for operation at reduced voltage levels. Significant accomplishments of the work were: fabrication of NMOS devices with an extrinsic transconductance of over 940μS/μm, clear observation of velocity overshoot, and reaching per-stage-delays of 13ps in ring oscillators. These results show that it is worthwhile to pursue miniaturization down into these dimensions.

Introduction

Historically, density and performance improvements in MOSFET integrated circuits have been achieved through miniaturization. However, questions have been raised about whether the limits of this path are in sight. For this reason an investigation has recently been undertaken regarding the feasibility of FETs in the 0.1μm gate-length regime. This paper summarizes the work, starting with the design aspects and ending with circuit performance results.

Design and Processing

Because of non-scalable parameters and noise margins, operating voltage cannot be indefinitely decreased with dimensions, and one is forced to work with higher voltage levels than dictated by ideal scaling. This becomes most detrimental once performance saturates at less than the operating voltage, due to velocity saturation type characteristics. In such an environment the measure of a good design lies in achieving operation at the lowest feasible voltage. This naturally leads to a low temperature (LT), 77K, design as primary approach. At LT the main obstacles in lowering the threshold, namely, subthreshold conduction and threshold shifts due to temperature changes, are drastically reduced. Other advantages to LT operation are: improved performance, better punch

through behavior, and the possibility of counteracting the junction potential by forward biasing the substrate.

The lowest practical threshold, even at 77K, was deemed not to be below 150mV. This choice then leads to a ~0.6-0.8V bias for both the drain and the substrate[1].

Seven different chips were designed for the investigations. Three of the chips served speed measurements, while the other four housed individual devices and various parametric test sites. Devices and circuits were designed in multiple versions with varying gate-lengths. These spanned from just above 0.25μm, down to 0.07μm. Gate widths ranged from 0.5μm to 11μm. (The 0.07&x0.5μm² devices are probably the smallest FETs made to date.) Small devices were complemented by large ones, of 100μm length and 50μm width.

As a first feasibility study the process was directed toward NMOS. Eleven wafers were fabricated each with different processing variations, the details of which can be found elsewhere[2,3]. The main features were as follows. Five lithographic levels, including one for alignment, were all done by e-beam exposure. The e-beam system was able to achieve[4] a few tens of nm features, with an overall overlay accuracy of better than 30nm.

The chosen isolation, semi-recessed oxide, allowed only $0.25\mu m$ ground rules at the diffusion level. Except in the case of the gate oxide, where for reliability reasons it was relaxed, the vertical scaling was fully consistent with the nominal, $0.1\mu m$, gate-length. On all the wafers but two, 4.5nm gate oxide was grown. One had 3.3nm gate oxide, and it was fully operational when completed. The last wafer was made with very thick gate oxide for special measurements. A schematic cross section of the device structure, drawn to scale, is shown in Fig. 1.
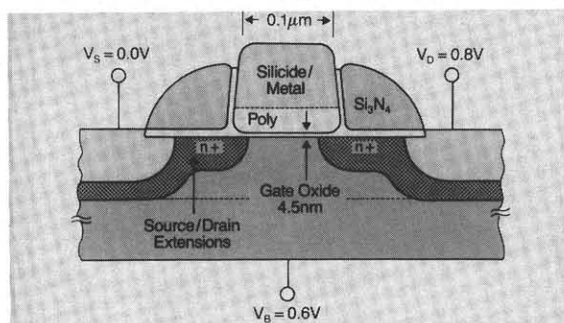

Fig. 1. Schematic device cross section and bias levels.

