Monolithic 3D CMOS Using Layered Semiconductors


Monolithic 3D integration, in which stacks of circuits are sequentially fabricated on top of each other, is a promising pathway to continuously increase device density. In addition, it can potentially reduce power consumption and increase operating frequency in integrated circuits (ICs) by breaking the chip input/output (I/O) bottleneck.1–4 We demonstrate monolithic 3D integration of high performance MoS2 and WSe2 layered semiconductor metal oxide semiconductor field effect transistors (MOSFETs) with electron and hole mobility of 38 and 238 cm2 V–1 s–1, respectively. Using these layered semiconductors, we implement digital circuits such as inverters, NAND and NOR; and analog circuits like differential amplifier, common-source amplifier, and signal mixer. We show experimental 3D complementary metal oxide semiconductor (CMOS) inverter operation at a record low supply voltage (VDD) of 150 mV and voltage gain of about 10 V/V at VDD = 3 V. Our work opens up the potential of monolithic 3D integration using layered semiconductors for high-density, ultra-low-voltage, and low-power applications.

ICs are so pervasive in modern human life that they are continuously shaping the very way we interact with the world. ICs enable a wide range of capabilities in information processing, transmission and storage, to name a few. N-channel MOSFET (NMOS) and P-channel MOSFET (PMOS) devices are ubiquitous active components in ICs. They are connected using metal interconnects to form digital, analog, and radio-frequency (RF) circuits. Scaling down the size of the MOSFET increases the areal density of the devices, thus enabling higher functionality, performance improvement, and cost reduction.5 To effectively turn on and turn off an extremely scaled future device, the channel thickness can only be a few nanometers,6 with a monolayer thickness being the limit. Layered transition metal dichalcogenide (TMD) semiconductors like MoS2 and WSe2 are promising channel materials for extremely scaled transistors due to their availability in uniform monolayer thin films. To this effect, several demonstrations of near-ideal 60 mV per decade switching behavior of monolayer and few layer TMD MOSFETs were reported.7,8 Layered materials like graphene and TMDs can be formed into monolayer sheets. Unlike graphene, which does not have an intrinsic bandgap, TMDs like MoS2 and WSe2 have sufficient bandgap to keep the transistor off-state leakage current low. Once the devices and interconnects reach the ultimate scaling limit governed by lithography and nanopatterning, integration density can no longer be increased by conventional methods.9,10 This paper demonstrates a “scaling” path to improve integration density by sequentially stacking the TMD MOSFETs by taking advantage of the third dimension. We demonstrate a variety of monolithic 3D integrated digital circuits such as inverters, NAND and NOR logic gates, and analog circuits such as differential amplifier, common-source amplifier, and frequency mixer.

Modern IC chips contain billions of transistors, of which those used for digital CMOS logic form the majority. NMOS and PMOS transistors are the smallest units (Figure 1a). They are connected to form CMOS logic circuits (Figure 1c). To integrate NMOS and PMOS transistors on the same plane requires the respective channel materials to be placed side-by-side and form the gate stack and the contacts. In one of the 3D integration methods, NMOS and PMOS transistors are stacked on top of each other separated by a thick isolation oxide.11–13 We propose and demonstrate further density improvement through sharing of the NMOS and PMOS transistors’ gates. In all CMOS logic circuits, a pair of NMOS and PMOS transistors shares the same gate that accepts an input signal, and NMOS and PMOS transistors always exist in pairs. Taking inverter as
an example, NMOS and PMOS transistors can be folded on top of each other along the line of symmetry (M–N) so that each pair of these transistors shares a common gate (Figure 1d). The source and drain contacts for NMOS and PMOS transistors cannot be always shared and hence should be electrically isolated from each other and connected using metal lines as required. The resulting unit cell for 3D CMOS has an NMOS and a PMOS stacked on top of each other sharing a common gate metal and the source and drain contacts for the two transistors are electrically isolated as shown in Figure 1b. Additional layers of logic gates and memory can be stacked over this layer separated with a thick dielectric called the interlayer dielectric to further increase the integration density. ICs also contain a significant number of analog and radio frequency circuits which require more flexible layouts and hence need transistors with independent gates. Monolithic 3D integration scheme for such circuits is comprised of stacking the transistors on different levels that are isolated using a thick insulating oxide (Figure S1, Supporting Information). Monolithic 3D integration over two levels nearly halves the area occupied by a wide variety of circuits such as inverter (INV), 2-input NAND (NAND2), static random access memory cell (SRAM), 1-bit Full Adder (FA), and 8-bit adder (ADD8) (Figure 1e). To calculate the circuit area, layouts were designed for single-layer (1 layer) and monolithic 3D (2 layer) CMOS using design rules for the 14 nm technology node.[14]

