A Bottom-Gate Metal–Oxide Thin-Film Transistor With Self-Aligned Source/Drain Regions

Zhihe Xia, Lei Lu, Member, IEEE, Jiapeng Li, Hoi-Sing Kwok, Fellow, IEEE, and Man Wong, Senior Member, IEEE

Abstract—Proper driving of a large-area, high-resolution, and high-frame-rate active-matrix display can be hindered by excessive delay along a signal path, such as a scan line. Depending on the transistor architecture, such delay could be dominated by the parasitic overlap capacitance between the gate electrodes and the source/drain (S/D) regions of the address-transistors attached to a scan line. While the capacitance can be minimized by employing a transistor with the edges of the regions self-aligned (SA) to those of the electrode, actual implementation is easier with a top-gate rather than the more commonly deployed bottom-gate architecture. Presently reported is the realization and characterization of a bottom-gate, SA indium–gallium–zinc oxide transistor. The extent of the overlap between the S/D and the gate is determined by a thermal “activation” process, similar to how it is controlled in a conventional top-gate, SA transistor.


I. INTRODUCTION

A MATRIX display is commonly driven by sequentially addressing its rows of scan lines. The delay of a signal propagating along a scan line is determined largely by the product of the associated resistance ($R$) and capacitance ($C$), with a larger RC time-constant corresponding to longer delay. With increasing display size, resolution, and frame rate, $R$ and $C$ must be adjusted accordingly. The two main components making up $C$ are the capacitance of the interconnect and that of the attached address transistors. The latter can be further decomposed to include an intrinsic gate-channel capacitance and a parasitic gate–source/drain (S/D) overlap capacitance.

The bottom-gate, staggered thin-film transistor (TFT) architecture is commonly used [1], [2] in a matrix display. An inevitable overlap exists between the gate electrode and the S/D regions or electrodes of a TFT, due to the requirement of finite-alignment margins. The resulting parasitic overlap capacitance [3], [4] contributes to larger RC. Such parasitic capacitance, hence also its undesirable effects, can be reduced by employing a TFT with the edges (junctions) of the S/D regions self-aligned (SA) to those of the gate electrode. It is for this reason that top-gate, SA TFT, an architecture not commonly used in large-area matrix displays, was deployed in recently commercialized large size (>65 in), ultrahigh definition (such as 8K), and high frame-rate (>120 Hz) television sets [5].

Attributed to their compatible fabrication process with the incumbent amorphous silicon TFTs and their relatively higher field-effect mobility ($\mu_{FE}$), metal–oxide (MO) TFTs [6]–[8] such as those based on indium–gallium–zinc oxide (IGZO) are being deployed in the back planes of a new generation of active-matrix displays [9]–[12]. Recently, SA MO TFT structures have also been investigated [13]–[18], addressing the issues of parasitic capacitance.

MO can be made conductive by reaction with aluminum [13], plasma [14] or wet etching treatment [15], ion implantation [16], deep ultraviolet irradiation [17], laser annealing [18], hydrogen doping [19], or nonoxidizing thermal annealing [20]. When applied to a top-gate architecture, some of these techniques could be used to realize an SA TFT. However, common gate-electrode materials such as copper (Cu) and molybdenum (Mo) are gas-impermeable. Since defects are generated rather than annihilated in a transistor channel during a thermal treatment subsequent to the formation of the impermeable gate electrode [21], it has been pointed out [5] that a top-gate MO TFT exhibits poorer performance and reliability than a bottom-gate TFT—particularly if the latter is protected by an etch-stop (ES) layer and incorporates a fully oxidized channel. To maintain process continuity and to obtain better TFT reliability, one would desire a technology enabling self-alignment of the junctions with a bottom-gate architecture. Though self-aligned, bottom-gate structures based on a back-side exposure process have been proposed [18], [22]; these methods incur a penalty in terms of process complexity (hence cost) because the channel and the S/D regions are separately formed.
characterization of EMMO TFTs with SA S/D junctions. The effects of the thermal treatment on the location of the induced S/D junctions were exploited to realize an SA, bottom-gate TFT with self-alignment of the thermally induced gas-permeabilities [23], the elevated metal–oxide semiconducting thin films under covers of different gas-permeabilities [23], the elevated metal–oxide (EMMO) TFT architecture has been proposed and demonstrated [24], [25]. A distinguishing feature of the EMMO (EMMO) TFT architecture has been proposed and demonstrated [24], [25]. A distinguishing feature of the EMMO architecture is the self-alignment of the thermally induced S/D regions to a gas-impermeable cover layer. Presently this architecture is the self-alignment of the thermally induced gas-imperables [23], the elevated metal–oxide (EMMO) TFT architecture has been proposed and demonstrated [24], [25].