To observe the smallest features it is necessary to remove the layers above the gate level. Fig. 2 shows an SEM micrograph of such a stripped region.
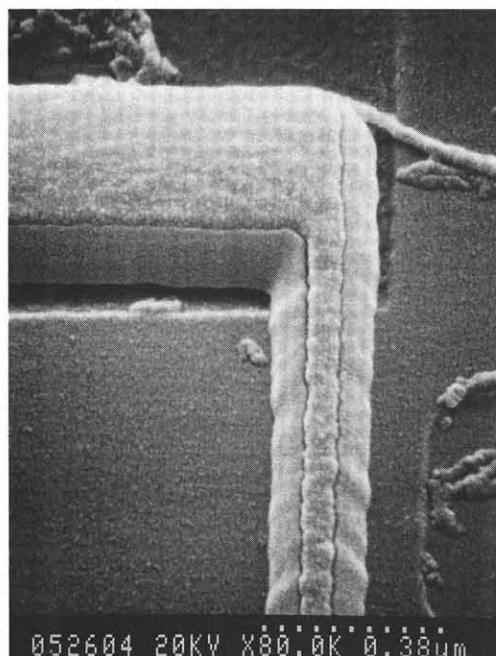

Fig. 2. SEM micrograph of a poly pad leading into a $0.07\mu m$ long gate. Metal and oxide has been removed. The nitride sidewall spacer is clearly visible.

Device operation and velocity overshoot

Approximately 75% of all tested structures were operational. Not surprisingly, no "generic" $0.1\mu m$ device behavior was found. Characteristics strongly depended on processing details. However, the common feature of the highest performance devices was an abrupt source/drain (S/D) junction-edge, resulting from the use of antimony. The device and circuit results to be presented are all from the same wafer, thus eliminating variations arising from processing differences. Fig. 3 shows the LT terminal characteristics of $0.1\mu m$ and $0.07\mu m$ gate-length devices.
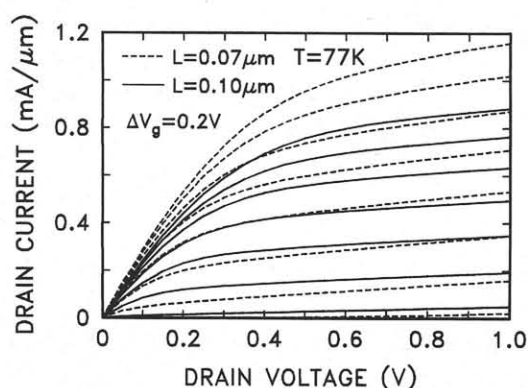

Fig. 3. Device terminal characteristics. Maximum V_G is 1.5V, substrate is biased at 0.6V. The $0.07\mu m$ device at $V_D = 1V$ has a maximum g_m of over $940\mu S/\mu m$, while the $0.1\mu m$ has $770\mu S/\mu m$. The room temperature g_m of these devices with 0V substrate bias is 590 and 505 $\mu S/\mu m$ respectively.

The transconductance, g_m, of these devices is the highest measured in FETs to date.

Fig. 4 gives g_m as function of gate-length. The plot was made for $V_D = 0.8V$, and for each gate-length at that V_G where g_m peaked. This typically occurred at $(V_G -V_T) \sim 0.6V$. The figure provides evidence for velocity overshoot[5] in the devices. This effect, by allowing carriers to surpass the steady-state saturation velocity, v_{sat}, leads to extraordinary g_m, provided parasitic resistances are under control[6]. The evidence for overshoot is twofold. First, there is simply the size of g_m in the shortest devices. If carriers were not capable of exceeding v_{sat}, then there would exist an unattainable upper limit for the intrinsic g_m : $v_{sat} \cdot C_{ox}$. This limit is marked on Fig. 4, and as can be seen, even the measured g_m exceeds this limit set for the intrinsic value. The second manifestation of velocity overshoot is in the trend

of g_m with gate-length. The solid line in Fig. 4 is g_m obtained by a conventional two-dimensional simulator (FIELDAY). The curve thus shows how g_m would have behaved in the absence of velocity overshoot.
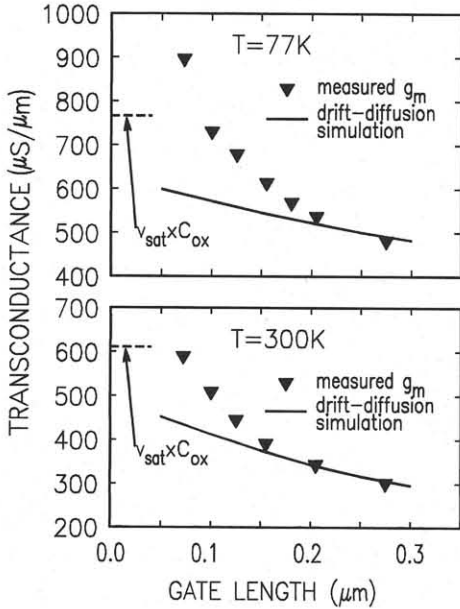
Fig. 4. Measured and calculated g_m at two temperatures. The solid lines show results of conventional two-dimensional simulations. The limit on the intrinsic g_m, $v_{sat} \cdot C_{ox}$, is also indicated.