Figure 2 shows the fabrication process scheme for monolithic 3D integration using TMDs. To fabricate the first layer MOSFET, MoS2 thin flakes are exfoliated on to Si/SiO2 (260 nm thick) substrate. Flakes with appropriate thicknesses (3–7 nm) are etched into rectangular patterns using xenon difluoride (XeF2)[15] (Figure 2a) followed by source/drain (Figure 2b), gate oxide (Figure 2c), and metal gate formation (Figure 2d). Two types of devices are implemented as second layer MOSFETs. First, for digital logic, a second layer thin gate oxide is deposited (Figure 2e). WSe2 is transferred on to the device stack using pick-and-place transfer method.[16] WSe2 films are patterned into rectangular patterns using XeF2 gas (Figure 2f), followed by source/drain metallization (Figure 2g). Figure 2h shows the high-resolution, cross-sectional transmission electron microscopy image of the structure. Energy dispersive
Figure 2. Process flow and device cross-sectional view of shared gate CMOS monolithic 3D architecture. a) Etching MoS₂ flake. b) Source and drain metallization on MoS₂. c) ZrO₂ oxide deposition using ALD. d) Gate electrode metallization. e) Second layer ZrO₂ deposition using ALD. f) Dry transfer and etching of WSe₂ flake. g) Source and drain metallization on WSe₂. h) HR-X-TEM images of the shared gate monolithic 3D CMOS structure, MoS₂, and WSe₂ regions. EDS mapping of the MoS₂ and WSe₂ regions. Scale bar for (a)–(g) is 7 µm.
X-ray spectroscopy data show the regions where MoS$_2$ and WSe$_2$ are present. Second, analog circuits require independent gates. The first layer of MoS$_2$ MOSFETs is fabricated as described earlier comprising MoS$_2$ film etch, source and drain metallization, ZrO$_2$ gate oxide deposition, and gate metallization (Figure S2a–c, Supporting Information). Thick layers of SiO$_2$ (10 nm) and ZrO$_2$ (40 nm) are deposited on top of the bottom layer MoS$_2$ transistors to reduce the interlayer capacitive coupling between bottom and top layer devices followed by MoS$_2$ dry transfer and etch (Figure S2d, Supporting Information). The interlayer capacitance is inversely proportional to the oxide thickness between the top and bottom layer transistors and increasing the oxide thickness decreases the interlayer capacitance. Next, source/drain metallization, ZrO$_2$ gate oxide deposition, and gate metallization are formed (Figure S2e–g, Supporting Information). The resulting analog circuit is a common source amplifier on two layers and its circuit schematic is shown in Figure S2h (Supporting Information).

Quasistatic DC $I$–$V$ measurements performed on the MoS$_2$ and WSe$_2$ MOSFETs show transfer and output characteristics of representative NMOS (Figure 3a,b) and PMOS (Figure 3c,d) devices with gate length/width of $2/3.3$ μm and $2/5.9$ μm, respectively. Both devices show an off-state leakage current ($I_{OFF}$) of about a few pA, on-state drain current ($I_{ON}$) in the μA range, and $I_{ON}/I_{OFF}$ ratio greater than $10^7$. The series resistance and mobility corrected for series resistance as extracted from drain current$^{[17]}$ is about $1.3$ kΩ $\mu$m and $38$ cm$^2$ V$^{-1}$ s$^{-1}$, respectively for MoS$_2$ MOSFET and $1.1$ kΩ $\mu$m and $238$ cm$^2$ V$^{-1}$ s$^{-1}$, respectively for WSe$_2$ MOSFET, which agree well with those reported in the literature$^{[17,18]}$. Both the devices show good electrostatic control as indicated by very low drain-induced barrier lowering (DIBL) of about $80$ mV V$^{-1}$ and $50$ mV V$^{-1}$ for MoS$_2$ and WSe$_2$ MOSFETs, respectively. The devices also show good output saturation. Output saturation, along with good transconductance, is important as they improve the gain of transistors and hence circuits, which in turn improves their noise immunity.$^{[19]}$ The MoS$_2$ MOSFET (Figure S3a, Supporting Information) is encapsulated when the second layer of ZrO$_2$ is deposited by atomic layer deposition (ALD) (Figure S3b, Supporting Information) which helps in reducing hysteresis (Figure S3c, Supporting Information). Forming gas annealing at $120$ °C for $30$ min improves mobility by passivating the dangling bonds at the MoS$_2$/ZrO$_2$ interface (Figure S3d, Supporting Information). Low temperature processing is a key feature in fabricating monolithic 3D structures presented here. Low-budget thermal processing is essential for monolithic 3D integration, as it reduces degradation of TMD, gate stack, and contacts. MoS$_2$ MOSFETs fabricated on two layers show nearly identical performance with similar drain current, excellent switching behavior, mobility of about $35$ cm$^2$ V$^{-1}$ s$^{-1}$, and good output saturation (Figure S4, Supporting Information). This shows the promise of (a) heterogeneous integration of TMD materials on existing device layers without significantly affecting the performance of the devices, and (b) building more TMD layers on top of the existing layers. WSe$_2$ PMOS device shows ambipolar conduction, which is not ideal for CMOS circuit operation. Ambipolar conduction increases leakage current and hence leakage power. Ambipolar conduction in WSe$_2$ can be controlled by choosing appropriate contact metal$^{[20]}$ and doping.$^{[7]}$ WSe$_2$ is exposed to the environment and can lead to stability issues. Device encapsulation using a suitable insulating oxide as followed in a state-of-the-art CMOS process will be required for commercial implementation of this structure.