II. ANALYSIS OF RC DELAY OF A SCAN LINE

For a display of respective horizontal and vertical dimensions of \( S_H \) and \( S_V \), the present analysis is based on the assumption of a “square” pixel of pitch \( P \) and a technology with minimum lithography resolution of \( \lambda \). Shown in Fig. 1(a) and (b) are the respective schematics of a scan line modeled as a distributed resistor–capacitor network and the layout of a subpixel. Defined as the time taken to reach 90% of the strength of a signal under transmission, the time constant \( T_{RC} \) of the scan line is given approximately by [26]

\[
T_{RC} \approx 3N_H R_{PX} \times 3N_H C_{PX} = 9N_H^2 R_{PX} C_{PX} \tag{1}
\]

where \( N_H \equiv S_H/P \) is the number of pixels per scan line. The factor of 9 accounts for the fact that each pixel consists of three subpixels of the primary colors. In this nomenclature, \( R_{PX} \) and \( C_{PX} \) are the respective resistance and capacitance per subpixels.

If the scan line is addressed using drivers from its two ends, then only half of the pixels \( N_H/2 \) needs to be addressed from each end, the resulting reduced time constant, denoted by \( T_{RC}^* \), is given by

\[
T_{RC}^* \approx \frac{9}{4} N_H^2 R_{PX} C_{PX} \tag{2}
\]

With \( W_{SC} \) and \( R_s' \) denoting, respectively, the width and the sheet resistance of the scan line, the one obtains

\[
R_{PX} = \frac{P}{3W_{SC}} R_s' = \frac{S_H}{3N_H W_{SC}} R_s'. \tag{3}
\]

If the width of the data line is denoted by \( W_{DT} \), then \( W_{SC}/W_{DT} \) are the area of the crossover between the scan line and data line. This contributes a crossover capacitance of \( W_{SC} W_{DT} C_X' \) to \( C_{PX} \), where \( C_X' \) is the crossover capacitance per unit area. The capacitive loading on the scan line attributed to the address TFT with width \( W \) and gate length \( L_g \) is \( W L_g C_{GI} \), where \( C_{GI} \) is the gate insulator capacitance per unit area. Consequently,

\[
C_{PX} = W_{SC} W_{DT} C_X' + W L_g C_{GI}. \tag{4}
\]

Substituting (3) and (4) into (2), one obtains

\[
T_{RC}^* \approx \frac{3}{4} R_s' (N_H S_H) \left( W_{DT} C_X' + \frac{W}{W_{SC}} L_g C_{GI} \right). \tag{5}
\]

For the popular display format of 16:9, the number of scan lines is \( N_V = 9/16 N_H \). Denoting the frame rate by \( f \), one obtains an upper bound for \( T_{RC}^* \)

\[
T_{RC}^* < \eta \frac{1}{f N_V} \tag{6}
\]

where \( \eta \approx 1 \) is included to account for the finite time required to pass the signal on the data line to the pixel by the address-TFT.

Adopting \( \eta = 1/4 \) [27] and substituting (5) into (6), one obtains after rearranging terms

\[
N \equiv \frac{R_s' \left( W_{DT} C_X' + \frac{W}{W_{SC}} L_g C_{GI} \right)}{\text{Material selection and pixel design}} \leq \Xi \equiv \frac{16}{27} \left( \frac{1}{N_H^2 R_{PX}} \right). \tag{7}
\]

The LHS of (7) is denoted by \( N \) and related to the pixel design through various widths, the transistor architecture through \( L_g \) and the technology through the selection of materials. The RHS is denoted by \( \Xi \) and related to the display specification.
Fig. 2. Schematic cross sections of (a) ES, (b) BCE, and (c) TSA address-TFT.