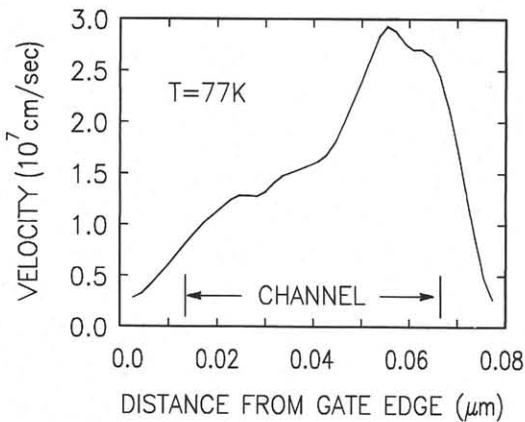The deviation of the data from the steady state transport curve is a clear sign of velocity overshoot.

Fig. 5. Carrier drift velocity as obtained by a two-dimensional self-consistent Monte Carlo simulator. Biases are: $V_D = 0.6V$, $V_G = 0.8V$, 0V on substrate.

The two-dimensional modeling also shows that the details of the S/D lateral junction edge, and the extent of the gate overlap of this region, are the most important factors in determining g_m. In spite of the abrupt junctions, these small devices inevitably become similar in behavior to LDD devices. Such characteristic arises

because even a ~20nm junction edge is not negligible when the channel is as short as 50 to 100nm. Of the total, ~250Ω•µm parasitic resistance, ~200Ω•µm was due to the S/D edges.

In the overshoot regime a non-local approach to transport is necessary. Fig. 5 shows the velocity as obtained in a self-consistent Monte Carlo simulation, which is non-local, and accounts for the full band structure of silicon[7] . According to the modeling, the carriers reach speeds that are more than double v_{sat} . It should be noted in general, that to achieve large g_m the carriers must reach a high velocity already at the metallurgical source edge. This means that they either must be able to accelerate inside the source, or to be injected into the channel with considerable energy by some novel mean.

Circuit performance

Of the three chip designed for switching time measurements two had inverter chains, while the third contained 21-stage ring-oscillators. The load elements in these circuits were enhancement mode devices with their gates tied to an independent power supply. The widths of the active and load devices were 5 and 1.25µm respectively. An SEM picture of one oscillator section is shown in Fig. 6.
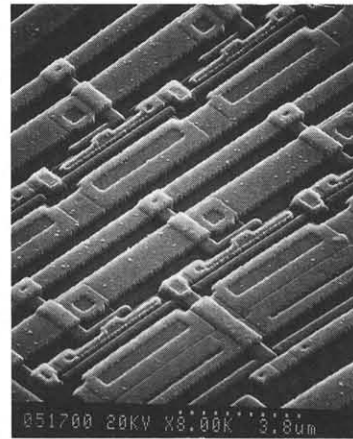
Fig. 6. SEM micrograph of a ring oscillator section.

The output of the ring-oscillators and chains were fed on-chip into either a push-pull driver, or a source follower. A trace of a 0.1µm gate-length ring-oscillator is shown in Fig. 7. The 13.1ps per-stage-delay is more than twice as fast as the previously obtained best value for FETs, and the fastest switching time for any type of device, including bipolars, that are made from silicon.
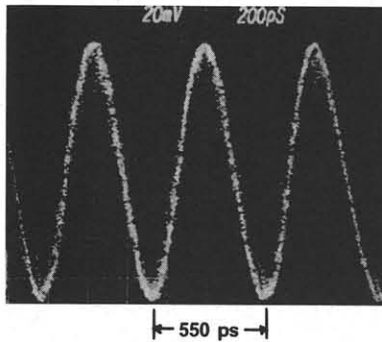
Fig. 7. Oscilloscope trace of a 21 stage $0.1\mu m$ gate-length unloaded ring oscillator. T=77K, V_{DD} =1.7V, 0.6V on substrate.