The monolithic 3D CMOS structure shown in Figure 2g was used to implement digital circuits like inverter, 2-input NAND, and 2-input NOR (Figure S5a, Supporting Information). To implement 3D CMOS inverter, only one half of the structure shown in Figure S5b (Supporting Information) is required. For NAND and NOR gates (Figure S5c,d), the source and drain terminals of the NMOS and PMOS devices are to be connected appropriately as shown in Figure S5e (Supporting Information). Voltage transfer characteristics of the inverter show that

![Figure 3. Quasi-static DC $I$–$V$ measurements of bottom layer MoS$_2$ NMOS and top layer WSe$_2$ PMOS devices. a) $I_D$–$V_G$ characteristics of MoS$_2$ N-MOSFET. b) $I_D$–$V_G$ characteristics of MoS$_2$ N-MOSFET. c) $I_D$–$V_G$ characteristics of WSe$_2$ P-MOSFET. d) $I_D$–$V_G$ characteristics of WSe$_2$ P-MOSFET.](ww.MaterialsViews.com)
the output voltage ($V_{\text{OUT}}$) is high when the input voltage ($V_{\text{IN}}$) is low and $V_{\text{OUT}}$ transitions to a low voltage when $V_{\text{IN}}$ is increased (Figure 4a). The switching is observed for supply voltages from large to small values (3 down to 0.5 V). The transition region is narrow and the resulting noise margins are high due to good output saturation in MoS$_2$ and WSe$_2$ MOSFETs. The highest gain occurs at the midpoint in the transition region. Figure 4a shows the voltage transfer characteristics of the CMOS inverter with gain of about 10 V/V for a supply voltage of 3 V, which is similar to those reported for TMD CMOS inverters fabricated on a single plane.[20,21] Note that the data shown in Figures 3 and 4a are from different devices. For an inverter to be useful in a logic circuit, its gain should be greater than 1. [19] The minimum supply voltage at which gain is greater than 1 is about 150 mV (Figure 4b), which shows promise for ultra-low-voltage, low-power circuits. The total power ($P_{\text{tot}}$) consumption in a CMOS circuit is comprised of dynamic power ($P_{\text{dyn}}$), static power ($P_{\text{stat}}$), and short-circuit power ($P_{\text{sc}}$): $P_{\text{tot}} = P_{\text{dyn}} + P_{\text{stat}} + P_{\text{sc}}$.[19] For a CMOS inverter, dynamic power consumption occurs due to the charging and discharging of the load capacitor ($C_I$). $P_{\text{dyn}} = \alpha C_I V_{\text{DD}}^2 f$, where $\alpha$ is the activity factor of the circuit. $V_{\text{DD}}$ is the supply voltage, $f$ is the frequency of operation. In most circuits, $P_{\text{dyn}}$ is the dominant component. Since $P_{\text{dyn}}$ has a quadratic relationship with $V_{\text{DD}}$, reduction of $V_{\text{DD}}$ from 3 V to 150 mV will result in an estimated 400× lower power consumption. Figure 4c shows the truth table of the CMOS digital circuits implemented in this work. Functionality of monolithic 3D CMOS digital logic circuits is verified by applying square waveforms to their gates using a signal generator and measuring the output waveforms using a digital storage oscilloscope. Figure 4d shows the output is high when the input is low, and output is low when the input is high for the inverter. NAND gate shows a low output only when both the inputs are high, and in other cases the output is low. NOR gate shows a high output only when both the inputs are low, and in other cases the output is low. The testing conditions for all input conditions of monolithic 3D CMOS logic gates are shown using schematic circuit diagram in Figure S6 (Supporting Information). Threshold voltage engineering will be required to form large interconnected circuits. Threshold voltages of NMOS and PMOS devices should be positive and negative, respectively.[20]