The $RC$ delay attributed to the crossover between the scan line and data line is proportional to $WDT$, and independent of $WSC$; that attributed to the address-TFT depends on the TFT architecture primarily through $Lg$. For a given interconnect geometry, both are reduced by employing a conductor (such as Cu) with lower resistivity. To the extent possible to maximize the aperture ratio of a pixel, $WDT$ and $WSC$ are set to $\lambda$ in the present analysis. In practice and if required, each could be suitably enlarged at the expense of smaller aperture ratio to better accommodate, respectively, the data line delay and the charging of the address-TFT. Depending on the value of $Lg$, $W$ is adjusted accordingly to achieve the same current drive [28].

Since $C_{GI} \geq CX$ in practice, $RC$ delay is often dominated by that attributed to the address-TFT. It is for this reason that a TFT architecture allowing a small $Lg$ is advantageous in reducing the overall signal delay. Shown in Fig. 2 is the cross-sectional schematics of the conventional back-channel-etched (BCE) and ES bottom-gate TFT architectures, and the SA, top-gate (TSA) TFT architecture.

It is noted that because of alignment requirement, the respective minimum $Lg$ is $5\lambda$, $5\lambda$, and $\lambda$ for the BCE, ES, and TSA TFT (Fig. 2). The TSA architecture further allows the placement of the data line above the planarization layer, resulting in a thicker insulation (hence smaller $C_X$) between the data line and scan line. The parameters used in the present analysis are summarized in Table I.

Shown in Fig. 3 is the dependence of on display diagonal size with $N_H = 4K$ or 8K and $f = 120$ Hz. K for different TFT architectures are also indicated.

The relatively large feature size of the ES architecture precludes its application in an 8K display. The BCE architecture fares better, but only for diagonal size up to about 30 in. Though this size limit could be raised by adopting more advanced driving techniques [32], it is clear that the BCE architecture is not the most versatile candidate for 8K displays. The TSA TFT technology is clearly superior: enabled by the small size of the address-TFT and the small $C_X$, the upper bound for diagonal size is beyond 100 in for $f = 120$ Hz and $N_H = 8K$.

While retaining the size advantage of a TSA TFT, there is a desire to overcome its poorer reliability [5] by developing a process to realize an SA, bottom-gate TFT. This is presently implemented based on the EMMO TFT (Fig. 4) architecture.
III. DEPENDENCE OF THE LOCATION OF AN ANNEALING-INDUCED JUNCTION ON THERMAL TREATMENT

A windowed bridge structure (Fig. 5) was used to investigate the relationship between the thermal treatment and the location of an annealing-induced junction. A window was opened in the top layer of gas-impermeable silicon nitride (SiN$_x$), thus allowing oxidation of the IGZO layer covered under the exposed layer of gas-permeable silicon oxide (SiO$_y$).

Fabrication of the structure started with a heavily doped, n-type silicon substrate covered with 100-nm-thick thermal oxide. A stack of insulating films consisting of 75-nm-thick SiO$_y$ on a 50-nm-thick SiN$_x$ was deposited by plasma-enhanced chemical vapor deposition (PECVD). The respective carrier gases were 8-SCCM silane/400-SCCM nitrous oxide and 40-SCCM silane/40-SCCM ammonia; the temperature was 300 °C, and the pressure was 0.9 Torr.

A 20-nm-thick IGZO active layer was subsequently deposited at room temperature in a radio frequency magnetron sputtering machine at a process pressure of 3 mTorr in an atmosphere of 5% oxygen (O$_2$) and 95% argon. The molar ratio of the IGZO target was In$_2$O$_3$:Ga$_2$O$_3$:ZnO = 1:1:1, and the base pressure was ~1 µTorr. An active island of width $W_{IS}$ was patterned using 1/1000 molar aqueous hydrofluoric acid solution, before a 300-nm-thick gas-impermeable SiO$_y$ passivation layer and an 80-nm-thick gas-impermeable SiN$_x$ cover layer were sequentially deposited using the same PECVD processes.

The SiN$_x$ cover layer was patterned and a window of length $L_A$ and width > $W_{IS}$ was opened by etching in an inductively coupled plasma etcher running a sulfur hexafluoride chemistry, thus exposing the underlying SiO$_y$ passivation layer in the windows. Contact holes were subsequently opened before the metal electrodes, consisting of sputtered and patterned stacks of 300-nm-thick aluminum (Al) on a 50-nm-thick Mo, were defined. Finally, the devices were annealed in a conventional resistively heated horizontal atmospheric pressure furnace at 400 °C for 2 h with an N$_2$ flow rate of 8 SLPM.