The wafer on which this data was taken had a higher threshold than the nominal design called for. For this reason the ring-oscillator performed best at high power levels. But, the delay versus power curve was rather flat, as expected[2]. Fig. 8. shows switching time against power for a $0.1\mu m$ gate-length ring-oscillator.
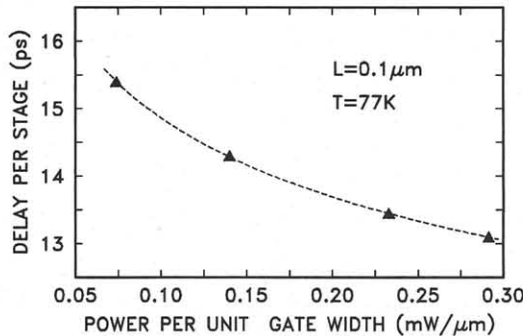

Fig. 8. Delay versus power plot. At the lowest power point V_{DD} =0.9V

The following delay times were measured as function of gate-length[8] 19.5ps at $0.20\mu m$, 17.8ps at $0.16\mu m$, 16ps at $0.13\mu m$, and 13.1ps at $0.10\mu m$. For the latter the room temperature value was 17.7ps. Measurements on inverter chains gave consistent results with the ring oscillators.

Circuit simulations based on the measured device characteristics show that for gate-lengths below ~$0.13\mu m$ the performance was limited by a particular processing problem, and does not represent the capability of the devices. For the $0.1\mu m$ gate-length the delay should be in the 7ps range. With an advanced isolation, for instance with shallow trenches, and with properly scaled diffusion level, the delay could be below 5ps.

Summary and conclusions

Although evaluation is still not complete, it appears that with proper design, and with processing along established avenues, an FET technology at, or even below, the $0.1\mu m$ gate-length level is possible. Performance enhancement due to velocity overshoot was clearly in manifest. The measured and projected delay times give confidence that LT MOSFETs can be contenders even for the highest speed circuits. Together these results show that the value of miniaturization extends to dimensions that are smaller than what was commonly assumed to be the case.

Acknowledgements

References

1) G. Baccarani, M.R. Wordeman, and R.H. Dennard; IEEE Trans. on Electron Dev., ED-31, 452 (1984).

2) G. A. Sai-Halasz, M. R. Wordeman, D. P. Kern, E. Ganin, S. Rishton, D. S. Zicherman, H. Schmid, M. R. Polcari, H. Y. Ng, P. J. Restle, T. H. P. Chang, and R. H. Dennard; IEEE Electron Device Lett. EDL-8, 463 (1987).

3) G. A. Sai-Halasz, M. R. Wordeman, D. P. Kern, E. Ganin, S. A. Rishton, H. Y. Ng, D. Zicherman, D. Moy, T. H. P. Chang, and R. H. Dennard; IEDM Tech. Dig., 397 (1987).

4) S. A. Rishton, H. Schmid, D. P. Kern, H. E. Luhn, T. H. P. Chang, G. A. Sai-Halasz, M. R. Wordeman, E. Ganin, and M. Polcari; J. Vac. Sci. Technol. B6, 140 (1988).

5) S.Y. Chou, D. A. Antoniadis, and H. I. Smith; IEEE Electron Device Letters, EDL-6, 665 (1985)

6) G. A. Sai-Halasz, M. R. Wordeman, D. P. Kern, S. Rishton, and E. Ganin; IEEE Electron Device Letters, Sept. (1988)

7) S. E. Laux and M. V. Fischetti; IEEE. Electron Device Lett. Sept. (1988)

8) G. A. Sai-Halasz, M. R. Wordeman, D. P. Kern, S. Rishton, E. Ganin, H. Y. Ng, D. Moy, T. H. P. Chang, and R. H. Dennard; IEEE Electron Device Letters, Submitted