For monolithic 3D analog circuits, the first layer MoS$_2$ devices are fabricated (Figure 5a), followed by deposition of interlayer dielectric and fabrication of second layer MoS$_2$ devices (Figure 5b), as described earlier. Figure 5c shows a differential amplifier as implemented in the bottom layer (layer 1) is operated in single-ended input and output mode. When an input peak-to-peak voltage of 50 mV is applied, an output peak-to-peak voltage of about 270 mV is observed resulting in a voltage gain of about 5.4 V/V as shown in Figure 5d. Output is connected to a common-source amplifier with current source load fabricated on the second layer. Output peak-to-peak voltage

---

**Figure 4.** Switching characteristics of static CMOS digital circuits using shared gate monolithic 3D CMOS architecture. a) Voltage transfer characteristics ($V_{\text{OUT}}$ vs $V_{\text{IN}}$) of inverter for supply voltage of 0.5–3 V. b) Voltage gain ($dV_{\text{OUT}}/dV_{\text{IN}}$) of inverter as a function of supply voltage. Minimum supply voltage at which voltage gain is greater than 1 is 150 mV. c) Truth table of inverter, 2-input NAND, and 2-input NOR. d) Switching waveforms for inverter, 2-input NAND, and 2-input NOR.
of this circuit is about 300 mV resulting in a voltage gain of about 1.1 V/V. Figure 5e shows common-source amplifiers with triode load transistors implemented on the bottom and top layers. The output of the bottom layer circuit is connected to the input of the top layer circuit. For a 50 mV peak-to-peak input signal, the output shows a peak-to-peak voltage of about 230 mV and hence a voltage gain of 4.7 V/V (Figure 5f). The bottom circuit drives an amplifier with voltage gain of about 1.2 V/V. Signal mixer is implemented on two levels using MoS2 MOSFETs. Mixer is a nonlinear circuit with two inputs of frequencies $f_1$ and $f_2$ and multiplies the inputs resulting in output of frequency $m f_1 \pm n f_2$, where $m$ and $n$ are integers. Small signals with frequencies ($f_1 = 1$ kHz and $f_2 = 0.05$ kHz) are applied to gate and source terminals of the top and bottom layer MoS2 MOSFETs. The output shows mixing operation of the two inputs. The number of MoS2 MOSFETs integrated to demonstrate monolithic 3D integration of analog circuits is between two to twelve.

Summarizing, using low-temperature processing we have demonstrated monolithic 3D integration of high performance TMDs like MoS2 and WSe2 with $I_{ON}/I_{OFF}$ ratio greater than $10^7$, and high electron and hole mobility of 38 cm$^2$ V$^{-1}$ s$^{-1}$ and 238 cm$^2$ V$^{-1}$ s$^{-1}$, respectively. We implemented complex digital CMOS logic circuits like inverter, NAND, and NOR using a shared-gate architecture, and analog circuits like amplifiers and mixers. The minimum supply voltage at which the monolithic 3D CMOS inverter can operate with a voltage gain of at least 1 V/V is 150 mV. Such an ultra-low-voltage circuit operation has neither been shown for circuits built with novel materials nor for any 3D-integrated circuit using any material system. The minimum voltage for circuit operation is comparable to the minimum supply voltage demonstrated in mature silicon CMOS technologies$^{[22,23]}$ that enables computation with ultralow energy consumption. This is a major breakthrough result which opens up the possibility of highly-integrated 3D, ultralow voltage, ultralow power applications. Our low-temperature, monolithic 3D integration platform can potentially enable highly integrated and massively interconnected computing platforms with ultra-low energy consumption such as neuromorphic computing.$^{[24–26]}$ A key requirement will be large area deposition or growth of TMDs with atomically controlled uniform thickness$^{[27–29]}$ sequentially on multiple levels. We envision that our demonstration here will enable TMD-enabled ultralow power systems with circuits for digital, analog, RF, memory, and sensors as stand-alone systems or integrated on conventional semiconductor platforms such as FinFET and Nanowire FETs by virtue of its very low-temperature processing.