Electrical characteristics were measured using an Agilent 4156C semiconductor parameter analyzer. Recognizing the similarity between the bridge structure and a bottom-gate TFT, one can modulate the conductance of the IGZO in the window by applying a “gate” bias $V_g$ to the heavily doped silicon substrate. In this analogy, the conductive regions on the two sides of the IGZO active layer in the window are equivalent to the S/D regions of the TFT.

Consistent with the formation mechanism of annealing-induced donor defects under the condition of nonoxidizing heat treatment, the edge (thus junction) of a conductive region is largely aligned to the corresponding edge of the gas-impermeable SiN$_x$ cover layer. With the misalignment between these edges denoted by $\Delta L_A$, the total resistance $R_T$ measured between the metal-to-S/D contacts is given by [33], [34]

$$R_T = \frac{L_A - 2\Delta L_A}{W_{IS}} \left( R_A' - R_{S/D} \right) + \frac{L_C}{W_{IS}} R_{S/D}^C \quad (8)$$

where $L_C = 300 \mu m$ is the separation between the contacts; $R_A'$ and $R_{S/D}^C$ are the respective sheet resistance of the IGZO active layer in the window and in the adjoining equivalent S/D regions. The dependence of $R_T$ on $L_A$ and $V_g$ after the initial 2 h, N$_2$-annealing at 400 °C is displayed in Fig. 6(a). The relatively constant $R_T$ implies $R_A' \approx R_{S/D}'$. This is reasonable since the IGZO inside the window was also subjected to a nonoxidizing anneal, resulting in the creation of donor defects, the large population of which is relatively insensitive to further modulation by $V_g$.

This population decreases after a subsequent oxidizing anneal in the same furnace and when it becomes negligible compared to the field-effect-induced population, one obtains

$$R_A' \approx \frac{1}{\mu_{FE} C_{GL} (V_g - V_T)} \quad (9)$$

It should be noted that $(V_g - V_T) \gg V_d$ is assumed in (9), where $V_d$ is the bias applied to the D electrode (Fig. 5).
is “exposed” to O2 diffusing through the SiOy passivation layer. Since the entire surface of the channel was fixed at 0.5 V.

Typical dependence of $R_T$ on $L_A$ and $V_g$ after a series of oxidizing thermal treatments with increasing duration is displayed in Fig. 6. $V_g$ varied between 18 and 22 V, and $V_d$ was fixed at 0.5 V.

For the initial state after the N2-annealing, the entire IGZO active layer was populated by donor defects [inset of Fig. 6(a)] without any junctions separating the “channel” in the window and the adjacent equivalent “S/D” regions. For each duration of O2-annealing at 400 °C, a unique $\Delta L_A$ can be extracted from the common intersection of the least-square fitting lines at various values of $V_g$. The dependence of $\Delta L_A$ on the O2-annealing time (Fig. 7) can be roughly divided into two regimes, with $\Delta L_A$, respectively, changing more quickly for annealing time less than 4 h and slower for longer annealing time.

During the subsequent short-duration oxidizing anneal, O2 entered the thin 20-nm IGZO in the window and annihilated the donor defects. Since the entire surface of the channel is “exposed” to O2 diffusing through the SiOy passivation layer, this is a regime characterized by “vertical” oxidation. The annihilation in the edge regions is incomplete during this short-duration anneal due to the supply of donor defects from the SiNx-covered equivalent S/D regions, resulting in a small but positive $\Delta L_A$ (inset I). Upon further annealing beyond 3.5 h, the defect-annihilation fronts (hence junctions) were pushed into the SiNx-covered regions. $\Delta L_A$ turned negative (Inset II) and its time rate of change slowed down appreciably. This is a regime characterized by “lateral” oxidation, hence annihilation, of the donor defects.

### IV. Self-Align IGZO TFTs Using Back-Side Exposure

The deeper knowledge gained on how a thermal treatment affects the alignment between an induced junction and the corresponding edge of a SiNx gas-impermeable cover was applied to the construction of an SA, bottom-gate TFT based on the EMMO architecture [35]. The fabrication of the TFT started with the sputter deposition and patterning of 120-nm-thick Mo gate electrode on a transparent glass substrate. Identical to the process sequence used to construct the bridge structure, the same stack of SiOy/SiNx gate insulator, IGZO semiconductor, gas-permeable SiOy passivation layer, and gas-impermeable SiNx cover layer were deposited and patterned.

A negative-tone photoresist layer was subsequently coated on the cover layer before it was flood exposed from the back side [Fig. 8(a)], thus capturing the image of the opaque gate electrode. A window was etched through the exposed cover layer [Fig. 8(b)] to reveal the underlying SiOy passivation layer. The edges of the window along the channel direction were self-aligned to those of the gate electrode. Contact regions exposing the IGZO active layer on the two sides were opened [Fig. 8(c)], before the S/D electrodes consisting of the same stack of Al on Mo were patterned. Finally, conductive S/D regions self-aligned to the edges of the cover layer, hence also to those of the gate electrode, were formed after a thermal treatment at 400 °C in O2 [Fig. 8(d)].

With an overlap of $\sim$4 µm between the gate and the S/D electrode, regular EMMO TFTs with lithographically defined S/D electrodes were also constructed for comparison. $V_g$-dependent gate capacitance of the regular and SA TFTs is compared in Fig. 9 for TFTs with different channel length $L$.

The values at sufficiently negative $V_g$ are insensitive to $L$ and provide estimates of the parasitic gate-junction capacitance of $\sim$241 and $\sim$30 fF, respectively, for the regular and the SA TFTs. Clearly the parasitic capacitance of the latter, at merely $\sim$13% of that of the former, is much smaller.

Shown in the Fig. 10 is drain current ($I_d$) versus $V_g$ transfer characteristics of SA EMMO TFTs with a different $W$ at the same $L = 5$ µm. From the peak transconductance ($g_m$) measured at a $V_d$ of 0.5 V, a $\mu_{FE} (\equiv L_{g_m} / W C_{GI} V_d)$ of 5.8 $\pm$ 0.4 cm²/Vs was extracted. A maximum subthreshold slope of $372 \pm 27$ mV/decade was extracted at $V_g$ larger than the turn-on voltage $V_{ON} (\equiv V_g$ to induce an $I_d$ of $L/W \times 10$ nA at $V_d = 5$ V). All devices exhibit relatively low OFF-state current of $\sim 10^{-13}$ A, hence a high ON/OFF-current ratio of at least $10^9$. This is consistent with the effectiveness of thermal oxidation in reducing the defect concentration in an
MO semiconductor. The performance of the SA EMMO TFTs is comparable to that previously reported for bottom-gate SA MO TFTs [18, 22].

Shown in the inset of Fig. 10 is the degradation of the normalized “ON” current ($I_{ON}$) of TFTs subjected to different annealing durations. Compared with $I_{ON}$ after 1 h of O$_2$-annealing at 400 °C, that after 2 h of annealing was slightly reduced. This is related to the continuous migration of the junctions toward to the edges defined by the S/D SiNx cover layer, hence a slight increase in the effective $L$. Upon further extending the annealing duration to 4 h and then 6 h, a precipitous drop in $I_{ON}$ was observed. This is caused by the migration of the junctions to beyond the edges of the gate electrode, thus breaking the connection between the field-induced channel and the annealing-induced S/D regions. In the present demonstration, annealing at 400 °C in O$_2$ between 1 and 3 h appears acceptable, reflecting a fairly wide process window.

V. Conclusion

It is shown that the gate-junction overlap capacitance of a TFT makes a significant contribution to the capacitive loading of a scan line in a matrix display employing the traditional bottom gate, staggered transistor architectures. The increased signal delay along the line, if not mitigated, also constrains the construction of large-area, high-resolution, and high-frame-rate displays. The process dependence of the migration of an annealing-induced junction in IGZO is studied and characterized. The insight acquired is applied to the construction of a bottom-gate transistor with annealing-induced junctions self-aligned to the edges of the gate electrode, employing merely an additional maskless, back-side flood exposure step. While offering the same small parasitic capacitance as that of an SA, top-gate transistor, this SA, bottom-gate technology further allows full oxidation of the channel region—thus improving the performance and reliability of the transistor.

ACKNOWLEDGMENT

The authors would like to thank the assistance of the Nanosystem Fabrication Facility, The Hong Kong University of Science and Technology, Hong Kong, in device fabrication.

REFERENCES