Experimental Section

Device Fabrication: Thin films from MoS2 crystal purchased from Structure Probe, Inc. (SPI) were transferred on to a Si/SiO2 (260 nm) substrate using mechanical exfoliation. The substrate was soaked
in a hot acetone bath for 1 h followed by rinsing with acetone and iso-propyl alcohol (IPA) to minimize tape residue. MoS2 films with 3–7 nm thicknesses were identified using color contrast and confirmed using atomic force microscopy (AFM). Poly methyl methacrylate (PMMA) C4 electron beam resist was used. Using electron-beam lithography (EBL) (PMMA) C4 electron beam resist was spin coated onto the substrate using atomic force microscopy (AFM). Poly methyl methacrylate layer MoS2 MOSFETs, a thick layer of SiO2–7 nm thicknesses were identified using color contrast and AFM. The source and drain regions were defined using electron-beam evaporation followed by metal liftoff process. For analog circuits, after forming the first metalization, gate oxide deposition, and gate metal deposition as described, followed by electron-beam evaporation of 10 nm platinum and 30 nm gold and metal liftoff process. For digital CMOS circuits, this metal layer also acts as the shared gate electrode for the WSe2 MOSFET which will be placed on top of this stack. The gate oxide region for the WSe2 MOSFET was defined using EBL. 1 nm of SiO2 was deposited using electron-beam evaporated followed by deposition of 20 nm of ZrO2 using ALD. The top gate region for the MoS2 MOSFET was defined using EBL so that the gate overlaps with the source and drain regions. 40 nm of Ni was deposited using thermal evaporation followed by metal liftoff process. For digital CMOS circuits, this metal layer also acts as the shared gate electrode for the WSe2 MOSFET which will be placed on top of this stack. The gate oxide region for the WSe2 MOSFET was defined using EBL. 1 nm of SiO2 and 20 nm of ZrO2 were deposited using electron-beam evaporation and ALD, respectively, followed by oxide liftoff. WSe2 films were etched from their crystals purchased from HQ Graphene. WSe2 films with 3–7 nm thicknesses were identified using color contrast and AFM. The required WSe2 films were transferred onto the device stack using dry transfer method followed by etching the films into rectangular shapes using XeF2 gas. The source and drain regions were defined using EBL followed by electron-beam evaporation of 10 nm platinum and 30 nm gold and metal liftoff process. For analog circuits, after forming the first layer MoS2 MOSFETs, a thick layer of SiO2 (10 nm) and ZrO2 (40 nm) was deposited using electron-beam evaporation and ALD, respectively. Thin films of MoS2 were transferred using dry transfer method followed by etching using XeF2 gas. This was followed by source and drain metallization, gate oxide deposition, and gate metal deposition as explained earlier.

Electrical Measurements: Quasistatic measurements were done using Hewlett-Packard 4155B semiconductor parameter analyzer. The transient operation of the integrated circuits was performed using Agilent 81150A arbitrary waveform generator as the input source and the output signals were captured using Agilent digital oscilloscope 1024A. All the measurements were performed in air inside a shielded probe station at room temperature.

Supporting Information
Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements
A.B.S. and M.T. contributed equally to this work. A.B.S. and C.H. conceived the idea. A.B.S., M.T., and A.J. formulated the fabrication flow. A.B.S., M.T., S.B.D., C.-Y.H., D.-H.L., S.R.M., M.H., J.S.K., and Y.Z. fabricated the devices. A.B.S. and M.T. performed electrical measurements. A.B.S. analyzed the data. Y.-Z.C. and Y.-L.C. did transmission electron microscopy. All the authors were involved in preparing the manuscript. A.B.S. was funded by Applied Materials, Inc. and Entegris, Inc. under the I-RiCE Program. M.T. was funded by the Director, Office of Science, Office of Basic Energy Sciences, and Materials Sciences and Engineering Division of the U.S. Department of Energy under Contract No. DE-AC02-05Ch11231.

Received: October 16, 2015
Revised: November 11, 2015
Published online: DOI: 10.1002/adma.201505113